IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 October 2024, accepted 23 November 2024, date of publication 27 November 2024, date of current version 9 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3507133

METHODS

The Impact of Random Models on Standardized Clustering Similarity

KAI KLEDE[1], THOMAS R. ALTSTIDL[1], DARIO ZANCA[1], AND BJÖRN M. ESKOFIER[1,2], (Senior Member, IEEE)

[1]Machine Learning and Data Analytics (MaD) Laboratory, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91052 Erlangen, Germany
[2]Translational Digital Health Group, Institute of AI for Health, Helmholtz Zentrum München, 85764 Oberschleißheim, Germany

Corresponding author: Kai Klede (kai.klede@fau.de)

This work was supported by the Digital Europe Grant Testing and Experimentation Facility for Health Artificial Intelligence (AI) and Robotics (TEF-Health) under Project 101100700.

ABSTRACT Clustering similarity measures are essential for evaluating clustering results and ensuring diversity in multiple clusterings of the same dataset. Common indices like the Mutual Information (MI) and Rand Index (RI) are biased towards smaller clusters and are often adjusted using a random permutation model. Recent advancements have standardized these measures to further correct biases, but the impact of different random models on these standardized measures has not yet been studied. In this work, we introduce equations for standardizing the MI/RI under non-permutation models, specifically focusing on a uniform model over all clusterings and a model that fixes the number of clusterings. Our results show that while standardization improves performance for the fixed number of clusters model, its benefits are limited in the more general uniform model. We validate our findings with gene expression data, highlighting the importance of choosing the right similarity metric for clustering comparison.

INDEX TERMS Clustering comparison, external evaluation metrics, mutual information, rand index, random model, machine learning.

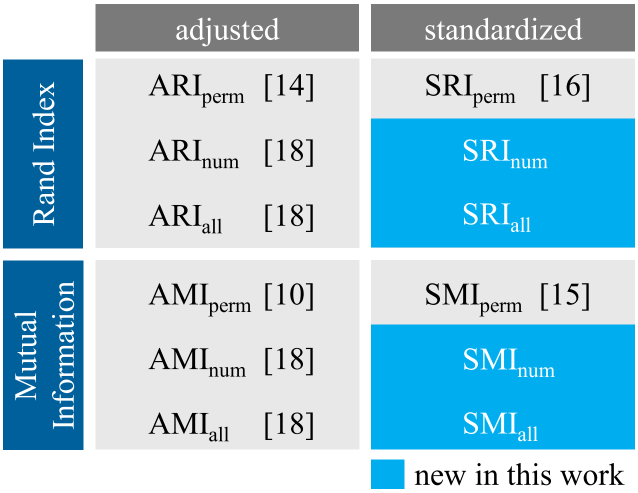
I. INTRODUCTION

Clustering describes the process of partitioning a set into meaningful subsets and is a fundamental technique of unsupervised learning. Once a clustering (set partition) is found, its quality can be assessed by comparing it to a reference clustering via a similarity index [1]. This procedure is called external validation and is common to determine the best clustering algorithm for a given dataset [2]. In that way, clustering similarity measures are employed in numerous domains, like topic modeling [3], image segmentation [4] or medical applications [5]. Besides their use in external validation, clustering similarity indices guide multiple-clustering algorithms to generate meaningfully diverse solutions [6], [7]. Recently, clustering similarity indices were also employed in loss functions for deep learning methods for community detection on graphs [8], [9].

Two of the most popular clustering similarity indices are the Mutual Information (MI) [10] and the Rand Index (RI) [11]. A well-known problem with these clustering similarity indices is that they are biased towards smaller clusters [12], [13]. The Adjusted Mutual Information (AMI) [10] and Adjusted Rand Index (ARI) [14] achieve a constant baseline by correcting the indices by their expected value under random permutation of the cluster labels. However, a more subtle bias towards smaller clusters remains when multiple candidate solutions are compared to a single reference to determine the best candidate [15]. To counter that second bias, the indices are further corrected by their second statistical moment under random permutation as the Standardized Mutual Information (SMI) [15] and the Standardized Rand Index (SRI) [16]. A recent study even demonstrated the theoretical benefits of a *p*-value correction, although its steep computational cost limits practical applications [17].

All these corrections use the random permutation model with fixed cluster sizes as a baseline, but a different random

The associate editor coordinating the review of this manuscript and approving it for publication was Mehedi Masud.



	adjusted	standardized
Rand Index	ARI_{perm} [14]	SRI_{perm} [16]
	ARI_{num} [18]	🟦 SRI_{num}
	ARI_{all} [18]	🟦 SRI_{all}
Mutual Information	AMI_{perm} [10]	SMI_{perm} [15]
	AMI_{num} [18]	🟦 SMI_{num}
	AMI_{all} [18]	🟦 SMI_{all}

🟦 new in this work

FIGURE 1. Gates and Ahn [18] extended the Adjusted Rand Index (ARI) and the Adjusted Mutual Information (AMI) to non-permutation models, to better reflect constraints of popular clustering methods like k-means. In a parallel development Romano et al. [15], [16] extended the adjustment to the second statistical moment, introducing the Standardized Rand Index (SRI) and Standardized Mutual Information (SMI) to remove a bias in the adjusted variants. In this work, we combine these approaches and introduce standardization for non-permutation models.

model could be more appropriate for many popular clustering algorithms [18]. In k-means clustering, for example, the number of clusters is fixed, but the individual cluster sizes can fluctuate [19]. Hence, random clusterings with a fixed number of clusters of any size are a more appropriate baseline for k-means. In other clustering algorithms like DBSCAN [20], not even the number of clusters is fixed, such that the uniform distribution over all clusterings given the dataset size is a better baseline. Gates and Ahn [18] studied these random models for the first-moment-corrected ARI and AMI.

In this work, we extend these random models to the second-order corrected SRI and SMI as illustrated in Figure 1. We further derive a general formulation of these corrections for any element-symmetric distribution, providing a framework for studying custom random models beyond those presented here. Our synthetic experiments validate the theoretical improvements offered by the second-order corrections in the fixed number of clusters model. An experiment on gene expression data highlights the importance of the random model choice.

II. BACKGROUND

A clustering $U : S \to \{1, 2, \dots, k_U\}$ is a surjection that partitions a dataset S with $N := |S|$ data points into k_U subsets. We denote the size of the i-th cluster with $u_i := |U^{-1}(i)|$. Clustering similarity indices like the MI and RI quantitatively compare two clusterings U, V. For the MI on a finite dataset, the cluster assignments U, V are understood as discrete random variables and hence we use the mutual information for discrete distributions [10], [12]. The RI counts the number of pairs of data points that agree in both clusterings, more on that in Section III, Equation (5).

To correct a bias in these indices that always favors smaller clusters, the AMI and ARI were introduced with the random permutation model [10], [14]. Gates and Ahn [18] generalized these adjustments to other random models as follows

$$\text{AMI}_{\text{model}}(U, V) = \frac{\text{MI}(U, V) - \mathbb{E}_{\text{model}}[\text{MI}]}{\max_{\text{model}}[\text{MI}] - \mathbb{E}_{\text{model}}[\text{MI}]} \tag{1}$$

$$\text{ARI}_{\text{model}}(U, V) = \frac{\text{RI}(U, V) - \mathbb{E}_{\text{model}}[\text{RI}]}{1 - \mathbb{E}_{\text{model}}[\text{RI}]}. \tag{2}$$

Note that $\max_{\text{model}}[\text{MI}]$ used to normalize the AMI, must be chosen appropriately for the random model [12], [18]. In the following, three random models are studied in more depth:

- In the **random permutation** model (perm), the expectation value is calculated over all permutations of the cluster labels, while the cluster sizes $u_1, \dots, u_{k_U}; v_1, \dots, v_{k_V}$ are fixed. It is the most widely used model for clustering similarity index adjustment.
- In the **fixed number of clusters** model (num), the expectation value is over all set partitions with k_U, k_V parts. This model can provide a better baseline for clustering methods with a predetermined number of clusters like k-means, spectral clustering, or Gaussian mixture methods.
- The uniform model over **all clusterings** (all) of N data points is the most general. It can serve as a starting point when the internal workings of the clustering algorithm are unknown.

Observe that all three models assign equal probability to all permutations of clustering, i.e., they are fully defined by a distribution over the integer partitions (cluster sizes). We call such distributions element-symmetric:

Definition 1 (Gösgens et al. [21]): A distribution over clusterings $\mathcal{U}$ is *element-symmetric* if it assigns the same probability to every two clusterings U and U' with the same cluster sizes.

While the permutation model is equivalent whether one or both clusterings are permuted, one- and two-sided corrections must be distinguished for the other random models [18]. We denote one-sided adjustments with a superscript 1, e.g., $\text{AMI}^1_{\text{num}}$ for the one-sided, fixed number of clusters adjusted Mutual Information.

The AMI and ARI satisfy the constant baseline property, but they are still prone to a selection bias, i.e. the tendency to select clusterings with more clusters than a reference, when comparing multiple candidate solutions [15]. Romano et al. [15], [16] introduced standardization under the random permutation model to mitigate that bias.

