MethylSight: a Computational Approach to Solving the Methyllysine Proteome

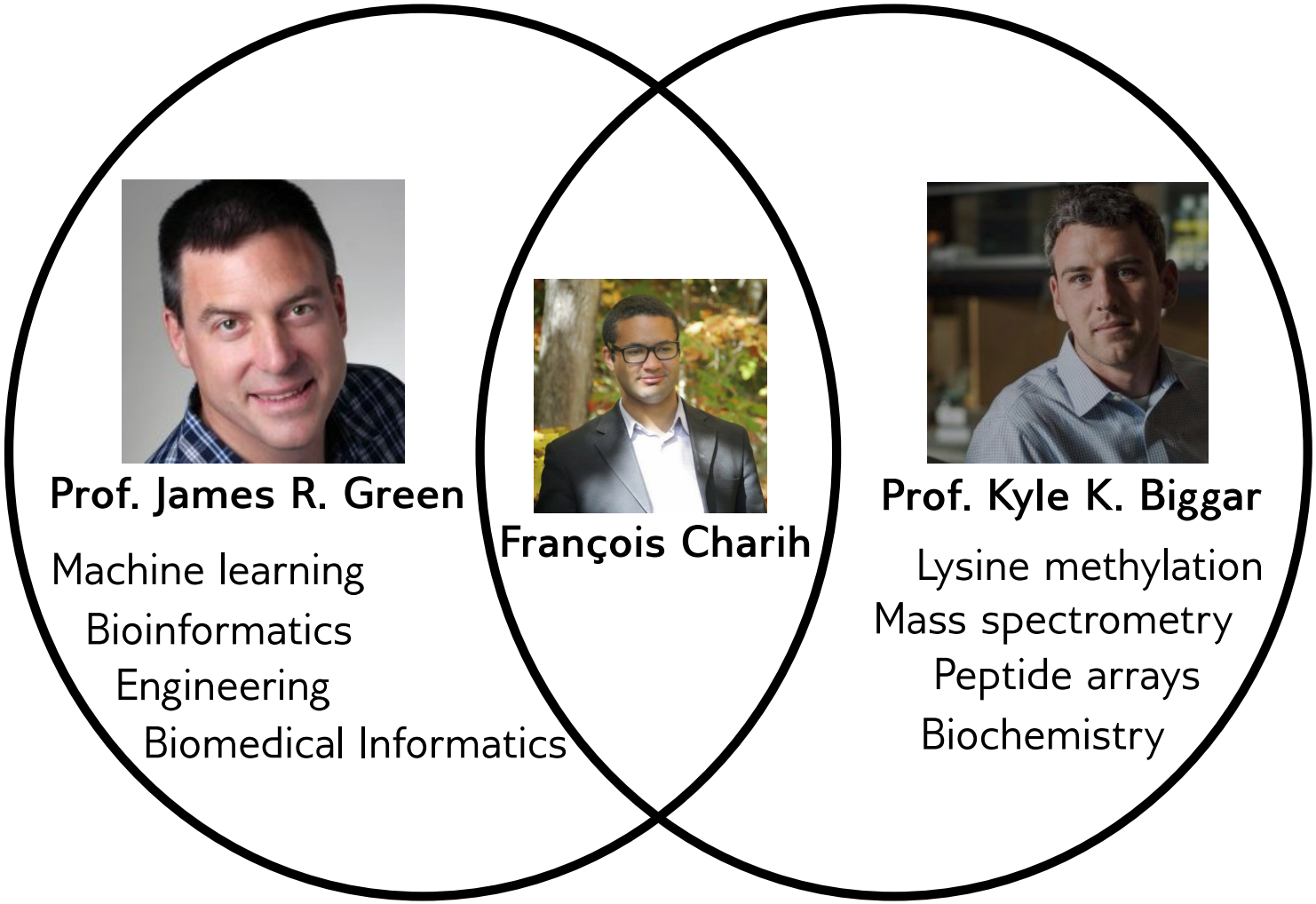
François Charih, Yasser B. Ruiz-Blanco, Kyle K. Biggar, James R. Green

Chemistry and Biochemistry Graduate Research Conference

CU BIC CARLETON UNIVERSITY BIOMEDICAL INFORMATICS COLABORATORY

Carleton UNIVERSITY Canada's Capital University

Who we are


Prof. James R. Green
Machine learning
Bioinformatics
Engineering
Biomedical Informatics
François Charih
Prof. Kyle K. Biggar
Lysine methylation
Mass spectrometry
Peptide arrays
Biochemistry

