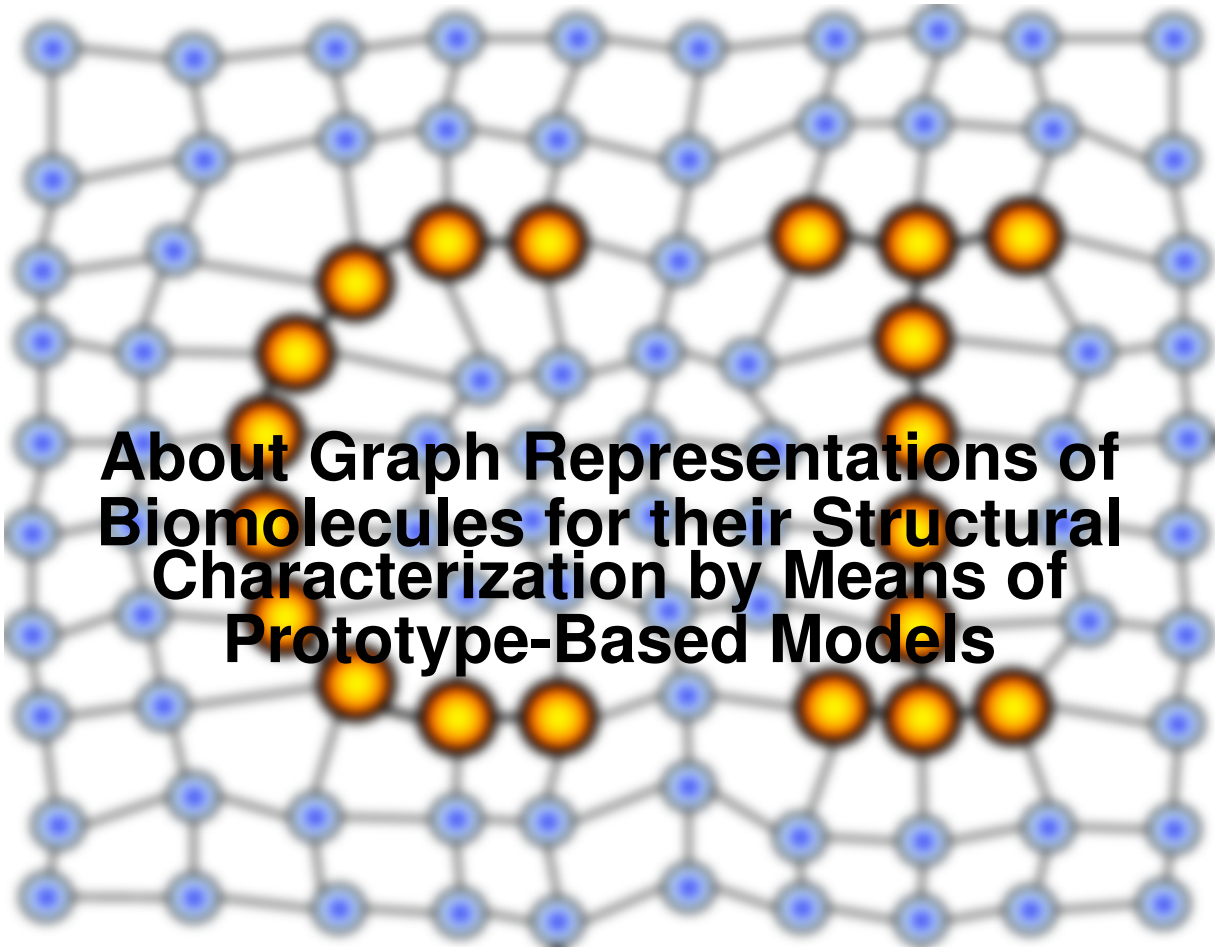
MACHINE LEARNING REPORTS



About Graph Representations of Biomolecules for their Structural Characterization by Means of Prototype-Based Models