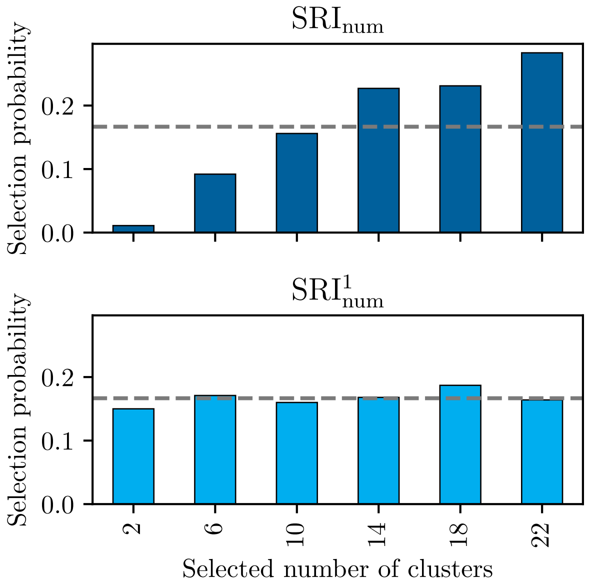


FIGURE 2. We compare a fixed reference clustering with $k_U = 10$ even clusters, to random clusterings with $k_V \in \{2, 6, 10, 14, 18, 22\}$ clusters and select the most similar random clustering. The plot shows the selection probabilities of each k_V for the standardized Rand index under the fixed number of clusters model with two- and one-sided adjustment. The two-sided $\mathrm{SRI}_{\mathrm{num}}$ is clearly biased towards more clusters, whereas the one-sided $\mathrm{SRI}^1_{\mathrm{num}}$ mostly mitigates that bias. In scenarios when there is one fixed reference clustering, the one-sided adjustment is more suitable.

III. STANDARDIZATION IN NON-PERMUTATION MODELS

In this work, we generalize the SMI and SRI to other random models as follows

$$\mathrm{SMI}_{\mathrm{model}}(U, V) = \frac{\mathrm{MI}(U, V) - \mathbb{E}_{\mathrm{model}}[\mathrm{MI}]}{\sqrt{\mathrm{Var}_{\mathrm{model}}[\mathrm{MI}]}} \tag{3}$$

$$\mathrm{SRI}_{\mathrm{model}}(U, V) = \frac{\mathrm{RI}(U, V) - \mathbb{E}_{\mathrm{model}}[\mathrm{RI}]}{\sqrt{\mathrm{Var}_{\mathrm{model}}[\mathrm{RI}]}}, \tag{4}$$

with $\mathrm{Var}_{\mathrm{model}}[X] = \mathbb{E}_{\mathrm{model}}[X^2] - \mathbb{E}_{\mathrm{model}}[X]^2$. Calculating the statistical moments is the crucial step for achieving standardization. Particularly the second moment is of interest, since the AMI and ARI already correct for the first statistical moment, whereas standardization also takes the second moment into account.

A. RAND INDEX

The Rand Index is the fraction of pairs of data points that agree in two clusterings U and V. Given a pair of points i, j, the clusterings agree when they map the points to either the same cluster or different clusters.

$$\begin{aligned}&\mathrm{RI}(U, V)\\ &:= \mathbb{E}_{i,j\sim\mathrm{Unif}\binom{S}{2}}\left[\mathbf{1}_{(U(i)=U(j)\wedge V(i)=V(j))\vee(U(i)\neq U(j)\wedge V(i)\neq V(j))}\right]\\ &= \mathbb{E}_{i,j\sim\mathrm{Unif}\binom{S}{2}}\left[\mathbf{1}_{U(i)=U(j)}\mathbf{1}_{V(i)=V(j)} + \mathbf{1}_{U(i)\neq U(j)}\mathbf{1}_{V(i)\neq V(j)}\right],\end{aligned} \tag{5}$$

with i, j a pair of distinct elements of the dataset S chosen uniformly at random and $\mathbf{1}$ the indicator function.

For the Rand Index, we consider the permutation model, the fixed number of clusters model, and the uniform model over all clusterings. The pairwise transposition model is omitted, as its only purpose is efficiently approximating the permutation model. Yet, the $\mathrm{SRI}_{\mathrm{perm}}$ was shown to be as efficient as the RI itself $\mathcal{O}(k_u k_v)$ [17].

1) EXPECTED RAND INDEX

Theorem 1: Let $\mathcal{U}$ and $\mathcal{V}$ denote two element-symmetric clustering distributions. The expected Rand Index under such distributions is

$$\mathbb{E}_{U\sim\mathcal{U}, V\sim\mathcal{V}}[\mathrm{RI}(U, V)] = p^{=}_{\mathcal{U}} p^{=}_{\mathcal{V}} + p^{\neq}_{\mathcal{U}} p^{\neq}_{\mathcal{V}}, \tag{6}$$

with $p^{=}_{\mathcal{U}} := \mathbb{E}_{U\sim\mathcal{U}; i,j\sim\mathrm{Unif}\binom{S}{2}}\left[\mathbf{1}_{U(i)=U(j)}\right]$ the probability that a random pair of datapoints i, j share the same cluster in a clustering drawn from $\mathcal{U}$. The probability that i and j are in different clusters is $p^{\neq}_{\mathcal{U}} = 1 - p^{=}_{\mathcal{U}}$.

All formal proofs are relegated to the appendix. Essentially, the restriction to element-symmetric distributions allows for factorization into individual probabilities since everything is fully determined by the cluster sizes (integer partitions), leaving $\mathbf{1}_{U(i)=U(j)}$ and $\mathbf{1}_{V(i)=V(j)}$ uncorrelated.

Note that the expected Rand Index is fully determined by $p^{=}_{\mathcal{U}}$. Following Gates and Ahn [18], we derive expressions for $p^{=}_{\mathcal{U}}$ for each of the random models.

a: PERMUTATION MODEL

When the cluster sizes u_i are fixed, the number of pairs where both points are inside the same cluster is $\sum_{i=1}^{k_U}\binom{u_i}{2}$. Divided by the total number of pairs, we get the desired probability

$$p^{=}_{\mathrm{perm}} = \sum_{i=1}^{k_U}\binom{u_i}{2} \Big/ \binom{N}{2} \tag{7}$$

b: FIXED NUMBER OF CLUSTERS

If we pick a random pair of data points from a random clustering with k clusters, $p^{=}_{\mathrm{num}}$ is the probability that these points are in the same cluster. The total number of clusterings of a dataset of size N into exactly k clusters is $S(N, k)$, the Stirling Number of the second kind [22]. Given a pair of data points i, j, we count the number of clusterings for which i and j are in the same cluster. To that end, we first hold out i and cluster all data points but i into k clusters. Then we add i to the same cluster as j. Hence, there are $S(N-1, k)$ clusterings where i and j are in the same cluster [23]. Note that this number is completely independent of i and j and therefore the probability $p^{=}_{\mathrm{num}}$ is

$$p^{=}_{\mathrm{num}} = \frac{S(N-1, k)}{S(N, k)}. \tag{8}$$

c: ALL CLUSTERINGS

The number of clusterings of any size, where a uniformly chosen pair i, j is in the same cluster, is simply the sum of that count for a given number of clusters k over all possible numbers of clusters (from a single large cluster to N singleton clusters)

$$p_{\text{all}}^{=} = \frac{\sum_{k=1}^{N} S(N-1,k)}{\sum_{k=1}^{N} S(N,k)} = \frac{B_{N-1}}{B_N}, \tag{9}$$

where B_N denotes the N-th Bell number [24].

2) VARIANCE RAND INDEX

Given the expected Rand Index from above, we lack the second moment of the Rand Index to compute the Variance.

Theorem 2: Let $\mathcal{U}$ and $\mathcal{V}$ denote two element-symmetric clustering distributions. Let

$$p_{\mathcal{U}}^{=,=} := \mathbb{E}_{U\sim\mathcal{U};(i,j),(i',j')\sim\binom{\binom{S}{2}}{2}} \left[\mathbf{1}_{U(i)=U(j)}\mathbf{1}_{U(i')=U(j')}\right] \tag{10}$$

denote the joint probability that two distinct pairs of distinct data points chosen uniformly at random, each lie inside a cluster in a clustering U drawn from $\mathcal{U}$. Further, let

$$p_{\mathcal{U}}^{=,\neq} = p_{\mathcal{U}}^{\neq,=} = p_{\mathcal{U}}^{=} - p_{\mathcal{U}}^{=,=} \tag{11}$$

denote the probability of one pair falling inside a cluster while the other falls in two distinct clusters, with $p_{\mathcal{U}}^{=}$ as in Theorem 1. Finally, let the probability of both pairs being split across clusters be

$$p_{\mathcal{U}}^{\neq,\neq} = 1 - p_{\mathcal{U}}^{=,=} - p_{\mathcal{U}}^{=,\neq} - p_{\mathcal{U}}^{\neq,=}. \tag{12}$$

