

# Single Model Quality Estimation of Protein Structures via Non-negative Tensor Factorization

Kazi Lutful Kabir<sup>1</sup>, Manish Bhattarai<sup>2</sup>, Boian S. Alexandrov<sup>3</sup>, and Amarda Shehu<sup>1</sup>

<sup>1</sup> Department of Computer Science, George Mason University,  
Fairfax VA 22030, USA  
{kkabir, ashehu}@gmu.edu

<sup>2</sup> Physics and Chemistry of Materials (T-1), Los Alamos National Laboratory,  
Los Alamos NM 87545, USA

<sup>3</sup> Fluid Dynamics and Solid Mechanics (T-3), Los Alamos National Laboratory,  
Los Alamos NM 87545, USA  
{boian, ceodsp spectrum}@lanl.gov

**Abstract.** Finding the inherent organization in the structure space of a protein molecule is central in many computational studies of proteins. Grouping or clustering tertiary structures of a protein has been leveraged to build representations of the structure-energy landscape, highlight stable and semi-stable structural states, support models of structural dynamics, and connect them to biological function. Over the years, our laboratory has introduced methods to reveal structural states and build models of state-to-state protein dynamics. These methods have also been shown competitive for an orthogonal problem known as model selection, where model refers to a computed tertiary structure. Building on this work, in this paper we present a novel, tensor factorization-based method that doubles as a non-parametric clustering method. While the method has broad applicability, here we focus and demonstrate its efficacy on the estimation of model accuracy (EMA) problem. The method outperforms state-of-the-art methods, including single-model methods that leverage deep neural networks and domain-specific insight.

**Keywords:** Protein Tertiary Structure · Single Model Quality Estimation · Tensor Factorization.

## 1 Introduction

The tertiary structure in which the amino acids constituting a protein molecule position themselves in three dimensions determines to a great extent the activities of a protein in the cell [9]. That is why the latest achievement of AlphaFold2 [15], which has already been employed to position about 58% of the total amino acids in human protein sequences [32], has been heralded as a great advance to obtain a mechanistic understanding of protein function. This achievement nonetheless ignores protein structure plasticity [23].

Methods that expand our view beyond one structure and account for the intrinsic ability of proteins to populate different functionally-relevant structures now abound [23]. Some methods sample the protein structure-energy landscape. Others use physics-based simulation to additionally reveal transitions between structures. Grouping/clustering computed structures of a protein has been used to build informative representations of the structure-energy landscape and expose the stable and semi-stable structural states that support the various activities of a protein in the cell. Work in [22] employs level sets to expose such states, whereas work in [16] leverages graph clustering and identifies macrostates by detecting communities in a graph embedding computed structures.

Over the years, we have proposed methods for grouping structures and supporting models of landscape-governed dynamics [16, 18]. These methods have also been competitive in an orthogonal application setting, where the goal is to evaluate the quality of tertiary structures computed by one or more methods for a given protein and select a *best* structure [3, 17]. This problem is also known as estimation of model accuracy (EMA) (model refers to a computed structure) or quality assessment (QA) and motivates much computational research [11].

In this paper we present a novel, tensor factorization-based method that organizes tertiary structures of a protein into groups. The method doubles as a non-parametric clustering method and so can broadly support various application settings. Here we focus and demonstrate its efficacy on EMA. As we expand further in Section 2, the method falls in the category of multi-model methods, as it extracts information from multiple structures/models. The method additionally computes an individual score for each structure that serves as a proxy of structure quality/accuracy. We show that the proposed method outperforms many state-of-the-art (SOTA) methods, including single-model methods. We first proceed with a review of related works.

## 2 Related Works

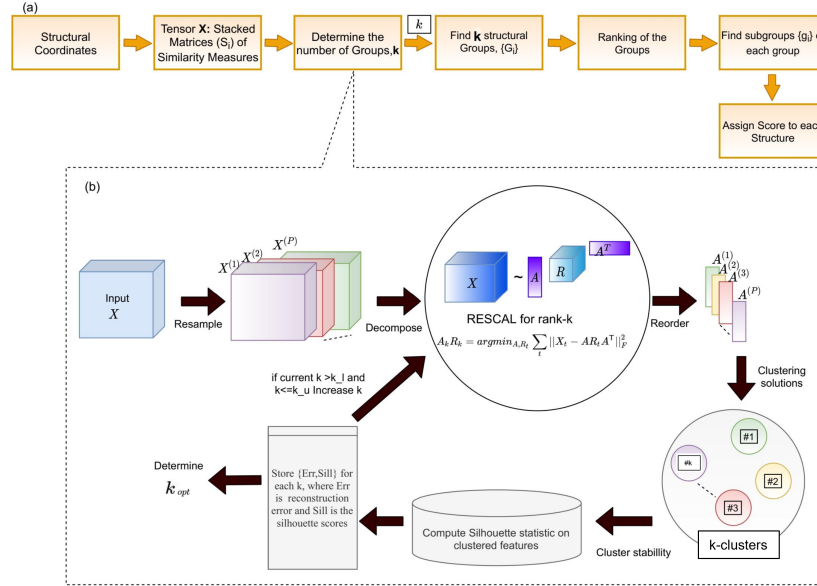
Early EMA methods were single-model. They arose before the term *model* in machine learning (ML) was popularized; model referred to a computed structure. Single-model methods provide one score per structure and first utilized molecular energy functions. These were not good proxies, though adding statistical terms offered some improvement [13]. Multi-model methods ignored scoring and clustered structures by similarity [37]. A combination of strategies were used to select a top cluster and then a best structure to offer as prediction. The landscape of multi-model methods is rich and now includes methods that consider structure energies to improve clustering [2]. Multi-model methods were superior for some time [26], until growth in structure databases facilitated single-model ML methods. Shallow ML methods focused on predicting a score for a given structure [25]. Then deep neural networks predicted increasingly accurate scores and are now SOTA for single-model methods in EMA [10].

