Applications of automated data quality curation in machine learning based modelling of immunoinformatic systems – literature review

Dominik Rafacz
Michał Burdukiewicz – supervisor

Warsaw University of Technology

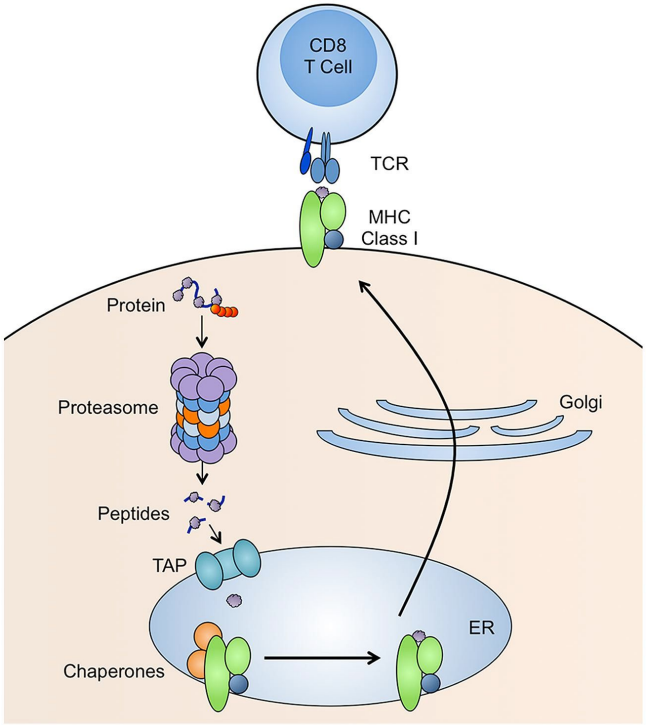
Diploma Seminar, 02.12.21

Agenda

1. Background of the epitope-receptor interaction problem.
2. Existing algorithms for interaction prediction.
3. Data quality concerns.
4. Existing solutions and study directions.
5. Aim of the thesis.

MHC, T cells and TCR

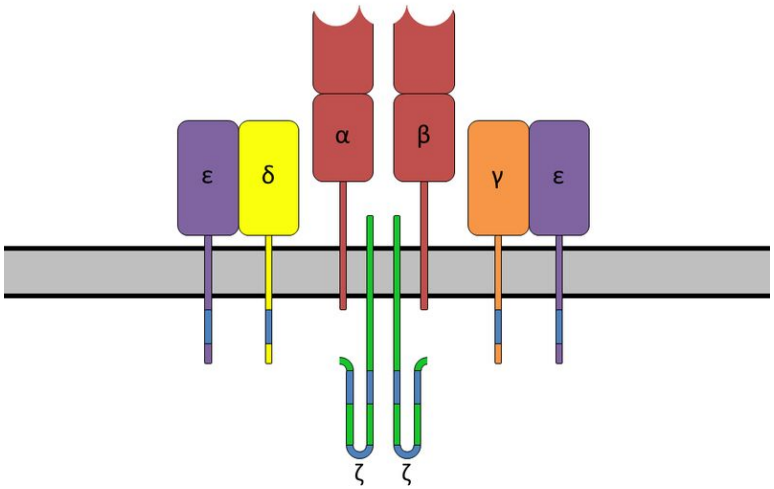
- Major histocompatibility complex, class I (MHC): recognizes and exposes epitopes (peptides present on antigens surfaces) to T cells
- T cell: an immune system element responsible for triggering immune system reaction
- T cell receptor (TCR): a peptide interacting with specific MHCs



MHC class I antigen presentation pathway (source: [1])

TCR specificity

- During development some variance is introduced into TCRs – this results in their diversity and ability to capture vast space of different pathogens.
- α and β chains of amino acid residues identify TCR.
- One TCR may bind even a few millions of peptide-MHC complexes [1, 2]



α-β T cell receptor complex (source: Wikimedia)

Interaction

One data point is one pair epitope sequence-TCR sequence.

Example:

Epitope:

```
ENPVVDFFKNIVTPR
```

TCR sequence (α and β chain):

```
TRA: AASSFGNEKLT
TRB: ATSALGDTQY
```

(source: www.iedb.org/receptor/265, [6])

The dream of immunobioinformaticians

- Understanding the rules governing interaction between peptide-MHC complexes and TCRs represents a paramount step in both personalized immune treatment and development of targeted vaccines. [3]
- These interactions are called protein-protein interactions (PPI) in general.
- Experimentally studying which pairs interact is laborious and time-consuming – scientists are looking for a model that describes this well.

