Scientific Reports

https://doi.org/10.1038/s41598-026-39136-9

Article in Press

Leveraging learned representations and multitask learning for lysine methylation site discovery

Received: 2 September 2025

Accepted: 3 February 2026

Published online: 23 February 2026

Cite this article as: Charih F., Boulter M., Biggar K.K. *et al.* Leveraging learned representations and multitask learning for lysine methylation site discovery. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-39136-9

François Charih, Mullen Boulter, Kyle K. Biggar & James R. Green

Leveraging learned representations and multitask learning for lysine methylation site discovery

François Charih[1,2,3,*], Mullen Boulter[2], Kyle K. Biggar[2,3,+] and James R. Green[1,3,+]

[1] Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada
[2] Institute of Biochemistry, Department of Biology, Carleton University, Ottawa, ON, Canada
[3] NuvoBio Corp., Ottawa, ON, Canada
[+] These authors contributed to the work equally

* Corresponding author: François Charih (francoischarih@cmail.carleton.ca)

Abstract

Lysine methylation is a dynamic and reversible post-translational modification of proteins carried out by lysine methyltransferase enzymes. The role of this modification in epigenetics and gene regulation is relatively well understood, but our understanding of the extent and the role of lysine methylation of non-histone substrates remains somewhat limited. Several lysine methyltransferases which methylate non-histone substrates are overexpressed in a number of cancers and are believed to be key drivers of cancer progression. There is great incentive to identify the lysine methylome, as this is a key step in identifying drug targets. While numerous computational models have been developed in the last decade to identify novel lysine methylation sites, the accuracy of these models has been modest, leaving much room for improvement. In this work, we leverage the most recent advancements in deep learning and present a transformer-based model for lysine methylation site prediction which achieves state-of-the-art accuracy. In addition, we show that other post-translational modifications of lysine are informative and that multitask learning is an effective way to integrate this prior knowledge into our lysine methylation site predictor, MethylSight 2.0. Finally, we validate our model by means of parallel reaction monitoring mass spectrometry experiments and identify 68 novel lysine methylation sites. This work constitutes another contribution towards the completion of a comprehensive map of the lysine methylome by providing a revised estimate of its extent to approximately 155,000 sites. Of those, MethylSight 2.0 is expected to correctly detect ~47,000, which is substantially more than expected with competing methods, which we show to be less sensitive on a subset of experimentally validated novel methylation sites. We foresee that MethylSight 2.0, whose performance significantly surpasses that of competing models, will facilitate the discovery of a large number of novel methylation sites.

Keywords: Lysine methylation, lysine methylome, deep learning, transformers, multitask learning

Introduction

Lysine methylation extends far beyond the realm of histone proteins and it may be more prevalent than previously believed[1]. Studies have uncovered the involvement of non-histone lysine methylation in oncogenic processes including chemoresistance and cancer cell proliferation[2-4], making it a very attractive target for anti-cancer therapies. Therapies targeting non-histone lysine methyltransferases (KMTs) are emerging, with some showing promising signs of efficacy at the clinical trial stage[5,6]. For instance, Tazemetostat, an inhibitor of EZH2, was trialed and received approval for the treatment of blood and solid malignancies[5]. EZH2 promotes tumorigenesis in glioblastoma and prostate cancer models via STAT3 methylation[7,8] and in diffuse large B-cell and follicular lymphomas via methylation of the PRC2 complex[5]. Considering that many cancers are driven by KMT overexpression, uncovering the human lysine methylome and the associated KMTs would have profound implications in drug discovery, as it could facilitate the identification of drug targets for therapeutic intervention. In addition, it promises to broaden our general understanding of the lysine methylation-dependent biological processes.

Identification of novel lysine methylation sites is possible through experimental methods such as mass spectrometry (MS)[9]. That said, the process is sufficiently resource-intensive that identifying new sites at the proteome scale experimentally is impractical. For that reason, a number of machine learning prediction models have been developed over the years to tackle the lysine methylation prediction problem. The rationale behind these models is that computation can guide our efforts so that time and resources can be invested on validating the most promising potential lysine methylation sites.

