

# GNNMetrics, a package for optimizing and understanding MMD parametrization for evaluating generative GNNs on protein datasets.

---

Philip Hartout

February 1, 2022

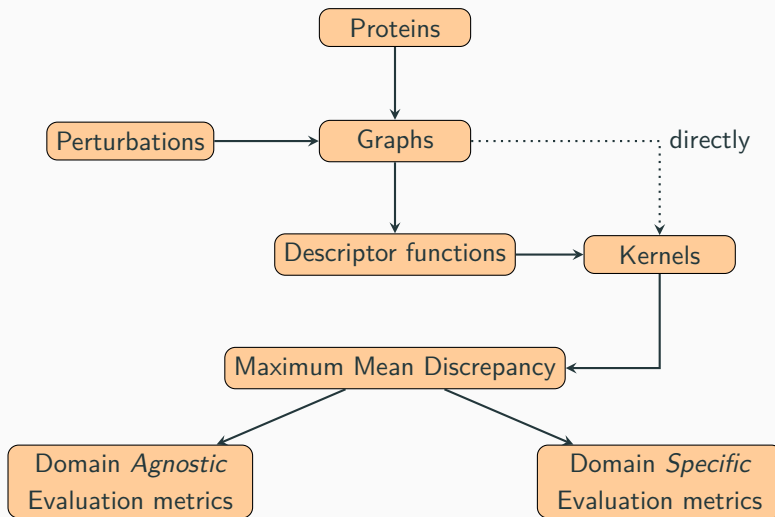


**DBSSE**

**ETH** zürich

- The following provides an overview of the plan for the thesis
- Aspirational/low priority items are [in blue](#).

# General workflow



**Figure 1:** Overview of the library

# Protein source datasets

- Start with human proteome from AlphaFold (23390 samples, [source](#))
- Expand to other experimentally datasets & add cleaning handlers

# Graph extraction from pdb file

Granularity:

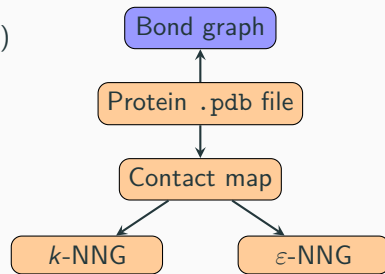
- CA-atom (1 point per amino acid)
- all atoms (loop through each residue to get atom coordinates)
- CB, C, O, ... see `Bio.PDB.Residue`

Graph extraction:

- Contact map (fully connected weighted graph).
- $\epsilon$ -neighborhood graph. [1]
- k-nearest neighbor graph. [9]
- Graph of all atoms with their bond type (incl. e.g. disulfide bond)

Dependencies

- only depend on Biopython to manipulate .pdb files.



**Figure 2:** Order of extraction of graphs from .pdb files.

# Descriptor functions & Kernels

## Descriptor functions

Domain agnostic graph descriptors [5]:

- Degree distribution histogram
- Clustering coefficient histogram
- Laplacian spectrum histogram

Topological descriptors (using the weighted, fully connected contact map):

- Persistence diagram, converted to Betti curves, persistence image, persistence landscapes. [7]
- Also possible to obtain vector representation from persistence diagram by applying a heat kernel to it. [6]

Domain specific: t.b.d

## Kernels

Conditions for selection: p.s.d & fast

General kernels [5]:

- RBF kernel
- Laplacian kernel
- Linear
- Neighborhood Subgraph Pairwise Distance graph kernel [3]

The last can be used when discrete edge features are employed.

Domain specific: t.b.d

# Maximum Mean Discrepancy

The Maximum Mean Discrepancy (MMD) [4, 2] is defined as follows:

$$\text{MMD}(X, Y) := \frac{1}{n^2} \sum_{i,j=1}^n k(x_i, x_j) + \frac{1}{m^2} \sum_{i,j=1}^m k(y_i, y_j) - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(y_i, y_j)$$

where:

- $x_i, x_j \sim \mathcal{X}$ ,  $n$  is the number of samples from non-empty set  $\mathcal{X}$ ;
- $y_i, y_j \sim \mathcal{Y}$ ,  $m$  is the number of samples from non-empty set  $\mathcal{Y}$ ;
- $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is a valid kernel.

