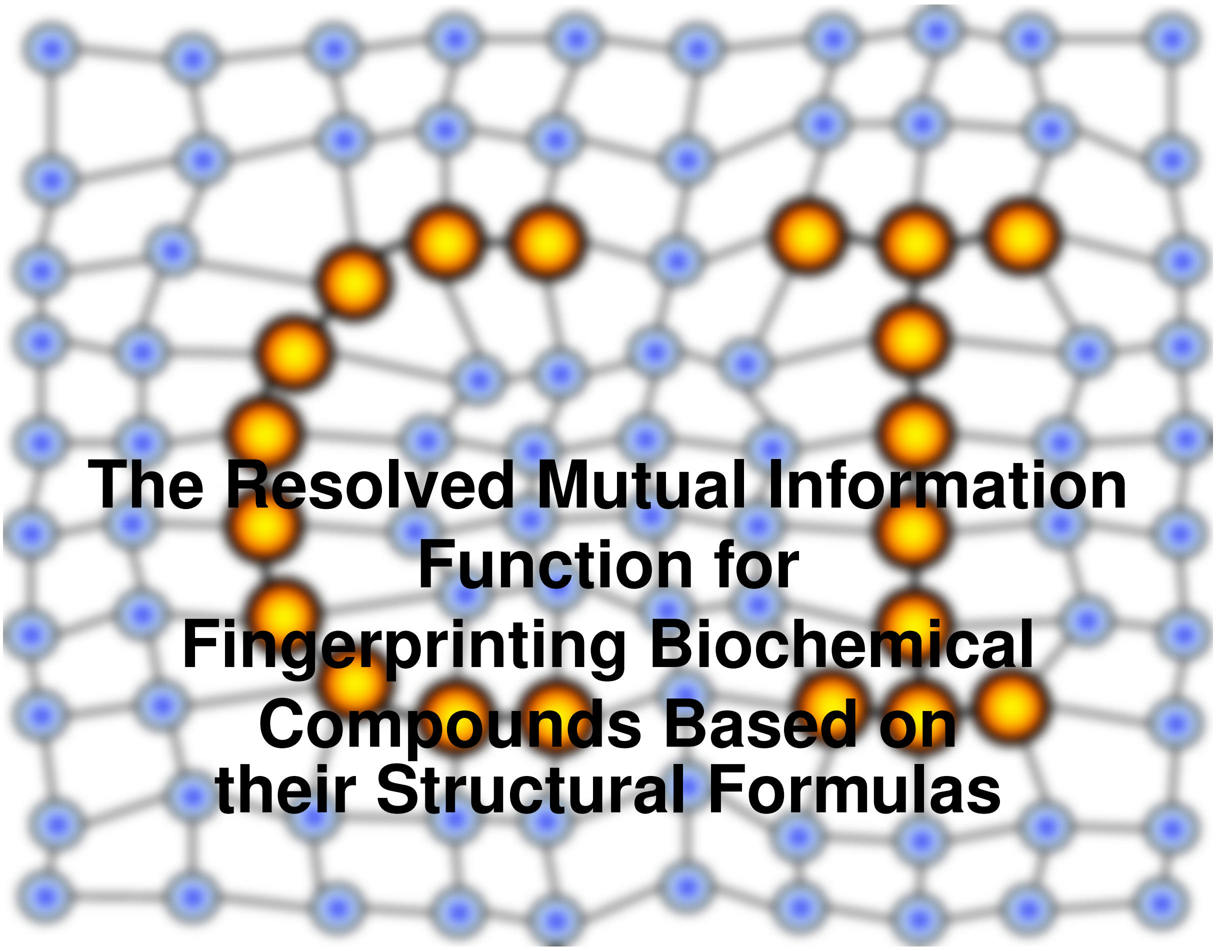
MACHINE LEARNING REPORTS



The Resolved Mutual Information Function for Fingerprinting Biochemical Compounds Based on their Structural Formulas

Report 01/2022
Submitted: 23.12.2021
Published: 31.12.2021

K. S. Bohnsack [1,*], M. Kaden [1], T. Villmann [1]
(1) Saxon Institute for Computational Intelligence and Machine Learning (SICIM)
Hochschule Mittweida
Mittweida, Germany

Machine Learning Reports
http://www.techfak.uni-bielefeld.de/∼fschleif/mlr/mlr.html