Report 01/2025
Submitted: 23.01.2025
Published: 21.02.2025

Katrin Sophie Bohnsack[1], Marika Kaden [1] and Thomas Villmann[1,2]

(1) University of Applied Sciences Mittweida, 09648 Mittweida, Technikumplatz 17, Saxon Institute for Computational Intelligence and Machine Learning (SICIM), 09648 Mittweida - Germany

(2) Technical University Freiberg, 09599 Freiberg, Akademiestraße 6, Germany

Abstract

The aim of this technical report is to highlight some approaches for graph representations of 3d-information of biomolecules which can be used for their structural and information theoretic characterization. Particularly, the report deals with the determination of graph representations reflecting the internal structure by means of prototype-based methods known from machine learning.

1 Introduction

The automatic analysis of databases for biochemical molecules and structures is a rapidly growing field in bioinformatics, which is accelerated by the increased number of machine learning tools and approaches. Frequently, this involves the computational comparison of respective molecule structures. In [10], an information theoretic tool was proposed for sequence analysis and comparison – the mutual information function. Recently, this approach was extended for 2d-structures represented by graphs.

In this contribution we present a further extension of this method for molecules given by their 3d structure information. For this purpose, we adapt machine learning algorithms - the so-called topology representing neural network (TRN) also known as neural gas (NG) vector quantizer proposed in [30, 29] and a k-nearest neighbor approach.

1.1 Fingerprinting structures

The comparison of molecule structures is the foundation for many data analysis and machine learning applications in the chem- and bioinformatics domain [51, 25, 36]. Direct use of raw 3d coordinates is unfavourable, as translation, rotation and permutation events cannot be captured adequately [19] and no task-specific domain knowledge may be integrated in the comparison process. Therefore, the molecule's 3d-information needs to be transformed into a representation suitable for machine learning models [47].

There are numerous approaches in order to do so as listed in [19]: It is distinguished between approaches that consider the full structure in terms of voxelization [43], torsion angles [1] or protein graphs [49] and approaches that extract structural features regarding the proteins surface [18], secondary structure or inter-residue distances [2]. In analogy to the Word2vec embedding [34, 35] in natural language processing, procedures like Mol2vec [24] learn a suitable molecule representation by means of machine learning models.

1.2 3d-Structure-Graphs by means of Delaunay tesselation

Foremost, the handling of proteins as graphs has gained interest [49], especially in the context of graph neural networks [40]. Thereby, the explicit design of the graph (in terms of node and edge attributes) is subject to the user and the given task. In general, a connectivity relation needs to be defined between the components of the 3d structure, i.e. the measured (relative) positions in $\mathbb{R}^3$ of their components.

To avoid arbitrarily chosen cutoffs to define neighbourhood in terms of the Euclidean distance, the Delaunay tessellation (or more precisely the unique Delaunay graph with respect to the Voronoï tessellation) of protein structures is well-known since its first mentioning in [44]. Applications involve structural alignment [23], structural topology comparison [11], studying the functional impact of mutations [33] and exploring graph theoretic properties of residue contact networks [46]. An overview of its applications for studying protein-related interactions and protein structure is given in [52]. What serves as nodes in this graph is up to the user: evident are the C_α atoms or the center of mass of the amino acids. Alternatively, individual atoms or collections of them may be considered [46].

While the full Delaunay graph captures internal connections, it also comes with the disadvantage of unwanted edges, especially on the protein surface. That is, it does not reconstruct the expected intuitive surface of the protein. In order to do so, post-processing of the obtained full Delaunay graph is required, i.e. removal of edges the lengths of which exceeds an arbitrary positive threshold (usually measured in Å) is common [11, 32].