Lysine methylation: a dynamic process

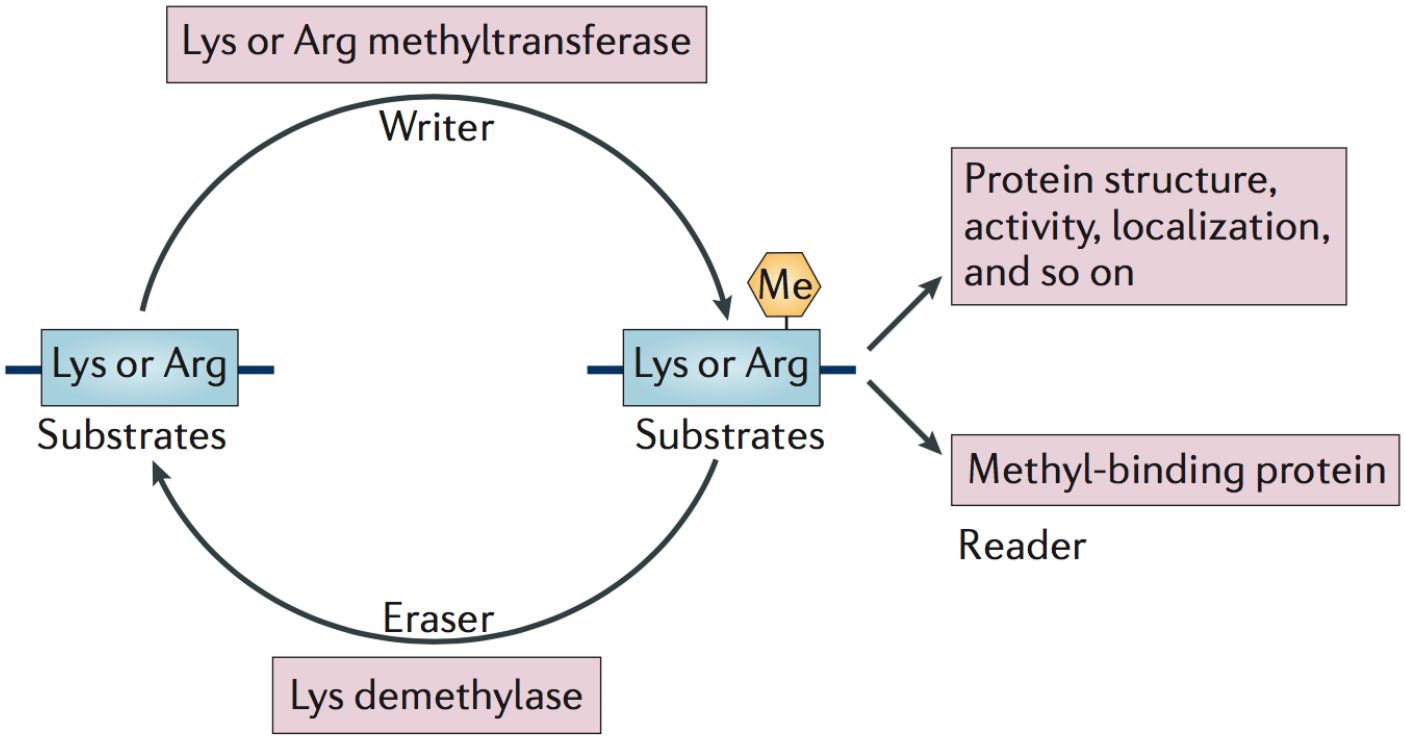
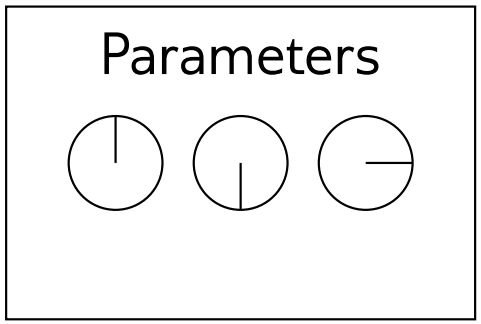


Figure 1: Dynamic methylation/demethylation cycle (Biggar and Li, 2015)