Impressum

Publisher:	University of Applied Sciences Mittweida Technikumplatz 17, 09648 Mittweida, Germany
Editor:	Prof. Dr. Thomas Villmann Prof. Dr. Frank-Michael Schleif
Techical-Editor:	Prof. Dr. Frank-Michael Schleif
Contact:	fschleif@techfak.uni-bielefeld.de
URL:	http://techfak.uni-bielefeld.de/∼ fschleif/mlr/mlr.html
ISSN:	1865-3960

The Resolved Mutual Information Function for Fingerprinting Biochemical Compounds Based on their Structural Formulas

Katrin Sophie Bohnsack, Marika Kaden and Thomas Villmann

University of Applied Sciences Mittweida

Saxon Institute for Computational Intelligence and Machine Learning (SICIM)

Abstract

In this technical report we present an extension of the (resolved) mutual information function, which can be applied to graph structures. In particular, it can be used to extract information theoretic features to characterize, i.e. fingerprint biochemical compounds given as structural formulas.

1 Introduction

Molecular fingerprinting is an essential step in ligand-based virtual screening approaches which aims at identifying similar compounds from a data set of known active molecules [6]. These molecules, usually are given as structural formulas in databases of chemical compounds like DrugBank [40] or ChEMBL [18]. Hence, the comparison of molecules for similarity search as well as for advanced data analyses including machine learning methods requires their characterization by numerical features reflecting physico-chemical properties. This feature extraction is known as molecular fingerprinting in this context.

Several possibilities to categorize fingerprints have been proposed [14, 6, 10]. According to [6] one can distinguish at least three basic principles:

1. noting the presence of certain substructures or features within the compound like e.g. MACCS (Molecular ACCess System) keys [15] or PubChem fingerprints [5],

2. analyzing the compounds topology in terms of possible paths like e.g. daylight fingerprints,

3. considering the environment of atoms in terms of a defined radius like e.g. ECFPs (Extended Connectivity Fingerprints) [33].

Apart from these chemical descriptors, compounds can also be referred to by line notations [10] such as SMILES (Simplified Molecular Input Line Entry System) [39] or InChi (International Chemistry Identifier) [22].

Generally, structural formulas can be processed mathematically as planar graphs such that a comparison can be made in terms of mathematical graph isomorphism. Respective algorithms are known to be linear in time [25], i.e. the time complexity is $O(|V|)$ where $|V|$ denotes the number of the vertices in the graphs to be compared. In general, the graph isomorphism problem neither provides any similarity measure for graph comparison nor it is solved for weighted graphs. Otherwise, many similarity measures based on mathematical properties of graphs are known [36].

A promising alternative for molecule graphs is to use information theoretic features characterizing molecule inherent relations, context information and topological attributes. In [19] an information theoretic concept for fingerprinting individual molecules based on their topological feature distributions is introduced: The shortest paths between all nodes of a structural formula graph are calculated, followed by determining the entropy of the path length distribution for each feature pair. Besides the well-known Shannon entropy, a Rényi variant fingerprint is introduced in a follow-up publication [12].

Another information theoretic application in [8] involves use of the point-wise mutual information (PMI) known from computational linguistics [7] to study the interrelation between structural features within molecules from a compound set. This interrelation profiling builds on available structure keys and aims at characterizing chemical databases, rather then characterizing individual molecules.

In [37] the Shannon entropy-based fingerprint similarity search is introduced. The idea behind is to transform each compound of a reference set into a binary fingerprint representation and to calculate the sets entropy position-wise. Subsequently, the entropy is recalculated for individual test compounds added to the set and from the resulting differences in entropy conclusions about molecule similarity can be drawn.

In nucleotide and protein *sequence analysis* the mutual information function (MIF) has gained attraction as an useful tool [11, 20, 26, 29], which originally was introduced in [28] for general symbolic sequences. The MIF reflects the short and long term correlations in sequences based on Shannon's entropic information [23, 3, 35]. In the bioinformatics context, MIF is also known as average mutual information (AMI) profile [1]. Recently, the resolved mutual information function (rMIF) was proposed as a variant of MIF for both the Shannon and the Rényi definition of entropy [4] providing more fine-grained structure correlation insights than MIF. The above mentioned PMI is closely related and was applied to molecular sequences, too [32].

