Decoy Selection in Protein Structure Determination via Symmetric Non-negative Matrix Factorization

Kazi Lutful Kabir
Dept of Computer Science
George Mason University
Fairfax, VA, USA
kkabir@gmu.edu

Gopinath Chennupati
Information Sciences (CCS-3) Group
Los Alamos National Laboratory
Los Alamos, NM, USA
gchennupati@lanl.gov

Raviteja Vangara
Fluid dynamics & Solid mechanics (T-3)
Los Alamos National Laboratory
Los Alamos, NM, USA
rvangara@lanl.gov

Hristo Djidjev
Information Sciences (CCS-3) Group
Los Alamos National Laboratory
Los Alamos, NM, USA
djidjev@lanl.gov

Boian S. Alexandrov
Physics & Chemistry of Materials (T-1)
Los Alamos National Laboratory
Los Alamos, NM, USA
boian@lanl.gov

Amarda Shehu
Dept of Computer Science
George Mason University
Fairfax, VA, USA
amarda@gmu.edu

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313299

***Abstract*—The so-called dark proteome, referring to regions of the protein universe that remain inaccessible by either wet- or dry-laboratory methods, continues to spur computational research in protein structure determination. An outstanding challenge relates to the ability to discriminate relevant tertiary structure(s) among many structures, also referred to as decoys, that are computed for a protein of interest. The problem is known as decoy selection. While prime for investigation as an inference problem, the decoy datasets generated in silico are sparse and highly imbalanced towards the negative class (irrelevant structures). These characteristics continue to challenge both supervised and unsupervised learning approaches to this problem. In this paper, we propose a novel decoy selection method based on symmetric non-negative matrix factorization in a graph clustering setting. The method is evaluated on two datasets, a benchmark dataset of ensembles of decoys for a varied list of protein molecules, and a dataset of decoy ensembles for targets drawn from the recent CASP competitions. The evaluation demonstrates that the proposed method outperforms several state-of-the-art decoy selection methods. This performance, as well as the method's computational expediency, suggest that the proposed method advances the state of the art in decoy selection and, in particular, our the ability to tackle inherent challenges related to imbalanced datasets.**

***Keywords*—decoy selection, eigen-gap heuristic, graph clustering, protein structure determination, symmetric NMF.**

I. Introduction

The tertiary (three-dimensional) structure is recognized to be central to the biological activities of a protein in the living cell [1]. Currently, however, many regions of the protein universe are inaccessible by either wet- or dry-laboratory methods [2]. It is estimated that about 44–54% of the proteome in eukaryotes and viruses (about $546,000$ proteins) and about 14% of the proteome in archaea and bacteria is dark.

The dark proteome continues to spur computational research in protein structure determination [3]. When no structural template exists for a protein target, the task of structure determination is beyond the scope of homology modeling [4]. Addressing this task with template-free methods involves first generating many physically-realistic tertiary structures. A designed molecular energy/score function is used as a *proxy* to evaluate the physical relevance, or the nativeness of a tertiary structure. The emphasis on proxy is due to the fact that, while useful, all such functions fall short and result in many irrelevant structures dominating the generated dataset [5], [6]; hence, the popular term *decoy* is used when referring to a generated structure in this setting.

It is imperative that the generated decoys be further assessed so as to determine one or more that are native. Determining this is known as model accuracy/quality assessment, model selection, or decoy selection. The terms *model* and *decoy* are used interchangeably. We will use the term decoy from now on, as model has a different meaning in machine learning (ML). It is also worth noting that there are differences between assessment and selection that are not explicitly stated in related literature. Decoy assessment involves assessing each given decoy via a quantity/score that evaluates its "nativeness". Decoy selection involves selecting from a given set of decoys one or few and predicting them as (near-)native.

One can utilize assessment for selection, for instance, by relying on a ranking-based approach, but not necessarily. Indeed, unsupervised learning approaches based on clustering remain popular. A summary of related work on decoy assessment and selection is provided in Section II. As our exposition relates, assessment methods are challenged by the accuracy of scoring functions, whereas selection methods that rely on clustering, and to a great extent even current methods that leverage supervised learning, struggle with data sparsity and imbalance. A recent method published by our laboratories offered non-negative matrix factorization as a novel framework for unsupervised decoy selection [7].

In this paper, we propose a novel decoy selection method, SNMF-DS, that utilizes symmetric non-negative matrix factorization (NMF) in the graph clustering setting for decoy selection. The method is fully non-parametric and employs the eigen-gap statistic to automatically determine the number of components for matrix factorization.

SNMF-DS proceeds in stages, first organizing the decoys into groups, identifying a best group, and then drawing a best decoy from this group. The potential energy of decoys is used to determine the best group. A decoy weighting scheme is employed to find the best decoy from the best group. Section III describes the proposed SNMF-DS in greater detail.

Extensive experiments and evaluation via rigorous metrics, related in Section IV, show that SNMF-DS outperforms several state-of-the-art methods, suggesting that matrix factorization advances the state of the art in decoy selection. Section V summarizes the performance and offers some further directions of work in this thread of research.

II. Related Works

In the very early days of decoy selection being recognized as central to protein structure determination, energy-based methods were prominent. These methods, also referred to as single-model methods (recall that model in this context is interchangeable with decoy), were based on the hypothesis that better scores or energies translated to closer proximity to the unknown native structure. This turned out not to be the case, particularly for the early energy functions [8].