Machine learning for PTM prediction

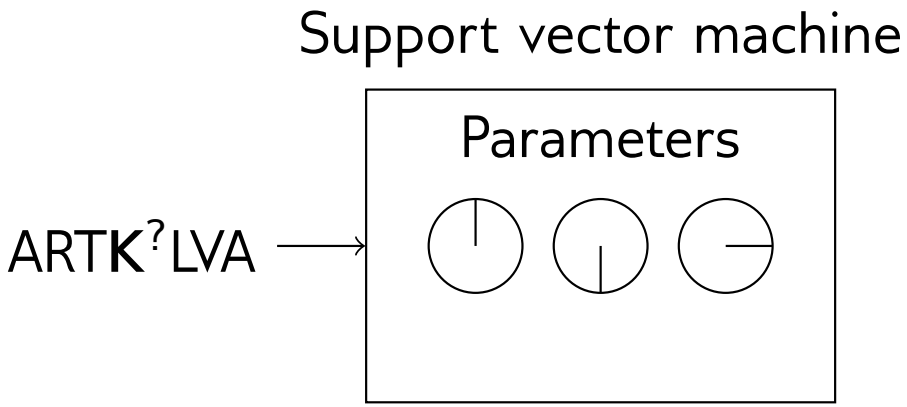
Machine learning algorithms **recognize patterns** in datasets that are too large to be analyzed by humans.

