

MethylSight: a Computational Approach to Solving the Methyllysine Proteome

François Charih, Yasser B. Ruiz-Blanco, Kyle K. Biggar, James R. Green

Chemistry and Biochemistry Graduate Research Conference

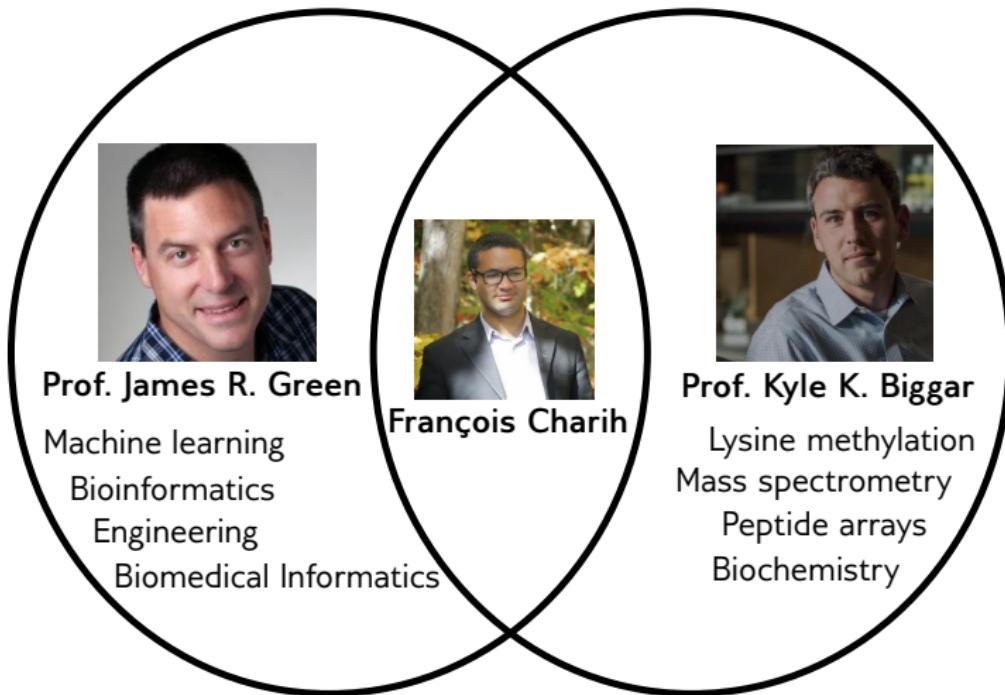


CARLETON UNIVERSITY
BIOMEDICAL INFORMATICS
COLABORATORY



Carleton
UNIVERSITY
Canada's Capital University

Who we are



Lysine methylation: a dynamic process

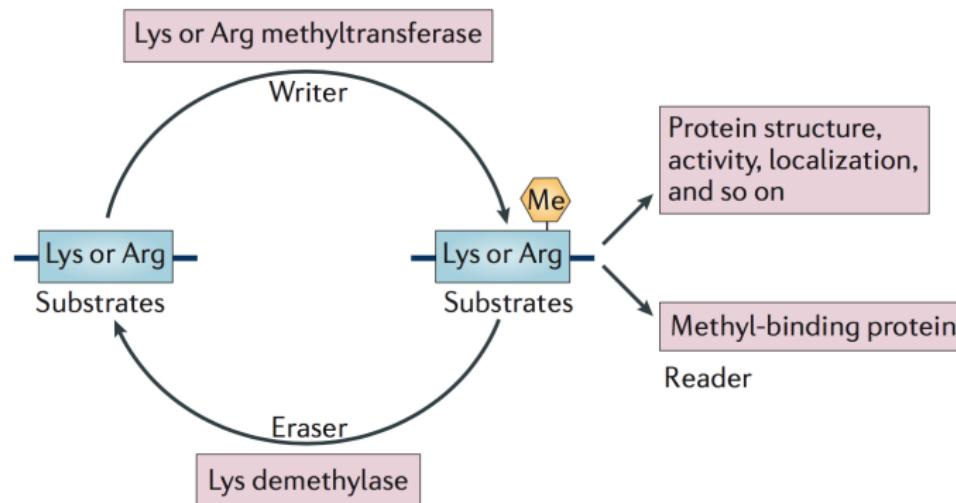


Figure 1: Dynamic methylation/demethylation cycle (Biggar and Li, 2015)

Machine learning for PTM prediction

Machine learning algorithms **recognize patterns** in datasets that are too large to be analyzed by humans.