Then the second statistical moment of the Rand Index under $\mathcal{U}$ and $\mathcal{V}$ is

$$\begin{aligned}&\mathbb{E}_{U\sim\mathcal{U},V\sim\mathcal{V}}[\text{RI}(U,V)^2]\\ &= \binom{N}{2}^{-1}(p_{\mathcal{U}}^{=}p_{\mathcal{V}}^{=} + p_{\mathcal{U}}^{\neq}p_{\mathcal{V}}^{\neq}) + \left(1-\binom{N}{2}\right)^{-1}\\ &\left(p_{\mathcal{U}}^{=,=}p_{\mathcal{V}}^{=,=} + 2\cdot p_{\mathcal{U}}^{=,\neq}p_{\mathcal{V}}^{=,\neq} + p_{\mathcal{U}}^{\neq,\neq}p_{\mathcal{V}}^{\neq,\neq}\right).\end{aligned} \tag{13}$$

Similarly to the expected Rand Index, the variance is now fully determined by $p_{\mathcal{U}}^{=}$ and $p_{\mathcal{U}}^{=,=}$. In the following we derive explicit expressions for $p_{\mathcal{U}}^{=,=}$ for each of the random models.

a: PERMUTATION MODEL

We first pick a pair where both points are inside the same cluster with probability $p_{\text{perm}}^{=}$ (Equation (7)). Then there are $\binom{N}{2} - 1$ pairs left out of which $\sum_{i=1}^{k_U} \binom{u_i}{2} - 1$ are in the same cluster, hence

$$p_{\text{perm}}^{=,=} = p_{\text{perm}}^{=}\left(\sum_{i=1}^{k_U}\binom{u_i}{2} - 1\right)\Big/\left(\binom{N}{2} - 1\right). \tag{14}$$

b: FIXED NUMBER OF CLUSTERS

Similar to $p_{\text{num}}^{=}$, we hold out two distinct data points i, i' and assign the remaining data points to k_U clusters. Then we assign i to the same cluster as j and i' to the same cluster as j', giving

$$p_{\text{num}}^{=,=} = S(N-2,k)/S(N,k). \tag{15}$$

Note that $j = j'$ is possible, but since the expectation value in Equation (10) is over two distinct pairs, we can always pick $i \neq i'$ in the first step.

c: ALL CLUSTERINGS

As for the expected Rand Index, we take the sum of $p_{\text{num}}^{=,=}$ over all possible values for k

$$p_{\text{all}}^{=,=} = B_{N-2}/B_N. \tag{16}$$

B. MUTUAL INFORMATION

Nguyen et al. [10] found that the expected mutual information under random permutation can be expressed using the hypergeometric distribution

$$\text{Hyp}(n|u,v,N) = \frac{\binom{u}{n}\binom{N-u}{v-n}}{\binom{N}{v}}, \tag{17}$$

that has finite support on the integers in $[\max(u+v-N,0), \min(u,v)]$. Here, we generalize this result to other random models.

1) EXPECTED MUTUAL INFORMATION

Theorem 3: Let $\mathcal{U}, \mathcal{V}$ be two element-symmetric clustering distributions. Let $p_{\mathcal{U}}(u)$ denote the probability of a random datapoint $i \sim \text{Unif}(S)$ lying in a cluster of size u in a clustering U drawn from $\mathcal{U}$. Then, the expected mutual information is

$$\begin{aligned}&\mathbb{E}_{U\sim\mathcal{U},V\sim\mathcal{V}}[\text{MI}(U,V)]\\ &= \sum_u\sum_v p_{\mathcal{U}}(u)p_{\mathcal{V}}(v)\sum_n \frac{nN}{uv}\log\left(\frac{nN}{uv}\right)\text{Hyp}(n|u,v,N).\end{aligned} \tag{18}$$

We define $n\log n = 0$ for $n = 0$ for notational simplicity.

The idea of factoring out the probabilities $p_{\mathcal{U}}(u)$ is not new and has proven beneficial for an efficient Monte Carlo approximation of the expected MI for the permutation model [25]. Here, we find $p_{\mathcal{U}}(u)$ also for the other random models, following the argumentation in [18].

a: PERMUTATION MODEL

In the random permutation model, the probability of picking a datapoint in a cluster of size u is simply u/N times the number of clusters of size u in the clustering U

$$p_{\text{perm}}^{U}(u) = \frac{|\{i : u_i = u\}|u}{N}. \tag{19}$$

b: FIXED NUMBER OF CLUSTERS

$p_{\text{num}}^{k,N}(u)$ denotes the probability that for a random clustering of N points with k clusters, a uniformly chosen data point lies in a cluster of size u. We can adapt the strategy from section III-A1b, yet here we are interested in the size of the cluster of a single point in contrast to a pair of points being in the same cluster as for the Rand Index. The total number of clusterings of N data points into k clusters is again $S(N,k)$ [22]. First, we search the number of clusterings that contain at least one cluster of size u. To ensure that such a cluster exists, we first pick u out of the N data points and put them in a cluster - there are $\binom{N}{2}$ ways to do that. Then there are $S(N-u,k-1)$ ways to put the remaining $N-u$ data points

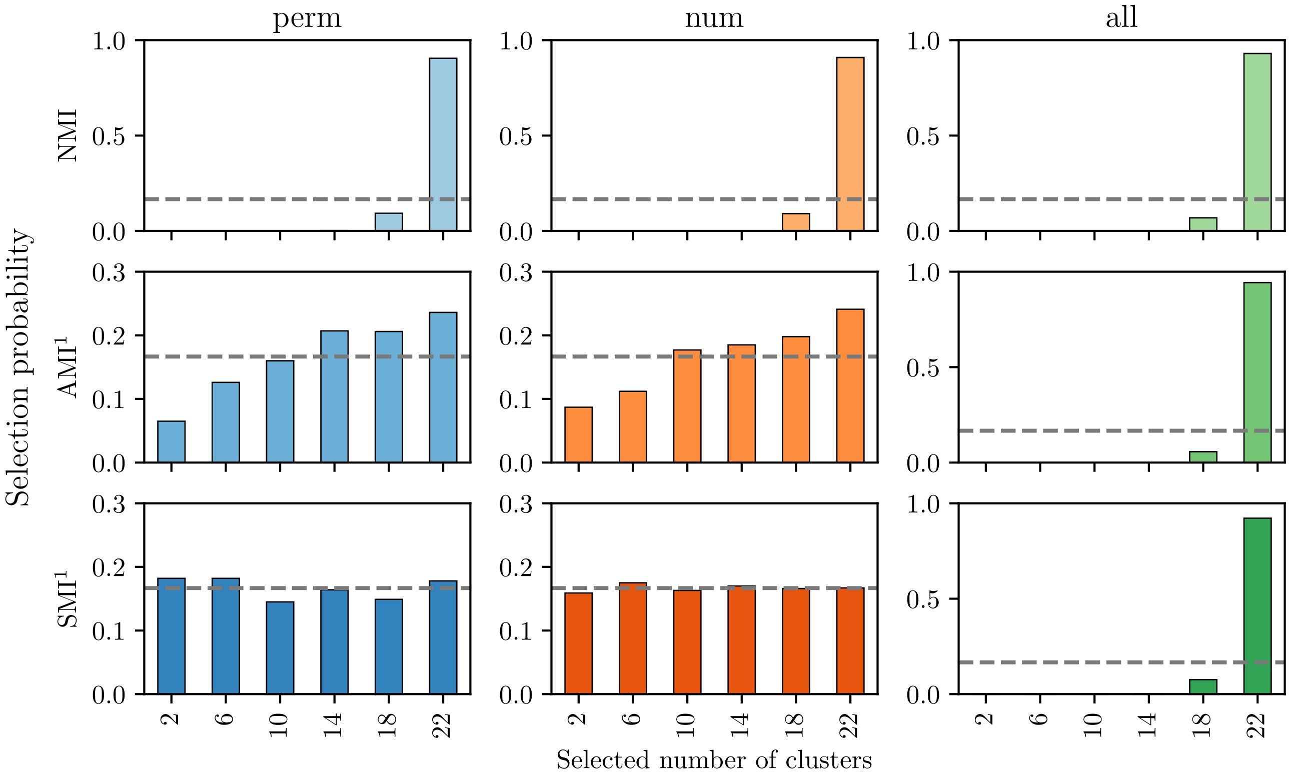


FIGURE 3. We compare a fixed reference clustering with $k_{\mathcal{U}} = 10$ even clusters, to random clusterings with $k_{\mathcal{V}} \in \{2, 6, 10, 14, 18, 22\}$ clusters. The plot shows the selection probabilities of each $k_{\mathcal{V}}$ for the $\mathrm{NMI}_{\mathrm{model}}$, $\mathrm{AMI}^1_{\mathrm{model}}$, $\mathrm{SMI}^1_{\mathrm{model}}$ for the permutation model (perm), the fixed number of clusters model (num) and the all clusterings model (all). For the permutation model and the fixed number of clusters model, standardization effectively removes the bias towards smaller clusters. Since the number of clusters N is constant throughout this experiment, the adjustments using the all clusterings model are effectively a constant scaling factor and standardization has no impact. This highlights the importance of not choosing a random model that is too general for the comparison scenario.