Support vector machine



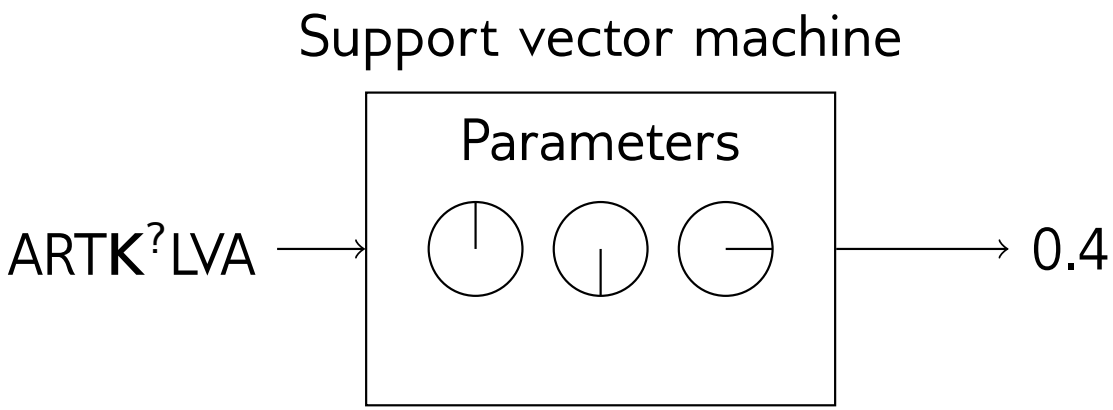
Machine learning for PTM prediction

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



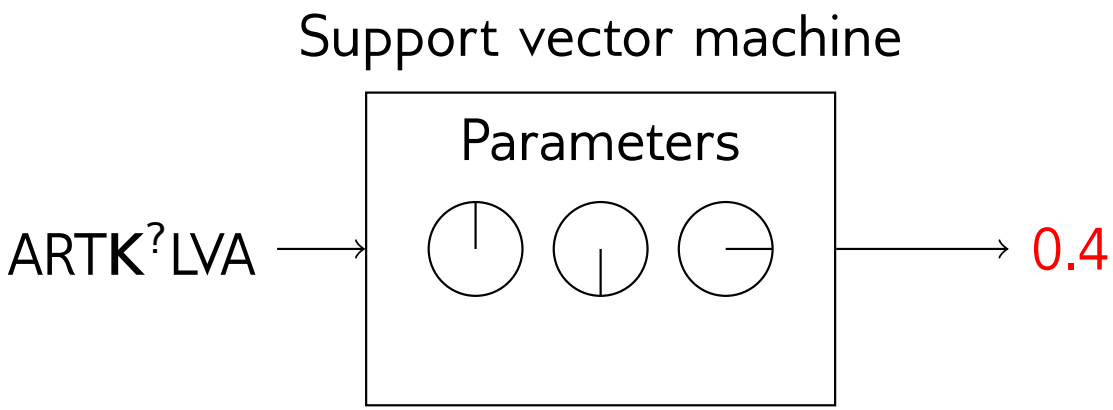
Machine learning for PTM prediction

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



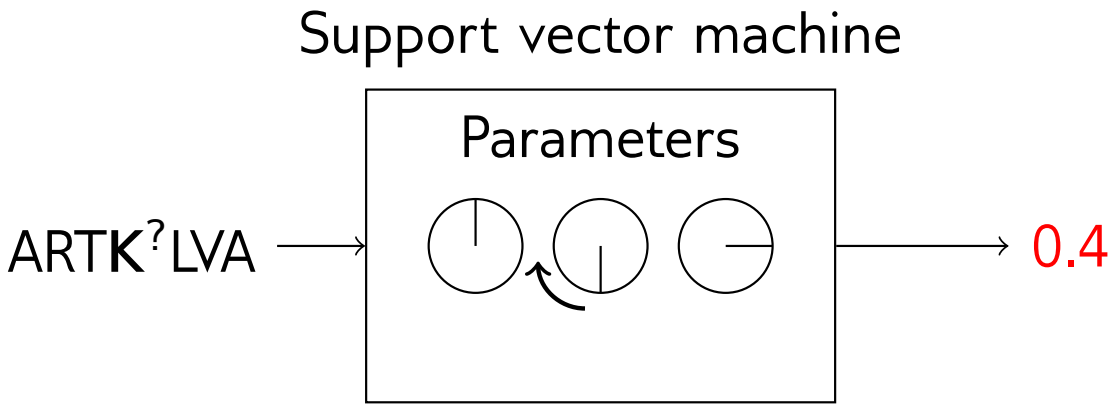
Machine learning for PTM prediction

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



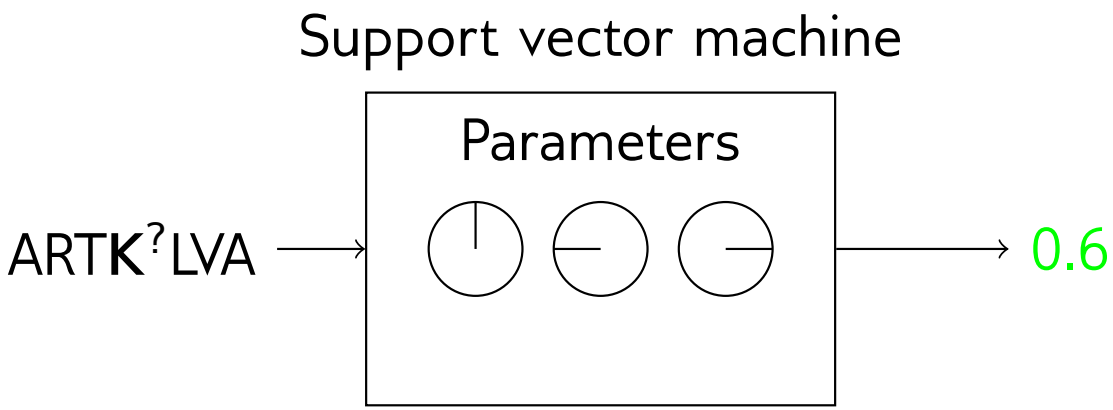
Machine learning for PTM prediction

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



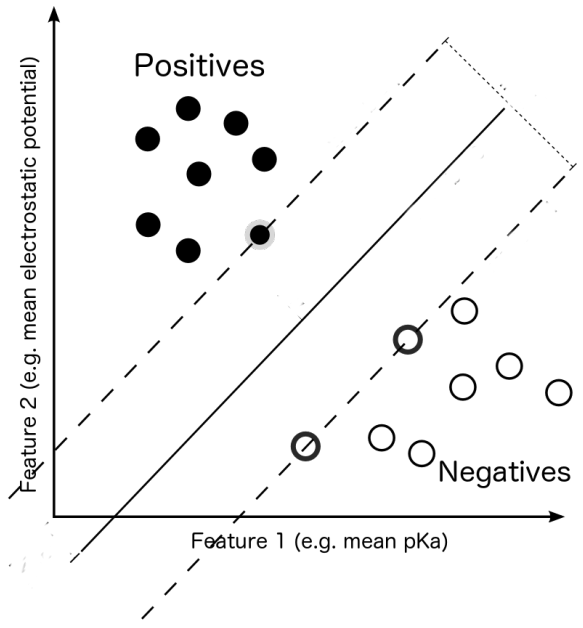
