Innovations in enzyme-substrate discovery: Decoding the SET8 substrate network using a hybrid approach to machine learning

Nashira H. Ridgeway, Anand Chopra, and Kyle K. Biggar
Insitute of Biochemistry, Carleton University, Ottawa ON Canada

Carleton University

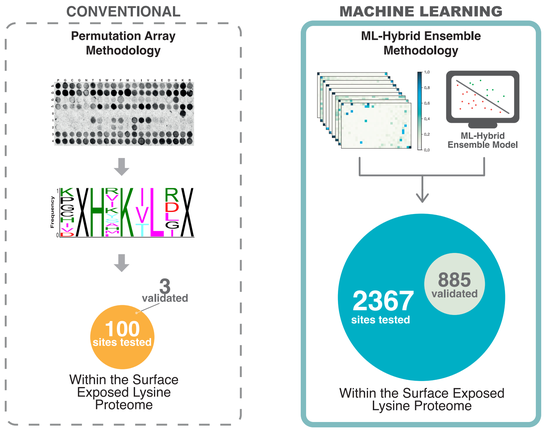
ABSTRACT

Identifying novel sites of post-translational modifications (PTMs) amidst the vast array of potential candidates within the proteome presents a formidable challenge. Through the integration of machine learning (ML) models with high-throughput *in vitro* peptide synthesis, we have pioneered a ML-hybrid search methodology to better predict enzyme-substrate selection. The resulting ensemble model demonstrates a true, experimentally validated precision of 37.4% and has uncovered 885 novel active SET8 methylation sites within the human proteome *in vitro*. Targeted mass spectrometry analysis has provided evidence that SETD1B-K41me1, KAT6A-K314me1 and PRDM12-K269me1 postively respond to SET8 overexpression in mammalian cells. This innovative approach has also illuminated critical insights into the breast cancer proteome, showcasing the gain (376) and loss (62) of SET8 substrates arising from missense mutations. This computationally driven methodology recaptures the essential biochemical properties representing enzyme substrates and stands as a transformative approach with the potential to redefine the landscape of enzyme-substrate discovery.

PROJECT OVERVIEW

In this study, we transcend traditional techniques by adopting a ML-hybrid ensemble approach to enzyme-substrate identification, using SET8 as a model. The success of the proposed method hinges on the following aims:

1. Successfully training a high-performance ML ensemble model using carefully curated peptide array data of the entire methyllysine proteome for SET8.
2. Application of the model to the human proteome to predict novel sites of SET8 methylation and evaluate predictions *in vitro* with peptide arrays and MS.
3. Characterise the effect of breast-cancer mediated missense mutations on SET8 pathways by the change in predicted ML-hybrid ensemble score.



METHODS

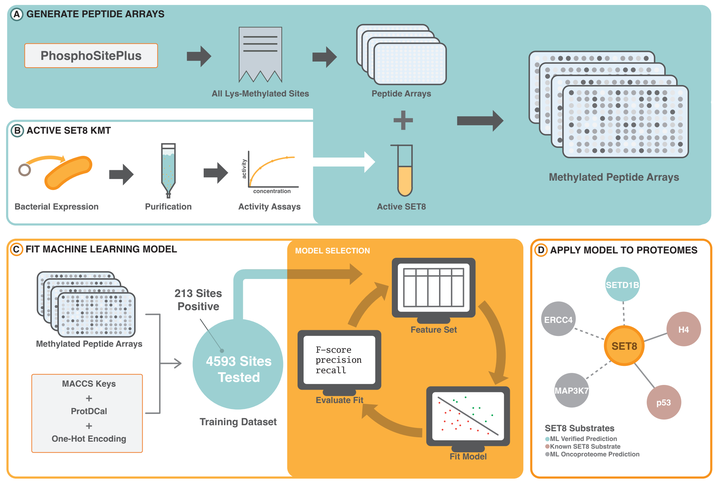


Figure 1. Graphical overview of the *in vitro* and *in silico* methods employed to generate a ML-hybrid ensemble model for the prediction of SET8 substrates. The generation of data is completed with peptide arrays consisting of the lysine methylome and a SET8 construct. ML models and data balancing methods are applied and assessed for performance on the imbalanced SET8 dataset with metrics such as F-score, precision and recall. Finally, the resulting model is applied to the proteome to search for novel substrates of SET8 within normal and cancerous cells.

RESEARCH IMPACT

1. The proposed ML-hybrid ensemble method is 37.4% accurate at identifying substrates of SET8, and identified 885 novel sites of SET8 methylation activity *in vitro*.
2. Per MS-monitoring, predicted substrates SETD1BK41me1, KAT6AK314me1, and PRDM12K269me1 were found to be elevated with SET8 overexpression.
3. Analysis of breast cancer-mediated missense mutations with the ML-hybrid ensemble model discovered sites that gained (376) and lost (62) SET8 methylation activity.
4. This method can seamlessly be applied to a range of lysine methyltransferases and other enzymes responsible for PTMs.