NetTCR-2.0 [3]

- Predicts which pair of epitope-TCR interacts.
- “Shallow” 1D convolutional neural network.
- The problem is addressed better with usage of both α and β chains.
- Trained on data from a single study – biased towards interactions occurring only in specific donors.
- Prone to overfitting.

TCRMatch [4]

- Estimates probability of given epitope interacting with each of the TCR sequences present in the training data.
- Based on sequence similarity calculation.
- ML used for tuning parameters.

Data used in studies

- Most commonly data used in studies comes from three sources [3]:
  - Immune Epitope Database, IEDB (~424 000 records, most redundant) [6],
  - VDJdb (~21 000 records) [7]
  - McPAS-TCR (5100 records) [8]
- Additional data sources are described in [9, 10]

Data insufficiency

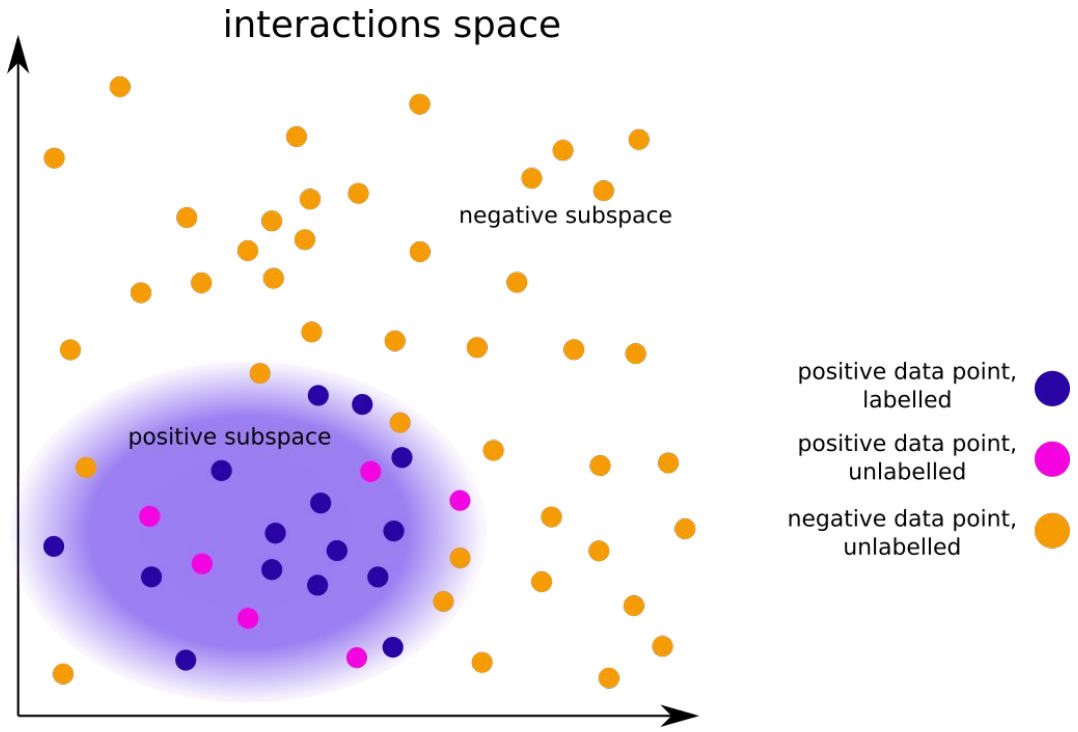
"The overall conclusion from earlier works is that while the prediction of TCR specificity is feasible, the volume and accuracy of current data limit the performance of the developed models" [3]

Problems concerning the data

- Negative datasets are non-existent
- Sampled data is non-representative for the whole dataset
- Data is clustered