We have investigated non-negative matrix factorization (NMF) to group structures for EMA. The NMF-MAD method in [3] decomposes a matrix of energy-based features of structures and outperforms multi-model methods. Work

in [17] proposes SNMF-DS, which removes the reliance on features and factorizes a structure similarity matrix, utilizes symmetric NMF, and employs the eigen-gap statistic to automatically determine the number of groups. SNMF-DS outperforms NMF-MAD, MUFOLD-CL [37], and single-model methods, such as SBROD [19]. Matrix-factorization methods have been effective due to two reasons. First, they can handle sparse and highly imbalanced datasets, unlike most clustering methods. Second, the quality of computed tertiary structures has increasingly improved. In this paper we propose a factorization-based method but expand from matrix to tensor factorization. This proves more powerful, as we demonstrate its performance in Section 4. We now describe the method in greater detail.

### 3 Methodology

We refer to the proposed method as NTF-REL (non-negative tensor factorization with RESCAL) from now on. NTF-REL proceeds in four stages. Stage I organizes given structures into groups  $\{G_i\}$  via tensor factorization. Stage II utilizes energies to rank the groups. Stage III partitions each group into subgroups and ranks them. Stage IV utilizes all this information to compute a score for each structure. A schematic is related in Fig. 1(a).



**Fig. 1.** (a) Schematic of the proposed method. (b) Finding the number of latent features with non-negative RESCAL factorization.

#### 3.1 Stage I: From Structures to Groups

Different metrics for comparing two structures capture different aspects and often provide complementary information [28]. We form a tensor  $X$  by stacking

(symmetric) similarity matrices  $S_{n,n}^i$  obtained on  $n$  structures, where  $i$  refers to a particular metric. Entry  $(a, b)$  in  $S^i$  measures the similarity according to metric  $i$  between two structures at positions  $a, b$  in a list of  $n$  structures and  $S_{a,a}^i = 1$ . We use 5 popular metrics, RMSD, TM-score, GDT-TS, GDT-HA, and MaxSub score [28, 30]. Since RMSD is a dissimilarity measure, we turn it into a similarity one as in  $S_{a,b} = \frac{1}{\text{RMSD}(a,b)}$ . Each  $S^i$  is a slice of the tensor. We use PyRosetta to obtain structures for a given protein, due to its popularity and ease of use; AlphaFold2 provides very few structures (about 45) in our experimentation with it. Each computed structure is stripped down to its main-chain carbon atoms (the CA atoms). This reduction brings down the cost of computing the tensor.

As Fig. 1(b) shows, the tensor  $X$  is then decomposed. We use the RESCAL tensor factorization approach [27] integrated with an automatic latent dimension determination method [31]. RESCAL was developed for extracting latent communities in relational data. Specifically, RESCAL factorizes tensors formed by a set of  $m$  stacked matrices of graphs (each graph has  $n$  nodes),  $X^{n \times n \times m}$  into a factor matrix  $A^{n \times k}$  and a core tensor  $R^{k \times k \times m}$ , where  $k$  is the latent dimension (or number of the latent communities/groups). The factorization solves the following optimization problem:

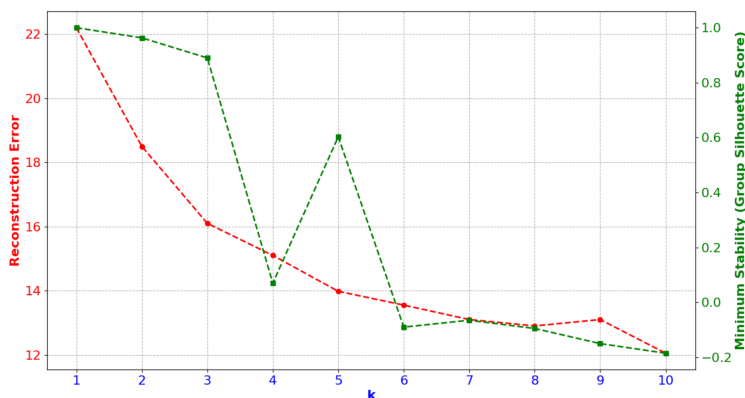
$$\operatorname{argmin}_{A,R} \|X - R \times_1 A \times_2 A\|_F^2 \quad (1)$$

where  $\times_i$  denotes the mode- $i$  product [20]. The extracted factors are interpretable; each column of  $A$  represents a latent community/group of objects, and each slice  $R_m$  of the core tensor  $R$  captures the relations among the groups at instance  $m$ . Considering the non-negativity of the data, we employ non-negative RESCAL [21]. The optimization with non-negativity constraints is given by,

$$\begin{aligned} & \operatorname{argmin}_{A,R_m} \sum_m \|X_m - AR_m A^\top\|_F^2 \\ & \text{subject to} \quad \sum_j A_{ij} = 1, \text{ for } 1 \leq j \leq k; A, R \geq 0 \end{aligned}$$

Fig. 1(b) shows our adaptation of RESCAL integrated with an algorithm to find the  $k$  latent groups, to which we refer as RESCAL- $k$  [31]. RESCAL- $k$  consists of the following components: (1) Custom Resampling: We generate an ensemble of  $X$  tensors,  $[X^{(q)}]_{q=1,\dots,P}$ , where the means of these matrices equal to the original tensor  $X$ . Each of these tensors  $X^{(q)}$  is built by perturbing each of the elements using random uniform noise, such that  $X^{(q)} = X(\odot)\Delta_q$  (for details see [6]). (2) RESCAL Minimization: We use Frobenius norm-based multiplicative updates [27] to explore various numbers of latent features;  $k$  in an interval  $[k_{min}, k_{max}]$ , for each of the  $P$  generated random tensors  $X^{(q)}$ . The decomposed component  $A$  corresponds to the samples in reduced latent dimension  $n \times k$  denoting the groups, whereas  $R$  is the  $k \times k \times m$  relational tensor representing the group interactions. (3) Custom Clustering: For each  $k \in [k_{min}, k_{max}]$ , we cluster the set of the  $n \times k$  latent components. To extract the latent dimension, we determine the dependency of the stability of the obtained clusters and the improvement of the reconstruction error on the latent dimension  $k$ . The final latent components,  $\hat{A}$ , are the medoids of the obtained stable clusters, with  $\hat{R}$  denoting the corresponding mixing coefficients. The latent group estimation pipeline is based on the pyDRESCALk toolbox [7, 8].