A wide range of machine learning models trained on lysine methylation datasets to predict lysine methylation sites from sequence only have been published over the last two decades. Most of them have relied on the use of “traditional” machine learning algorithms such as support vector machines (SVMs) or random forests (RFs) and human-crafted numerical features. The first predictor, MeMO[10], was built using an SVM classifier using what is now referred to as “one-hot” encoding for a 15 amino acids lysine-centered window. The training set used in that study consisted of a total of 156 positive lysine methylation sites, which represents only an infinitesimal fraction of the space of all lysine-centered 15-mers ($156/14^{20} \approx 10^{-19}$% of the space of possible windows). More recent models used different strategies to represent lysine-centered windows. For example, iMethyl-PseAAC[11] used a SVM model in conjunction with a representation the authors termed “pseudo amino acid composition” (PseAAC). This representation combines evolutionary information from the position-specific scoring matrix (PSSM), physicochemical properties of individual amino acids from the AAIndex[12], and disorder scores to generate a 346-dimensional feature vector. Another well-cited method is GPS-MSP (Group-based Prediction System Methyl-group Specific Predictor), an algorithm published in 2017[13], was trained on 1,521 methyllysine sites and used an alignment-based custom scoring function to measure window similarity in conjunction with □-means clustering (unsupervised learning) to predict not only methylation sites, but also the methylation state (mono-, di-, or tri-), an ambitious task given the scarcity of data available to train models to this level of granularity. Met-Predictor[14] is another method which was trained to predict all three lysine methylation states from interpretable sequence and structural features. In 2020, we described the original MethylSight lysine methylation predictor[15], an SVM-based model trained on 28 protein features generated with ProtDCal[16], a toolkit which computes and aggregates the physicochemical properties of amino acid sequences into thousands of

machine learning-compatible numerical descriptors of proteins. MethylSight 1.0 set itself apart from its competitors as one of the few post-translational modifications (PTM) predictors to have been subjected to experimental validation. In fact, MethylSight 1.0 enabled the identification of H3BK43 as a novel methylation site, which we later identified as a potentially important site for the differentiation of mouse embryonic stem cells[15]. While MethylSight 1.0 outperformed the state-of-the-art at the time, prediction accuracy remained somewhat modest.

Attempts to leverage deep learning methods to address the challenging task of accurately predicting the lysine methylome remain scarce as of the time of writing. Recently, Spadaro *et al.* applied convolutional neural networks (CNNs) to representations combining phylogenetic, physicochemical, structural, and binary encodings to predict lysine methylation sites[17]. The PTM-Mamba model[18] makes use of the Mamba architecture, an attention-free state-space model with architectural optimizations designed to enhance computational efficiency over long sequences. More specifically, PTM-Mamba fuses Mamba-generated sequence embeddings with embeddings generated by the ESM-2 protein language model (pLM) to predict a wide array of PTMs which do not include lysine methylation.

Bepler and Berger have shown that multitask language models better capture the semantic organization of proteins by training a bidirectional long short-term memory (LSTM) to complete three tasks simultaneously: masked language modeling, residue-residue contact prediction, and structural similarity prediction[19].

It is a reasonable supposition that there might be some partial overlap in the physico-chemical environments that promote lysine methylation and other PTMs, such as solvent accessibility, surrounding amino acid properties, and steric constraints. This idea is supported by the fact that numerous sites are subject to more than one PTM (Figure 1). As a result, the lysine methylation prediction problem is amenable to a multitask learning formulation, wherein lysine methylation prediction is one task among several PTM prediction tasks, and that jointly training a single model on several such tasks concomitantly could lead to better prediction accuracy through knowledge transfer.

To our knowledge, this idea has been exploited only once for the uncommon *propionylation* PTM of lysines[21]. In that work, the authors trained a recurrent neural network (RNN) on lysine malonylation sites and *fine-tuned* it using a dataset of lysine propionylation sites to extract features that are then fed into an SVM classifier. That work did exploit transfer learning, but did not use a multitask learning scheme, as the training was not performed for both tasks (*i.e.* propionylation and malonylation prediction) *simultaneously*. A model was trained for the malonylation prediction task first, and subsequently trained to predict propionylation sites, of which there were fewer known instances (431 as opposed to 9,584).

Currently, no one has proposed a lysine methylation site prediction model that leverages 1) state-of-the-art neural architecture, *i.e.* the transformer *and* 2) domain adaptation by means of transfer learning techniques such as multitask learning. In addition, very few groups have proven with *in vitro* experiments that the estimates of accuracy of their models translate into the lab upon deployment.