The aim of this paper is to extend the rMIF as an information theoretic fingerprint of chemical compounds, i.e. extracting the correlation information of the corresponding molecule graphs. For this purpose, a molecule graph is treated as a weighted graph such that atom correlations are evaluated based on the respective shortest path lengths between them.

The outline of the paper is as follows: First, we describe the graph-based rMIF (grMIF). Thereafter, we relate the grMIF to molecule graphs, i.e. to the context of chemical compounds, followed by implications for future research.

2 Variants of the Mutual Information Function for Weighted Labeled Graphs

We consider a connected undirected *graph* $G = (V, E)$ with the set of vertices (nodes) $V = \{v_i\}_{i \in 1 \ldots N}$ and the corresponding set $E \subseteq V \times V$ of edges. In the view of molecular graphs we suppose graphs without self-loops.

Each vertex v_i is equipped with a label $l(v_i) \in \mathcal{A}$ by the label function $l : V \to \mathcal{A}$, where $\mathcal{A}$ is as discrete finite label set. This set can be seen as available class labels for the object to be considered, e.g. the classes of molecular units for molecule graphs.

A weight $w(e_i) \in \mathbb{R}$ is assigned to each $e_i \in E$ by means of the weighting function

$w : E \to \mathbb{R}$.[1] Unweighted graphs can be taken as weighted graphs with equal weights for all edges usually set to be one. Thus, the graph is taken as a tuple $G^{V,E,\mathcal{A}}_{l,w} = (V, E, \mathcal{A}, l, w)$.

The calculation of the rMIF for molecular sequences assumes discrete spatial distances between the contained molecular units (object classes). This is achieved by determination of shortest paths with corresponding shortest path lengths, which may be followed by an explicit discretization and normalization. Afterwards, the resulting discretized distances are used for the calculation of the joint and marginal probability distributions of unit pairs in dependence on their spatial relation. Finally, the resulting probabilities serve as input for the calculation of the rMIF, which is denoted as graph resolved mutual information function (grMIF) in this context. In the following subsections we explain these steps.

2.1 Identification of shortest paths between nodes in graphs

A *path* is an alternating series of vertices and edges of G, starting and ending with a vertex, in which each edge is incident with the vertices immediately preceding and immediately following it and each vertex is traversed only once. A *shortest path* is a path between two vertices minimizing the sum of its edge weights.

Here, we are faced with the all-pairs shortest paths problem, i.e. the shortest path determination between every two nodes. The resulting matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ contains the minimum lengths s_{ij} between the nodes v_i and v_j.

For unweighted graphs, the value s_{ij} is just the number of edges n_{ij} separating the nodes v_i and v_j, which can be determined using a breadth-first search algorithm (BFS) [31, 9]. This strategy fails for weighted graphs.

For those graphs, depending on the weight values determined by the weight function w and the topological properties of the considered graph, restrictions for efficient calculation of the matrix $\mathbf{S}$ apply: The Dijkstra algorithm is popular because of its simplicity [13]. If runtime is a major concern, the A* algorithm should be considered as an heuristic extension of Dijkstra [21]. If negative weights are to be considered, the slower Bellman-Ford algorithm has to be applied, which can even detect negative cycles [2, 17, 34], while the Floyd-Warshall algorithm fails at the latter [16, 38].

Note that the the value s_{ij} in weighted graphs is sometimes denoted as *flow*.

[1] In case of directed graphs, a second weighting function is necessary to weight edges in dependence on their direction. Note that for molecular graphs only undirected graphs are considered.

2.2 Dicretization of shortest paths lengths in weighted graphs

The calculation of the rMIF for *molecular sequences* takes discrete spatial distances between the contained molecular units (in terms of the label set $\mathcal{A}$), e.g. nucleotides or amino acids. Thus, the utilization of rMIF for weighted graphs requires a respective discretization of the shortest path lengths s_{ij}. This can be achieved by various strategies explained in the following without any claim to completeness but motivated by the biochemical applications in mind.

- **Number of edges corresponding to a least flow:** A simple approach is to define the discretized distances $d_{ij} \in \mathbb{N}$ as the minimum number of edges in a shortest path corresponding to the flow value s_{ij}. All distances d_{ij} form the discretized distance matrix $\mathbf{D}$.