in $k-1$ remaining clusters. Finally the chance of picking one of the u elements when randomly choosing a data point is u/N, and hence

$$p^{k,N}_{\mathrm{num}}(u) = \binom{N}{u}\frac{S(N-u,k-1)}{S(N,k)}\frac{u}{N}. \tag{20}$$

c: ALL CLUSTERINGS

For the all clusterings model, we take the sum over all possible numbers of clusters $\sum_k p^{k,N}_{\mathrm{num}}$, giving

$$p^{N}_{\mathrm{all}}(u) = \binom{N}{u}\frac{B_{N-u}}{B_N}\frac{u}{N}. \tag{21}$$

2) VARIANCE MUTUAL INFORMATION

Theorem 4: Let $\mathcal{U}, \mathcal{V}$ be two element-symmetric distributions and $p_{\mathcal{U}}(u)$ as in Theorem 3. We further denote the probability of two random datapoints $i, i' \sim \mathrm{Unif}(S)$ being in clusters of size u and u' in a clustering drawn from $\mathcal{U}$ as $p_{\mathcal{U}}(u, u')$. For convenience, we further define the conditional probability of cluster size u', given u, but reduced by the case where both i and i' are in the same cluster

$$q_{\mathcal{U}}(u'|u) = \frac{p_{\mathcal{U}}(u,u')}{p_{\mathcal{U}}(u)} - \mathbf{1}_{u=u'}\frac{u'}{N}. \tag{22}$$

The second moment of the Mutual Information is then

$$\begin{aligned}
&\mathbb{E}_{U\sim\mathcal{U},V\sim\mathcal{V}}[\mathrm{MI}(U,V)^2] = \\
&\sum_u\sum_v p_{\mathcal{U}}(u)p_{\mathcal{V}}(v)\sum_n \frac{nN}{uv}\log\left(\frac{nN}{uv}\right)\mathrm{Hyp}(n|u,v,N) \\
&\left[\frac{n}{N}\log\left(\frac{nN}{uv}\right)+\sum_{u'}q_{\mathcal{U}}(u'|u)\sum_{n'}\frac{n'}{u'}\log\left(\frac{n'N}{u'v}\right)\right. \\
&\mathrm{Hyp}(n'|u',v-n,N-u)+\sum_{v'}q_{\mathcal{V}}(v'|v)\sum_{n'} \\
&\mathrm{Hyp}(n'|v',u-n,N-v)\left(\frac{n'}{v'}\log\left(\frac{nN}{uv'}\right)\right. \\
&\quad+\sum_{u'}q_{\mathcal{V}}(u'|u)\sum_{n''}\frac{n''N}{u'v'}\log\left(\frac{n''N}{u'v'}\right) \\
&\left.\left.\mathrm{Hyp}(n''|u',v'-n',N-u)\right)\right].
\end{aligned} \tag{23}$$

In the following we derive $q_{\mathcal{U}}(u'|u)$ for the three random models.

a: PERMUTATION MODEL

In the case of the random permutation model, finding a cluster of size u and another of size u' is independent since there is only one fixed clustering U, and the two data points are chosen independently. From Equation (22) follows

$$q^U_{\mathrm{perm}}(u'|u) = p^{\mathcal{U}}_{\mathrm{perm}}(u') - \mathbf{1}_{u=u'}\frac{u}{N}. \tag{24}$$

For non-permutation models, we first pick one of multiple possible clusterings and only then pick two cluster sizes in that clustering, leading to statistical dependence in general.

b: FIXED NUMBER OF CLUSTERS

Given a clustering of size u, the probability of finding a clustering of size u' in the remaining $k-1$ clusters of $N-u$ data points is

$$q^{k,N}_{\mathrm{num}}(u'|u) = p^{k-1,N-u}_{\mathrm{num}}(u'). \tag{25}$$

Note that this excludes the case of picking the same cluster twice as desired in the definition of $q_{\mathcal{U}}$.

c: ALL CLUSTERINGS

Similar to the fixed number of clusters case, there are $N-u$ data points left, given a cluster of size u, such that

$$q^N_{\mathrm{all}}(u'|u) = p^{N-u}_{\mathrm{all}}(u'). \tag{26}$$

IV. COMPUTATIONAL COMPLEXITY AND MONTE CARLO APPROXIMATION

While the ARI and SRI retain the same asymptotic time complexity as the RI itself $\mathcal{O}(k_U k_V)$ in the random permutation model [17], the AMI has complexity $\mathcal{O}(\max(k_U, k_V)N)$ and the SMI is even worse with $\mathcal{O}(k_U k_V N^3)$ [15], [16]. The other random models essentially increase the cardinality of the support for $p_{\mathcal{U}}$ (See Equation (18) and (23)), making these adjustments impractical for larger datasets. For the random permutation model, Lazarenko and Bonald [26] address this by iterating only over pairwise permutations, however strictly speaking this is a different random model and it is unclear how to generalize this to the fixed number of clusters and the all clusterings models. Instead, we implement an efficient Monte Carlo estimator for cases where the exact SMI exceeds a timeout, similar to the approach described in [25] and [27]. For the fixed number of clusters model a brute force sampling technique based on the recurrence relation $S(N+1,k) = kS(N,k) + S(n,k-1)$ is sufficient, whereas we leverage Dobiński's formula for the all clusterings model [28].

An experimental comparison of the runtimes obtained using this Monte Carlo approach is shown in Figure 4.[1] The results confirm that while the exact computation of the SMI is computationally infeasible for even small datasets ($N \simeq 100$), the Monte Carlo approximation significantly reduces the computational cost, making the SMI applicable to datasets with sizes on the order of $N \simeq 10,000$. The exact SRI is the fastest method overall, yet the calculation of the Bell numbers makes the **all** variant slightly less efficient than the other random models. For the Monte Carlo SMI, the **num** model stands out, as the brute force approach for generating clusterings with a fixed number of clusters could be improved. A more sophisticated approach could borrow ideas from probabilistic divide and conquer methods [29], [30], though this improvement is beyond the scope of the current work.

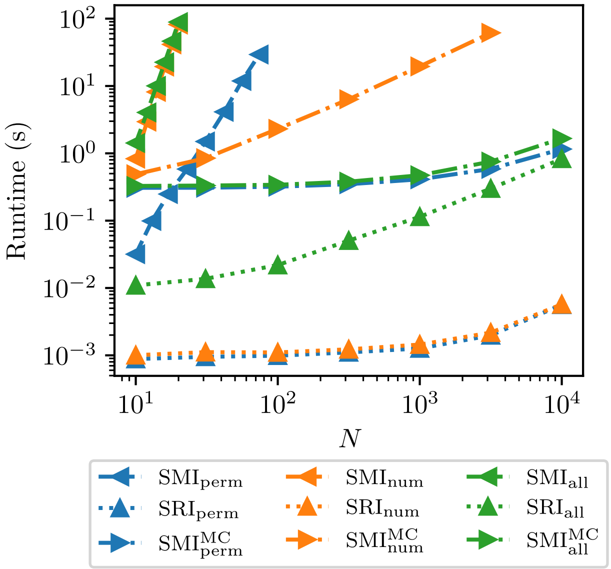


FIGURE 4. Runtime comparison of the SMI, SRI and Monte Carlo SMI at a precision of 0.1 for the random models perm, num, and all, averaged over 100 random trials on an AMD Ryzen 9 5950X. The SRI is the fastest method, followed by the Monte Carlo SMI. The exact SMI is computationally inaccessible, even for small datasets $N \simeq 100$. The brute force sampling technique used for the num model takes a toll on the runtime of the Monte Carlo approximation, whereas Knuth's algorithm for the all model adds minimal overhead compared to the perm model [28]. For the SRI, the calculation of the Bell numbers slows the all model down compared to the more efficient calculation of Stirling numbers of the second kind for small k in the num model.