In response, researchers pursued clustering-based methods (also known as multi-model methods). These methods ignored decoy energies/scores and instead clustered decoys based only on structural similarity [9], [10]. MUFOLD-CL [11] represents a state-of-the-art clustering method to which we compare the proposed SNMF-DS method in this paper.

Clustering-based methods were shown superior to energy-based methods for some time, as evaluated in the Critical Assessment of protein Structure Prediction (CASP) series of biannual community-wide experiments [12]. However, better energy-based methods have emerged recently, with energy functions of ever-increasing accuracy. SBROD [13] represents the state-of-the-art in such methods, and for this reason we include it in the comparative evaluation of the SNMF-DS method we propose in this paper.

The landscape of decoy selection methods is rich in its diversity. The landscape includes quasi-single and supervised learning methods. Quasi-single combine (and improve upon) the aspects of energy- and clustering-based methods by first picking up some high-quality structures that are then compared with the rest of the decoys [14]. Recent methods leverage the concept of the energy landscape and offer basins as more effective clusters [15], [16].

A recent thread of research introduced NMF for decoy selection [7], [17]. NMF has shown promise in various computational biology applications [18]–[20]. However, NMF remains largely unexplored for protein decoy selection. Work in [7] debuted an NMF-based method for decoy selection that leveraged several features pre-computed for decoys and can be explored further at a large scale [21]. In this paper, we debut a symmetric NMF framework for decoy selection. The framework is feature-agnostic, non-parametric, computationally expedient, and can handle sparse and highly imbalanced datasets. We now describe it in greater detail.

III. Methodology

We first provide a conceptual summary of the proposed SNMF-DS before relating further methodological details.

A. Proposed Framework for Decoy Selection

The framework that SNMF-DS operationalizes for decoy selection proceeds in three stages. In the first stage, given decoys, which we recall are tertiary structures of a given target protein, are organized into groups $\{G_i\}$. The second stage utilizes decoy energies to discriminate among the groups and select a best group G^* from $\{G_i\}$. In the third stage, a weighting scheme associates weights with decoys in the best group to select a best decoy from the best group. The best decoy is the one offered for prediction. Conceptually, the framework is related in Figure 1.

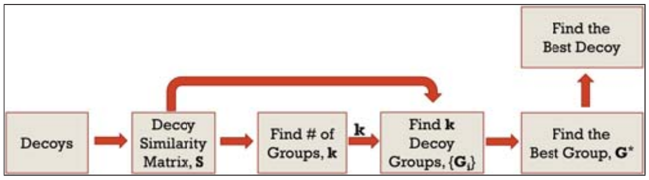

Fig. 1. The framework operationalized by the proposed SNMF-DS method is shown schematically here.

Figure 1 shows that the input to SNMF-DS are the Cartesian coordinates of the decoys. These are utilized to construct a decoy similarity matrix, which is then subjected to an EigenGap heuristic in order to determine k, the number of groups. We use this information of k to perform symmetric non-negative matrix factorization. The resultant factor (*W*) is used to elucidate k groups in which decoys are organized by finding the group membership from the factor matrix, *W*. The method is non-parametric, as the value for k is determined automatically by exploiting the EigenGap heuristic. We now proceed to relate each of these steps in greater detail.

B. From Decoys to Decoy Similarity Matrix

SNMF-DS takes as input the decoys of a given protein target. Each decoy is a tertiary structure computed via a template-free method. Given the popularity and public availability of the Rosetta platform and, in particular, the ease of using the open-source PyRosetta, the decoys of a given protein target in this paper are generated via the Rosetta AbInitio protocol [22]. Each computed tertiary structure is stripped down to its main-chain carbon atoms (the CA atoms), discarding side-chain atoms and other backbone atoms. This reduction improves the cost of computing the similarity matrix S in Figure 1.

The matrix S is symmetric and contains at entry $S_{i,j}$ the similarity between two decoys i, j in the given decoy set. Specifically, $S_{i,j} = \frac{1}{\mathrm{RMSD(i,j)}+\epsilon}$. In this equation, RMSD refers to the root-mean-squared-deviation (RMSD) [23] metric that measures the dissimilarity between two decoys, and ϵ refers to an infinitesimally-small constant set to $1e-12$. While different metrics other than RMSD can be used, we elect to use RMSD due to its popularity in comparing molecular structures. RMSD averages the Euclidean distance over the number of atoms (CA atoms in our case). In order to remove differences due to rigid-body motions in three dimensions (whole-body translation and rotation), the decoys are first optimally superimposed over an arbitrarily-chosen decoy (we use the first in the set of given decoys) to minimize differences due to rigid-body motions. In this way, the decoy pairwise RMSD values capture internal structural differences rather than differences due to whole-body motions in space. In the above equation, the reason for using ϵ is to guard against, in principle, a division by 0 in the case of two identical decoys.

C. From Decoy Similarity Matrix to Number of Decoy Groups

In this step, SNMF-DS finds the number of decoy groups in a non-parametric manner using EigenGap heuristic [24], [25]. The pairwise decoy similarity matrix S is used to search for the m nearest neighbors of each decoy. Some suggestions about the value of m are available in literature [26], such as $log(n) + 1$, $\sqrt{n}$, $2n^{1/d}$ where, n is the number of decoys in our setting, and d is the number of coordinates in a decoy. For our implementation, we pick $m = \sqrt{n}$.