MMD is a kernelized proxy of the distance between two graph distributions  $G$  and  $G^*$  computed as  $d_{\text{MMD}}(G, G^*) = \text{MMD}(f(G), f(G^*))$ , where  $f$  is the descriptor function of the graph. A lower  $d_{\text{MMD}}(G, G^*)$  indicates a greater similarity between  $G$  and  $G^*$ . [Potentially look at other metrics for GNNs. \[8\]](#)

# Perturbations to graphs

Domain agnostic [5]:

- edge insertion
- edge removal
- edge rewiring
- node addition

Domain specific:

- t.b.d. (e.g. add edge “disrupting” binding pocket, how to do this?)



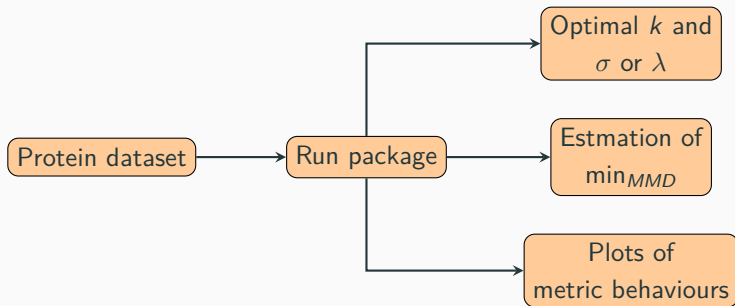
Domain agnostic [5]:

- Correlation with perturbation
- Correlation with graph edit distance

Domain specific:

- Alignment
- (estimated) folding energy
- Investigate what makes a “good” protein.





## Workflows enabled by the package






**Figure 3:** Workflows enabled by the package.

# Plots to understand the behaviour of MMD

- Distribution of  $\min_{MMD}$  for different random test/train splits.
- Correlation perturbation with MMD values with different parameters
- MMD vs. parameter

-  D. C. Anastasiu.  
***Algorithms for Constructing Exact Nearest Neighbor Graphs.***  
PhD thesis, University of Minnesota, 2016.
-  K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola.  
**Integrating structured biological data by kernel maximum mean discrepancy.**  
*Bioinformatics*, 22(14):e49–e57, 2006.
-  F. Costa and K. De Grave.  
**Fast neighborhood subgraph pairwise distance kernel.**  
In *ICML*, 2010.
-  A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola.  
**A kernel two-sample test.**  
*The Journal of Machine Learning Research*, 13(1):723–773, 2012.

-  L. O’Bray, M. Horn, B. Rieck, and K. Borgwardt.  
**Evaluation metrics for graph generative models: Problems, pitfalls, and practical solutions.**  
*arXiv preprint arXiv:2106.01098*, 2021.
-  J. Reininghaus, S. Huber, U. Bauer, and R. Kwitt.  
**A stable multi-scale kernel for topological machine learning.**  
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4741–4748, 2015.
-  G. Tauzin, U. Lupo, L. Tunstall, J. B. Pérez, M. Caorsi, A. M. Medina-Mardones, A. Dassatti, and K. Hess.  
**giotto-tda:: A topological data analysis toolkit for machine learning and data exploration.**  
*J. Mach. Learn. Res.*, 22:39–1, 2021.



R. Thompson, B. Knyazev, E. Ghalebi, J. Kim, and G. W. Taylor.

**On evaluation metrics for graph generative models.**

*arXiv preprint arXiv:2201.09871*, 2022.



W.-L. Zhao, H. Wang, and C.-W. Ngo.

**Approximate k-nn graph construction: a generic online approach.**

*IEEE Transactions on Multimedia*, 2021.