Further, the numerical complexity of the full Delaunay graph determination is costly: Suppose N vectors in $\mathbf{x}_k \in \mathbb{R}^n$. The Delaunay triangulation of them can be extracted from the convex hull of a suitable set obtained by the lifting operation $\mathbf{x}_k \mapsto \tilde{\mathbf{x}}_k = (\tilde{x}_{k,1}, \ldots, \tilde{x}_{k,n+1})^\top \in \mathbb{R}^{n+1}$ with $\tilde{x}_{k,j} = x_{k,j}$ for $1 \leq j \leq n$ and $\tilde{x}_{k,n+1} = \sum_{i=1}^{n} x_{k,i}^2$. Then the edges of the convex hull of those points in the $n+1$-dimensional space connect the images of those points that are connected in the Delaunay triangulation, which takes time complexity $O\left(N^{\lceil n/2 \rceil + 1}\right)$ to calculate them and requires $O\left(N^{\lfloor n/2 \rfloor}\right)$ memory capacity for intermediate results [14].

Hence, in the light of this numerical complexity, efficient approximations are appreciated keeping in mind that for 3d-structure estimation the full Delaunay graph has to be reduced applying heuristic strategies. .

1.3 Graph features

For applicability in distance-based machine learning approaches like [39, 17, 48], methods to compare graphs by quantifying their shared or discriminating features, i.e. object commonalities or differences, are in need. This may be achieved by considering spectral distances or distances based on node affinities [50]. Alternatively, features like the average path length, betweenness, closeness centrality, density or transitivity [13] are promising in combination with a suitable distance measure.

Another approach is to draw on information theoretic features to characterize graphs in terms of inherent relations, context information and topological attributes, especially in the cheminformatics context [21, 16].

Information theoretic features such as the mutual information have excelled in the analysis and characterization of *molecular sequences* [15, 26, 27]. Thereby, the mutual information function (MIF), also known as average mutual information (AMI) [4], is promising to characterize short and long term correlations [22, 6, 45]. In [10], a resolved version of MIF was introduced which steps beyond the Shannon notion of entropic information [41] by considering a respective Rényi variant [38]. This is to be distinguished from the prevailing use of mutual information in the context of multiple sequence alignments [20, 42, 8].

The MIF was extended in [9] for fingerprinting chemical compounds in the shape of structural formula graphs for molecular screening and machine learning applications.

1.4 Outline of the paper

In this contribution we consider two strategies for converting 3d molecules into graph representations such that graph feature based method like graph-MIF become available for in depth-feature analysis [9]:

In section 2.1, we present a method to capture internal connections related to the Delaunay graph, which, however, avoids the problem of long surface edges in the first

place by an alternative calculation of the tessellation based on the competitive Hebb rule and prototype based learning (neural gas)[31]. Further, we introduce in section 2 an intuitive nearest neighbours approach related to neural gas but incorporating a k-nearest-neighbor strategy.

2 Graph representations of biomolecules

Structural information of molecules can be retrieved from respective databases like the Protein Data Bank (PDB) [5]. This section describes the transformation of 3d coordinates into graphs covering different aspects of the molecule: either their surface information or their inner relations. As addressed in the previous section, these original coordinates may correspond to a molecule's individual atoms, C_α atoms or the center of mass of the amino acids.

All of the following approaches yield a graph representation. As there are various equivalent notations of graphs, we simply focus on those by maybe (non-negatively) weighted adjacency matrices, denoting whether pairs of vertices are adjacent or not or are related by the respective weight.

2.1 Molecule raphs by Delaunay tessellations obtained by prototype-based models

We start with the definition of Delaunay tessellations of the $\mathbb{R}^d$ given a finite subset $S \subset \mathbb{R}^d$:

Definition 1 *Let $\mathbf{s}_1, \ldots, \mathbf{s}_n$ be the points in $S \subset \mathbb{R}^d$. Then the* Voronoï cell of $\mathbf{s}_i$ *is denoted by V_i. It is defined as the set of all points in $\mathbb{R}^d$, which are closest to $\mathbf{s}_i$:*

$$V_i = \left\{ \mathbf{p} \in \mathbb{R}^d \mid d(\mathbf{p}, \mathbf{s}_i) \leq d(\mathbf{p}, \mathbf{s}_j),\ 1 \leq j \leq n,\ j \neq i \right\}$$

where $d(\mathbf{p}, \mathbf{s}_i)$ is a given dissimilarity measure.