Finding the m nearest neighbors of each decoy in the decoy set is instantiating a nearest-neighbor graph (nngraph), where decoys are vertices, and edges connect decoys to their nearest-neighbors. We note that this graph is not explicitly constructed. Instead, SNMF-DS constructs an adjacency matrix A and a degree matrix D. The entries of A indicate whether pairs of vertices/decoys are adjacent or not in the nngraph. Since the nngraph is a finite simple graph, A is a $(0, 1)$-matrix with $zeros$ on its diagonal (we do not allow a decoy to be considered a nearest neighbor of itself). The degree matrix D is a diagonal matrix that contains the degree of each vertex; $D_{i,j} = deg(i, j)$ if $i = j$ and 0 otherwise. A and D are used together to construct the Laplacian matrix, $L = D - A$ of the nngraph and obtain the optimal number k of groups in which to organize decoys.

D. Organizing Decoys into Groups via Symmetric Non-negative Matrix Factorization

In this step, the proposed SNMF-DS method approximates the similarity matrix S by a lower-rank factorization WW^T. The matrix W is interpreted by SNMF-DS as the cluster membership indicator matrix, which reveals the groups to which decoys belong.

We recall that NMF is an unsupervised method which approximates a given non-negative data-matrix, $X \in \mathbb{R}_+^{n\times d}$ by factoring it into two non-negative factor matrices, $W \in \mathbb{R}_+^{n\times k}$, and $H \in \mathbb{R}_+^{k\times d}$ such that, $X = WH$ [27]; note that k is identified via the EigenGap heuristic. Symmetric NMF is a special case of NMF having completely positive and identical non-negative factor matrices [28]. In symmetric NMF [29], we solve the following equation for the cluster membership indicator matrix $W \in \mathbb{R}_+^{n\times k}$,

$$min_{W\geq 0}\ f(W) = ||S - WW^T||_F^2 \tag{1}$$

where similarity matrix $S \in \mathbb{R}_+^{n\times n}$ (note that S is symmetric, so $S = S^T$), and $||.||_F$ indicates the *Frobenius norm*-based minimization. Typically, $k << n$. To solve the optimization problem in Eq. (1) for W, we apply alternating non-negative least squares (ANLS) optimization (with block principal pivoting) that converges to stationary points [30].

Symmetric NMF is a graphical clustering framework, which exhibits enhanced clustering characteristics on non-linear manifolds of the data as well, than regular NMF or distance based clustering techniques like k-means which are reliable for only linear manifolds that exists in the data [31]. Traditional graph-based clustering like spectral clustering relies on the additional initialization-sensitive methods like k-means on spectral embeddings of the data, generated with eigen vectors of graph Laplacian. Symmetric NMF, in fact, overcomes this problem with non-negative constrained optimization to generate stationary point solutions [31].

We apply non-negative double singular value decomposition (NNDSVD) [32] which is based on approximations of positive sections of the partial SVD factors of the similarity matrix [32], to initialize the factor W for SymmNMF optimization to have better convergence. The largest entry in each row of the W matrix indicates the clustering assignments [31].

E. Determining the Best Group of Decoys

After determining the group composition using the matrix W as described above, we then identify the best group as follows. Each group of decoys is associated a score that is computed as the average over the potential energies of decoys in the group. We recall that our decoy datasets have been computed with the Rosetta AbInitio protocol, which provides all-atom detail and evaluates each decoy with its all-atom *score12* scoring function [33]. Thus, each decoy, even when stripped down to its CA atoms by SNMF-DS, is associated with this score. The groups are then ranked, and the one with the lowest score is selected as the best group.

F. Determining the Best Decoy in a given Group

Once the best group of decoys is determined, the decoys in the group are evaluated so as to determine a best decoy. We make use of the strategy recently proposed in [17], which employs a decoy density score [34]. Specifically, let a decoy x_i belong to a group comprised of l decoys. The density score ds_i of decoy x_i is given by $ds_i = \frac{\sum_{j=1}^{l} r_{ij}}{l}$; where r_{ij} denotes the pairwise root-mean-squared-deviation (RMSD) between decoy x_i and decoy x_j $(1 \leq i, j \leq l)$. The decoy density scores are normalized to be in the range -1 and 1. The normalized density score $ds_i^{'}$ is given by

$$ds_i^{'} = \begin{cases} \frac{(ds_i - ds_{median})}{ds_{median} - ds_{min}} & \text{if } ds_i < ds_{median} \\ 0 & \text{if } ds_i = ds_{median} \\ \frac{(ds_i - ds_{median})}{ds_{max} - ds_{median}} & \text{if } ds_i > ds_{median} \end{cases}$$

where ds_{min}, ds_{max}, and ds_{median} denote the minimum, maximum, and median density scores respectively. Using these normalized scores, we then assign weight w_i to each decoy as in: $w_i = e^{-ds_i^{'}}$. Once the decoys in a group are weighted in this manner, the maximum-weigh decoy is then selected as the best decoy and offered for prediction.