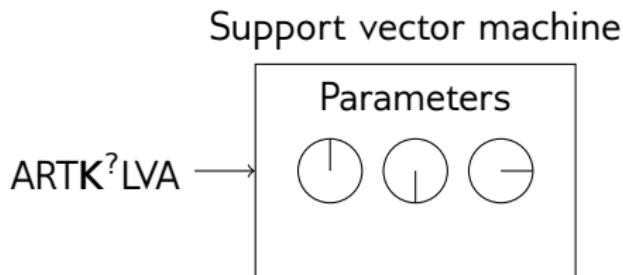
Support vector machine

Parameters



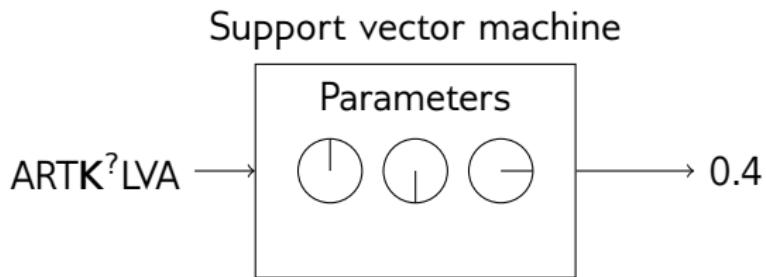
Machine learning for PTM prediction

Machine learning algorithms **recognize patterns** in datasets that are too large to be analyzed by humans.



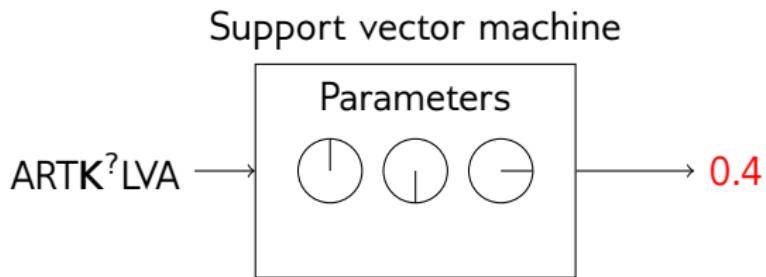
Machine learning for PTM prediction

Machine learning algorithms **recognize patterns** in datasets that are too large to be analyzed by humans.



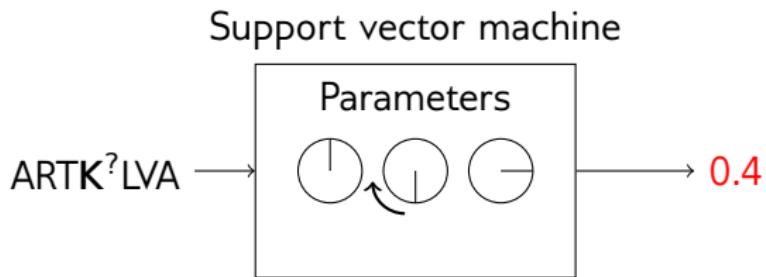
Machine learning for PTM prediction

Machine learning algorithms **recognize patterns** in datasets that are too large to be analyzed by humans.