- **Dominating distance of shortest path:** Let $s_{\max} = \max_{ij}(s_{ij})$ and $s_{\min} = \min_{ij}(s_{ij})$ be the maximum and the minimum flows in G, respectively.

  A *partition* $\mathcal{P}_n$ of the closed interval $[s_{\min}, s_{\max}]$ is determined by a set $P = \{\zeta_0, \zeta_1, \zeta_2, \ldots, \zeta_n\}$ of values $\zeta_i < \zeta_j$ for $i < j$ such that $s_{\min} = \zeta_0$ and $s_{\max} = \zeta_n$. Consequently, the interval $[s_{\min}, s_{\min}]$ can be seen as the union

  $$\cup_{k=1}^{n} I_k = [\zeta_0, \zeta_1] \cup (\zeta_1, \zeta_2] \cup \ldots \cup (\zeta_{n-1}, \zeta_n]$$

  of the intervals I_k.

  We define the discrete distance d_{ij} regarding the flow s_{ij} as the dominating distance with respect to $\mathcal{P}_n$, i.e. d_{ij} takes the value ζ_k for which $s_{ij} \in I_k$ holds. As before, all d_{ij} generate the discretized distance matrix $\mathbf{D}$.

  The granularity of the partition taken as the number n of intervals I_k is subject to choice as well as the distribution of the boundary values ζ_i under the constraints of $\zeta_i < \zeta_j$ for $i < j$ being valid. Frequently, equi-distances are used.

We remark at this point that $\mathbf{D}$ is symmetric with zero diagonals in case of undirected graphs. Further, it is worth to be mentioned that the final calculation of the (discretized) distances based on the flow values may incorporate explicit domain knowledge of the application in mind.

Let the set $T(\mathbf{D}) = \{\tau_1, \ldots, \tau_M\}$ be the minimum ordered set of values τ_k with $\tau_k < \tau_{k+1}$ such that all matrix entries d_{ij} of the distance matrix $\mathbf{D}$ are contained in $T(\mathbf{D})$, i.e. $d_{ij} \in T(\mathbf{D})$ holds. We denote $T(\mathbf{D})$ as the *support* of $\mathbf{D}$.

2.3 Calculation of the joint and marginal probabilities and the mutual information function variants

Let $G = G_{l,w}^{V,E,\mathcal{A}}$ be a given graph. The quantity

$$p_{\tau}\left(l_i, l_j\right) = p\left(l_i = l\left(v_k\right), l_j = l\left(v_m\right) | d_{km} = \tau,\ k, m = 1, \ldots, N\right)$$

denotes the joint probability of co-occurence of the labels $l_i, l_j \in \mathcal{A}$ at distance $\tau \in T\left(\mathbf{D}\right)$ regarding the graph G with the corresponding discretized distance matrix $\mathbf{D}$ and its support $T\left(\mathbf{D}\right)$, as defined previously. Let $p\left(l_i\right)$ be the overall probability of the label $l_i \in \mathcal{A}$ in the graph G. Further, let $p\left(l_i, \tau\right) = \sum_j p_{\tau}\left(l_i, l_j\right)$ be the marginal probability of the label l_i in front of a node pair with distance τ between the nodes. Analogously, the marginal probability $q\left(l_j, \tau\right) = \sum_i p_{\tau}\left(l_i, l_j\right)$ is the probability of label l_j in the back of a node pair with distance τ.

Using the τ-dependent marginals $p\left(l_i, \tau\right)$ and $q\left(l_j, \tau\right)$ the Shannon Mutual information function (MIF) is given as

$$F(G, \tau) = \sum_{l_i \in \mathcal{A}} \sum_{l_j \in \mathcal{A}} p_{\tau}\left(l_i, l_j\right) \cdot \log\left(\frac{p_{\tau}\left(l_i, l_j\right)}{p\left(l_i, \tau\right) \cdot q\left(l_j, \tau\right)}\right)$$

whereas the Rényi mutual information function (RMIF) is obtained as

$$F_{\alpha}^{\mathrm{R}}(G, \tau) = \sum_{l_i \in \mathcal{A}} \sum_{l_j \in \mathcal{A}} \frac{\left(p_{\tau}\left(l_i, l_j\right)\right)^{\alpha}}{\left(p\left(l_i, \tau\right) \cdot q\left(l_j, \tau\right)\right)^{\alpha - 1}}$$