Machine learning for PTM prediction

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

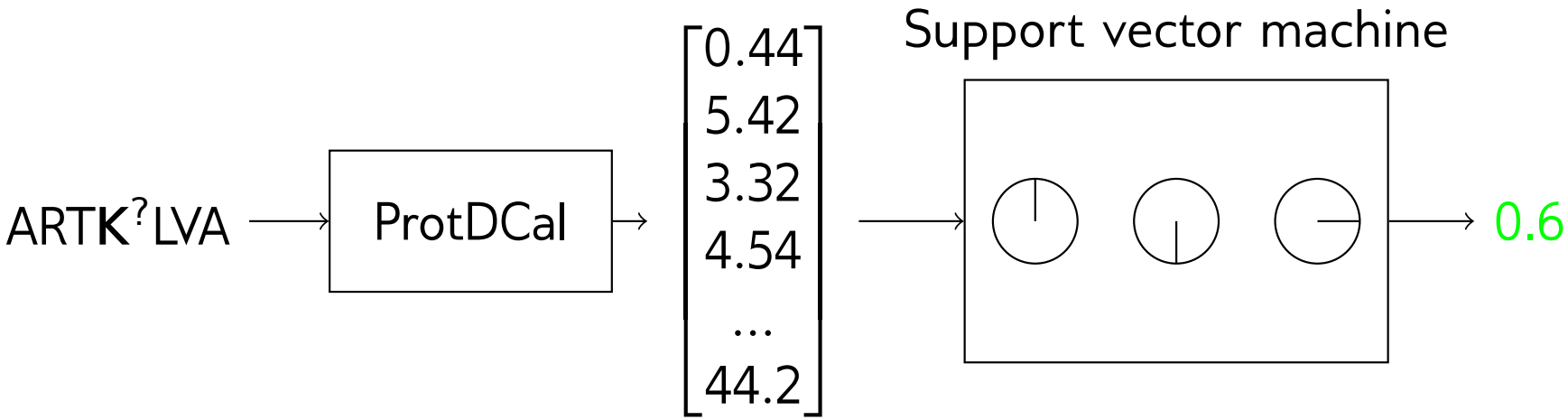
- Model expects numerical input
  - Must turn a lysine-centered window of a.a. to numbers
- ProtDCal (Ruiz-Blanco et al., 2015) feature extraction software
  - Produces thousands of numerical *features* from sequence

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

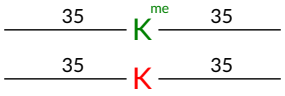
Building a protein representation

- 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

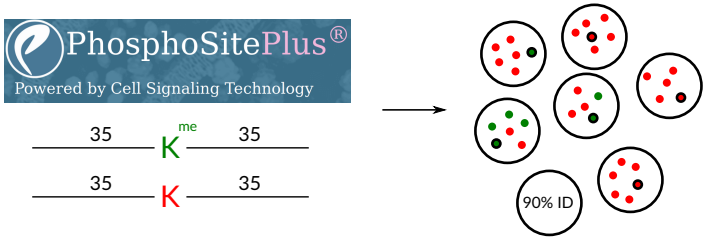
1. Build a dataset of positive and negatives