Clearly, any two Voronoï cells have disjoint interiors but they may intersect along their boundaries and share a common border, in which case they are called neighbours. Clearly, no two Voronoï cells can share more than one side and together, the n Voronoï cells cover the entire $\mathbb{R}^d$. The *Voronoï diagram* of S, denoted by $Vor(S)$ is the set of all Voronoï cells of points of S: $Vor(S) = \{V_i \mid 1 \leq i \leq n\}$. It is also denoted as Delaunay or Voronoï tessellation of $\mathbb{R}^d$.

Definition 2 *Let $Vor(S)$ be a Voronoï tessellation of $\mathbb{R}^d$ for a finite set $S \subset \mathbb{R}^d$. The corresponding* Delaunay graph *takes the points $\mathbf{s}_i \in S$ as nodes. Two nodes $\mathbf{s}_i$ and $\mathbf{s}_j$ are connected in this graph iff their Voronoï cells are neighbours.*

Remark 1 *The Delaunay graph is the mathematical* dual *to the Voronoï diagram of S. If a Voronoï tessellation of the Euclidean plane is considered, the Delaunay graph is also denoted as the Delaunay triangulation of S.*

Interestingly, the Delaunay triangulation of a set S of points in $\mathbb{R}^d$ is related to the set's convex hull in $\mathbb{R}^{d+1}$ via the so-called lifting transform [12, 3, 37].

Applied to 3-dimensional Euclidean data, it yields an aggregate of non-overlapping (except the shared borders for neighboured Voronoï cells) space-filling irregular tetrahedra [44].

In the next step we consider a (continuous) manifold $M \subset \mathbb{R}^d$ with the Euclidean norm as the underlying topological assumption for the tessellation with respect to a finite set $S \subset M$.

Definition 3 *Let M be a (continuous) manifold $M \subset \mathbb{R}^d$ and $S \subset M$ be a finite subset. Let $Vor(S)$ be the Voronoï diagram of $\mathbb{R}^d$ with respect to S with Voronoï cells V_i. The* masked Voronoï cells *R_i are defined as the intersection $R_i = V_i \cap M$.*

The tessellation of M using masked Voronoï cells was introduced by T. MARTINETZ in [30] for the investigation of Topology Representing Networks (TRNs). We denote those tessellations as *masked tessellations*. The *masked Delaunay graph* is the subgraph of the Delaunay graph as the dual of the masked tessellation.

In the next step we explain a method to determine the masked Delaunay graph proposed in [30]. To do so, we need the following definition:

Definition 4 *The set (distribution) S of the points $\mathbf{s}_i \in M, i = 1 \dots n$ is* dense *on the manifold M with topology induced by a given norm, if for each $\mathbf{v} \in M$ the triangle $\Delta(\mathbf{v}, \mathbf{s}_{i_0}, \mathbf{s}_{i_1})$ is completely contained in M whereby $\mathbf{s}_{i_0}$ and $\mathbf{s}_{i_1}$ denote the first and second closest point to $\mathbf{v}$ w.r.t. the norm, respectively.*

Remark 2 *Obviously but worth to be mentioned: If M is convex with respect to the given topology, any arbitrarily chosen set $W \subset M$ constitutes a dense subset.*

Now we can state the following theorem proposed in [30]:

Theorem 1 *If the distribution S of the points $\mathbf{s}_i$ is dense on M with respect to the given norm, the part of the Delaunay triangulation that is formed by a competitive Hebb Learning algorithm is the* induced Delaunay triangulation.