There is virtually no data marked explicitly as negative

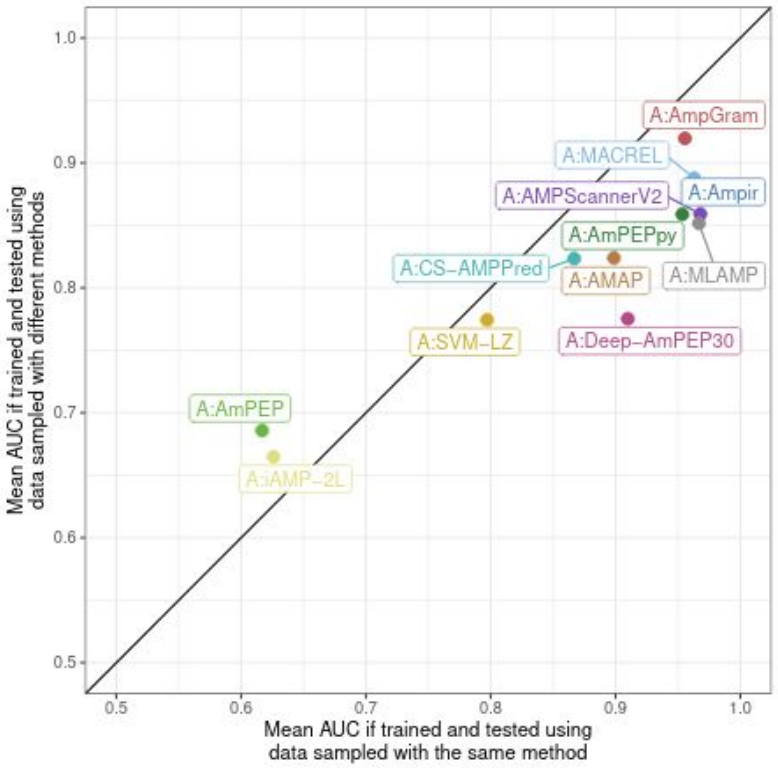
- It is impossible to mark a pair as non interacting with certainty.
- Solutions [11]:
  - one class classification
  - positive vs unlabelled classification
  - negative data sampling



Visualization of negative space problem (source: own work)

AMP problem: sampling bias [12]

- In area of predicting specific properties of peptides where labelled negative dataset is lacking, unfair benchmark results may appear.
- 10 out of 12 algorithms perform significantly better when training and testing data built using the same sampling method.
- Similar problem may be present in area of PPI prediction [13].



Comparison of AUC for various AMP-prediction algorithms (source: [12])

Sampling of PPI negative space [13]

- Common approach: random sampling of unlabelled data
- Selecting random protein pairs from different cellular locations – yields better results, but is less representative

SVM method [14]

x – training data vectors
l – input data size
ν – upper bound on the fraction of outliers
ξ – slack variables
ρ – offset
ω – instance weight
φ, k – transformation functions
α – coefficients of support vectors

primal problem:

$$\min_{\omega,\xi,\rho} \frac{1}{2}||\omega||^2 + \frac{1}{vl}\sum_{i=1}^{l}\xi_i - \rho$$

$$subject\ to\ (\omega,\phi(x_i)) \geq \rho - \xi_i, \xi_i \geq 0$$

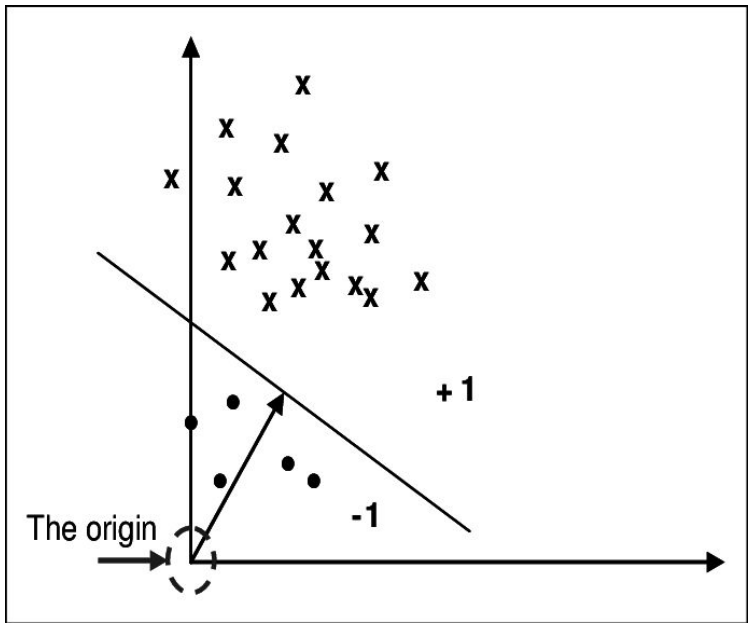
dual problem:

$$\min_{\alpha} \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j k(x_i,x_j)$$

$$subject\ to\ 0 \leq \alpha_i \leq \frac{1}{vl}, \sum_i \alpha_i = 1$$

SVM method, ctd.