RESCAL- $k$  employs Silhouette statistics [29] to determine the cluster stability for each  $k$ . The Silhouette parameter that quantifies the cluster stability is in the range  $[-1, 1]$ , where  $-1$  corresponds to a bad clustering and  $1$  to perfect clustering. Fig. 2 shows how using both Silhouette and reconstruction norm, we can determine the optimal  $k$ . The final representative  $A$  corresponding to the RESCAL factorization of  $X$  for  $k_{opt}$  estimated by RESCAL- $k$  is then used to identify the best composition of the  $k$  groups identified, as illustrated in Fig. 2.



**Fig. 2.** We illustrate here the stability analysis for one of our protein targets, CASP target T1008-D1. Candidate values of  $k$  contain considerably larger gaps between the relative reconstruction error and the silhouette statistics with silhouette score  $\geq 0.6$ . Out of the candidates (2, 3, and 5), we choose the one with the lowest reconstruction error (i.e.,  $k_{opt} = 5$  in this example).

### 3.2 Stage II: Ranking Groups

After determining the group composition using the matrix  $A$  identified as described above, we then rank the groups. Each group (of structures) is associated the average value over the Rosetta all-atom (score12) energies of the structures in the group. The groups are then ranked in ascending order of the group energy score; the group with the lowest score is the best-ranked group. The rank of a group  $G$  is denoted by  $R_G$ .

### 3.3 Stage III: Partitioning Groups into Subgroups

We hybridize the tensor-based approach above with graph clustering. We utilize work in [2] which embeds structure-energy pairs in a nearest-neighbor graph (using RMSD to identify nearest neighbors), over which it identifies local energy minima representing different energy basins and groups vertices into basins. In the interest of space, we spare the methodological details of how the minima and groupings are identified and refer the interested reader to work in [2]. Our adaptation here is not to apply this approach over all  $n$  structures, but instead

to apply it to each group identified via tensor factorization in order to partition each group into “basins”; to which we refer more generally as subgroups.

### 3.4 Stage IV: Scoring each Structure

To score each structure, we modify the strategy proposed in [1], which employs a model density score [36]. Let a structure  $x_i$  belong to a group  $G$  comprised of  $l$  structures. As in [1], we associate a density score  $ds_i$  with  $x_i$  as:  $ds_i = \frac{\sum_{j=1}^l \text{TM-Score}_{i,j}}{l}$ , where  $\text{TM-Score}_{i,j}$  is the TM-Score between  $x_i$  and  $x_j$  ( $1 \leq i, j \leq l$ ). In principle, different metrics can be measured; unlike work in [1], which uses RMSD, we use TM-score; its  $[0, 1]$  range easily translates via the above formula into density scores in the range  $[0, 1]$ .

Our density score additionally utilizes information from Stages II-III. Let the number of subgroups in group  $G$  be  $z$ . The subgroups are first sorted and ranked in descending order of size. Let the rank of each subgroup  $g \in G$  be  $r_g$ . The last  $d$  ( $d = \lceil \frac{z}{3} \rceil$ ) subgroups are further sorted (in ascending order) by the average potential energy of the structures in a subgroup, resulting in  $r_g'$ .

The modified structure density score  $ds_i'$  is then:

$$ds_i' = \begin{cases} \frac{ds_i}{\max(R_G) + r_g'} & \text{if } x_i \in \text{last } d \text{ subgroups} \\ \frac{ds_i}{R_G + r_g} & \text{otherwise} \end{cases}$$

Using these modified scores, we then assign weight/score  $w_i$  to each structure as in:  $w_i = e^{ds_i'}$ . Once the structures are weighted in this manner, the highest-weight structure is considered as the *best* structure.

### 3.5 Experimental Setup

We compare NTF-REL to representative SOTA methods: (1) Single-structure methods ProQ2 [35], ProQ3 [34], ProQ3D [33], and ProQ4 [24], and (2) recent NMF-based methods [3, 4, 17]. These NMF-based methods were shown to outperform MUFOLD-CL and other multi-model methods.

### 3.6 Dataset

As in [1, 17], methods are evaluated on two datasets. The first, shown in Table 1 (left panel), contains 18 benchmark proteins of different folds and lengths (number of amino acids). The second dataset, shown in Table 1 (right panel), contains 10 targets selected from the free modeling category in CASP12 and CASP13; the list includes several hard targets. The Rosetta AbInitio protocol is used to generate 12,000 all-atom structures for each protein target. We note here that the goal is not to just generate one or a few structures, as now possible through AlphaFold2, but to obtain a broader view of the structure space containing alternative structures.

Tables 1 provide additional details for each dataset. Table 1 (left panel) shows the entry id of an experimentally known structure (ground truth) for each target in the Protein Data Bank (PDB) [5], the fold of the known structure, and the

number of amino acids in the corresponding target. The minimum RMSD to the known structure in a dataset is shown in Column 6 and used to estimate the difficulty of a dataset for EMA. 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. Table 1 (right panel) lists similar information for the CASP targets. We note that in two cases, marked by asterisks, the known structure is only available on the CASP website.

**Table 1. Left Panel:** Benchmarks dataset (\* denotes proteins with a predominant  $\beta$  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 a dataset to corresponding experimental structure are shown for each target. **Right Panel:** CASP dataset. CASP target IDs are shown in Column 2. PDB ID, Length, and Min RMSD over dataset to corresponding experimental structure are shown for each target. CASP targets with no experimentally-available structure in the PDB but only in the CASP website are marked by asterisks.

Difficulty	#	PDB ID	Fold	Length	Min (Å)	RMSD
Easy	1	1ail	$\alpha$	70	0.573	
	2	1dtd(B)	$\alpha + \beta$	61	0.565	
	3	1wap(A)	$\beta$	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	$\beta$	53	1.308	
	10	1hhp	$\beta^*$	99	1.761	
	11	1fwp	$\alpha + \beta$	69	1.568	
	12	1sap	$\beta$	66	2.031	
	13	2h5n(D)	$\alpha$	123	2.053	
Hard	14	2ezk	$\alpha$	93	3.475	
	15	1aoy	$\alpha$	78	3.496	
	16	1aly	$\beta$	146	9.179	
	17	1cc5	$\alpha$	83	4.654	
	18	1isu(A)	<i>coil</i>	62	5.912	

#	Target ID	PDB ID	Length	Min (Å)	RMSD
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	

### 3.7 Evaluation Metrics

Since NTF-REL assigns a score to each structure and so can also select a single structure as the best one, we evaluate its performance as an EMA method, as well as a single-structure selection method. First, we evaluate the quality of the scores assigned to structures by measuring the Pearson correlation between these scores and the true TM-Score from the *ground truth* (the experimentally-known structure for each target). We also measure loss as the difference in quality between the structure selected by a method and the best-quality structure in a dataset, with quality assessed by any of the three following metrics, RMSD, TM-Score, and GDT-TS, with respect to the experimentally-known structure.