Algorithm 1 Determination of the induced Delaunay graph by competitive Hebb learning for a given norm

procedure INDUCED DELAUNAY GRAPH(M, S)
  initialize the adjacency matrix $\mathbf{C}$ by $C_{ij} = 0 \; \forall \; i, j$
  repeat
    randomly select a sample (stimulus) $\mathbf{v} \in M$
    competition: determine $||\mathbf{v} - \mathbf{s}_{i_0}|| \leq ||\mathbf{v} - \mathbf{s}_{i_1}|| \leq \ldots \leq ||\mathbf{v} - \mathbf{s}_{i_{n-1}}||$ using the given norm
    if $C_{i_0 i_1} = 0$ **then**
      set $C_{i_0 i_1} = 1$ (Hebb-learning)
  until convergence
  return adjacency matrix $\mathbf{C}$

The pseudo-code of the competitive Hebb Learning algorithm is given by algorithm Alg. 1, returning the adjacency matrix $\mathbf{C}$. This matrix can be taken as the connectivity matrix of the points (coordinates). It should be mentioned that this procedure/algorithm corresponds to a variant of the Neural Gas algorithm [31] to learn topological manifolds – the above mentioned TRN.

In the biochemical context, we consider the set S as given atom coordinates $\mathbf{s}_i \in M, i = 1 \dots n$ of a molecule's 3d structure with $M \subset \mathbb{R}^3$. Thereby, in order to execute the algorithm consistently, we need to define M in agreement with the given set of points. In particular, we require in the assumptions of the Theorem 1 that S is dense on M. Here M is an appropriately chosen subset of $\mathbb{R}^d$ reflecting the 3d-shape of

the considered molecule. For example, M could be chosen as the minimum cuboid containing S.

Another possibility is to approximate M by so-called β-ball masks: for each $\mathbf{s}_i \in S$, the β-ball is the sphere $B_i^\beta \in \mathbb{R}^d$ with radius β surrounding $\mathbf{s}_i$. Then M is obtained as $M = \cup_{\mathbf{s}_i \in S} B_i^\beta$. For the above algorithm, one has to choose a sufficient number of stimuli uniformly distributed in M.

Last, we have can state that the convergence time of the approach to approximate the true graph structure grows roughly as $O\left(N_S^{\lceil n^*/2 \rceil + 1}\right)$ in analogy to full triangulation but where n^* is the intrinsic data dimension (Hausdorff-dimension), for which usually $n^* \ll n$ is valid. The capacity requirements are $O\left(N_S^2 + N_S \cdot n\right)$ and, hence, much less compared to the full approach.

2.2 The k-nearest neighbours approach for constructing the adjacency matrix

We present here an alternative way to construct the adjacency matrix for molecules based on a heuristic combining the *Neural Gas* [31] and graph theory's *k-nearest neighbours* algorithm. Unlike the previous method, this approach does not explicitly rely on the Delaunay graph. We suppose a finite set $S \subset \mathbb{R}^d$ which would be the set of atom coordinates in the above mentioned biochemical context. We suppose n atoms in the following.

The algorithm is presented as Alg. 2.

Algorithm 2 k-nearest neighbor algorithm for adjacency matrix generation

procedure $k-$NEAREST NEIGHBOR(S)
  initialize $t = 0$, $\Delta t > 0$ and a randomly selected set of points W $\hookrightarrow$ ($W \subset \mathbb{R}^d$, $|W| = N \leq n, 1 < k << N$)
  repeat
    select a point $\mathbf{s} \in S \subseteq \mathbb{R}^d$ randomly
    find $(k+1)$-nearest points $\mathbf{w}_j \in W$ to $\mathbf{s}$ using the dissimilarity measure $\hookrightarrow$ $d(\mathbf{s}, \mathbf{w}_j)$ in S
    add edges in W between the closest point to every other k-nearest point $\hookrightarrow$ (modified Hebb)
    move $(k+1)$-nearest points $\mathbf{w}_j \in W$ towards s, proportionally to their $\hookrightarrow$ distance from it
    if there are any crossing edges and $N \leq n$ is valid **then**
      add a point to W such that the crossing is resolved
    remove those edges which are not refreshed in the last Δt steps
    time increment $t = t + 1$
  until convergence
  return adjacency matrix $\mathbf{C}$ of W

Finally, the adjacency matrix of W represents the adjacencies (topological relations) in S whereby the vectors $\mathbf{w}_j \in W$ are approximations of the atom coordinates $\mathbf{s}_j$. The set W could be initialized to be a finite uniform representation of the manifold $M = \cup_{\mathbf{s}_i \in S} B_i^\beta$ using the β-balls. It is worth mentioning that this method generates

the adjacency matrix representing the inner structure of molecules, rather than their boundaries.