G. Experimental Setup

We compare SNMF-DS with three representative, state-of-the-art methods, (1) our most recent NMF-based method [7] which was shown to outperform the basin-based [16] decoy selection methods (which outperform community-based graph-clustering methods [17], [35]), (2) SBROD, an energy-based method, and (3) MUFOLD-CL, a clustering-based method. The comparative evaluation is carried out on two datasets and via rigorous metrics, as described below.

Dataset: SNMF-DS is evaluated on two datasets. The first, shown in Table I, contains 18 benchmark proteins of different folds and lengths (number of amino acids). The second dataset, shown in Table II, contains 10 targets selected from the free modeling category in CASP12 and CASP13; the list includes several hard targets [36], [37].

As described above, for each protein target, we use the Rosetta AbInitio protocol to generate $12,000$ decoys. Tables I-II provide additional detail for each decoy dataset. For instance, Table I shows the entry id of a known native structure (ground truth) for each target in the Protein Data Bank (PDB) [38]. The fold of the native structure, and the number of amino acids in the corresponding target are shown, as well. The minimum RMSD to the native structure in a decoy dataset is shown in Column 6. This value is utilized to estimate the difficulty of a decoy dataset for decoy selection. Targets where this value does not exceed 1Å are considered easy; those where this value does not exceed 3Å are considered medium; the rest are considered hard. Moreover, Table II lists similar information for the CASP targets. We note that in two cases, marked by asterisks, the native structure has not been deposited yet in the PDB and is only available on the CASP website. The minimum RMSDs shown in Column 5 convey the higher difficulty Rosetta experiences on the CASP targets and generally convey the variability of the quality of decoy datasets over which a decoy selection method has to perform.

TABLE I
BENCHMARKS DATASET (* DENOTES PROTEINS WITH A PREDOMINANT β FOLD AND A SHORT HELIX). THE CHAIN EXTRACTED FROM A MULTI-CHAIN PDB ENTRY IS SHOWN IN PARENTHESES. PDB ID, FOLD, LENGTH, AND MIN RMSD OVER DECOY DATASET TO CORRESPONDING NATIVE STRUCTURE ARE SHOWN FOR EACH TARGET.

Difficulty	#	PDB ID	Fold	Length	RMSD (Å)
Easy	1	1ail	α	70	0.573
	2	1dtd(B)	$\alpha+\beta$	61	0.565
	3	1wap(A)	β	68	0.568
	4	1tig	$\alpha+\beta$	88	0.623
	5	1dtj(A)	$\alpha+\beta$	74	0.701
	6	1hz6(A)	$\alpha+\beta$	64	0.827
Medium	7	1c8c(A)	$\beta*$	64	1.331
	8	2ci2	$\alpha+\beta$	65	1.581
	9	1bq9	β	53	1.308
	10	1hhp	$\beta*$	99	1.761
	11	1fwp	$\alpha+\beta$	69	1.568
	12	1sap	β	66	2.031
	13	2h5n(D)	α	123	2.053
Hard	14	2ezk	α	93	3.475
	15	1aoy	α	78	3.496
	16	1aly	β	146	9.179
	17	1cc5	α	83	4.654
	18	1isu(A)	$coil$	62	5.912

Evaluation Metrics: Since SNMF-DS selects a best group and a best decoy, we evaluate the quality of each.

a) Group Purity: The quality of a group is assessed via a metric we have introduced in earlier work on decoy selection [16]. The metric, known as *purity* is related to the concept of precision in ML, as it counts the fraction of near-native decoys in a given group over the total number of decoys in the group. If we relate near-native decoys to true positives (TP), then purity $p(G_i)$ of a group G_i measures $\frac{TP}{TP+FP}$, where FP, false positives, corresponds to non-native decoys. We delay details on how a decoy is determined to be near-native in the presence of a known native structures (the ground truth) later in this section in the interest of clarity.

TABLE II
CASP DATASET. CASP TARGET IDS ARE SHOWN IN COLUMN 2. PDB ID, LENGTH, AND MIN RMSD OVER DECOY DATASET TO CORRESPONDING NATIVE STRUCTURE ARE SHOWN FOR EACH TARGET. NATIVE STRUCTURES ONLY AVAILABLE IN THE CASP WEBSITE ARE MARKED BY ASTERISKS.

#	Target ID	PDB ID	Length	RMSD (Å)
1	T1008-D1	6msp	77	1.542
2	T0886-D1	5fhy	69	5.102
3	T0953s1-D1	6f45	67	6.344
4	T0960-D2	6cl5	84	6.402
5	T0898-D2	**	55	6.598
6	T0892-D2	5nv4	110	6.950
7	T0953s2-D3	6f45	77	7.607
8	T0957s1-D1	6cp8	108	7.677
9	T0897-D1	**	138	9.638
10	T0859-D1	5jzr	113	10.268

b) Decoy Loss: The quality of a decoy is assessed via the loss metric we have introduced in [7]. While work in [7] utilizes only RMSD loss, here we additionally make use of TM-Score loss and GDT-TS loss. TM-Score [39] and GDT-TS [40] vary in $[0, 1]$, capture the similarity between two tertiary structures, and are popular in CASP. We measure loss as the difference in quality between the decoy selected by a decoy selection method and the best-quality decoy in a dataset, with quality assessed by any of the metrics. For instance, when RMSD is used, loss is measured as the difference between the selected decoy and the decoy with the lowest RMSD to the known native structure in a given dataset. When TM-Score or GDT-TS are used, the best decoy in a dataset is the one with the highest TM-Score (alternatively, highest GDT-TS score).