## 4 Results

We present three sets of results, comparison with SOTA methods on target-wise correlation with respect to the true TM-Score, structure loss, and an analysis of statistical significance. We compare NTF-REL, SNMF-DS [17], NMF-MAD [3], ProQ2 [35], ProQ3 [34], ProQ3D [33], and ProQ4 [24] on the CASP and benchmark datasets.

#### 4.1 Comparative Evaluation on Correlation with TM-Score

Table 2 relates the comparison on the benchmark and CASP targets. The top two predictions on each target are highlighted in boldface font. Table 2 shows that NTF-REL and ProQ4 outperform ProQ2, ProQ3, and ProQ3D on all benchmark and CASP targets. NTF-REL performs comparably to ProQ4, with differences often observed in the third digit after the decimal sign. In particular, both NTF-REL and ProQ4 achieve a Pearson correlation higher than 0.7 on 12/18 of the benchmark targets, respectively, and 8/10 and 10/10 of the CASP targets, respectively (with strictly no rounding). In many targets, both methods achieve or come very close to a Pearson correlation of 0.8.

**Table 2.** Target-wise Pearson correlation with respect to true TM-Score. Top two values are highlighted in boldface font.

Benchmark Targets					
Target-ID	NTF-REL	ProQ2	ProQ3	ProQ3D	ProQ4
1a1l	<b>0.7821</b>	0.683	0.699	0.743	<b>0.787</b>
1dtj(A)	<b>0.802</b>	0.701	0.707	0.762	<b>0.807</b>
1dtd(B)	<b>0.7713</b>	0.675	0.683	0.733	<b>0.776</b>
1wap(A)	<b>0.7432</b>	0.658	0.665	0.713	<b>0.747</b>
1tig	<b>0.715</b>	0.624	0.634	0.689	<b>0.719</b>
1hz6(A)	<b>0.741</b>	0.647	0.657	0.714	<b>0.745</b>
1bg9	<b>0.688</b>	0.616	0.607	0.655	<b>0.697</b>
1c8c(A)	<b>0.728</b>	0.636	0.643	0.693	<b>0.733</b>
1fvp	<b>0.733</b>	0.641	0.646	0.702	<b>0.728</b>
1hhp	<b>0.679</b>	0.602	0.605	0.645	<b>0.683</b>
1sap	<b>0.7066</b>	0.617	0.621	0.673	<b>0.711</b>
2ci2	<b>0.746</b>	0.655	0.661	0.709	<b>0.741</b>
2h5n(D)	<b>0.7204</b>	0.623	0.637	0.685	<b>0.724</b>
1aoy	<b>0.686</b>	0.599	0.604	0.652	<b>0.677</b>
1aly	<b>0.652</b>	0.596	0.592	0.639	<b>0.661</b>
1cc5	<b>0.709</b>	0.627	0.625	0.684	<b>0.714</b>
1isu(A)	<b>0.6938</b>	0.607	0.611	0.679	<b>0.698</b>
2ezk	<b>0.667</b>	0.602	0.607	0.653	<b>0.675</b>

CASP Targets					
Target-ID	NTF-REL	ProQ2	ProQ3	ProQ3D	ProQ4
T0859-D1	<b>0.7031</b>	0.619	0.642	0.689	<b>0.717</b>
T0886-D1	<b>0.6972</b>	0.624	0.634	0.684	<b>0.704</b>
T0892-D2	<b>0.7044</b>	0.643	0.638	0.691	<b>0.719</b>
T0897-D1	<b>0.6896</b>	0.637	0.628	0.676	<b>0.703</b>
T0898-D2	<b>0.7203</b>	0.638	0.656	0.707	<b>0.734</b>
T0953a1-D1	<b>0.701</b>	0.632	0.603	0.652	<b>0.708</b>
T0953a2-D3	<b>0.718</b>	0.627	0.617	0.678	<b>0.725</b>
T0957a1-D1	<b>0.738</b>	0.602	0.635	0.696	<b>0.745</b>
T0960-D2	<b>0.7161</b>	0.624	0.646	0.667	<b>0.731</b>
T1008-D1	<b>0.7533</b>	0.643	0.648	0.701	<b>0.761</b>

#### 4.2 Loss-based Comparison

The above analysis suggests that ProQ3D and ProQ4 decidedly outperform ProQ2 and ProQ3, confirming findings reported in [12]. Therefore, we narrow further comparisons to ProQ3D and ProQ4. Since NTF-REL is a decomposition-based methods, like SNMF-DS and NMF-MAD, we add the latter two to the comparative evaluation on loss. As described in Section 3, we compute RMSD, TM-Score, and GDT-TS loss and relate these results in Table 3 on both benchmark and CASP targets.

Table 3 shows the superiority of NTF-REL over other methods. For instance, on the benchmark targets, the RMSD loss incurred by NTF-REL is below 1Å for 12/18 of the benchmark targets and under 2Å for 6/10 of the CASP targets. The structure selected by NTF-REL has the minimum loss compared to the structure selected by other methods in terms of at least one of the three metrics (RMSD, TM-Score, and GDT-TS) on 8/18 of the benchmark targets and 6/10 of the CASP targets.

#### 4.3 Statistical Significance Analysis

We carry out a statistical significance analysis on both TM-Score loss and GDT-TS loss combined over the benchmark and CASP datasets. We report the results



**Table 3.** NTF-REL, SNMF-DS, ProQ3D, ProQ4, and NMF-MAD are compared on RMSD, TM-Score, and GDT-TS loss. Lowest loss per metric per target is highlighted in boldface font.