2.3 Characterizing molecule graphs by their mutual information function

In machine learning approaches for molecule investigations, efficient methods for molecule comparisons are required. The graph mutual information function introduced in [9] gives the possibility to extract topological information about the molecules from their graph representations. For this purpose, the graphs are considered as labeled graphs, where the node labels are determined as the categories of atoms/components of the investigated molecules. The mutual information function reflects the spatial correlations in an information theoretic manner and can be used as characteristic feature vectors for machine learning methods not restricted to graph neural networks.

In particular, by means of this approach, vector quantization methods [7] become applicable which belong to the class of interpretable machine learning approaches [28].

3 Conclusion and Future Work

In this contribution, we propose strategies for representation of 3d molecules as graphs that put focus on either the components inner relations or molecule surface properties. From these graphs, information theoretic concepts and quantities can be derived in order to characterize spatial correlations between the atoms. Particularly, graph-MIF as described in [9] allows the generation of those features for data analysis and machine learning applications.

The surveyed approaches for graph generation raise several starting points for future research:

First of all, the presented methods are flexible with regard to the data basis for generating the graphs. This could be the 3d information of whole biomolecules, as assumed here so far. Alternatively, only parts known for functional relevance such as binding sites may be taken into account for graph generation, allowing a different perspective on the molecule.

Furthermore, in the context of internal structure graphs, the two concepts for excluding non-relevant long outer edges generated by simple algorithms must be related or contrasted more closely: on the one hand side, preventing the emergence of these edges in advance, realized by topology representing networks with a suitable mask, and on the other hand, removing these edges from the standard Delaunay graph afterwards.

The advantages and disadvantages of the TRN method compared to the Delaunay triangulation for the reconstruction of protein graphs should be investigated in real world applications for reliability. We see possible applications in the context of graph-based structural alignments [23].

The versatility of the MIF approach allows a comprehensive understanding of biomolecules: We can investigate the same molecule on different levels and thereby consider different information, i.e. that of the protein sequence (1d) and of the structure (3d). It is to be investigated whether a mapping can be made/learned between the two.

Acknowledgement

M.K. is supported by the DLR-project AIMS.

References

[1] Mohammed AlQuraishi. End-to-End Differentiable Learning of Protein Structure. *Cell Systems*, 8(4):292–301.e3, April 2019.

[2] Namrata Anand and Possu Huang. Generative modeling for protein structures. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[3] C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, December 1996.

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

[5] H. M. Berman. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.

[6] M. J. Berryman, A. Allison, and D. Abbott. Mutual information for examining correlations in DNA. *Fluctuation and Noise Letters*, 04(02):L237–L246, June 2004.

[7] Michael Biehl, Barbara Hammer, and Thomas Villmann. Prototype-based models in machine learning. *WIREs Cognitive Science*, 7(2):92–111, 2016.

[8] Anne-Florence Bitbol. Inferring interaction partners from protein sequences using mutual information. *PLOS Computational Biology*, 14(11):e1006401, November 2018.

[9] K S Bohnsack, M Kaden, and T Villmann. The Resolved Mutual Information Function for Fingerprinting Biochemical Compounds Based on their Structural Formulas. *Machine Learning Reports*, page 14, 2022.

[10] 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.

[11] David L. Bostick, Min Shen, and Iosif I. Vaisman. A simple topological representation of protein structure: Implications for new, fast, and robust structural classification. *Proteins: Structure, Function, and Bioinformatics*, 56(3):487–501, May 2004.