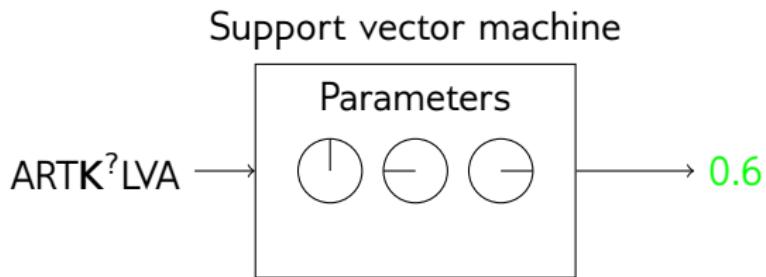
Machine learning for PTM prediction

Machine learning algorithms **recognize patterns** in datasets that are too large to be analyzed by humans.



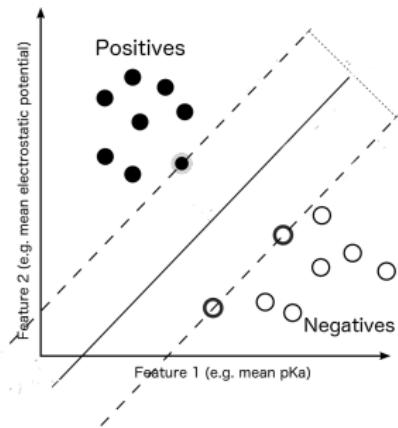
Machine learning for PTM prediction

Machine learning algorithms **recognize patterns** in datasets that are too large to be analyzed by humans.



Machine learning for PTM prediction

Machine learning algorithms **recognize patterns** in datasets that are too large to be analyzed by humans.



Goal: Find the model parameters for which the model's predictions are most often correct on a “training set” of lysine-centered windows.

Building a protein representation

- ▶ Model expects numerical input
 - ▶ Must turn a lysine-centered window of a.a. to numbers

Building a protein representation

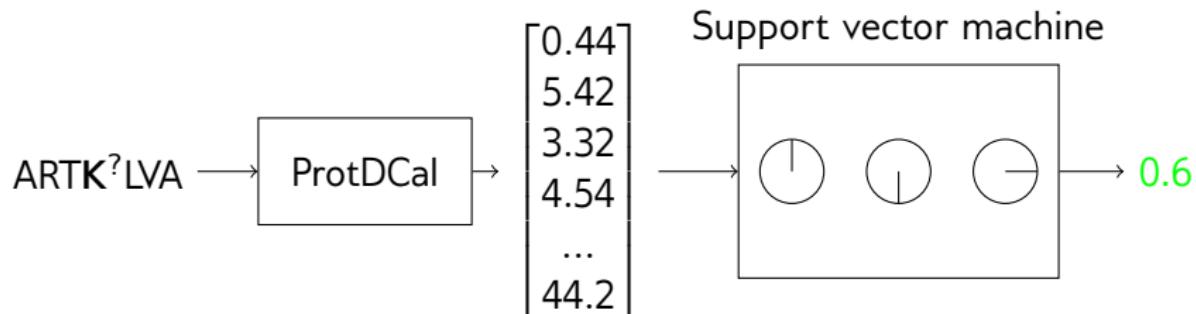
- ▶ Model expects numerical input
 - ▶ Must turn a lysine-centered window of a.a. to numbers
- ▶ ProtDCal (Ruiz-Blanco et al., 2015) feature extraction software
 - ▶ Produces thousands of numerical *features* from sequence

Building a protein representation

- ▶ Model expects numerical input
 - ▶ Must turn a lysine-centered window of a.a. to numbers
- ▶ ProtDCal (Ruiz-Blanco et al., 2015) feature extraction software
 - ▶ Produces thousands of numerical *features* from sequence
- ▶ Must select “best” features
 - ▶ Genetic algorithm for feature selection
 - ▶ Final representation has 28 features

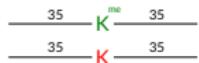
Building a protein representation

- ▶ Model expects numerical input
 - ▶ Must turn a lysine-centered window of a.a. to numbers
- ▶ ProtDCal (Ruiz-Blanco et al., 2015) feature extraction software
 - ▶ Produces thousands of numerical *features* from sequence
- ▶ Must select “best” features
 - ▶ Genetic algorithm for feature selection
 - ▶ Final representation has 28 features