ID	Benchmark Targets				
	RMSD Loss, TM-Score Loss, GDT-TS Loss				
	SNMF-DS	ProQ3D	ProQ4	NMF-MAD	NTF-REL
1a1l	0.5084, 0.0655, 0.072	0.357, 0.042, 0.034	0.486, 0.063, 0.057	0.971, 0.1604, 0.1357	<b>0.1527, 0.03, 0.012</b>
1dtj(A)	0.1941, 0.0048, 0.0296	<b>0.125, 0.0118, 0.0057</b>	0.21, 0.0179, 0.0089	0.3345, 0.0782, 0.1081	0.166, <b>0.0043, 0.026</b>
1dtd(B)	0.3528, 0.0042, 0.0041	0.245, 0.0026, 0.0022	0.75, 0.0061, 0.0091	0.5915, 0.0329, 0.0451	<b>0.117, 0.0015, 0.0016</b>
1wap(A)	0.3425, 0.0288, 0.0166	0.352, 0.0277, 0.0245	0.423, 0.0311, 0.029	0.6219, 0.0531, 0.04	<b>0.2285, 0.021, 0.0123</b>
1tig	<b>0.0717, 0.003, 0.0053</b>	0.496, 0.0035, 0.0065	0.479, 0.0032, 0.0061	0.6569, 0.0469, 0.0483	0.72, 0.13, 0.091
1hz6(A)	<b>0.0936, 0.002, 0.0034</b>	0.397, 0.0031, 0.0042	0.291, 0.0025, 0.0039	0.809, 0.0415, 0.0352	0.1248, 0.005, 0.006
1bg9	1.1992, 0.1677, 0.1389	0.745, 0.112, 0.0896	<b>0.673, 0.103, 0.0749</b>	1.3089, 0.1167, 0.0755	1.6362, 0.226, 0.1875
1c8c(A)	0.7991, 0.1092, 0.086	0.521, 0.0953, 0.077	<b>0.444, 0.0887, 0.069</b>	1.092, <b>0.0596, 0.0429</b>	0.7991, 0.1092, 0.086
1fvp	0.5085, 0.0034, 0.0036	0.473, 0.0048, <b>0.0013</b>	0.491, 0.0059, 0.0019	0.5319, 0.0471, 0.0616	<b>0.2658, 0.0019, 0.0023</b>
1hhp	2.1971, 0.0601, 0.0707	0.928, 0.073, 0.069	<b>0.77, 0.0467, 0.0503</b>	2.6835, 0.2939, 0.2828	2.3188, 0.0634, 0.075
1sap	0.5592, 0.074, 0.0417	0.719, 0.0637, 0.0398	0.875, 0.0714, 0.0416	2.075, 0.0989, 0.125	<b>0.3229, 0.045, 0.0248</b>
2c12	0.3118, 0.007, 0.006	<b>0.213, 0.0056, 0.0042</b>	0.831, 0.013, 0.015	1.7897, 0.3246, 0.3462	0.3656, 0.01, 0.008
2h5n(D)	3.7028, 0.3178, 0.3215	0.917, 0.0475, 0.0276	<b>0.839, 0.0465, 0.0315</b>	3.3498, 0.0805, 0.0732	2.987, 0.267, 0.2708
1aoy	2.7896, 0.1136, 0.093	1.265, 0.0567, <b>0.0428</b>	<b>1.074, 0.0511, 0.0431</b>	2.9788, 0.2918, 0.2788	2.346, 0.107, 0.089
1aly	5.7842, 0.0167, 0.0368	<b>2.674, 0.0162, 0.027</b>	2.733, 0.0193, 0.0325	7.9939, 0.1411, 0.1635	2.912, <b>0.015, 0.0345</b>
1cc5	<b>0.4732, 0.048, 0.0452</b>	1.161, 0.0791, <b>0.0388</b>	1.045, 0.0602, 0.0441	2.1843, 0.0565, 0.0573	0.539, 0.054, 0.0509
1isu(A)	2.9928, 0.2182, 0.2299	1.036, <b>0.072, 0.0705</b>	1.124, 0.0733, 0.0717	2.5552, 0.081, 0.0887	<b>0.8689, 0.112, 0.135</b>
2ezk	2.9154, 0.0188, 0.0177	0.729, 0.0027, 0.0063	<b>0.625, 0.0019, 0.0042</b>	3.5136, 0.0229, 0.0296	2.986, 0.021, 0.023

Target ID	CASP Targets				
	RMSD Loss, TM-Score Loss, GDT-TS Loss				
	SNMF-DS	ProQ3D	ProQ4	NMF-MAD	NTF-REL
T1008-D1	0.3656, 0.007, <b>0.0011</b>	0.2838, 0.04, 0.035	0.326, 0.091, 0.088	1.0238, 0.0156, 0.0162	<b>0.2717, 0.006, 0.005</b>
T0886-D1	3.6714, <b>0.03, 0.0362</b>	0.983, 0.12, 0.112	<b>1.147, 0.172, 0.153</b>	2.5984, 0.0331, <b>0.029</b>	2.9813, 0.038, 0.034
T0953s1-D1	2.9398, <b>0.02, 0.0112</b>	<b>0.564, 0.053, 0.041</b>	1.179, 0.022, 0.019	2.613, 0.0225, 0.0223	3.3869, 0.0293, 0.0289
T0960-D2	1.8595, 0.0307, 0.0268	0.765, 0.13, 0.125	<b>0.634, 0.067, 0.081</b>	2.6181, <b>0.0182, 0.0178</b>	1.519, 0.031, 0.033
T0898-D2	1.4889, <b>0.003, 0.0071</b>	1.186, 0.019, 0.078	<b>0.917, 0.0159, 0.068</b>	2.3824, 0.0108, 0.0181	1.0799, 0.004, <b>0.0053</b>
T0892-D2	<b>0.9038, 0.0119, 0.004</b>	1.257, 0.189, 0.076	1.391, 0.263, 0.097	2.8416, 0.0242, 0.009	1.5471, 0.021, <b>0.0038</b>
T0953s2-D3	1.4223, <b>0.01, 0.011</b>	0.954, 0.161, 0.136	<b>0.818, 0.0685, 0.078</b>	1.8621, 0.0256, 0.0153	1.3667, 0.0218, <b>0.0108</b>
T0897-D1	3.471, 0.0263, 0.0108	0.973, 0.033, 0.013	<b>0.849, 0.029, 0.011</b>	2.9413, <b>0.0158, 0.009</b>	3.1845, 0.025, 0.0102
T0957s1-D1	1.18, 0.0027, 0.0047	1.161, 0.031, 0.096	1.294, 0.078, 0.173	1.6803, 0.018, 0.0076	<b>0.7426, 0.002, 0.0045</b>
T0859-D1	2.3755, 0.056, 0.045	<b>1.925, 0.0752, 0.0853</b>	1.972, 0.0734, 0.0771	3.5967, 0.0329, 0.0132	2.3317, <b>0.0265, 0.0124</b>