[12] Kevin Q. Brown. Voronoï diagrams from convex hulls. *Information Processing Letters*, 9(5):223–228, December 1979.

[13] Liam Childs, Zoran Nikoloski, Patrick May, and Dirk Walther. Identification and classification of ncRNA molecules using graph properties. *Nucleic Acids Research*, 37(9):e66–e66, May 2009.

[14] P. Cignoni, C. Montani, and R. Scopigno. DeWall:A fast divide and conquer Delaunay triangugaltion algorithm in $\mathbb{E}^d$. *Computer-Aided Design*, 30(5):333–341, December 1998.

[15] 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.

[16] 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.

[17] B. J. Frey and D. Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, February 2007.

[18] P. Gainza, F. Sverrisson, F. Monti, E. Rodolà, D. Boscaini, M. M. Bronstein, and B. E. Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, February 2020.

[19] Wenhao Gao, Sai Pooja Mahajan, Jeremias Sulam, and Jeffrey J. Gray. Deep Learning in Protein Structural Modeling and Design. *Patterns*, 1(9):100142, December 2020.

[20] Gregory B. Gloor, Louise C. Martin, Lindi M. Wahl, and Stanley D. Dunn. Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions. *Biochemistry*, 44(19):7156–7165, May 2005.

[21] 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.

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

[23] Valentin A. Ilyin, Alexej Abyzov, and Chesley M. Leslin. Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. *Protein Science*, 13(7):1865–1874, July 2004.

[24] Sabrina Jaeger, Simone Fulle, and Samo Turk. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *Journal of Chemical Information and Modeling*, 58(1):27–35, January 2018.

[25] T. Kawabata. MATRAS: A program for protein 3D structure comparison. *Nucleic Acids Research*, 31(13):3367–3369, July 2003.

[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] 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.

[28] P. Lisboa, S. Saralajew, A. Vellido, and T. Villmann. The coming of age of interpretable and explainable machine learning models. In M. Verleysen, editor, *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2021), Bruges (Belgium)*, pages 547–556, Louvain-La-Neuve, Belgium, 2021. i6doc.com.

[29] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten. 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569, July 1993.

[30] Thomas Martinetz. Topology Representing Networks. *Neural Networks*, 7(3):507–522, 1994.

[31] Thomas Martinez and Klaus Schulten. A "Neural-Gas" Network Learns Topologies. *Artificial Neural Networks*, pages 397–402, 1991.

[32] Majid Masso. Modeling functional changes to Escherichia coli thymidylate synthase upon single residue replacements: A structure-based approach. *PeerJ*, 3:e721, January 2015.

[33] Ewy Mathe, Magali Olivier, Shunsuke Kato, Chikashi Ishioka, Iosif Vaisman, and Pierre Hainaut. Predicting the transactivation activity of p53 missense mutants using a four-body potential score derived from Delaunay tessellations. *Human Mutation*, 27(2):163–172, February 2006.

[34] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv*, September 2013.

[35] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[36] Minh N. Nguyen, Adelene Y. L. Sim, Yue Wan, M. S. Madhusudhan, and Chandra Verma. Topology independent comparison of RNA 3D structures using the CLICK algorithm. *Nucleic Acids Research*, 45(1):e5–e5, January 2017.

[37] V. T. Rajan. Optimality of the Delaunay triangulation in $\mathbb{R}^d$. *Discrete & Computational Geometry*, 12(2):189–202, December 1994.

[38] A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 1961. University of California Press.

[39] Atsushi Sato and Keiji Yamada. Generalized Learning Vector Quantization. In *Advances in Neural Information Processing Systems*, pages 423–429, Cambridge, MA, USA, 1996. MIT Press.

[40] F. Scarselli, M. Gori, Ah Chung Tsoi, M. Hagenbuchner, and G. Monfardini. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 20(1):61–80, January 2009.