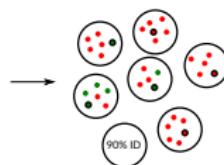
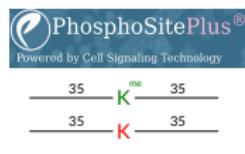
Experimental design

- ① Build a dataset of positive and negatives

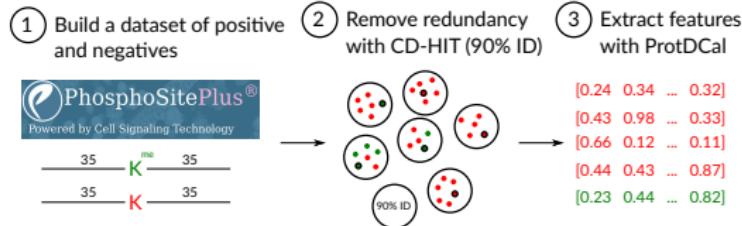


Experimental design

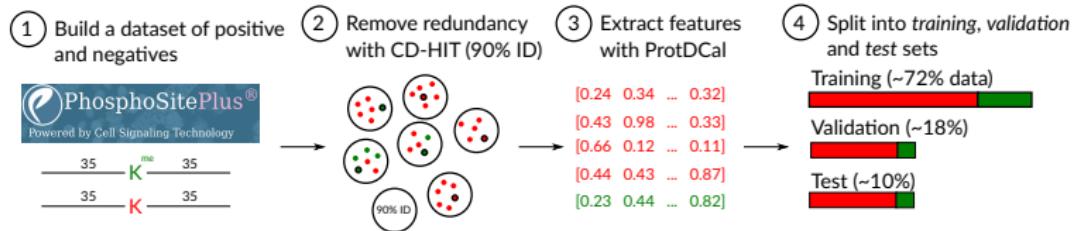
- ① Build a dataset of positive and negatives
- ② Remove redundancy with CD-HIT (90% ID)