of Friedman statistical tests with Hommel’s post-hoc analysis [14]. We note that Friedman’s test is ideal for conducting statistical significance of multiple methods contending over multiple test cases. The test is non-parametric and evaluates the null hypothesis (The null hypothesis states that there is negligible difference between the contending methods). Then, we conduct Hommel’s post-hoc analysis to fully evaluate the performance of NTF-REL in comparison to other methods. The statistical tests are performed on all the 28 (benchmark and CASP) targets at  $\alpha = 0.05$ . The results are related in Table V. The lowest average rank are reported in Columns 2 and 5 for TM-Score loss and GDT-Score loss, respectively. The best rank is achieved by NTF-REL on either TM-Score loss or GDT-TS loss. These results conclusively demonstrate the superiority of NTF-REL.

**Table 4.** Statistical significance of various methods over all 28 targets (benchmark and CASP) determined through Friedman’s tests with Hommel’s post-hoc analysis at  $\alpha = 0.05$ . The best rank on either TM-Score or GDT-TS loss is highlighted in boldface.

Method	TM-Score Loss			GDT-TS Loss		
	Avg. Rank	p value	p Hommel	Avg. Rank	p value	p Hommel
NMF-MAD	3.107	0.063	0.0167	3.607	0.0425	0.0125
SNMF-DS	3.249	0.0562	0.0125	2.911	0.7037	0.025
ProQ3D	2.923	0.151	0.025	2.768	0.9663	0.05
ProQ4	2.893	0.177	0.05	2.965	0.6121	0.0167
NTF-REL	<b>2.322</b>	—	—	<b>2.75</b>	—	—

## 5 Conclusion

This paper presents a complete EMA framework that leverages a novel, tensor factorization-based method. The framework associates a score with an individual structure, so it has attributes of single-model EMA method. In addition, the method organizes structures into groups, so it has attributes of a multi-model method. The hybrid framework is shown to outperform various SOTA methods, including distance-based methods currently considered to be the most accurate.

The proposed tensor factorization method doubles as a non-parametric clustering method and so can support various structure-function studies requiring identification of structural macrostates. In this paper, for the purpose of a rigorous and targeted evaluation, we have restricted our attention to EMA and so many of our metrics of performance consider one, experimentally-available structure as the *ground truth*. However, as we make the case in Section 1, it is important to extend our attention beyond the single-structure view of proteins and evaluate, for instance, how the method proposed in this paper and others, can detect in the data multiple alternative, functionally-relevant structures. Our future work will investigate this setting, and it will additionally contribute curated benchmark datasets for rigorous evaluation.

In future work we will investigate additional settings, such as summarization of protein dynamics. The computation of the adjacency and degree matrices can be further expedited via proximity query data structures. Finding target-wise sub-spaces representative of a structure set may also prove informative.

## Acknowledgment

This work is supported in part by NSF Grant No. 1900061 to AS. Resources were partly provided by the Los Alamos National Laboratory (LANL) Institutional Computing Program, which is supported by the DOE National Nuclear Security Administration under Contract No. DE-AC52-06NA25396 and LANL LDRD Grant No. 20190020DR. High-performance computations were run on Darwin, a LANL heterogeneous cluster for research computing, 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 while serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

## References

1. Akhter, N., Chennupati, G., Kabir, K.L., Djidjev, H., Shehu, A.: Unsupervised and supervised learning over the energy landscape for protein decoy selection. *Biomolecules* **9**(1), 607 (2019)
2. Akhter, N., Shehu, A.: From extraction of local structures of protein energy landscapes to improved decoy selection in template-free protein structure prediction. *Molecules* **23**(1), 216 (2018)

3. Akhter, N., Vangara, R., Chennupati, G., Alexandrov, B., Djidjev, H., Shehu, A.: Non-negative matrix factorization for selection of near-native protein tertiary structures. In: IEEE Intl Conf on Bioinf and Biomed (BIBM). pp. 70–73. San Diego, CA (2019)
4. Akhter, N., Kabir, K.L., Chennupati, G., Vangara, R., Alexandrov, B., Djidjev, H.N., Shehu, A.: Improved protein decoy selection via non-negative matrix factorization. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2021)
5. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. Nucleic acids research **28**(1), 235–242 (2000), <https://www.rcsb.org/>
6. Bhattarai, M., Chennupati, G., Skau, E., Vangara, R., Djidjev, H., Alexandrov, B.S.: Distributed non-negative tensor train decomposition. In: 2020 IEEE High Performance Extreme Computing Conference (HPEC). pp. 1–10. IEEE (2020)
7. Bhattarai, M., Kharat, N., Skau, E., Truong, D., Eren, M., Rajopadhye, S., Djidjev, H., Alexandrov, B.: pyDRESCALk: Python distributed non negative rescal decomposition with determination of latent features (2021), <https://github.com/lanl/pyDRESCALk>
8. Bhattarai, M., Nebgen, B., Skau, E., Eren, M., Chennupati, G., Vangara, R., Djidjev, H., Patchett, J., Ahrens, J., Alexandrov, B.: pyDNMFk: Python distributed non negative matrix factorization (2021), <https://github.com/lanl/pyDNMFk>
9. Boehr, D.D., Wright, P.E.: How do proteins interact? Science **320**(5882), 1429–1430 (2008)
10. Chen, X., Liu, J., Guo, Z., et al.: Protein model accuracy estimation empowered by deep learning and inter-residue distance prediction in CASP14. Sci Reports **11**, 10943 (2021)
11. Cheng, J., Choe, M., Elofsson, A.S., et al.: Estimation of model accuracy in CASP13. Proteins **87**(12), 1361–1377 (2021)
12. Cheng, J., Choe, M.H., Elofsson, A., et al.: Estimation of model accuracy in casp13. Proteins: Structure, Function, and Bioinformatics **87**(12), 1361–1377 (2019)
13. Felts, A.K., Gallicchio, E., Wallqvist, A., Levy, R.M.: Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the opls all-atom force field and the surface generalized born solvent model. Proteins: Structure, Function, and Bioinformatics **48**(2), 404–422 (2002)
14. Garcia, S., Herrera, F.: An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. Journal of machine learning research **9**(Dec), 2677–2694 (2008)
15. Jumper, J., Evans, R., et al.: Highly accurate protein structure prediction with alphafold. Nature (2021)
16. Kabir, K.L., Akhter, N., Shehu, A.: From molecular energy landscapes to equilibrium dynamics via landscape analysis and markov state models. J Bioinf and Comput Biol **17**(6), 1940014 (2019)
17. Kabir, K.L., Chennupati, G., Vangara, R., Djidjev, H., Alexandrov, B., Shehu, A.: Decoy selection in protein structure determination via symmetric non-negative matrix factorization. In: IEEE Intl Conf on Bioinf and Biomed (BIBM). pp. 23–28. Virtual (2020)
18. Kabir, K.L., Hassan, L., Rajabi, Z., Akhter, N., Shehu, A.: Graph-based community detection for decoy selection in template-free protein structure prediction. Molecules **24**(5), 854 (2019)