- Train SVM model to fit positive data and mark all the other data as negative.
- May require extensive computational power.
- Bias of using the same method for sampling and training.
- Difficult to generalize, prone to manipulation with the difficulty of the problem.



Visualization of one-class SVM problem (source: www.researchgate.net/figure/Classification-in-one-class-SVM_fig1_242572058)

NIP-SS, NIP-RW [15]

- Two algorithms for generating negative datasets.
- NIP-SS is based on sequence similarity, NIP-RW on random walking on graph.
- Uses cross-validation to assess efficiency, which may not be a valid solution
- There is a risk of creating “too trivial task” – as in previous method.

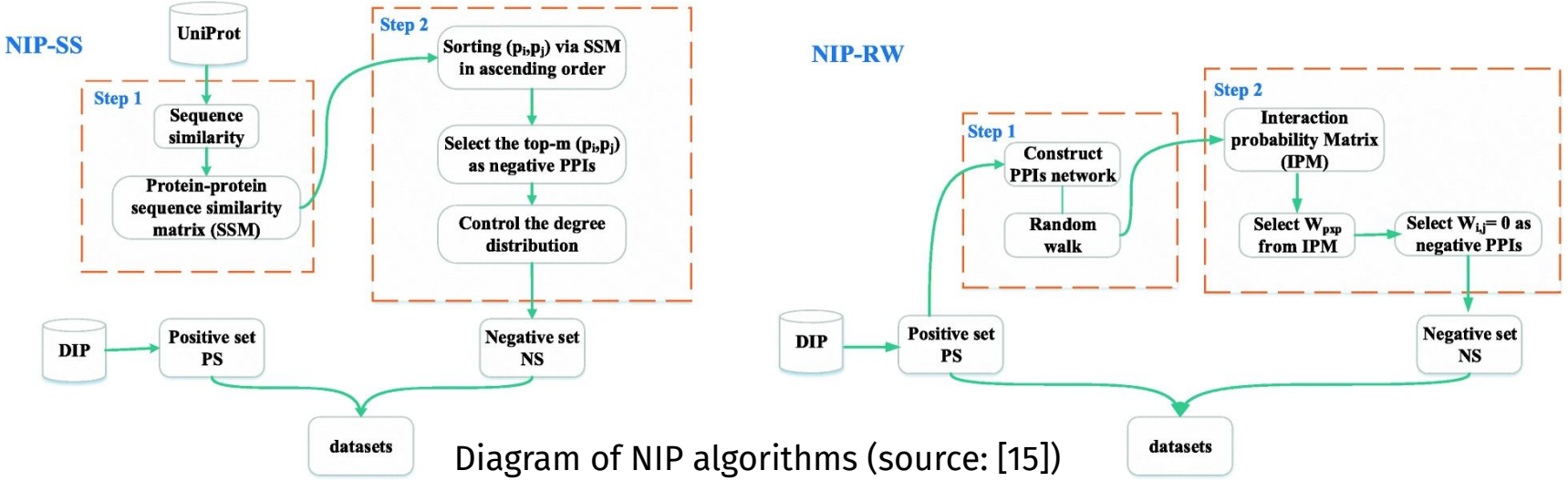


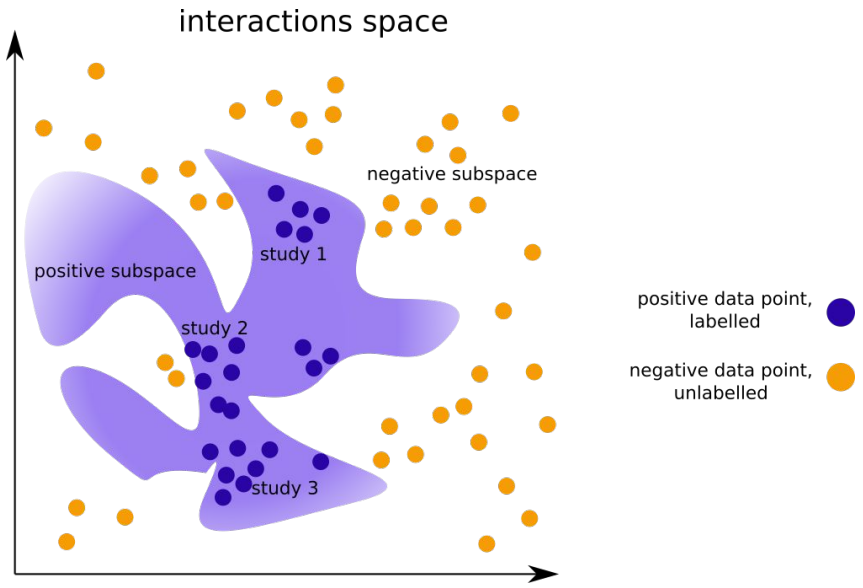
Diagram of NIP algorithms (source: [15])