IV. RESULTS

We present two sets of results, comparison with state-of-the-art methods in terms group purity, and analysis of decoy loss with the help of three popular measures.

A. Purity Comparison

We now compare the purity of the group/cluster selected by SNMF-DS, NMF-MAD, and MUFOLD-CL. We note that SBROD ranks decoys by energies and so does not organize them into groups. Table III compares purities over the benchmark targets, whereas Table IV does so over the CASP targets.

Tables III-IV show that SNMF-DS and NMF-MAD largely outperform MUFOLD-CL. Specifically, for the easy benchmark targets, the purity values obtained by MUFOLD-CL range from 17% to 62%, whereas NMF-MAD attains 78% to 100% purity, and SNMF-DS dominates with 100% purity in each target. For the medium benchmark targets, SNMF-DS achieves better purity than NMF-MAD on $4/7$ cases; MUFOLD-CL is inferior to SNMF-DS on all the medium benchmark targets (and with only two marginal wins over

TABLE III
THE PURITY(%) OF THE GROUP/CLUSTER SELECTED BY SNMF-DS, NMF-MAD, AND MUFOLD-CL FOR THE BENCHMARK TARGETS

Difficulty	#	PDB ID	SNMF-DS	NMF-MAD	MUFOLD-CL
Easy	1	1ail	99.75	94.6	17.15
	2	1dtd(B)	100	100	58.34
	3	1wap(A)	100	99.96	62.2
	4	1tig	100	92.31	38.6
	5	1dtj(A)	100	100	51.35
	6	1hz6(A)	100	78.27	22.72
Medium	7	1c8c(A)	70.1	85.4	20.8
	8	2ci2	85.14	100	62.14
	9	1bq9	44.44	86.8	10.05
	10	1hhp	88.54	50.22	0
	11	1fwp	64.95	21.05	14.96
	12	1sap	19.56	0	1.1
	13	2h5n(D)	17.54	2.55	5.1
Hard	14	2ezk	51	11.61	14.37
	15	1aoy	43.15	81.81	12.24
	16	1aly	3.35	0	3.05
	17	1cc5	28.75	66.66	19.55
	18	1isu(A)	5.15	71	1.1

TABLE IV
THE PURITY(%) OF THE GROUP/CLUSTER SELECTED BY SNMF-DS, NMF-MAD, AND MUFOLD-CL FOR THE CASP TARGETS

#	Target ID	SNMF-DS	NMF-MAD	MUFOLD-CL
1	T1008-D1	21.43	28.12	0
2	T0886-D1	17.7	33.33	5.15
3	T0953s1-D1	21.86	25.93	6.12
4	T0960-D2	30.7	18.18	8.99
5	T0898-D2	49.2	46.67	16.37
6	T0892-D2	14.3	50	3.65
7	T0953s2-D3	4.32	44.44	5.29
8	T0957s1-D1	23.88	21.95	13.15
9	T0897-D1	16.94	20	4.54
10	T0859-D1	3.14	13.64	3.2

NMF-MAD). On the hard benchmark targets, NMF-MAD does particularly well, reaching purity value from 11% to 81% (except for 1aly); SNMF-DS purities over these targets range from 3% to 51%. MUFOLD-CL does not perform better than SNMF-DS on any target; it only beats NMF-MAD on one target (2ezk). These observations are further confirmed over the CASP dataset. On the 10 CASP targets, MUFOLD-CL reaches purities ranging from 3% to 16% (on T1008-D1, purity is 0%) and is inferior to both NMF-MAD and SNMF-DS; it performs as well or slighly better than SNMF-DS on only $2/10$ targets. Specifically, in $7/10$ targets, NMF-MAD outperforms SNMF-DS with purities ranging from 13% to 50%; in the remaining 3 targets, SNMF-DS performs better than NMF-MAD with purities ranging from 3% to 49%. Altogether, these results demonstrate that SNMF-DS is as competitive as NMF-MAD in terms of the quality of the selected group.

B. Loss Comparison

We compare SNMF-DS, NMF-MAD, MUFOLD-CL, and SBROD in terms of RMSD loss, TM-Score loss, and GDT-TS loss of the selected decoy. This comparison is in Table V for the benchmark targets and in Table VI for the CASP targets.

Tables V-VI make clear the superiority of SNMF-DS over the other methods. For instance, Table V shows that the RMSD loss incurred by SNMF-DS is below $1\mathring{A}$ for $11/18$ of the benchmark targets. Table V also shows that for $14/18$ of these targets, the best decoy selected by SNMF-DS incurs the minimum loss compared to the other methods in terms of at least one of the three quantities (RMSD loss, TM-Score loss, and GDT-TS loss). Table VI shows that the RMSD loss incurred by SNMF-DS is below $2\mathring{A}$ for $6/10$ of the CASP targets. For $7/10$ CASP targets, the best decoy selected by SNMF-DS incurs the minimum loss compared to the other methods in terms of at least one of the three measures (RMSD loss, TM-Score loss, and GDT-TS loss).