In this work, we address these opportunities to develop a more accurate and robust lysine methylation predictor. MethylSight 2.0, unlike its SVM-based predecessor trained using human engineered protein representations, leverages the transformer architecture and protein representations learned by pLMs trained at metagenomic

scales, *i.e.* on millions of protein sequences. These advancements result in significant improvements in performance. Our contributions are summarized below:

Contribution 1 - Improved prediction accuracy: We bootstrap embeddings generated with pLMs trained on millions of protein sequences to train a model, MethylSight 2.0, which produces dramatically more accurate predictions than previous lysine methylation predictors;

Contribution 2 - Use of multitask learning to enhance model accuracy: We demonstrate that training a transformer-based neural network architecture with a multitask learning strategy can lead to more accurate predictions;

Contribution 3 - Experimental validation: We show, by means of MS validation experiments performed on sites predicted to be methylated by our model, MethylSight 2.0, that the accuracy of our model translates experimentally and identify 68 novel lysine methylation sites;

Contribution 4 - Bioinformatics analysis of the MethylSight 2.0-predicted lysine methylome: We deploy MethylSight 2.0 at the proteome scale to identify previously unknown methylation sites and conduct analyses to identify biological processes wherein lysine methylation sites are overrepresented.

Methods

Lysine methylation dataset preparation

With the intent of creating a dataset for multitask learning involving multiple PTM types, we retrieved PTM data for lysines occurring in human proteins by mining the PhosphoSitePlus database[20] (10/17/24 update) for methylation, ubiquitination, sumoylation, and acetylation, which are all known to be modifications of lysines. The composition of the PhosphoSitePlus dataset is summarized in Table 1.

Gathering positive lysine modification data is relatively straightforward, but identifying the “negative” sites to complete the training set needed to train a binary classifier is more arduous. It is difficult to ascertain with confidence that a lysine not known to be modified *never* is. In reality, sites taken to be “negative” may correspond to yet-to-be discovered methylation sites. Some groups simply take as negative training examples sites not known to be modified[14,26], which we argue might disproportionately bias the learning algorithm towards making negative predictions. For this reason, it is typical to only use a subset of sites without PTM annotations as negative in PTM site prediction challenges, following some heuristics[15]. A typical practice is to only label as negatives unlabeled sites that occur within a protein containing a known modified site elsewhere[10,15,27–29]. To train and test MethylSight[15], a lab-validated SVM-based model that achieved state-of-the-art performance upon publication, we applied two additional criteria to label potential lysine methylation as “negative” in addition to the latter. More specifically, sites were considered “negative” in the training set if they were not known to be substrates for another PTM (ubiquitination, sumoylation, or acetylation) *and* were predicted to be buried (relative solvent accessibility factor < 0.2, as predicted with NetSurfP v1.0[30]).

In this work, we applied the same curation method (Figure 2), but used an updated version of NetSurfP (v3.0[31]). This approach allowed us to build a dataset with high-confidence negatives.

To address the issue of redundancy in the data, which could cause overrepresentation of certain patterns in the dataset and data leakage, *i.e.* similar patterns in the training

and test data, we clustered the windows based on sequence identity at a similarity threshold of 70% with CD-HIT[32], as done previously[15], and selected one representative from each cluster at random, favouring a positive representative (methylation site) if one occurred within a cluster. Finally, 20% of the non-redundant sites were set aside for testing.

We used an identical workflow to assemble the training sets for ubiquitination, acetylation, and sumoylation sites required for multitask learning, but did not set any data aside for testing, given that we are only interested in methylation site prediction.

The composition of the final dataset is presented in Table 2.

Pre-trained protein-language model embeddings

pLMs have been shown to generate rich embeddings that capture physicochemical, phylogenetic and structural information that are extremely useful for a variety of downstream tasks, including structure prediction[33,34], property prediction (*e.g.* viscosity[35], stability[36], *etc.*), localization prediction[36,37], and peptide binder design[38–40], to cite a few.

Given that these representations were learned on massive collections of protein sequences and performed well on these tasks, we hypothesized that they may also contain useful information for the prediction of lysine methylation sites. Moreover, these embeddings capture more context about the potential methylation sites than traditional human-engineered representations. They consider a large portion (or all) of the protein, depending on the pLM's context length (window size) and the protein length.

We leveraged representations learned by three state-of-the-art foundational model pLMs: ProtT5[41], ESM-2[33], and Ankh[37] (Table 3) and fed the human proteome to all three models to generate embeddings for each lysine in the training and test sets – and to later predict the comprehensive human lysine methylome.

Training multilayer perceptrons leveraging pLM-generated embeddings

We trained 4 separate multilayer perceptrons (MLPs), each taking as an input the lysine residue-specific embeddings extracted with either ProtT5, ESM-2, Ankh, or all three (combined with concatenation). Hyperparameter tuning was performed using a random grid search strategy with Optuna[42] to determine the optimal combinations of learning rate, number and width of hidden layers, and the dropout rate used for regularization. The final selected hyperparameter sets were those that yielded the highest validation area under the precision-recall curve (AUPRC). The models were trained using PyTorch[43] with the Adam optimizer[44] with a batch size of 64, using binary cross-entropy as the loss function:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)$$

where $y_i = 1$ if the site i is methylated and $y_i = 0$ otherwise, while $\hat{y}_i \in [0,1]$ is the predicted probability of that the site i is methylated.