19. Karasikov, M., Pagès, G., Grudin, S.: Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. *Bioinformatics* **35**(16), 2801–2808 (2019)
20. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. *SIAM review* **51**(3), 455–500 (2009)
21. Krompaß, D., Nickel, M., Jiang, X., Tresp, V.: Non-negative tensor factorization with rescal. In: *Tensor Methods for Machine Learning, ECML workshop*. pp. 1–10 (2013)
22. Lei, J., Akhter, N., Qiao, W., Shehu, A.: Reconstruction and decomposition of high-dimensional landscapes via unsupervised learning. In: *ACM SIGKDD Intl Conf on Knowledge Discovery & Data Mining*. pp. 2505–2513. San Diego, CA (2020)
23. Maximova, T., Moffatt, R., Ma, B., Nussinov, R., Shehu, A.: Principles and overview of sampling methods for modeling macromolecular structure and dynamics. *PLoS Comp. Biol.* **12**(4), e1004619 (2016)
24. Menéndez Hurtado, D., Uziela, K., Elofsson, A.: A novel training procedure to train deep networks in the assessment of the quality of protein models (2019)
25. Mirzaei, S., Sidi, T., Keasar, C., Crivelli, S.: Purely structural protein scoring functions using support vector machine and ensemble learning. *IEEE/ACM Trans Comp Biol & Bioinf* (2016)
26. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Tramontano, A.: Critical assessment of methods of protein structure prediction (casp)—round x. *Proteins: Structure, Function, and Bioinformatics* **82**, 1–6 (2014)
27. Nickel, M., Tresp, V., Kriegel, H.P.: A three-way model for collective learning on multi-relational data. In: *Icml* (2011)
28. Olechnovič, K., Monastyrskyy, B., Kryshtafovych, A., et al.: Comparative analysis of methods for evaluation of protein models against native structures. *Bioinformatics* **35**(6), 937–944 (2019)
29. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53–65 (1987)
30. Siew, N., Elofsson, A., Rychlewski, L., Fischer, D.: Maxsub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics* **16**(9), 776–785 (2000)
31. Truong, D.P., Skau, E., Valtchinov, V.I., Alexandrov, B.S.: Determination of latent dimensionality in international trade flow. *Machine Learning: Science and Technology* **1**(4), 045017 (2020)
32. Tunyasuvunakool, K., Adler, J., Wu, Z., et al.: Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021)
33. Uziela, K., Menendez Hurtado, D., Shu, N., Wallner, B., Elofsson, A.: Proq3d: improved model quality assessments using deep learning. *Bioinformatics* **33**(10), 1578–1580 (2017)
34. Uziela, K., Shu, N., Wallner, B., Elofsson, A.: Proq3: Improved model quality assessments using rosetta energy terms. *Scientific reports* **6**(1), 1–10 (2016)
35. Uziela, K., Wallner, B.: Proq2: estimation of model accuracy implemented in rosetta. *Bioinformatics* **32**(9), 1411–1413 (2016)
36. Wang, K., Fain, B., Levitt, M., Samudrala, R.: Improved protein structure selection using decoy-dependent discriminatory functions. *BMC structural biology* **4**(1), 1–18 (2004)
37. Zhang, J., Xu, D.: Fast algorithm for population-based protein structural model analysis. *Proteomics* **13**(2), 221–229 (2013)

# Consent to Publish

## Lecture Notes in Computer Science

---

Title of the Book or Conference Name: .....

Volume Editor(s) Name(s): .....

Title of the Contribution: .....

Author(s) Full Name(s): .....

Corresponding Author's Name, Affiliation Address, and Email:

.....  
.....

When Author is more than one person the expression "Author" as used in this agreement will apply collectively unless otherwise indicated.

The Publisher intends to publish the Work under the imprint **Springer**. The Work may be published in the book series **Lecture Notes in Computer Science (LNCS, LNAI or LNBI)**.

### § 1 Rights Granted

Author hereby grants and assigns to **Springer Nature Switzerland AG, Gewerbestrasse 11, 6330 Cham, Switzerland** (hereinafter called **Publisher**) the exclusive, sole, permanent, world-wide, transferable, sub-licensable and unlimited right to reproduce, publish, distribute, transmit, make available or otherwise communicate to the public, translate, publicly perform, archive, store, lease or lend and sell the Contribution or parts thereof individually or together with other works in any language, in all revisions and versions (including soft cover, book club and collected editions, anthologies, advance printing, reprints or print to order, microfilm editions, audiograms and videograms), in all forms and media of expression including in electronic form (including offline and online use, push or pull technologies, use in databases and data networks (e.g. the Internet) for display, print and storing on any and all stationary or portable end-user devices, e.g. text readers, audio, video or interactive devices, and for use in multimedia or interactive versions as well as for the display or transmission of the Contribution or parts thereof in data networks or search engines, and posting the Contribution on social media accounts closely related to the Work), in whole, in part or in abridged form, in each case as now known or developed in the future, including the right to grant further time-limited or permanent rights. Publisher especially has the right to permit others to use individual illustrations, tables or text quotations and may use the Contribution for advertising purposes. For the purposes of use in electronic forms, Publisher may adjust the Contribution to the respective form of use and include links (e.g. frames or inline-links) or otherwise combine it with other works and/or remove links or combinations with other works provided in the Contribution. For the avoidance of doubt, all provisions of this contract apply regardless of whether the Contribution and/or the Work itself constitutes a database under applicable copyright laws or not.