TABLE V
SNMF-DS, MUFOLD-CL, SBROD, AND NMF-MAD ARE COMPARED IN TERMS OF RMSD, TM-SCORE, AND GDT-TS LOSS ON THE BENCHMARK TARGETS. LOWEST LOSS PER PDB ID IN ANY METRIC (RMSD, TM-SCORE, OR GDT-TS) IS HIGHLIGHTED IN BOLD.

PDB ID	RMSD Loss, TM-Score Loss, GDT-TS Loss			
	SNMF-DS	MUFOLD-CL	SBROD	NMF-MAD
1ail	**0.5084**, **0.0655**, **0.072**	1.447, 0.1676, 0.1336	2.937, 0.314, 0.3478	0.971, 0.1604, 0.1357
1dtj(A)	0.1941, **0.0048**, 0.0296	**0.036**, 0.0198, **0.0066**	0.69, 0.006, 0.0329	0.3345, 0.0782, 0.1081
1dtd(B)	0.3528, **0.0042**, **0.0041**	0.49, 0.0052, 0.0043	**0.12**, 0.005, 0.0082	0.5915, 0.0329, 0.0451
1wap(A)	0.3425, 0.0288, **0.0166**	**0.263**, **0.0242**, 0.0233	1.242, 0.1107, 0.1	0.6219, 0.0531, 0.04
1tig	**0.0717**, **0.003**, **0.0053**	0.749, 0.004, 0.008	0.709, 0.0134, 0.016	0.6569, 0.0469, 0.0483
1hz6(A)	**0.0936**, **0.002**, **0.0034**	0.405, 0.0037, 0.0036	0.191, 0.0145, 0.0382	0.809, 0.0415, 0.0352
1bq9	**1.1992**, 0.1677, 0.1389	2.02, 0.2115, 0.1759	1.337, 0.1331, 0.1065	1.3089, **0.1167**, **0.0755**
1c8c(A)	**0.7991**, 0.1092, 0.086	1.012, 0.135, 0.1016	1.531, 0.1465, 0.1328	1.092, **0.0596**, **0.0429**
1fwp	**0.5085**, **0.0034**, 0.0036	0.724, 0.0074, **0.0018**	1.039, 0.0589, 0.1332	0.5319, 0.0471, 0.0616
1hhp	**2.1971**, 0.0601, 0.0707	10.919, 0.6326, 0.6161	2.76, **0.0533**, **0.0606**	2.6835, 0.2939, 0.2828
1sap	**0.5592**, **0.074**, **0.0417**	1.61, 0.0831, 0.0492	1.873, 0.141, 0.1136	2.075, 0.0989, 0.125
2ci2	**0.3118**, **0.007**, **0.006**	3.202, 0.1155, 0.1114	3.083, 0.1334, 0.1175	1.7897, 0.3246, 0.3462
2h5n(D)	3.7028, 0.3178, 0.3215	7.806, 0.2479, 0.2576	3.883, 0.0856, 0.094	**3.3498**, **0.0805**, **0.0732**
1aoy	2.7896, **0.1136**, **0.093**	5.246, 0.1856, 0.1635	**2.047**, 0.1286, 0.1218	2.9788, 0.2918, 0.2788
1aly	5.7842, 0.0167, 0.0368	**3.467**, **0.0155**, **0.024**	4.373, 0.029, 0.0325	7.9939, 0.1411, 0.1635
1cc5	**0.4732**, **0.048**, 0.0452	1.159, 0.0831, **0.0392**	1.949, 0.0501, 0.0551	2.1843, 0.0565, 0.0573
1isu(A)	2.9928, 0.2182, 0.2299	6.357, 0.2106, 0.242	5.32, 0.1603, 0.2137	**2.5552**, **0.081**, **0.0887**
2ezk	2.9154, 0.0188, 0.0177	**1.172**, **0.003**, **0.0076**	3.142, 0.0178, 0.0244	3.5136, 0.0229, 0.0296

TABLE VI
SNMF-DS, MUFOLD-CL, SBROD, AND NMF-MAD ARE COMPARED IN TERMS OF RMSD, TM-SCORE, AND GDT-TS LOSS ON THE CASP TARGETS. LOWEST LOSS PER TARGET ID IN ANY METRIC (RMSD, TM-SCORE, OR GDT-TS) IS HIGHLIGHTED IN BOLD.

Target ID	RMSD Loss, TM-Score Loss, GDT-TS Loss			
	SNMF-DS	MUFOLD-CL	SBROD	NMF-MAD
T1008-D1	**0.3656, 0.007, 0.0011**	3.305, 0.0137, 0.065	0.398, 0.0086, 0.0032	1.0238, 0.0156, 0.0162
T0886-D1	3.6714, **0.03**, 0.0362	4.94, 0.0403, 0.0435	**2.12**, 0.034, 0.0326	2.5984, 0.0331, **0.029**
T0953s1-D1	2.9398, **0.02**, 0.0112	2.947, 0.055, 0.0187	3.032, 0.084, **0.0037**	**2.613**, 0.0225, 0.0223
T0960-D2	1.8595, 0.0307, 0.0268	**0.53**, 0.0384, 0.0328	0.67, 0.0505, 0.0417	2.6181, **0.0182**, **0.0178**
T0898-D2	1.4889, 0.003, **0.0071**	0.468, 0.008, 0.0091	**0.162**, **0.001**, 0.0137	2.3824, 0.0108, 0.0181
T0892-D2	**0.9038**, **0.0119**, **0.004**	1.787, 0.0129, 0.0069	1.51, 0.0134, 0.0114	2.8416, 0.0242, 0.009
T0953s2-D3	1.4223, **0.01**, 0.011	2.137, 0.0187, 0.0162	**0.326**, 0.0109, **0.0033**	1.8621, 0.0256, 0.0153
T0897-D1	3.471, 0.0263, 0.0108	1.137, 0.0064, 0.018	**0.236**, **0.0032**, **0.0055**	2.9413, 0.0158, 0.009
T0957s1-D1	1.18, **0.0027**, 0.0047	0.709, 0.008, 0.0023	**0.423**, 0.0079, **0.001**	1.6803, 0.018, 0.0076
T0859-D1	2.3755, 0.056, 0.045	**0.421**, 0.0094, **0.0023**	0.518, **0.0088**, 0.0044	3.5967, 0.0329, 0.0132