Clustering of negative space [16]

- Clustering of negative space based on physicochemical properties of proteins.
- Selecting representatives of each cluster.
- Reduces within- and between-class imbalance.
- Do not tackles with clustering positive class.
- Selects only one representative of each cluster – may also be biased.
- May result in selecting false negatives, as it favors selecting the most dissimilar ones.

Biased cross-validation [17]

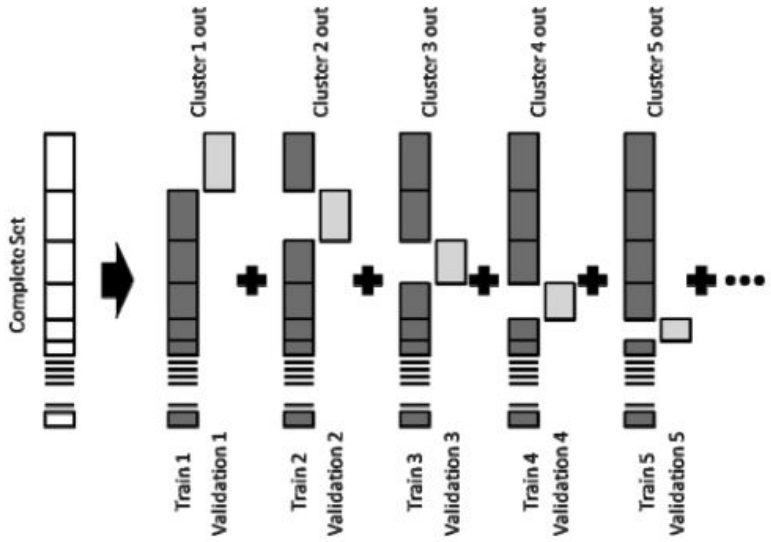
- Sufficient requirement for unbiasedness of the CV estimator: joint distribution of train and test same is the same as joint distribution of train fold and test fold for each fold.
- Knowing distributions covariances, we may introduce correction to the result.



Visualization of the bias of correlated data problem (source: own work)

Leaving cluster out [18]

- Large differences measured in R or R^2 depending on whether it is validated against the core set or validated in a leave-cluster-out cross-validation.
- If proteins from the same family are present in both the training and validation set, the estimated prediction quality from standard validation techniques looks too optimistic.



Schema of leave-cluster-out validation (source: [18])

Clustering methods for proteins

- CLANS (CLuster ANalysis of Sequences) [19]
- TRIBE-MCL (Markov CLuster) [20]

Aim of the thesis

Scope of the thesis:

- Perform cluster analysis on positive and negative dataset.
- Identify motifs typical for the data.
- Evaluate quality of the data.
- Propose an appropriate data sampling strategy (good generalization, good for non-trivial problems) – both for negative sampling and cross validation sampling.
- Implement a tool for automated building of the dataset.

Possible directions:

- Implement a ML model?

Thanks for your attention

Questions?

Biblography, pt. I

[1] : Mary K. McCarthy and Jason B. Weinberg. The immunoproteasome and viral infection: a complex regulator of inflammation. *Frontiers in Microbiology*, 6:21, 2015.

[2] : Dale I. Godfrey, Jamie Rossjohn, and James McCluskey. The delity, occasional promiscuity, and versatility of t cell receptor recognition. *Immunity*, 28(3):304314, 2008.

[3] : Alessandro Montemurro, Viktoria Schuster, Helle Rus Povlsen, Amalie Kai Bentzen, Vanessa Jurtz, William D. Chronister, Austin Crinklaw, Sine R. Hadrup, Ole Winther, Bjoern Peters, Leon Eyrich Jessen, and Morten Nielsen. NetTCR-2.0 enables accurate prediction of TCR peptide

binding by using paired TCRα and β sequence data. *Communications Biology*, 4(1), September 2021.