V. SYNTHETIC EXPERIMENTS

Similar to the experiments in the original SMI paper [15], we model an external validation scenario where a fixed ground truth reference is given, and multiple candidate solutions are compared to the same reference. We first split $N = 500$ data points into $k_U = 10$ evenly sized clusters, the clustering u. Then we generate random clusterings with $k_V \in \{2, 6, 10, 14, 18, 22\}$ clusters and compare them with u. The clustering with the highest similarity is selected and we record the selection probability for each k_V over 5000 trials. While the $\mathrm{AMI_{perm}}$ and $\mathrm{ARI_{perm}}$ are known to be biased in this scenario (Figure 3), the $\mathrm{SMI_{perm}}$ and $\mathrm{SRI_{perm}}$ were introduced to mitigate that bias [16]. While there is no difference between a one- and two-sided adjustment in the permutation model, it does make a difference as we generalize the experiment to the other random models. It could be argued that the one-sided correction is the appropriate choice for

[1] Code available at https://github.com/mad-lab-fau/standardized-clustering-comparison

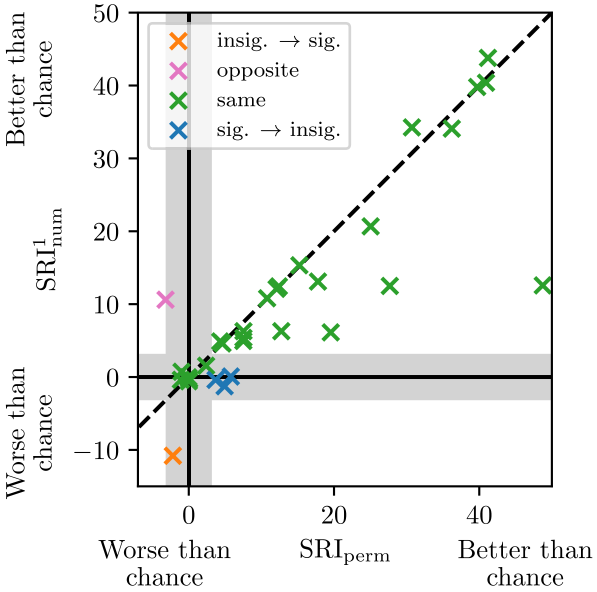


FIGURE 5. **Comparison of agglomerative hierarchical clusterings to the reference clustering using $\mathrm{SRI}_{\mathrm{perm}}$ and $\mathrm{SRI}^1_{\mathrm{num}}$ for gene expression data from 35 studies on cancerous and healthy tissue samples. The absolute evaluation (better than chance) of most studies would not change had a one-sided fixed number of clusters correction been employed instead of the usual permutation model. However, the relative significance of the results varies and the assessment of significance changes for some studies.**

the proposed experiment since it involves a fixed reference clustering. Figure 2 confirms that argument and shows the results for the one- and two-sided $\mathrm{SRI}_{\mathrm{num}}$. While the two-sided variant has a clear tendency to select higher numbers of clusters, the one-sided variant is more balanced. The discrepancy from a perfectly balanced selection stems can be explained by the fact that standardization is just an approximation to the true p-value, yet a useful simplification for a computationally feasible adjustment [17].

In Figure 3 we compare the normalized MI (NMI), AMI^1, and SMI^1 across all three random models, otherwise using the same setup. For the random permutation model and the fixed number of clusters model, standardization removes the bias towards smaller clusters. The all clusterings model on the other hand amounts to a constant correction in this scenario, as the number of data points is kept constant and standardization does not remove the bias towards clusterings with large numbers of clusters. An analogous analysis with similar results for Rand Index based indices is in Appendix B.

VI. EXPERIMENTS ON REAL DATASETS

Expectation value adjustment indicates whether a clustering method is more similar to a desired reference than a random clustering, an assessment that crucially depends on the selected random model. In this work, we extended this adjustment to the second statistical moment, indicating not only whether a clustering method is better than chance, but also if it is significant. To analyze the impact of the choice of random model on standardized clustering comparison, we apply 12 different clustering algorithms to 35 datasets of gene expression in tissue samples from cancerous and healthy cells and compare the results with a reference clustering [31]. Figure 5 shows an in-depth comparison of the standardized Rand Index using the permutation model and the one-sided fixed number of cluster model for agglomerative hierarchical clustering. We categorize the changes in study assessment into four distinct categories:

TABLE 1. **The number of gene expression studies with the same, the opposite, or a change in significance of the assessment when choosing a non-permutation random model. Particularly when all possible clusterings of a dataset are taken into account, the assessment of the gene expression studies changes drastically.**

	same	opposite	sig. → insig.	insig. → sig.
$\mathrm{SRI}^1_{\mathrm{num}}$	27	4	3	1
$\mathrm{SRI}_{\mathrm{num}}$	24	7	3	1
$\mathrm{SRI}^1_{\mathrm{all}}$	19	10	3	3
$\mathrm{SRI}_{\mathrm{all}}$	3	21	4	7
$\mathrm{SMI}^1_{\mathrm{num}}$	35	0	0	0
$\mathrm{SMI}_{\mathrm{num}}$	35	0	0	0
$\mathrm{SMI}^1_{\mathrm{all}}$	13	4	14	4
$\mathrm{SMI}_{\mathrm{all}}$	0	25	1	9

- *Same assessment* (same): If the either the assessment remained better/worse than chance or it was deemed insignificant ($\mathrm{SRI} \leq 3$) by both variants.
- *Opposite asssessment* (opposite): If the assessment changed from better than chance to worse than chance or vice versa.
- *Significant to insignificant* (sig. → insig.): If the observation is deemed significant ($\mathrm{SRI} > 3$) in the permutation model but insignificant in the other model.
- *Insignificant to significant* (insig. → sig.): If the observation is deemed insignificant in the permutation model but significant in the other model.

Table 1 shows the number of studies that fall into these categories for different random models compared to their respective baseline $\mathrm{SRI}_{\mathrm{perm}}$ or $\mathrm{SMI}_{\mathrm{perm}}$. While the macro assessment remains the same for most studies in the fixed number of clusters model, the all clusterings model offers an entirely different perspective, underlining the importance of choosing an appropriate random model.

Figure 6 shows the Spearman correlation of the various clustering similarity indices when comparing the 12 clustering algorithms for each of the 35 studies. Essentially three different groups of similarity indices emerge: The all clustering adjusted measures, the fixed number of clusters adjusted RI variants, and the commonplace permutation variants with the fixed number of clusters adjusted MI variants. In practice, this means that the choice of MI or RI matters far less than the choice of random model, except in the case of the num model, where it should be carefully considered whether the MI or RI is used. In that case, the study by Romano et al. [16] indicates that RI-based measures

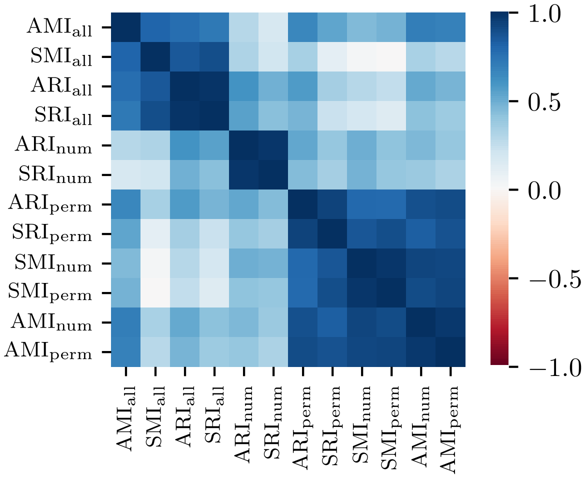


FIGURE 6. Spearman correlation of the adjusted and standardized MI, RI for the permutation, the fixed number of clusters, and the uniform random model on clusterings of gene expression data from 35 studies on cancerous and healthy tissue samples using 12 different clustering algorithms. The MI adjustments using the num model correlate with the perm adjusted measures, whereas the num-adjusted RI and the all-adjusted methods provide a novel interpretation of the results.

are favorable in scenarios with large equal-sized clusters, whereas MI-based measures should be used in scenarios with unbalanced clusterings. However, further research is required to understand if these guidelines hold in the same way for non-permutation based adjustments.

VII. DISCUSSION

This work introduces non-permutation models for the standardization of Mutual Information (MI) and Rand Index (RI). The standardization helps alleviate a bias towards smaller numbers of clusters, that prior non-permutation indices like the ARI and AMI had, particularly in the fixed number of clusters case. Compared to the existing permutation-based SMI, SRI, the adjustments in this work are more suitable for algorithms like k-means and DBSCAN, as they better align with these algorithms' characteristics by accounting for variable cluster sizes or an a priori unknown number of clusters. Other clustering comparison indices, such as the Jaccard Index or the Sokal and Sneath index, that are not based on the MI or RI, are known to exhibit similar biases [13]. Future work could investigate if these biases can be corrected using similar approaches to the ones presented in this work.

Our experiments on gene expression data show that the choice of a suitable random model can drastically influence the interpretability and utility of clustering evaluations. This has practical implications for the process of algorithm selection for clustering, where multiple clustering results are typically compared to a single ground truth reference. We showed that exchaning permutation model with for instance the uniform model over all clusterings, completely changes the interpretation of the clustering results and practitioners might chose a different clustering algorithm based on this analysis. This emphasizes how crucial the careful choice of a random model is and by providing a framework for standardization using any element-symmetric random model and two concrete example models, we provide a good starting point for such considerations.

Broader implications of this work might extend to the design of clustering algorithms themselves. By providing a method to accurately measure the similarity between clusterings under different assumptions about cluster configurations, we enable a more nuanced evaluation of clustering quality. This can guide the development of new algorithms or the refinement of existing ones to better fit the structural characteristics of specific types of data.

VIII. CONCLUSION

When performing multiple comparisons of clusterings against a reference, as is typically the case when comparing different clustering algorithms, standardization is vital to adjust for second-order bias. To this end both SRI and SMI are well-established metrics. However, these metrics are only designed for the random permutation model, which assumes label assignments are shuffled, but otherwise stay the same. Gates and Ahn [18] showed that this model assumption may not be optimal when using the AMI and ARI metrics for some clustering algorithms, such as k-means, spectral clustering, or Gaussian mixture models, all of which assume only that the number of clusters is fixed, but not necessarily their sizes. We extend their work to the SMI and SRI metrics by deriving adjustment equations for two additional random models: one that only fixes the number of clusters and one that allows all possible label assignments. We formulate the adjustment equations to be easily extensible to any other element-symmetric random model.