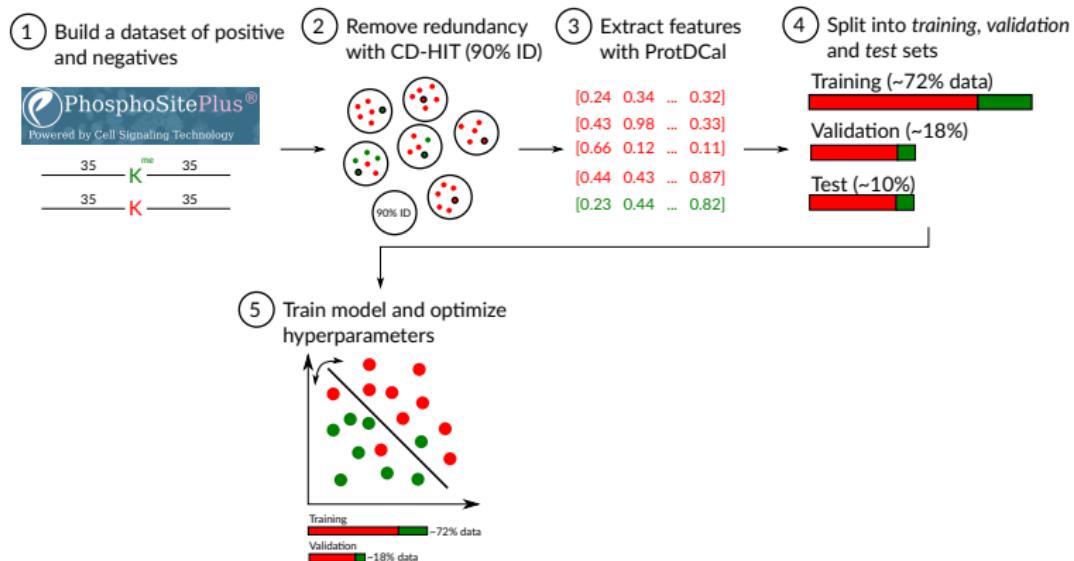
Experimental design



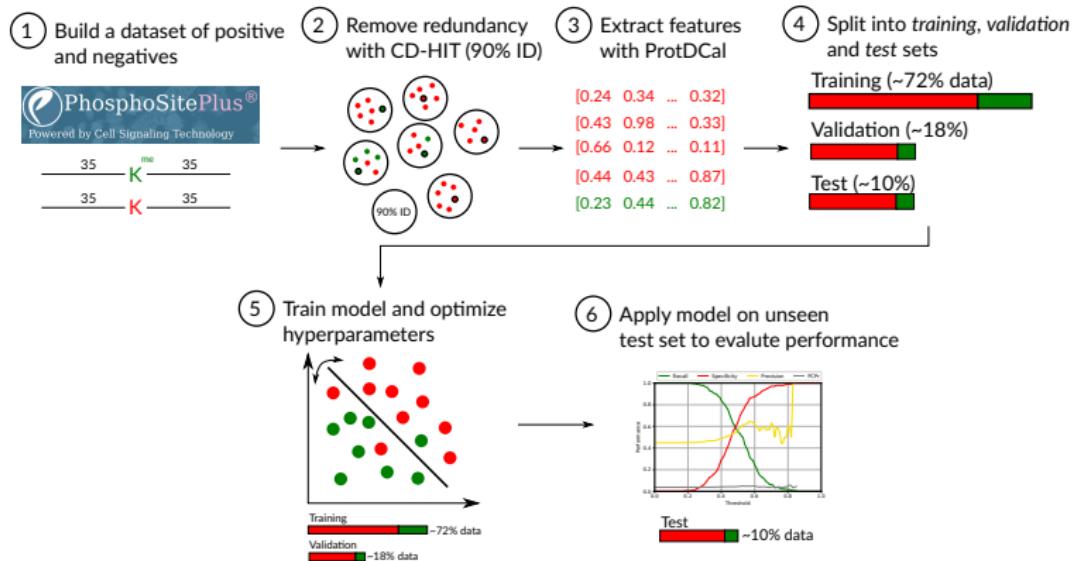
Experimental design



Experimental design



Experimental design



Results: MethylSight outperforms other methods

Table 1: Performance of top predictors

Method	Sensitivity	Precision	F1-score	Accuracy
BPB-PPMS	0.011	0.028	0.016	0.952
MEMO	0.0	0.0	N/A	0.955
iMethyl-PSEAAC	0.083	0.041	0.055	0.900
MethylSight	0.544	0.044	0.081	0.536

Prediction validation

- ▶ Selected at threshold of 0.7 to operate
 - ▶ Methyllysine proteome: 35,973 positive predictions
 - ▶ 95% specificity

Prediction validation

- ▶ Selected at threshold of 0.7 to operate
 - ▶ Methyllysine proteome: 35,973 positive predictions
 - ▶ 95% specificity
- ▶ Selected 57 high-scoring sites within histones for validation via MS, based on availability of purified substrate
 - ▶ 51/57 (89%) actually methylated

MethylSight: a web server

Retrieve predictions (web interface)

Usage

1 - Query the MethylSight database

Our database currently holds the predictions for all potentially methylated lysines in the human proteome. You may retrieve the sites for a given protein with a query. Or if you know what the target is for your favorite protein(s) & its ligand, you can find it here.

Feel free to try out the MethylSight lysine methylation predictor with one of these example proteins: [HUMAN HIF1A](#)-[APOLI](#).

2 - Submit your own protein sequence

The MethylSight lysine methylation predictor can provide predictions for arbitrary protein sequences. Please enter your query as a FASTA-formatted sequence below. Predictions will be returned for each sequence entry. If you have many sequences, consider using our bulk submission service.