[4] : William D. Chronister, Austin Crinklaw, Swapnil Mahajan, Randi Vita, Zeynep Koşaloğlu-Yalçn, Zhen Yan, Jason A. Greenbaum, Leon E. Jessen, Morten Nielsen, Scott Christley, Lindsay G. Cowell, Alessandro Sette, and Bjoern Peters. TCRMatch: Predicting t-cell receptor specificity based on sequence similarity to previously characterized receptors. *Frontiers in Immunology*, 12, March 2021.

[5] : Aijun Deng, Huan Zhang, Wenyan Wang, Jun Zhang, Dingdong Fan, Peng Chen, and Bing Wang. Developing computational model to predict protein-protein interaction sites based on the XGBoost algorithm. *International Journal of Molecular Sciences*, 21(7):2274, March 2020.

[6] : Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339D343, October 2018.

[7] : Dmitry V Bagaev, Renske M A Vroomans, Jerome Samir, Ulrik Stervbo, Cristina Rius, Garry Dolton, Alexander Greenshields-Watson, Meriem Attaf, Evgeny S Egorov, Ivan V Zvyagin, Nina Babel, David K Cole, Andrew J Godkin, Andrew K Sewell, Can Kesmir, Dmitriy M Chudakov, Fabio Luciani, and Mikhail Shugay. VDJdb in 2019: database extension, new analysis infrastructure and a t-cell receptor motif compendium. *Nucleic Acids Research*, 48(D1):D1057D1062, October 2019.

Biblography, pt. II

[8] : Nili Tickotsky, Tal Sagiv, Jaime Prilusky, Eric Shifrut, and Nir Friedman. McPAS-TCR: a manually curated catalogue of pathology-associated t cell receptor sequences. *Bioinformatics*, 33(18):29242929, May 2017.

[9] : 10x GENOMICS. A new way of exploring immunity: linking highly multiplexed antigen recognition to immune repertoire and phenotype. *(Application note)* (https://go.technologynetworks.com/new-way-of-exploring-immunity)

[10] : Sean Nolan, Marissa Vignali, Mark Klinger, Jennifer N. Dines, Ian M. Kaplan, Emily Svejnoha, Tracy Craft, Katie Boland, Mitch Pesesky, Rachel M. Gittelman, Thomas M. Snyder, Christopher J. Gooley, Simona Semprini, Claudio Cerchione, Massimiliano Mazza, Ottavia M. Delmonte, Kerry Dobbs, Gonzalo Carreño-Tarragona, Santiago Barrio, Vittorio Sambri, Giovanni Martinelli, Jason D. Goldman, James R. Heath, Luigi D. Notarangelo, Jonathan M. Carlson, Joaquin Martinez-Lopez, and Harlan S. Robins. A large scale database of t-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. *(PREPRINT)* , August 2020.

[11] : Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: a survey. *Machine Learning*, 109(4):719760, Apr 2020.

[12] : Katarzyna Sidorczuk, Przemysław Gagat, Filip Pietluch, Jakub Kała, Dominik Rafacz, Laura Bąkała, Jadwiga Słowik, Rafał Kolenda, Stefan Rödiger, Legana C H W Fingerhut, Ira R Cooke, Paweł Mackiewicz and Michał Burdukiewicz. The impact of negative data sampling on antimicrobial peptide prediction. *(ARTICLE IN PREPARATION)*

[13] : Asa Ben-Hur and William Staord Noble. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, 7(S1), March 2006.

[14] : Suyu Mei and Hao Zhu. A novel one-class SVM based negative data sampling method for reconstructing proteome-wide HTLV-human protein interaction networks. *Scientific Reports*, 5(1), January 2015.

[15] : Long Zhang, Guoxian Yu, Maozu Guo, and Jun Wang. Predicting protein-protein interactions using high-quality non-interacting pairs. *BMC Bioinformatics*, 19(S19), December 2018.

[16] : Abhigyan Nath and André Leier. Improved cytokine receptor interaction prediction by exploiting the negative sample space. *BMC Bioinformatics*, 21(1), October 2020.

Biblography, pt. III

[17] : Assaf Rabinowicz and Saharon Rosset. Cross-validation for correlated data. *Journal of the American Statistical Association*, pages 114, September 2020.

[18] : Christian Kramer and Peter Gedeck. Leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets. *Journal of Chemical Information and Modeling*, 50(11):19611969, October 2010.

[19] : T. Frickey and A. Lupas. CLANS: a java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, 20(18):37023704, July 2004.

[20] : A. J. Enright. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):15751584, April 2002.

Authorship note

I declare that this presentation is my original work prepared by me alone and does not contain any content obtained in a manner inconsistent with the applicable rules.