RESULTS

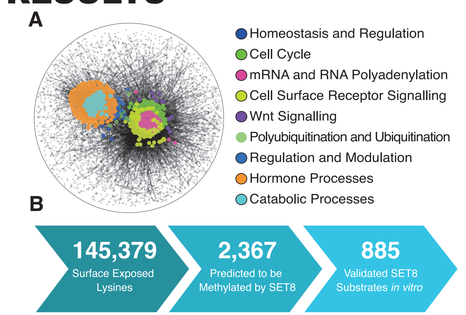


Figure 2. ML-hybrid ensemble model results of the proteome set containing surface exposed lysines. **A** SAFE mapping of the model's predictions to the HuRI interactome, clustered for associated biological processes. **B** Progression of predictions and validated SET8 substrates with peptide array experiments.

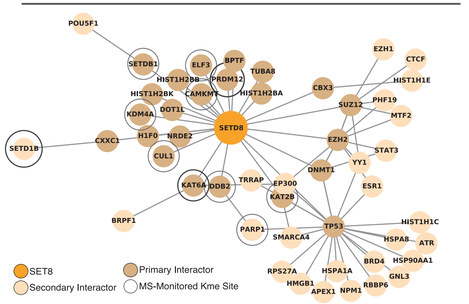


Figure 3. Primary and secondary interactors of SET8 predicted to be methylated by SET8 by the ML-hybrid ensemble model represented within the STRING database, to guide MS-monitoring. Circles represent MS-monitored sites.

Table 1. Comparison of methods for substrate discovery including a random search, the SET8 recognition motif as generated with permutation arrays, and the ML-hybrid ensemble model as applied to the proteome dataset comprising surface exposed lysines.

METHODS	VALIDATED ACCURACY
Random Search	0%
SET8 Recognition Motif	3.0%
SET8 ML-Hybrid Model	37.4%

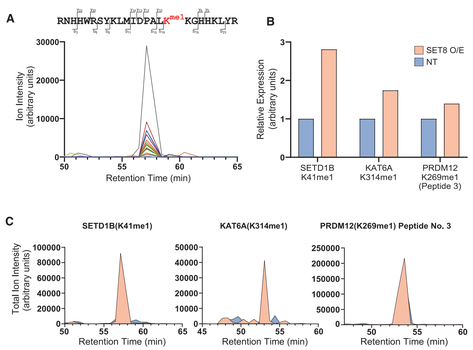
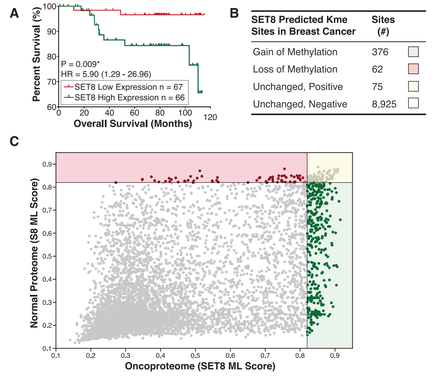


Figure 4. Targeted MS analysis of the validated predictions of SET8 methyltransferase activity, as generated by the ML-hybrid ensemble model. **A** Results of SETD1BK41me1, including the peptide sequence featured within the isolation list. **B** Average relative expression of SETD1BK41me1, KAT6AK314me1, and PRDM12K269me1 of SET8 overexpression (SET8 O/E) as compared to wild type (NT). **C** Total peak intensities of targeted MS analysis for the three substrates.



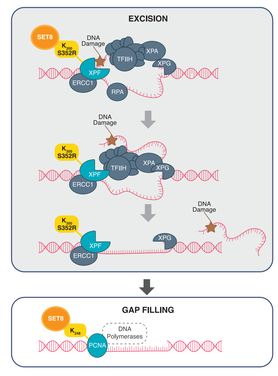


Figure 5. Oncoproteomic investigation of SET8 in breast cancer. **A** Kaplan-Meier plot depicting patient survival with high (green) and low (red) SET8 expression in breast cancer. Adapted from Liu B, Zhang X, Song F, et al. A functional single nucleotide polymorphism of SET8 is prognostic for breast cancer. *Oncotarget*, 2016. **B** Predicted SET8 sites of lysine methyltransferase activity, as compared between healthy and mutated cancerous states. Squares within each row correspond to color in **C** the visualization of cancer mutated and healthy SET8 predicted ML-hybrid ensemble scores for each site analyzed. **D** Nuclear excision repair pathway as initiated by DNA damage. Mutation of XPF (ERCC4) involved in critical excision step is predicted to invoke SET8 methyltransferase activity by the ensemble model. Gap filling by PCNA, a known substrate of SET8 in normal cells is included.

Author's Contact Information: gNashira@gmail.com | github.com/nashirag
Corresponding Author's Contact Information: kyle_biggar@carleton.ca | www.biggarlab.ca