in agreement with [4]. The resolved mutual information function (rMIF) is defined as

$$F(l_i, \tau) = \sum_{l_j \in \mathcal{A}} p_{\tau}\left(l_i, l_j\right) \cdot \log\left(\frac{p_{\tau}\left(l_i, l_j\right)}{p\left(l_i, \tau\right) \cdot q\left(l_j, \tau\right)}\right)$$

and the resolved Rényi mutual information function (rRMIF)

$$F_{\alpha}^{\mathrm{R}}(l_i, \tau) = \sum_{l_j \in \mathcal{A}} \frac{\left(p_{\tau}\left(l_i, l_j\right)\right)^{\alpha}}{\left(p\left(l_i, \tau\right) \cdot q\left(l_j, \tau\right)\right)^{\alpha - 1}}$$

is calculated accordingly.

If the τ-dependence is dropped, one has to replace $p\left(l_i, \tau\right)$ by $p\left(l_i\right)$ and $q\left(l_j, \tau\right)$ by $p\left(l_j\right)$, respectively.

Further, we can consider the quantities

$$F(l_i, l_j, \tau) = p_{\tau}\left(l_i, l_j\right) \cdot \log\left(\frac{p_{\tau}\left(l_i, l_j\right)}{p\left(l_i, \tau\right) \cdot q\left(l_j, \tau\right)}\right)$$

and

$$F_{\alpha}^{\mathrm{R}}(l_i, l_j, \tau) = \frac{\left(p_{\tau}\left(l_i, l_j\right)\right)^{\alpha}}{\left(p\left(l_i, \tau\right) \cdot q\left(l_j, \tau\right)\right)^{\alpha - 1}}$$

which are just the point-wise mutual informations proposed in [7] and used by [8, 32] in other contexts.

For a deeper derivation of the rMIF concepts and a placement in the larger context of information theory we refer to the paper [4].

3 Adaptation of grMIF to Structural Formulas

We consider a structural formula represented by the graph $G_{l,w}^{V,E,\mathcal{A}} = (V, E, \mathcal{A}, l, w)$ where the label function $l : V \to \mathcal{A}$ assigns IUPAC (International Union of Pure and Applied Chemistry) atom types to each node in the graph and $w : E \to \mathbb{R}$ is an edge weighting function incorporating domain knowledge. For the pure structural formula it is just a numerical representation of the bonding types (order), e.g. single (1), aromatic (1.5) and double (2). Alternatively, different weightings could be applied reflecting other relationships between the atoms. The label alphabet $\mathcal{A}$ has to be chosen in dependence on the given task. For example, investigations of simple organic compounds may use atoms from $\mathcal{A} = \{\,\text{S,C,H,N,O,P}\,\}$. Otherwise, the hydrogen atoms could be dropped because respective bounding information may be derived from the molecule graph. Alternatively, we could label the nodes of the molecule graph according to the atom's physico-chemical properties, e.g. as hydrophobic (H), aromatic (R), acceptor (A) and donor (D). These were used in [19, 12].

4 Conclusion and Future work

In this report, we have introduced a technique to extend the use of the mutual information function to graph-like entities with potential (real-world) applications in both, the chem- and bioinformatics domain.

The *interaction information* is a promising concept for deeper investigations. It is a generalization of the mutual information for more than two variables [24, 30]. In contrast to the MI, it can take on negative and positive values. Further, a value of zero does not correspond to no interaction, making the interpretation at least difficult [27]. However, in the graph context this might help to better capture more intricate correlations within molecules.

If a connectivity relation is ensured that includes spatial interrelations, the grMIF-approach can also be adapted to the application on 3D structures such as proteins.

Of course, the provided grMIF can be used for characterization of graphs in contexts other than chem- or bioinformatics. In particular, it would be interesting to investigate social networks in terms of those information theoretic features.

An open question is, whether it is possible to extend the grMIF-approach to the multi-labeling case for the nodes in a meaningful manner.

Acknowledgement

K.S.B. and M.K. were supported by a grant of the European Social Fund (ESF) for the young researcher group MaLeKITA at the Saxon Institute for Computational Intelligence and Machine Learning of the University of Applied Sciences Mittweida.

References

[1] Mark Bauer, Sheldon M Schuster, and Khalid Sayood. The Average Mutual Information Profile as a Genomic Signature. *BMC Bioinformatics*, 9(1):48, 2008.