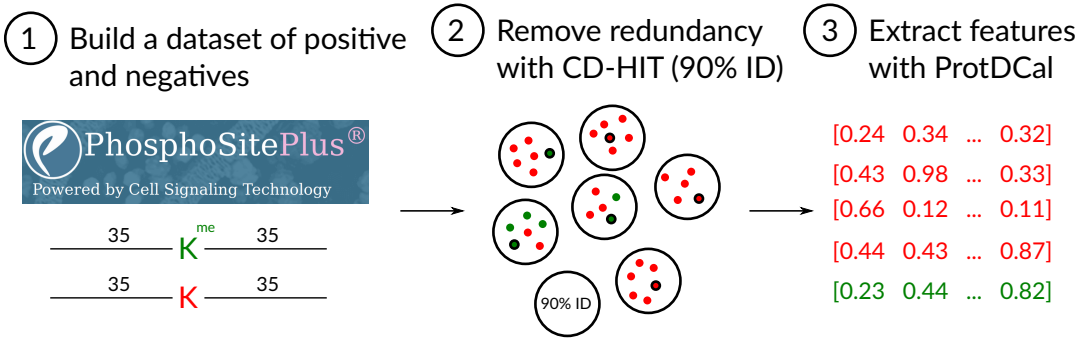
Experimental design

1 Build a dataset of positive and negatives

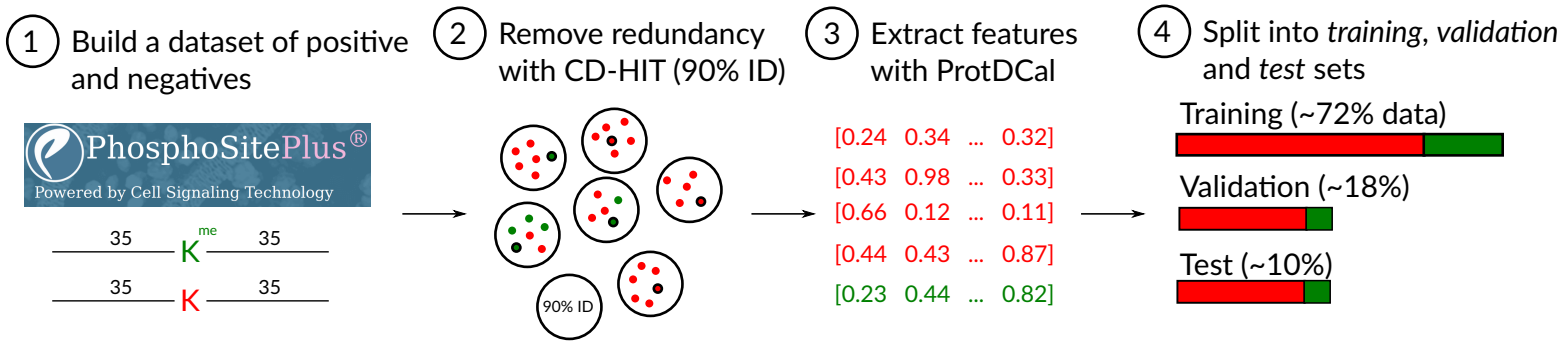
2 Remove redundancy with CD-HIT (90% ID)