While it is impossible to make general statements about which random model is the best, we hope to raise awareness with practitioners for the impact the choice of random models has on the evaluation of clustering algorithms and offer additional choices. Results on a gene expression dataset show that the evaluation of SRI and SMI with respect to better than chance outcomes and their significance can change drastically depending on the random model. In particular for clustering algorithms that assume a fixed number of clusterings, such as k-means, we thus postulate that the equivalent random model may be better suited and should at least be considered or compared against instead of the standard random permutation model. To this end we also make our code available such that it may easily be used in practice.

APPENDIX A
PROOFS FOR EXPECTATION VALUES AND VARIANCE

A. RAND INDEX

First we derive the expression for the expected Rand index for any pair of element-symmetric clustering distributions, which is a generalization of the results in [18].

Proof of Theorem 1: The expected Rand Index is by definition

$$
\begin{aligned}
&\mathbb{E}_{U\sim\mathcal{U},V\sim\mathcal{V}}[\mathrm{RI}(U,V)]\\
&\quad=\mathbb{E}_{U\sim\mathcal{U},V\sim\mathcal{V}}\Big[\mathbb{E}_{i,j\sim\mathrm{Unif}\binom{S}{2}}\big[\mathbf{1}_{U(i)=U(j)}\mathbf{1}_{V(i)=V(j)}\\
&\qquad+\mathbf{1}_{U(i)\neq U(j)}\mathbf{1}_{V(i)\neq V(j)}\big]\Big] \qquad (27)\\
&\quad=\mathbb{E}_{i,j\sim\mathrm{Unif}\binom{S}{2}}\big[\mathbb{E}_{U\sim\mathcal{U}}[\mathbf{1}_{U(i)=U(j)}]\mathbb{E}_{V\sim\mathcal{V}}[\mathbf{1}_{V(i)=V(j)}]\\
&\qquad+\mathbb{E}_{U\sim\mathcal{U}}[\mathbf{1}_{U(i)\neq U(j)}]\mathbb{E}_{V\sim\mathcal{V}}[\mathbf{1}_{V(i)\neq V(j)}]\big]. \qquad (28)
\end{aligned}
$$

Now, for any $i, j, i', j' \in S$, there exists a permutation σ such that

$$\mathbb{E}_{U\sim\mathcal{U}}[\mathbf{1}_{U(i')=U(j')}]=\mathbb{E}_{U\sim\mathcal{U}}[\mathbf{1}_{U\circ\sigma(i)=U\circ\sigma(j)}]. \qquad (29)$$

Since $\mathcal{U}$ is element-symmetric, $U' = U \circ \sigma$ has the same probability as U and the expectation value remains unchanged

$$\mathbb{E}_{U\sim\mathcal{U}}[\mathbf{1}_{U(i')=U(j')}]=\mathbb{E}_{U\sim\mathcal{U}}[\mathbf{1}_{U(i)=U(j)}]. \qquad (30)$$

Hence, the inner expectation values in Equation (28) are constants to the outer expectation value, and the statement follows. □

Similarly, we proceed for the variance of the RI under any pair of element-symmetric distributions.

Proof of Theorem 2: First, we introduce two pairs of distinct data points i, j and i', j' in the expectation value in $\mathrm{RI}(U, V)^2$

$$
\begin{aligned}
\mathrm{RI}(U,V)^2&=\mathbb{E}_{i,j\sim\mathrm{Unif}\binom{S}{2}}\big[\mathbf{1}_{U(i)=U(j)}\mathbf{1}_{V(i)=V(j)}\\
&\qquad+\mathbf{1}_{U(i)\neq U(j)}\mathbf{1}_{V(i)\neq V(j)}\big]^2\\
&=\mathbb{E}_{i,j;i',j'\sim\mathrm{Unif}\binom{S}{2}}\big[\big(\mathbf{1}_{U(i)=U(j)}\mathbf{1}_{V(i)=V(j)}\\
&\qquad+\mathbf{1}_{U(i)\neq U(j)}\mathbf{1}_{V(i)\neq V(j)}\big)\cdot\big(\mathbf{1}_{U(i')=U(j')}\mathbf{1}_{V(i')=V(j')}\\
&\qquad+\mathbf{1}_{U(i')\neq U(j')}\mathbf{1}_{V(i')\neq V(j')}\big)\big]. \qquad (31)
\end{aligned}
$$

Then in $\binom{N}{2}$ out of $\binom{N}{2}^2$ cases $i = i'$ and $j = j'$ and we can use Theorem 1 to obtain

$$
\begin{aligned}
&\mathbb{E}_{U\sim\mathcal{U},V\sim\mathcal{V}}[\mathrm{RI}(U,V)^2]\\
&\quad=\binom{N}{2}^{-1}\left(p_{\mathcal{U}}^{=}p_{\mathcal{V}}^{=}+p_{\mathcal{U}}^{\neq}p_{\mathcal{V}}^{\neq}\right)+\left(1-\binom{N}{2}\right)^{-1}\cdot\\
&\quad\mathbb{E}_{U\sim\mathcal{U};V\sim\mathcal{V};(i,j),(i',j')\sim\mathrm{Unif}\left(\binom{S}{2}\right)}\Big[\\
&\quad\mathbf{1}_{U(i)=U(j)}\mathbf{1}_{U(i')=U(j')}\mathbf{1}_{V(i)=V(j)}\mathbf{1}_{V(i')=V(j')}\\
&\qquad+\mathbf{1}_{U(i)=U(j)}\mathbf{1}_{U(i')\neq U(j')}\mathbf{1}_{V(i)=V(j)}\mathbf{1}_{V(i')\neq V(j')}\\
&\qquad+\mathbf{1}_{U(i)\neq U(j)}\mathbf{1}_{U(i')=U(j')}\mathbf{1}_{V(i)\neq V(j)}\mathbf{1}_{V(i')=V(j')}\\
&\qquad+\mathbf{1}_{U(i)\neq U(j)}\mathbf{1}_{U(i')\neq U(j')}\mathbf{1}_{V(i)\neq V(j)}\mathbf{1}_{V(i')\neq V(j')}\Big] \qquad (32)
\end{aligned}
$$

Now, with the same argument as in Proof of Theorem 1, since $\mathcal{U}, \mathcal{V}$ are element-symmetric, the expectation value factorizes, giving

$$
\begin{aligned}
&\mathbb{E}_{U\sim\mathcal{U},V\sim\mathcal{V}}[\mathrm{RI}(U,V)^2]\\
&\quad=\binom{N}{2}^{-1}\left(p_{\mathcal{U}}^{=}p_{\mathcal{V}}^{=}+p_{\mathcal{U}}^{\neq}p_{\mathcal{V}}^{\neq}\right)+\left(1-\binom{N}{2}\right)^{-1}\cdot\\
&\quad\left(p_{\mathcal{U}}^{=,=}p_{\mathcal{V}}^{=,=}+p_{\mathcal{U}}^{=,\neq}p_{\mathcal{V}}^{=,\neq}+p_{\mathcal{U}}^{\neq,=}p_{\mathcal{V}}^{\neq,=}+p_{\mathcal{U}}^{\neq,\neq}p_{\mathcal{V}}^{\neq,\neq}\right).
\end{aligned}
\qquad (33)
$$

The statement follows as $p_{\mathcal{U}}^{=,\neq} = p_{\mathcal{U}}^{\neq,=}$. □

B. MUTUAL INFORMATION

Similar to the proofs for the Rand Index, we derive the expressions for the first and second statistical moment of the mutual information under any pair of element-symmetric clustering distributions.

Proof of Theorem 3: Nguyen et al. [12] showed that the expected mutual information under the random permutation model is

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{perm}}[\mathrm{MI}(U,V)]\\
&\quad=\sum_{i,j}\sum_{n=\max(u_i+v_i-N,0)}^{\min(u_i,v_i)}\frac{n}{N}\log\left(\frac{nN}{u_iv_i}\right)\mathrm{Hyp}(n|u_i,v_i,N)\\
&\quad=\sum_{u,v}p_{\mathrm{perm}}^{U}(u)p_{\mathrm{perm}}^{V}(v)\sum_{n}\frac{nN}{uv}\log\left(\frac{nN}{uv}\right)\mathrm{Hyp}(n|u,v,N).
\end{aligned}
\qquad (34)
$$

Since $\mathcal{U}, \mathcal{V}$ are element-symmetric, we can introduce an expectation value under random permutation as follows

$$\mathbb{E}_{U\sim\mathcal{U},V\sim\mathcal{V}}[\mathrm{MI}(U,V)]=\mathbb{E}_{U\sim\mathcal{U},V\sim\mathcal{V}}[\mathbb{E}_{\mathrm{perm}}[\mathrm{MI}(U,V)]]. \qquad (35)$$