[2] Richard Bellman. On a routing problem. *Quarterly of Applied Mathematics*, 16(1):87–90, 1958.

[3] M.J. Berryman, A. Allison, and D. Abbott. Mutual information for examining correlataions in DNA. *Fluctuation and Noise Letters*, 4(2):237–246, 2004.

[4] Katrin Sophie Bohnsack, Marika Kaden, Julia Abel, Sascha Saralajew, and Thomas Villmann. The Resolved Mutual Information Function as a Structural Fingerprint of Biomolecular Sequences for Interpretable Machine Learning Classifiers. *Entropy*, 23(10):1357, October 2021.

[5] Evan E. Bolton, Yanli Wang, Paul A. Thiessen, and Stephen H. Bryant. Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities. In Ralph A. Wheeler and David C. Spellmeyer, editors, *Annual Reports in Computational Chemistry*, volume 4, pages 217–241. Elsevier, January 2008.

[6] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, January 2015.

[7] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics* -, pages 76–83, Vancouver, British Columbia, Canada, 1989. Association for Computational Linguistics.

[8] I. Čmelo, M. Voršilák, and D. Svozil. Profiling and analysis of chemical compounds using pointwise mutual information. *Journal of Cheminformatics*, 13(1):3, December 2021.

[9] Thomas H. Cormen, editor. *Introduction to Algorithms*. MIT Press, Cambridge, Mass, 3rd ed edition, 2009.

[10] Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in AI-driven drug discovery: A review and practical guide. *Journal of Cheminformatics*, 12(1):56, September 2020.

[11] Manuel Dehnert, Werner E. Helm, and Marc-Thorsten Hütt. Informational structure of two closely related eukaryotic genomes. *Physical Review E*, 74(2):021913, August 2006.

[12] Laura Delgado-Soler, Raul Toral, M. Santos Tomás, and Jaime Rubio-Martinez. RED: A Set of Molecular Descriptors Based on Rényi Entropy. *Journal of Chemical Information and Modeling*, 49(11):2457–2468, November 2009.

[13] E. W. Dijkstra. A Note on Two Problemsin Connexionwith Graphs. *Numerische Mathematik*, 1:269–271, 1959.

[14] Jianxin Duan, Steven L. Dixon, Jeffrey F. Lowrie, and Woody Sherman. Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *Journal of Molecular Graphics and Modelling*, 29(2):157–170, September 2010.

[15] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, November 2002.

[16] Robert W. Floyd. Algorithm 97: Shortest path. *Communications of the ACM*, 5(6):345, June 1962.

[17] L. R. Jr. Ford. Network flow theory. In *Paper-923*, page 17, Santa Monica, California, 1956.

[18] Anna Gaulton, Anne Hersey, Michał Nowotka, A. Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J. Bellis, Elena Cibrián-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, María Paula Magariños, John P. Overington, George Papadatos, Ines Smit, and Andrew R. Leach. The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954, January 2017.

[19] Elisabet Gregori-Puigjané and Jordi Mestres. SHED: Shannon Entropy Descriptors from Topological Feature Distributions. *Journal of Chemical Information and Modeling*, 46(4):1615–1622, July 2006.

[20] Ivo Grosse, Hanspeter Herzel, Sergey V. Buldyrev, and H. Eugene Stanley. Species independence of mutual information in coding and noncoding DNA. *Physical Review E*, 61(5):5624–5629, May 2000.

[21] P. E. Hart, N. J. Nilsson, and B. Raphael. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, July 1968.

[22] Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, 7(1):23, December 2015.

[23] Hanspeter Herzel and Ivo Große. Measuring correlations in symbol sequences. *Physica A: Statistical Mechanics and its Applications*, 216(4):518–542, July 1995.

[24] Vladimir Hnizdo, Jun Tan, Benjamin J. Killian, and Michael K. Gilson. Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods. *Journal of Computational Chemistry*, 29(10):1605–1614, July 2008.

[25] J. E. Hopcroft and J. K. Wong. Linear time algorithm for isomorphism of planar graphs. In *STOC '74: Proceedings of the sixth annual ACM symposium on Theory of computing*, pages 172–184, NY, USA, 1974. Association for Computing Machinery (ACM).

[26] B T Korber, R M Farber, D H Wolpert, and A S Lapedes. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 90(15):7176–7180, August 1993.