V. CONCLUSION

The evaluation presented in Section IV shows that the proposed SNMF-DS method is a powerful method for decoy selection, outperforming state-of-the-art methods. These results are very encouraging, as exploiting non-negative matrix factorization is a relatively new thread of research for decoy selection. Several directions of future work are warranted. One can pursue different metrics, such as TM-Score and GDT-TS in the construction of the similarity matrix. The computation of the adjacency and degree matrices can be further expedited by utilizing proximity query data structures, such as C-trees. Alternative techniques can be considered to automatically determine the optimal number of decoy groups. Finding target-wise sub-spaces of features representative of a decoy dataset could additionally prove informative.

ACKNOWLEDGMENT

This work is supported in part by NSF FET Grant No. 1900061 to AS. Resources utilized for this research were partly provided by the Los Alamos National Laboratory (LANL) Institutional Computing Program, which is supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. DE-AC52-06NA25396 and LANL laboratory directed research and development (LDRD) grant 20190020DR. High-performance computations were run on Darwin, a heterogeneous cluster for research computing at LANL, and on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University. This material is additionally based upon work by AS 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. Finally, we thank our lab. alumnus Dr. Nasrin Akhter for useful discussions on this work.

References

[1] D. D. Boehr and P. E. Wright, "How do proteins interact?" *Science*, vol. 320, no. 5882, pp. 1429–1430, 2008.

[2] N. Perdigao *et al.*, "Unexpected features of the dark proteome," *Proc Natl Acad Sci USA*, vol. 112, no. 52, pp. 15 898–1590, 2015.

[3] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu, "Principles and overview of sampling methods for modeling macromolecular structure and dynamics," *PLoS Comp. Biol.*, vol. 12, no. 4, p. e1004619, 2016.

[4] J. Lee, P. Freddolino, and Y. Zhang, "Ab initio protein structure prediction," in *From Protein Structure to Function with Bioinformatics*, 2nd ed., D. J. Rigden, Ed. Springer London, 2017, ch. 1, pp. 3–35.

[5] R. Das, "Four small puzzles that rosetta doesn't solve," *PLoS One*, vol. 6, no. 5, p. e20044, 2011.

[6] K. Molloy, S. Saleh, and A. Shehu, "Probabilistic search and energy guidance for biased decoy sampling in ab-initio protein structure prediction," *IEEE/ACM Trans Comput Biol and Bioinf*, vol. 10, no. 5, pp. 1162–1175, 2013.

[7] N. Akhter, R. Vangara, G. Chennupati, B. Alexandrov, H. Djidjev, and A. Shehu, "Non-negative matrix factorization for selection of near-native protein tertiary structures," in *Intl Conf on Bioinf and Biomed (BIBM)*. IEEE, 2019, pp. 70–73.

[8] Y. N. Vorobjev and J. Hermans, "Free energies of protein decoys provide insight into determinants of protein stability," *Protein Science*, vol. 10, no. 12, pp. 2498–2506, 2001.

[9] S. Lorenzen and Y. Zhang, "Identification of near-native structures by clustering protein docking conformations," *PROTEINS: Structure, Function, and Bioinformatics*, vol. 68, no. 1, pp. 187–194, 2007.

[10] Y. Zhang and J. Skolnick, "Spicker: a clustering approach to identify near-native protein folds," *Journal of computational chemistry*, vol. 25, no. 6, pp. 865–871, 2004.

[11] J. Zhang and D. Xu, "Fast algorithm for population-based protein structural model analysis," *Proteomics*, vol. 13, no. 2, pp. 221–229, 2013.

[12] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (casp)—round x," *Proteins: Structure, Function, and Bioinformatics*, vol. 82, pp. 1–6, 2014.

[13] M. Karasikov, G. Pagès, and S. Grudinin, "Smooth orientation-dependent scoring function for coarse-grained protein quality assessment," *Bioinformatics*, vol. 35, no. 16, pp. 2801–2808, 2019.

[14] X. Jing, K. Wang, R. Lu, and Q. Dong, "Sorting protein decoys by machine-learning-to-rank," *Scientific reports*, vol. 6, p. 31571, 2016.

[15] N. Akhter, J. Lei, W. Qiao, and A. Shehu, "Reconstructing and decomposing protein energy landscapes to organize structure spaces and reveal biologically-active states," in *IEEE Intl Conf on Bioinf and Biomed (BIBM)*. Madrid, Spain: IEEE, 2018, pp. 56–60.