The only U, V dependence in Equation (34) is in $p_{\mathrm{perm}}^{U}(u), p_{\mathrm{perm}}^{V}(v)$ and

$$
\begin{aligned}
&\mathbb{E}_{U\sim\mathcal{U}}[p_{\mathrm{perm}}^{U}(u)]\\
&\quad=\mathbb{E}_{U\sim\mathcal{U}}[\mathbb{E}_{\sigma\in S_N,i\sim\mathrm{Unif}(S)}[\mathbf{1}_{|\sigma^{-1}\circ U^{-1}\circ U\circ\sigma(i)|=u}]]\\
&\quad=\mathbb{E}_{U'\sim\mathcal{U},i\sim\mathrm{Unif}(S)}[\mathbf{1}_{|U'^{-1}\circ U'(i)|=u}]=p_{\mathcal{U}}(u), \qquad (36)
\end{aligned}
$$

since $\mathcal{U}$ is element-symmetric. □

Proof of Theorem 4: Theorem 1 in [15] gives an explicit expression for the second moment of the mutual information

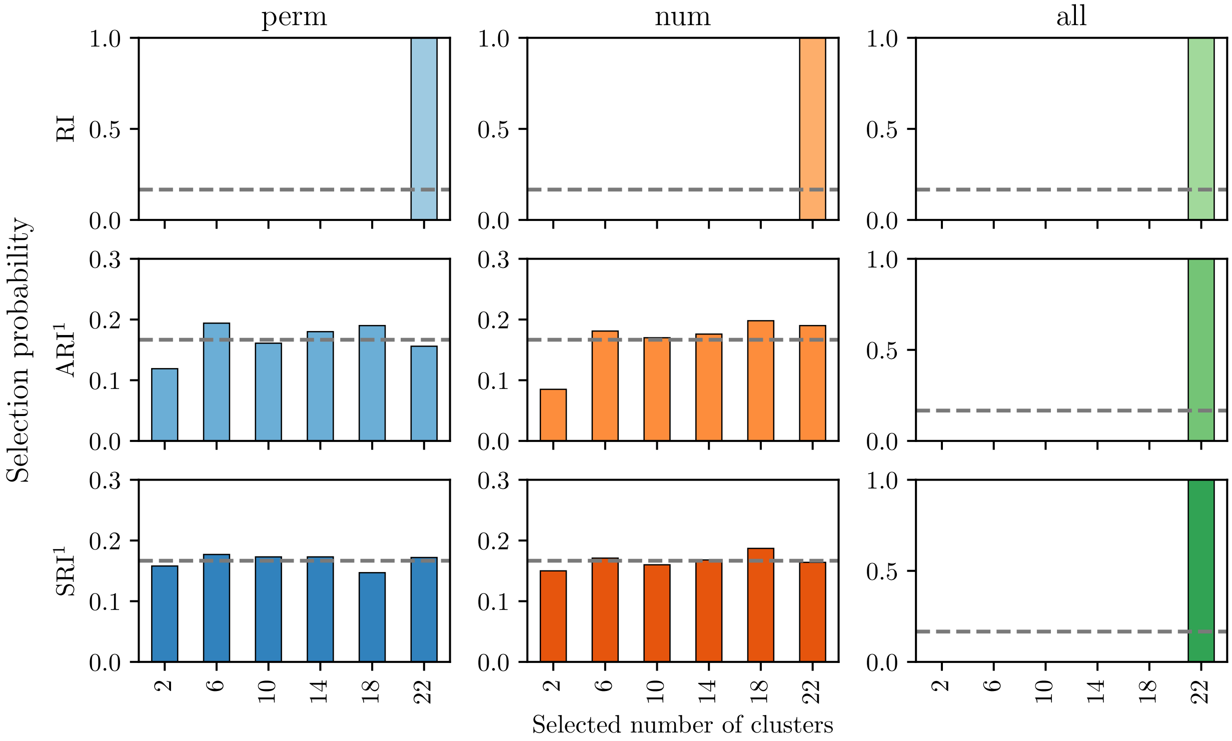


FIGURE 7. We compare a fixed reference clustering with $k_{\mathcal{U}} = 10$ even clusters, to random clusterings with $k_{\mathcal{V}} \in \{2, 6, 10, 14, 18, 22\}$ clusters. The plot shows the selection probabilities of each $k_{\mathcal{V}}$ for the RI, ARI^1 and SRI for the permutation model (perm), the fixed number of clusters model (num) and the all clusterings model. The results are in line with the observations for the respective MI variants in Figure 3.

under random permutation

$$\begin{aligned}
&\mathbb{E}_{\text{perm}}[\text{MI}(U,V)^2] \\
&= \sum_{i,j}\sum_{n} \frac{n}{N}\log\left(\frac{nN}{u_iv_i}\right)\text{Hyp}(n|u_i,v_j,N)\cdot\Bigg[\frac{n}{N}\log\left(\frac{nN}{u_iv_j}\right) \\
&\quad+\sum_{i'\neq i}\sum_{n'} \frac{n'}{N}\log\left(\frac{n'N}{u_{i'}v_j}\right)\text{Hyp}(n'|u_{i'},v_j-n,N-u_i) \\
&\quad+\sum_{j'\neq j}\sum_{n'}\text{Hyp}(n'|u_i-n,v_{j'},N-v_j)\Bigg(\frac{n'}{N}\log\left(\frac{n'N}{u_iv_{j'}}\right) \\
&\quad+\sum_{i'\neq i}\sum_{n''}\frac{n''}{N}\log\left(\frac{n''N}{u_{i'}v_{j'}}\right)\text{Hyp}(n''|u_{i'},v_{j'}-n',N-u_i)\Bigg)\Bigg].
\end{aligned} \tag{37}$$

Analogous to Equation (34) we can replace $\sum_i f(u_i)$ with $\sum_u p^U_{\text{perm}}(u)\frac{N}{u}f(u)$. For the sums where the i-th clustering is excluded, we can replace $\sum_{i'\neq i} f(u_i, u_{i'})$ with $\sum_{u'} q^U_{\text{perm}}(u'|u)\frac{N}{u'}f(u,u')$, as by definition it explicitly excludes the case where u and u' refer to the same cluster. Using the fact that $\mathcal{U}, \mathcal{V}$ are element-symmetric, we can then apply the same argument as in Proof of Theorem 3 Equations (35) - (36) and the statement follows. □

APPENDIX B
SYNTHETIC EXPERIMENTS FOR THE RAND INDEX

For completeness, we repeat the same experiment as in Figure 3 also for Rand Index based similarity indices, modelling the external validation setup. A fixed ground truth reference of $k_U = 10$ evenly sized clusters is compared to randomly generated clusterings with $k_V \in \{2, 6, 10, 14, 18, 22\}$ clusters. We then select the clustering with the highest similarity and record the selection probability for each k_V over 5000 trials for the RI, ARI^1, SRI^1 and all three random models. The results in Figure 7 are in line with the observations in section V. The fixed number of clusters model benefits from standardization, while the all number of clusters model is too general for this scenario.

REFERENCES

[1] S. Zhang, Z. Yang, X. Xing, Y. Gao, D. Xie, and H.-S. Wong, "Generalized pair-counting similarity measures for clustering and cluster ensembles," *IEEE Access*, vol. 5, pp. 16904–16918, 2017. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8012357

[2] C. Aggarwal and C. Reddy, *Data Clustering Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, Aug. 2013.

[3] M. Alhawarat and M. Hegazi, "Revisiting K-means and topic modeling, a comparison study to cluster Arabic documents," *IEEE Access*, vol. 6, pp. 42740–42749, 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8402221

[4] H. Jia, J. Ma, and W. Song, "Multilevel thresholding segmentation for color image using modified moth-flame optimization," *IEEE Access*, vol. 7, pp. 44097–44134, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8678762

[5] L. Antony, S. Azam, E. Ignatious, R. Quadir, A. R. Beeravolu, M. Jonkman, and F. D. Boer, "A comprehensive unsupervised framework for chronic kidney disease prediction," *IEEE Access*, vol. 9, pp. 126481–126501, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9525378

[6] E. Müller, S. Günnemann, I. Färber, and T. Seidl, "Discovering multiple clustering solutions: Grouping objects in different views of the data," in *Proc. Proc. 10th IEEE Int. Conf. Data Mining (ICDM)*, G. I. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, Eds. Sydney, NSW, Australia: IEEE Computer Society, Dec. 2010, p. 1220, doi: 10.1109/ICDM.2010.85.

[7] S. Wei, G. Han, R. Wang, Y. Yang, H. Zhang, and S. Li, "Inductive multi-view multiple clusterings," in *Proc. 7th Int. Conf. Big Data Inf. Anal. (BigDIA)*, Oct. 2021, pp. 308–315.

[8] X. Su, S. Xue, F. Liu, J. Wu, J. Yang, C. Zhou, W. Hu, C. Paris, S. Nepal, D. Jin, Q. Z. Sheng, and P. S. Yu, "A comprehensive survey on community detection with deep learning," 2021, *arXiv:2105.12584*.

[9] D. Jin, Z. Yu, P. Jiao, S. Pan, D. He, J. Wu, P. S. Yu, and W. Zhang, "A survey of community detection approaches: From statistical modeling to deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 2, pp. 1149–1170, Feb. 2023, doi: 10.1109/TKDE.2021.3104155.

[10] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, Dec. 2010.