[27] Klaus Krippendorff. Information of interactions in complex systems. *International Journal of General Systems*, 38(6):669–680, August 2009.

[28] Wentian Li. Mutual information functions versus correlation functions. *Journal of Statistical Physics*, 60(5-6):823–837, September 1990.

[29] Flavio Lichtenstein, Fernando Antoneli, and Marcelo R. S. Briones. MIA: Mutual Information Analyzer, a graphic user interface program that calculates entropy, vertical and horizontal mutual information of molecular sequence sets. *BMC Bioinformatics*, 16(1):409, December 2015.

[30] Hiroyuki Matsuda. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Physical Review E*, 62(3):3096–3102, September 2000.

[31] Edward F. Moore. The shortest path through a maze. In *Proceedings of the International Symposium on the Theory of Switching*, pages 285–292. Harvard University Press, 1959.

[32] Garin Newcomb and Khalid Sayood. Use of Average Mutual Information and Derived Measures to Find Coding Regions. *Entropy*, 23(10):1324, October 2021.

[33] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50:742–754, 2010.

[34] A. Shimbel. Structure in communication nets. In *Proceedings of the Symposium on Information Networks*, pages 199–203, New York, 1955. Polytechnic Press of the Polytechnic Institute of Brooklyn.

[35] D. Swati. Use of Mutual Information Function and Power Spectra for Analyzing the Structure of Some Prokaryotic Genomes. *American Journal of Mathematical and Management Sciences*, 27(1-2):179–198, January 2007.

[36] Mattia Tantardini, Francesca Ieva, Lucia Tajoli, and Carlo Piccardi. Comparing methods for comparing networks. *Scientific Reports*, 9(Article number: 17557), 2019.

[37] Yuan Wang, Hanna Geppert, and Jürgen Bajorath. Shannon Entropy-Based Fingerprint Similarity Search Strategy. *Journal of Chemical Information and Modeling*, 49(7):1687–1691, July 2009.

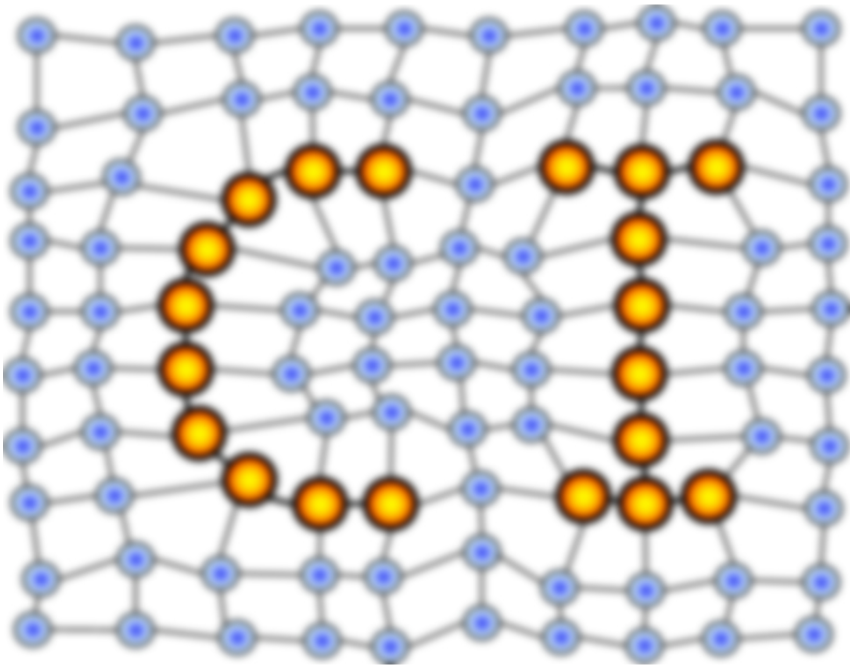
[38] Stephen Warshall. A Theorem on Boolean Matrices. *Journal of the ACM*, 9(1):11–12, January 1962.

[39] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36, February 1988.

[40] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, January 2018.

MACHINE LEARNING REPORTS

Report 01/2022



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany
• http://www.mni.hs-mittweida.de/

Prof. Dr. rer. nat. Frank-Michael Schleif
University of Bielefeld
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
• http://www.cit-ec.de/tcs/about

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.