More... [View help](#)

Sequence as FASTA-formatted sequence:

```
Q09463|Elongation factor 1-alpha 2 [Q09463] Mammalia; Elongation factor 1-alpha 2 (EF1alpha2) is a member of the EF1alpha family of elongation factors. It is involved in the initiation of protein synthesis at the 5' end of the ribosomal RNA. It is also involved in the recycling of the ribosomes after elongation has been completed. It is a homodimeric protein composed of two EF1alpha subunits. Each subunit contains a nucleotide binding domain and a dimerization domain. The nucleotide binding domain is responsible for the binding of GTP and the dimerization domain is responsible for the formation of the dimeric protein.
```

3 - Submit your own protein sequence

The MethylSight lysine methylation predictor can provide predictions for arbitrary protein sequences. Please enter your query as a FASTA-formatted sequence below. Predictions will be returned for each sequence entry. If you have many sequences, consider using our bulk submission service.

More... [View help](#)

Sequence as FASTA-formatted sequence:

```
Q09463|Elongation factor 1-alpha 2 [Q09463] Mammalia; Elongation factor 1-alpha 2 (EF1alpha2) is a member of the EF1alpha family of elongation factors. It is involved in the initiation of protein synthesis at the 5' end of the ribosomal RNA. It is also involved in the recycling of the ribosomes after elongation has been completed. It is a homodimeric protein composed of two EF1alpha subunits. Each subunit contains a nucleotide binding domain and a dimerization domain. The nucleotide binding domain is responsible for the binding of GTP and the dimerization domain is responsible for the formation of the dimeric protein.
```

Visualize KMT structures

Structure visualization (beta)

MethylSight currently holds the structures of 1000 structures, and we are now working on visualizing the structures of the human lysine methylases. Whether you are looking to see the structure of a specific lysine methylase, or just want to explore the structures of the human lysine methylases, this feature is for you.

Please note that this feature is still in beta and is not yet available for all structures. We are working on improving the visualization and adding more structures over time.

1 - Select a methylation and a representation

Region: N-terminal C-terminal Both ends

Region length (aa): 1000

Description: This visualization displays the structure of the selected lysine methylase in a ribbon representation. The structure is composed of alpha-helices and beta-sheets. The regions selected are highlighted in green. The structure is shown in a 3D space, allowing for rotation and zooming.

2 - Highlight regions of interest

This visualization highlights the regions of the selected lysine methylase that are predicted to be methylated. The regions are highlighted in green. The structure is shown in a 3D space, allowing for rotation and zooming.

Regions of interest: Histone methyltransferase domain Nucleotide binding domain Dimerization domain Other regions

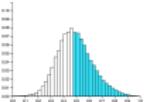
Annotations: Histone methylation activity Nucleotide binding activity Dimerization activity Other annotations

Start: 1 End: 1000 Step: 1000

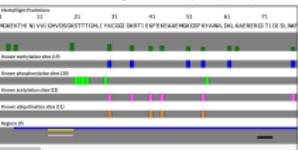
Retrieve predictions (REST API)

Visualize predictions

Distribution of scores for human lysines



Predictions for Elongation factor 1-alpha 2 (Q09463)



<http://www.methylsight.com>

Concluding remarks

Limitations of MethylSight

- ▶ Assumption about negative sites unrealistic
- ▶ Accuracy/precision still modest

Concluding remarks

Limitations of MethylSight

- ▶ Assumption about negative sites unrealistic
- ▶ Accuracy/precision still modest

Future work

- ▶ Investigate deep learning methods for lysine window representation (embeddings, auto-encoders)
- ▶ Determine whether PTM competition and cross-talk can inform a classifier
- ▶ Investigate methyltransferase prediction

References

- Kyle K. Biggar and Shawn S.-C. Li. Non-histone protein methylation as a regulator of cellular signalling and function. *Nature Reviews Molecular Cell Biology*, 16(1): 5-17, January 2015. ISSN 1471-0072, 1471-0080. doi: 10.1038/nrm3915.
- Yasser B. Ruiz-Blanco, Waldo Paz, James Green, and Yovani Marrero-Ponce. ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins. *BMC Bioinformatics*, 16(1):162, May 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0586-0.