[11] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, Dec. 1971, doi: 10.1080/01621459.1971.10482356.

[12] X. V. Nguyen, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Is a correction for chance necessary?" in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, vol. 382, montreal, QC, Canada, A. P. Danyluk, L. Bottou, and M. L. Littman, Eds., 2009, pp. 1073–1080, doi: 10.1145/1553374.1553511.

[13] Y. Lei, J. C. Bezdek, S. Romano, N. X. Vinh, J. Chan, and J. Bailey, "Ground truth bias in external cluster validity indices," *Pattern Recognit.*, vol. 65, pp. 58–70, May 2017, doi: 10.1016/j.patcog.2016.12.003.

[14] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985, doi: 10.1007/bf01908075.

[15] S. Romano, J. Bailey, V. Nguyen, and K. Verspoor, "Standardized mutual information for clustering comparisons: One step further in adjustment for chance," in *Proc. 31st Int. Conf. Mach. Learn.*, Jun. 2014, pp. 1143–1151. [Online]. Available: https://proceedings.mlr.press/v32/romano14.html

[16] S. Romano, N. X. Vinh, J. Bailey, and K. Verspoor, "Adjusting for chance clustering comparison measures," *J. Mach. Learn. Res.*, vol. 17, no. 134, pp. 1–32, 2016. [Online]. Available: http://jmlr.org/papers/v17/15-627.html

[17] K. Klede, T. Altstidl, D. Zanca, and B. Eskofier, "*p*-value adjustment for monotonous, unbiased, and fast clustering comparison," in *Proc. Adv. Neural Inf. Process. Syst. 36 (NeurIPS)*, vol. 36, Dec. 2023, pp. 27113–27128. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/-Abstract-Conference.html

[18] A. J. Gates and Y.-Y. Ahn, "The impact of random models on clustering similarity," *J. Mach. Learn. Res.*, vol. 18, pp. 1–28, 2017. [Online]. Available: http://jmlr.org/papers/v18/17-039.html

[19] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, Oakland, CA, USA, 1967, pp. 281–297.

[20] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Proc. KDD*, vol. 96, no. 34, pp. 226–231, 1996.

[21] M. Gösgens, A. Tikhonov, and L. Prokhorenkova, "Systematic analysis of cluster similarity indices: How to validate validation measures," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, vol. 139, M. Meila and T. Zhang, Eds., 2021, pp. 3799–3808. [Online]. Available: http://proceedings.mlr.press/v139/gosgens21a.html

[22] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics*. Reading, MA, USA: Addison-Wesley, 1988.

[23] J. L. DuBien, W. D. Warde, and S. S. Chae, "Moments of rand's c statistic in cluster analysis," *Statist. Probab. Lett.*, vol. 69, no. 3, pp. 243–252, Sep. 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167715204001646

[24] J. H. Conway and R. K. Guy, "Famous families of numbers: Bell numbers and stirling numbers," in *The Book of Numbers*. Cham, Switzerland: Springer, 1996, pp. 91–94.

[25] K. Klede, L. Schwinn, D. Zanca, and B. M. Eskofier, "FastAMI- aMonte Carlo approachto the adjustment for chancein clustering comparison metrics," in *Proc. AAAI Conf. Artif. Intell.*, Washington, DC, USA, B. Williams, Y. Chen, and J. Neville, Eds., Feb. 2023, pp. 8317–8324, doi: 10.1609/aaai.v37i7.26003.

[26] D. Lazarenko and T. Bonald, "Pairwise adjusted mutual information," 2021, *arXiv:2103.12641*.

[27] W. M. Patefield, "Algorithm AS 159: An efficient method of generating random R × C tables with given row and column totals," *J. Roy. Stat. Soc. C*, vol. 30, no. 1, pp. 91–97, 1981.

[28] D. Knuth, "The complexity of nonuniform random number generation," in *Proc. Symp. New Directions Recent Results Algorithms Complexity*. The Computer Science Department, Carnegie-Mellon University, Apr. 1976, pp. 357–428.

[29] R. Arratia and S. DeSalvo, "Probabilistic divide-and-conquer: A new exact simulation method, with integer partitions as an example," 2011, *arXiv:1110.3856*.

[30] S. DeSalvo and J. Zhao, "Random sampling of contingency tables via probabilistic divide-and-conquer," *Comput. Statist.*, vol. 35, no. 2, pp. 837–869, Jun. 2020, doi: 10.1007/s00180-019-00899-7.

[31] M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: A comparative study," *BMC Bioinf.*, vol. 9, no. 1, p. 497, Nov. 2008, doi: 10.1186/1471-2105-9-497.

KAI KLEDE received the B.Sc. and M.Sc. degrees in physics from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree in machine learning with the Machine Learning and Data Analytics Laboratory, Department Artificial Intelligence in Biomedical Engineering.

From January to March 2024, he was a Visiting Researcher with the Computer Laboratory, University of Cambridge, U.K. He has published papers in top-tier AI conferences, such as NeurIPS and AAAI. His research interests include clustering comparison measures, data cleaning, healthcare fraud detection, and the application of machine learning in various domains.

Mr. Klede received the Ohm Prize for an outstanding bachelor's thesis in 2019, a full scholarship from German Academic Scholarship Foundation from 2016 to 2021, and a DAAD IFI scholarship for his research visit to Cambridge in 2024.

THOMAS R. ALTSTIDL received the B.Sc. and M.Sc. degrees in computer science from Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree with the Machine Learning and Data Analytics Laboratory.

From 2017 to 2018, he was a Teaching Assistant with the Distributed Systems and Operating Systems Group, Friedrich-Alexander-Universität Erlangen-Nürnberg. From 2019 to 2021, he held the position as a Research Assistant with the Machine Learning and Data Analytics Laboratory, Friedrich-Alexander-Universität Erlangen-Nürnberg, and the Fraunhofer Society's Precise Positioning and Analytics Department. He has authored several conference papers, including publications at the International Joint Conference on Neural Networks (IJCNN), the IEEE Engineering in Medicine and Biology Conference (EMBC), and the Conference on Neural Information Processing Systems (NeurIPS). His work spans diverse areas, such as computer vision, biomedical engineering, and theoretical machine learning. His research interests include scale equivariance in convolutional neural networks, deep learning applications in medical imaging, clustering comparison methods, and machine learning for signal processing.

DARIO ZANCA received the B.S. degree in mathematics for scientific communication and the M.S. degree in mathematics from the University of Palermo, Italy, in 2013 and 2015, respectively, and the Ph.D. degree in smart computing from the University of Florence, Italy, in 2019.

From 2019 to 2020, he was a Postdoctoral Researcher with the NeuroSense Joint Laboratory, Department of Medicine, Surgery and Neuroscience, University of Siena, Italy. Since 2020, he has been a Postdoctoral Researcher and the Head of the Applied Machine Learning Group, Machine Learning and Data Analytics Laboratory, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany. He is the author of more than 30 articles. His research interests include computer vision, machine learning, and computational neuroscience, focusing on visual attention and eye movements.

Dr. Zanca received the Emergent Talents Initiative (ETI) grant for research on ‘‘Human-Inspired Computer Vision’’ in 2023 and received a special mention at the ‘‘Premio Marco Cadoli’’ 2020 for the Best Ph.D. Thesis in Artificial Intelligence.

BJÖRN M. ESKOFIER (Senior Member, IEEE) received the degree in electrical engineering from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), in 2006, and the Ph.D. degree in biomechanics from the University of Calgary, Calgary, AB, Canada, under the supervision of Prof. Dr. Benno Nigg.

In 2016, he was a Visiting Professor with the Prof. Paolo Bonato’s Motion Analysis Laboratory, Harvard Medical School (February–March). In 2018, he was a Visiting Professor with the Prof. Alex Sandy Pentland’s Human Dynamics Group, MIT Media Laboratory (March–August). Since April 2023, he has been an Associate Principal Investigator and the Leader of the Research Group Translational Digital Health, Helmholtz Zentrum München. From April 2023 to August 2023, he was a Visiting Professor with the Prof. Scott Delp’s NMBL Laboratory that is part of Stanford University’s Schools of Engineering and Medicine. He currently heads the Machine Learning and Data Analytics (MaD) Laboratory, FAU, Erlangen, Germany. He is also the founding Spokesperson of the Department Artificial Intelligence in Biomedical Engineering, FAU, and German Ministry of Economic Affairs and Climate Action GAIA-X usecase Project TEAM-X, and the Co-Spokesperson of German Research Foundation Collaborative Research Center EmpkinS. He authored more than 400 peer-reviewed articles, holds five patents, started three spinoff startup companies, and is in a supporting role for further startups.

Dr. Eskofier received several medical-technical research awards, including the Curious Mindsaward 2021 in Life Sciences by Manager Magazin and Merck. He is also active in the organization of several IEEE and ACM meetings (e.g., BSN, BHI, EMBC, IJCAI, ISWC, and UbiComp), most recently as the General Chair of BHI 2023. He was the Area Editor of the IEEE Open Journal of Engineering in Medicine and Biology and an Associate Editor of the IEEE Journal of Biomedical and Health Informatics.