[16] N. Akhter and A. Shehu, "From extraction of local structures of protein energy landscapes to improved decoy selection in template-free protein structure prediction," *Molecules*, vol. 23, no. 1, p. 216, 2018.

[17] N. Akhter, G. Chennupati, K. Kabir, H. Djidjev, and A. Shehu, "Unsupervised and supervised learning over the energy landscape for protein decoy selection," *Biomolecules*, vol. 9, no. 1, p. 607, 2019.

[18] K. Devarajan, "Nonnegative matrix factorization: an analytical and interpretive tool in computational biology," *PLoS Comput Biol*, vol. 4, no. 7, p. e1000029, 2008.

[19] M. Stražar, M. Žitnik, B. Zupan, J. Ule, and T. Curk, "Orthogonal matrix factorization enables integrative analysis of multiple rna binding proteins," *Bioinformatics*, vol. 32, no. 10, pp. 1527–1535, 2016.

[20] W. Kim, B. Chen, J. Kim, Y. Pan, and H. Park, "Sparse nonnegative matrix factorization for protein sequence motif discovery," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13 198–13 207, 2011.

[21] G. Chennupati, R. Vangara, E. Skau, H. Djidjev, and B. S. Alexandrov, "Distributed non-negative matrix factorization with determination of the number of latent features," *Journal of Supercomputing*, vol. 76, no. 9, pp. 7458–7488, 2020.

[22] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak *et al.*, "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules," *Methods Enzymol*, vol. 487, pp. 545–574, 2011.

[23] A. D. McLachlan, "A mathematical procedure for superimposing atomic coordinates of proteins," *Acta Crystallogr. A.*, vol. 26, no. 6, pp. 656–657, 1972.

[24] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in neural information processing systems*, 2005, pp. 1601–1608.

[25] K. Pelechrinis, *Spectral clustering with eigengap heuristic: A MATLAB implementation*, 2013 (Last Accessed: August 30, 2020). [Online]. Available: http://kokkodis.blogspot.com/2013/02/spectral-clustering-with-eigengap.html

[26] Y.-H. Kung, P.-S. Lin, and C.-H. Kao, "An optimal k-nearest neighbor for density estimation," *Statistics & Probability Letters*, vol. 82, no. 10, pp. 1786–1791, 2012.

[27] K. Huang, N. D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 211–224, 2013.

[28] Z. He, S. Xie, R. Zdunek, G. Zhou, and A. Cichocki, "Symmetric nonnegative matrix factorization: Algorithms and applications to probabilistic clustering," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2117–2131, 2011.

[29] D. Kuang, S. Yun, and H. Park, "Symnmf: nonnegative low-rank approximation of a similarity matrix for graph clustering," *Journal of Global Optimization*, vol. 62, no. 3, pp. 545–574, 2015.

[30] J. Kim and H. Park, "Fast nonnegative matrix factorization: An active-set-like method and comparisons," *SIAM Journal on Scientific Computing*, vol. 33, no. 6, pp. 3261–3281, 2011.

[31] D. Kuang, C. Ding, and H. Park, "Symmetric nonnegative matrix factorization for graph clustering," in *Proceedings of the 2012 SIAM international conference on data mining*. SIAM, 2012, pp. 106–117.

[32] C. Boutsidis and E. Gallopoulos, "Svd based initialization: A head start for nonnegative matrix factorization," *Pattern recognition*, vol. 41, no. 4, pp. 1350–1362, 2008.

[33] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel *et al.*, "The rosetta all-atom energy function for macromolecular modeling and design," *Journal of chemical theory and computation*, vol. 13, no. 6, pp. 3031–3048, 2017.

[34] K. Wang, B. Fain, M. Levitt, and R. Samudrala, "Improved protein structure selection using decoy-dependent discriminatory functions," *BMC structural biology*, vol. 4, no. 1, pp. 1–18, 2004.

[35] K. L. Kabir, L. Hassan, Z. Rajabi, N. Akhter, and A. Shehu, "Graph-based community detection for decoy selection in template-free protein structure prediction," *Molecules*, vol. 24, no. 5, p. 854, 2019.

[36] L. A. Abriata, G. E. Tamò, B. Monastyrskyy, A. Kryshtafovych, and M. Dal Peraro, "Assessment of hard target modeling in casp12 reveals an emerging role of alignment-based contact prediction methods," *Proteins: Structure, Function, and Bioinformatics*, vol. 86, pp. 97–112, 2018.

[37] J. Cheng, M.-H. Choe, A. Elofsson, K.-S. Han, J. Hou, A. H. Maghrabi, L. J. McGuffin, D. Menéndez-Hurtado, K. Olechnovič, T. Schwede *et al.*, "Estimation of model accuracy in casp13," *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 12, pp. 1361–1377, 2019.

[38] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000. [Online]. Available: https://www.rcsb.org/

[39] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins: Structure, Function, and Bioinformatics*, vol. 57, no. 4, pp. 702–710, 2004.

[40] A. Zemla, "Lga: a method for finding 3d similarities in protein structures," *Nucleic acids research*, vol. 31, no. 13, pp. 3370–3374, 2003.