[41] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[42] Franco L. Simonetti, Elin Teppa, Ariel Chernomoretz, Morten Nielsen, and Cristina Marino Buslje. MISTIC: Mutual information server to infer coevolution. *Nucleic Acids Research*, 41(W1):W8–W14, July 2013.

[43] Martin Simonovsky and Joshua Meyers. DeeplyTough: Learning Structural Comparison of Protein Binding Sites. *Journal of Chemical Information and Modeling*, 60(4):2356–2366, April 2020.

[44] R. K. Singh, A. Tropsha, and I. I. Vaisman. Delaunay tessellation of proteins: Four body nearest-neighbor propensities of amino acid residues. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 3(2):213–221, 1996.

[45] 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.

[46] Todd J. Taylor and Iosif I. Vaisman. Graph theoretic properties of networks formed by the Delaunay tessellation of protein structures. *Physical Review E*, 73(4):041925, April 2006.

[47] Jacob Townsend, Cassie Putman Micucci, John H. Hymel, Vasileios Maroulas, and Konstantinos D. Vogiatzis. Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nature Communications*, 11(1):3230, December 2020.

[48] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[49] Saraswathi Vishveshwara, K. V. Brinda, and N. Kannan. Protein Structure: Insights from Graph Theory. *Journal of Theoretical and Computational Chemistry*, 01(01):187–211, July 2002.

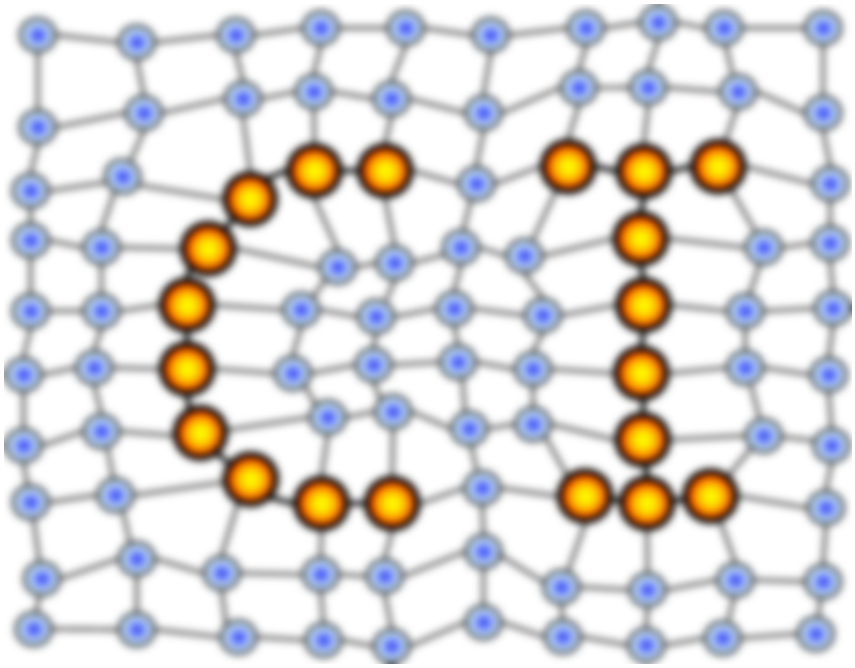
[50] Peter Wills and François G. Meyer. Metrics for graph comparison: A practitioner's guide. *PLOS ONE*, 15(2):e0228728, February 2020.

[51] Yuting Xu, Deeptak Verma, Robert P. Sheridan, Andy Liaw, Junshui Ma, Nicholas M. Marshall, John McIntosh, Edward C. Sherer, Vladimir Svetnik, and Jennifer M. Johnston. Deep Dive into Machine Learning Models for Protein Engineering. *Journal of Chemical Information and Modeling*, 60(6):2773–2790, June 2020.

[52] W. Zhou and H. Yan. Alpha shape and Delaunay triangulation in studies of protein-related interactions. *Briefings in Bioinformatics*, 15(1):54–64, January 2014.

MACHINE LEARNING REPORTS

Report 01/2025



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/

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.