We selected the model using an early stopping strategy, using the validation loss to monitor for overfitting. We repeated this procedure using the embeddings generated by all three aforementioned pLMs. In addition, we trained MLPs on a "combined"

representation resulting from a concatenation of all three embeddings, for a total of four final MLP models.

Training a transformer model

To determine whether training a transformer-based model could further improve the quality of the predictions, we implemented a model which leverages this architecture.

We used a context size of 31 amino acids, the ProtT5 embeddings as representations for the individual amino acids in the sequence (*i.e.* the tokens) and padded with a zero-filled 1,024-D vector if the lysine site was too close to the end of the protein chain (Figure 3A). We opted to use ProtT5 embeddings only and not the concatenation of ProtT5/ESM-2/Ankh embeddings to reduce computational requirements and because concatenation of embeddings only marginally improved performance when training MLPs on pLM embeddings. To capture the positional information of the individual tokens, we used the canonical positional embedding strategy described in the original transformer paper[45]. We used 4 heads in each attention block. The Adam optimizer was used, but with a batch size of 128.

Similarly to the approach used to train the MLPs, we conducted hyperparameter tuning in a randomized fashion and varied the number of encoder transformer blocks, the learning rate, the number of hidden layers in the classification module (*i.e.* the dense layers that follow the transformer layers), and the width of the "embedding layer" and trained a total of 50 models.

We used the same loss function and early stopping strategy as for the MLPs to select the final model for each run. The final transformer architecture selected was the one with the highest AUPRC on the validation set.

Multitask learning with a transformer model

To investigate whether a multitask learning strategy could enhance the quality of the predictions, we enriched our training set with sites and their annotation for the three other PTMs of interest: acetylation, ubiquitination, and sumoylation.

In this context, the "tasks" consist in predicting the four different PTMs of interest. We do not know or can't assume with a satisfying level of certainty the true label for each task for all sites. For example, we may know that a site is acetylated, but not know whether it is also ubiquitinated. Consequently, we chose to not associate each instance or site with 4 labels. Instead, each instance in the dataset is a site associated with a PTM and a label associated to that site and PTM. Consequently, a given site may appear up to 4 times in the training set, in the specific case where a label for each PTM is known.

We implemented another transformer model where the last transformer block is followed by a flattening layer whose output is sent to one of four classification heads, depending on the task, *i.e.* prediction of methylation, acetylation, *etc.* (Figure 3B). Each classification head is designed to predict whether a site is subjected to the corresponding PTM. For each instance in the training set, we only probe the probability output by the head corresponding to the PTM ("task") associated with the instance.

We use a custom batch sampling strategy to train the model wherein all instances in a batch are associated with only *one* of the four PTM. This ensures that the loss over a batch is only used to update the parameters of the classification head associated with the PTM (and the upstream parameters), but not the three classification heads which are used for the other tasks. In other terms, we use *partial* parameter sharing, *i.e.* only the parameters in the transformer layers and upstream are shared across tasks.

Given that the methylation sites are vastly outnumbered by other sites, we multiply the loss for methylation batches by a factor γ in order to produce larger updates for the shared model weights when methylation sites are misclassified relative to misclassified instances of other PTMs. We tried $\gamma \in \{1, 13.5, 20\}$, 1:13.5 being approximatly the methylation-to-other PTMs ratio.

The loss function for the multitask learning strategy effectively takes the form:

$$L = \gamma L_{CE,me} + L_{CE,ub} + L_{CE,ac} + L_{CE,su}$$

where

$$L_{CE,t} = \begin{cases} -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i), & \text{if batch if for task t} \\ 0, \text{ otherwise} \end{cases}$$

The rest of the model selection was done as for the transformer model without the multitask learning training strategy described in the previous section. We henceforth refer to this model as **MethylSight 2.0**.

Estimation of the expected imbalance

To accurately estimate the precision of MethylSight 2.0 upon deployment on the human proteome, an estimate of the class imbalance is required. The human proteome in UniProt/Swiss-Prot database (2025_01 release)[46] comprises 654,185 lysines residues in 20,417 unique proteins, of which an unknown fraction can be methylated under specific biological circumstances such as in response to a biological event, in a stage of development, or in specific tissue types.