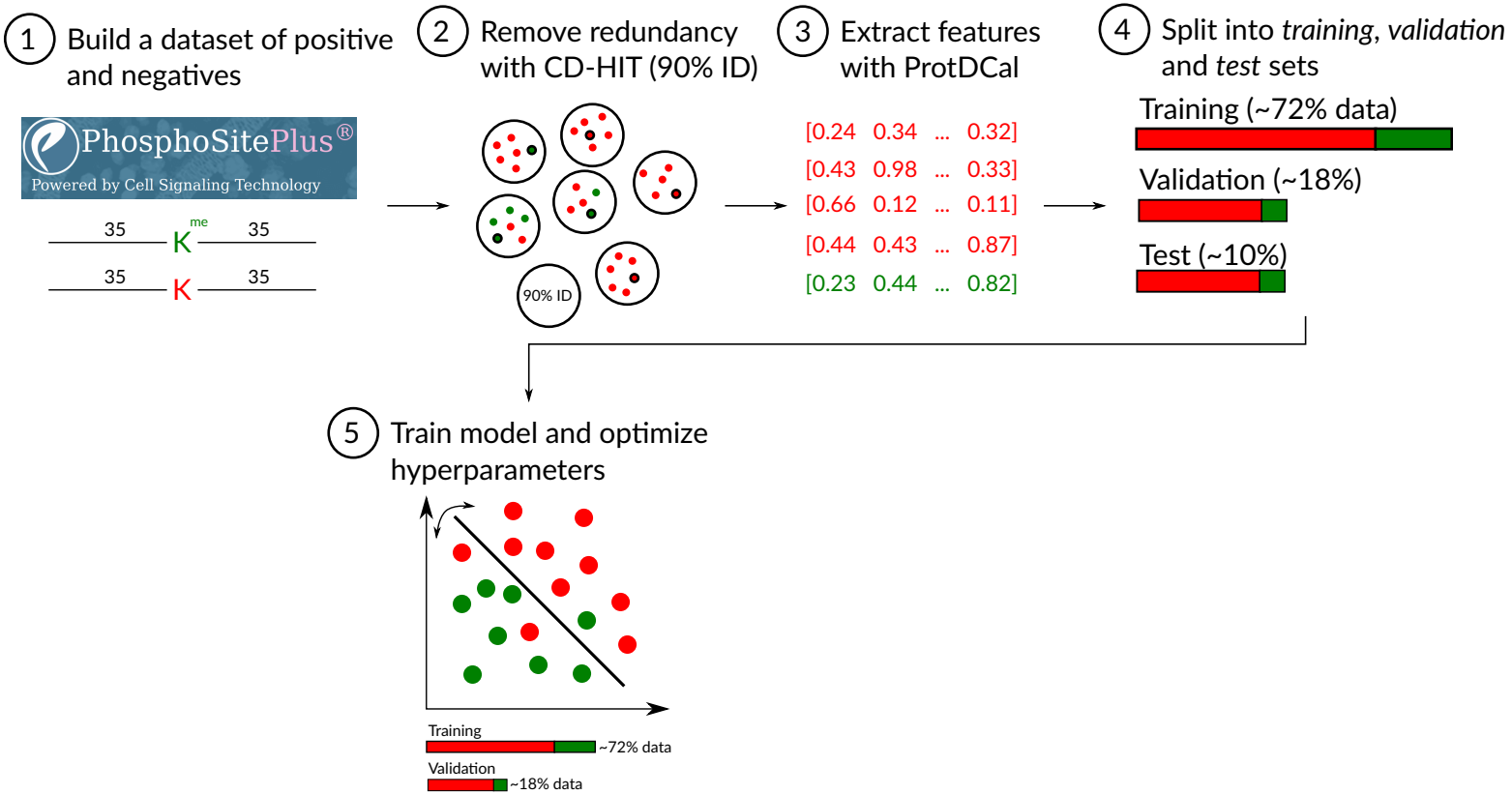
Experimental design



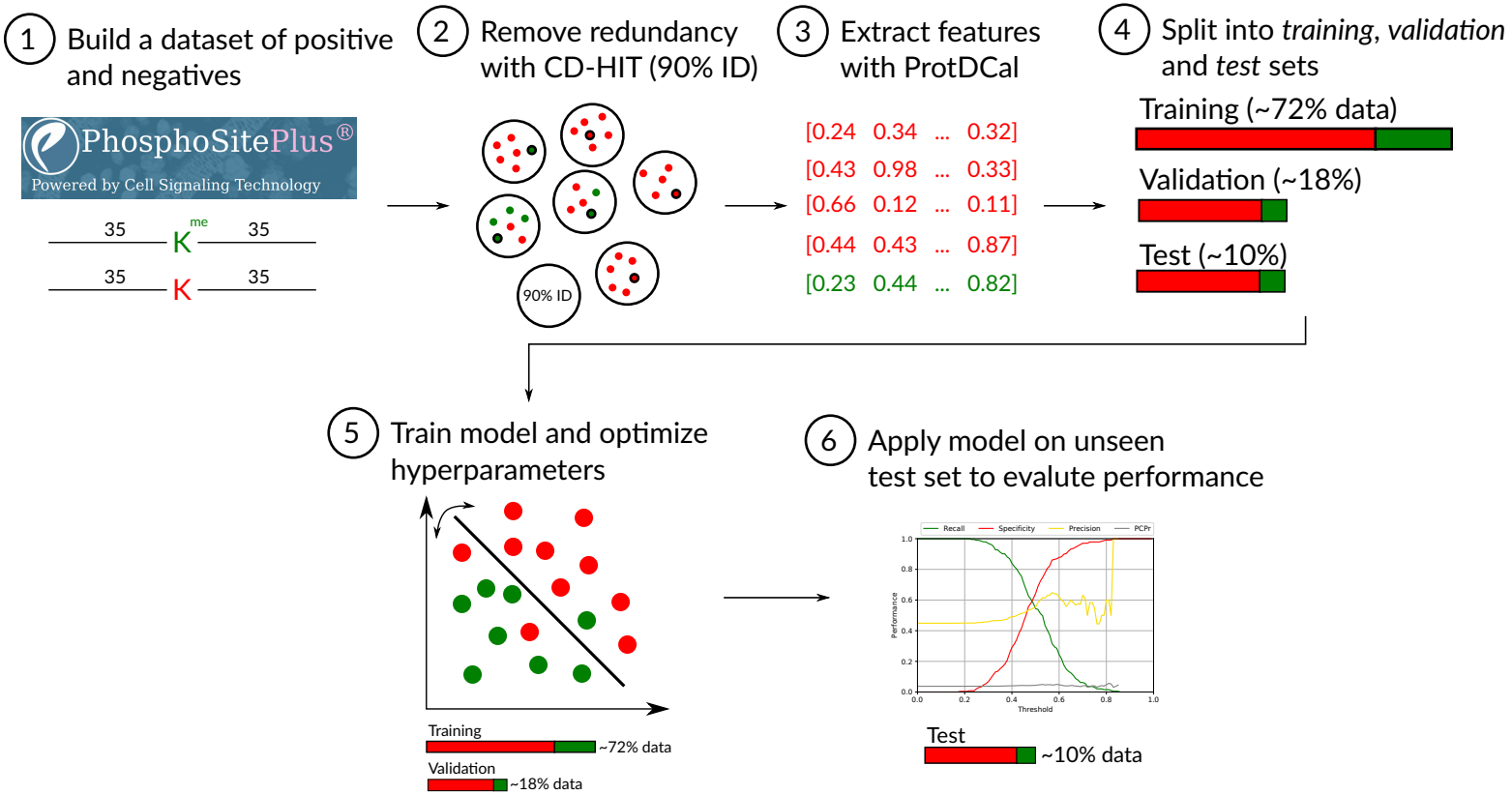
Experimental design



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

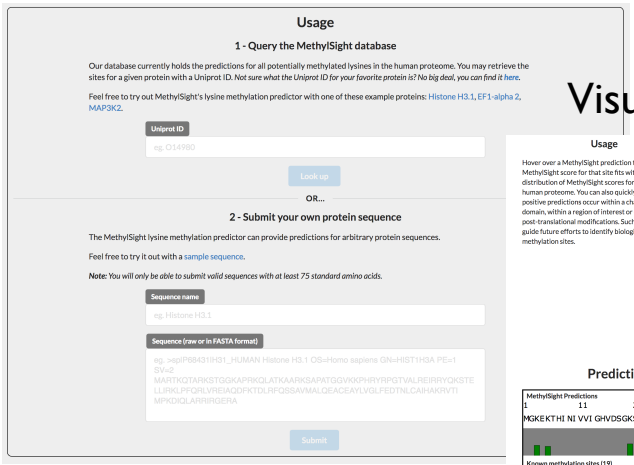
- Selected at threshold of 0.7 to operate
  - Methyllysine proteome: 35,973 positive predictions
  - 95% specificity

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