The copyright in the Contribution shall be vested in the name of Publisher. Author has asserted his/her right(s) to be identified as the originator of this Contribution in all editions and versions of the Work and parts thereof, published in all forms and media. Publisher may take, either in its own name or in that of Author, any necessary steps to protect the rights granted under this Agreement against infringement by third parties. It will have a copyright notice inserted into all editions of the Work and on the Contribution according to the provisions of the Universal Copyright Convention (UCC).

The parties acknowledge that there may be no basis for claim of copyright in the United States to a Contribution prepared by an officer or employee of the United States government as part of that person's official duties. If the Contribution was performed under a United States government contract, but Author is not a United States government employee, Publisher grants the United States government royalty-free permission to reproduce all or part of the Contribution and to authorise others to do so for United States government purposes. If the Contribution was prepared or published by or under the direction or control of the Crown (i.e., the constitutional monarch of the Commonwealth realm) or any Crown government department, the copyright in the Contribution shall, subject to any

agreement with Author, belong to the Crown. If Author is an officer or employee of the United States government or of the Crown, reference will be made to this status on the signature page.

## **§ 2 Rights Retained by Author**

Author retains, in addition to uses permitted by law, the right to communicate the content of the Contribution to other research colleagues, to share the Contribution with them in manuscript form, to perform or present the Contribution or to use the content for non-commercial internal and educational purposes, provided the original source of publication is cited according to the current citation standards in any printed or electronic materials. Author retains the right to republish the Contribution in any collection consisting solely of Author's own works without charge, subject to ensuring that the publication of the Publisher is properly credited and that the relevant copyright notice is repeated verbatim. Author may self-archive an author-created version of his/her Contribution on his/her own website and/or the repository of Author's department or faculty. Author may also deposit this version on his/her funder's or funder's designated repository at the funder's request or as a result of a legal obligation. He/she may not use the Publisher's PDF version, which is posted on the Publisher's platforms, for the purpose of self-archiving or deposit. Furthermore, Author may only post his/her own version, provided acknowledgment is given to the original source of publication and a link is inserted to the published article on the Publisher's website. The link must be provided by inserting the DOI number of the article in the following sentence: "The final authenticated version is available online at [https://doi.org/\[insert DOI\]](https://doi.org/[insert DOI])." The DOI (Digital Object Identifier) can be found at the bottom of the first page of the published paper.

Prior versions of the Contribution published on non-commercial pre-print servers like ArXiv/CoRR and HAL can remain on these servers and/or can be updated with Author's accepted version. The final published version (in pdf or html/xml format) cannot be used for this purpose. Acknowledgment needs to be given to the final publication and a link must be inserted to the published Contribution on the Publisher's website, by inserting the DOI number of the article in the following sentence: "The final authenticated publication is available online at [https://doi.org/\[insert DOI\]](https://doi.org/[insert DOI])".

Author retains the right to use his/her Contribution for his/her further scientific career by including the final published paper in his/her dissertation or doctoral thesis provided acknowledgment is given to the original source of publication. Author also retains the right to use, without having to pay a fee and without having to inform the Publisher, parts of the Contribution (e.g. illustrations) for inclusion in future work. Authors may publish an extended version of their proceedings paper as a journal article provided the following principles are adhered to: a) the extended version includes at least 30% new material, b) the original publication is cited, and c) it includes an explicit statement about the increment (e.g., new results, better description of materials, etc.).

## **§ 3 Warranties**

Author agrees, at the request of Publisher, to execute all documents and do all things reasonably required by Publisher in order to confer to Publisher all rights intended to be granted under this Agreement. Author warrants that the Contribution is original except for such excerpts from copyrighted works (including illustrations, tables, animations and text quotations) as may be included with the permission of the copyright holder thereof, in which case(s) Author is required to obtain written permission to the extent necessary and to indicate the precise sources of the excerpts in the manuscript. Author is also requested to store the signed permission forms and to make them available to Publisher if required.

Author warrants that Author is entitled to grant the rights in accordance with Clause 1 "Rights Granted", that Author has not assigned such rights to third parties, that the Contribution has not heretofore been published in whole or in part, that the Contribution contains no libellous or defamatory statements and does not infringe on any copyright, trademark, patent, statutory right or proprietary right of others, including rights obtained through licences. Author agrees to amend the Contribution to remove any potential obscenity, defamation, libel, malicious falsehood or otherwise unlawful part(s) identified at any time. Any such removal or alteration shall not affect the warranty given by Author in this Agreement.

## **§ 4 Delivery of Contribution and Publication**

Author agrees to deliver to the responsible Volume Editor (for conferences, usually one of the Program Chairs), on a date to be agreed upon, the manuscript created according to the Publisher's Instructions for Authors. Publisher will undertake the reproduction and distribution of the Contribution at its own expense and risk. After submission of the Consent to Publish form signed by the Corresponding Author, changes of authorship, or in the order of the authors listed, will not be accepted by the Publisher.

### § 5 Author's Discount for Books

Author is entitled to purchase for his/her personal use (if ordered directly from Publisher) the Work or other books published by Publisher at a discount of 40% off the list price for as long as there is a contractual arrangement between Author and Publisher and subject to applicable book price regulation.

Resale of such copies is not permitted.


### § 6 Governing Law and Jurisdiction

If any difference shall arise between Author and Publisher concerning the meaning of this Agreement or the rights and liabilities of the parties, the parties shall engage in good faith discussions to attempt to seek a mutually satisfactory resolution of the dispute. This agreement shall be governed by, and shall be construed in accordance with, the laws of Switzerland. The courts of Zug, Switzerland shall have the exclusive jurisdiction.

Corresponding Author signs for and accepts responsibility for releasing this material on behalf of any and all Co-Authors.

**Signature of Corresponding Author:**

**Date:**

.....  


- ☐ I'm an employee of the US Government and transfer the rights to the extent transferable (Title 17 §105 U.S.C. applies)
- ☐ I'm an employee of the Crown and copyright on the Contribution belongs to the Crown

*For internal use only:*

Legal Entity Number: 1128 Springer Nature Switzerland AG  
 Springer-C-CTP-07/2018