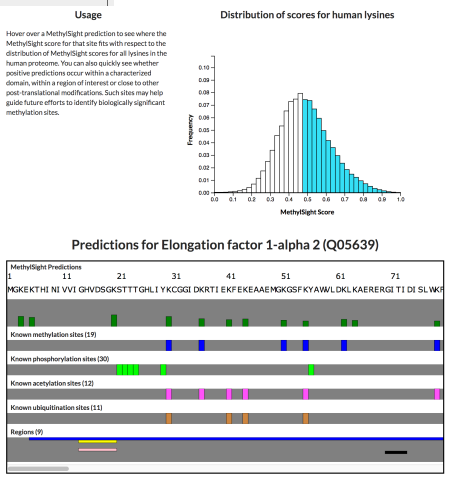
MethylSight: a web server

Retrieve predictions (web interface)

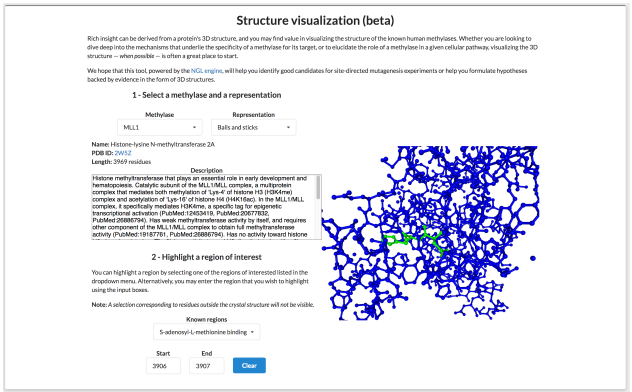


Retrieve predictions (REST API)

Visualize predictions



Visualize KMT structures



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: $10.1038/\mathrm{nrm3915}$.

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: $10.1186/\mathrm{s12859\text{-}015\text{-}0586\text{-}0}$.