Berryhill *et al.*[47] published a study which provides some useful insight into the ratio of methylated to unmethylated lysines observable through mass spectrometry experiments. In their study, they assessed the sequence bias of commercially available pan-methyllysine antibodies and performed global profiling of lysine methylation in HEK293T (human embryonic kidney) and U2OS (human osteosarcoma) cells with samples enriched with anti-Kme1, anti-Kme2, and anti-Kme3 antibodies, as well as with unenriched samples. They identified a total of 5,089 lysine methylation sites evenly distributed through the proteome, of which 4,862 are novel.

Using the data collected in this study, we made the assumption that the estimated the imbalance ratio of methylated-to-unmethylated lysines *detectable via mass spectrometry without and following enrichment* with the antibodies currently in use to be roughly 1:36. This corresponds to the ratio of methylated lysines to lysines not found to be methylated in the proteins that were pulled down in the samples (*i.e.* with at least one epitope for the anti-Kme antibodies used). It is difficult to speculate about what lysines are or are not methylated in proteins that were not pulled down, so we only estimate what one may observe in a global profiling experiment with mass spectrometry. We use this ratio to evaluate the anticipated precision of MethylSight 2.0, when coupled with a mass spectrometry experiment.

This figure is an approximation derived from samples extracted from two specific cell types, and as such, it may not apply uniformly across all tissue types and across the entire proteome.

Selection of predicted methylation sites for in vitro validation

We subsequently sought to estimate the actual precision of MethylSight 2.0 upon deployment onto the human proteome. To achieve this, we selected 100 sites predicted to be methylated by MethylSight 2.0, but which were not known methylation sites. Using a conservative threshold on predicted methylation probability (*i.e.* $PCPr_{1:36}$ = 0.75), we sampled 50 sites at random from each of the following two sets:

1. **Set 1:** Exposed lysine residues known to be acetylated, ubiquitinated, and/or sumoylated;
2. **Set 2:** Exposed lysine residues with no known modification.

Furthermore, under the hypothesis supported by the phenomena of PTM competition that these other modifications provide useful information for the identification of novel lysine methylation sites, one would expect to detect more methylation events within sites sampled from Set 1 than within sites sampled from Set 2. To allow for this comparison, we ensured that the methylation probabilities were similarly distributed in both samples.

As stated previously, in contrast with “positives”, it is very difficult to decisively confirm a site as “negative”. A negative observation for a given site in a PRM-MS is insufficient to rule out the possibility that the site could be methylated in a different cell type or in under different biological conditions. As such, sampling negative sites would be largely uninformative. Furthermore, tools like MethylSight 2.0 are most often used to detect novel lysine methylation sites, not to identify negative sites. For these reasons, we chose not to sample negative sites for experimental validation.

Validation of predicted methylation sites via mass spectrometry

Using the Pyteomics package for Python[48], we generated an isolation list tabulating the tryptic peptide fragments (assuming no missed cleavage) and their mass-to-charge ratios for the +2, +3 and +4 charged states and for all four methylation states (*i.e.* null, mono-, di-, tri-), resulting in a total of 1,200 predicted peaks. The isolation list can be found in the supplementary materials.

Parallel reaction monitoring mass spectrometry (PRM-MS) experiments were conducted at the John L. Holmes Mass Spectrometry Facility at the University of Ottawa with a Q Exactive™ Plus Hybrid Quadrupole-Orbitrap™ mass spectrometer, using the aforementioned isolation list to guide the scanning. The results were obtained from a single injection of a Thermo-Fisher Pierce™ HeLa protein digest standard pooled from several vials into a single master batch to maximize sample homogeneity. We intentionally opted to monitor for methylation in this sample because it is guaranteed to have a low missed tryptic cleavage rate (<10%) and minimal methionine oxidation and lysine carbamylation (<10%). Furthermore, these standards are thoroughly tested for quality, which improves the reproducibility of the results and eliminates the need for biological replicates. Because of very low biological noise and because PRM-MS is highly targeted and non-stochastic, a single run was judged sufficient for the purpose of validating a subset of positive predictions made with MethylSight 2.0.

Analysis of the spectra was carried out in Skyline[49] (version 24.1.0.414) and followed a protocol described previously[50]. Target ions with a minimum of 3 transition states were considered positive detections. Detection of an unmethylated target with no methylated target was considered a “negative” (in the sample). Detection of a methylated target of any degree, regardless of the detection of the corresponding

unmethylated target, were deemed positive results. Targets which were undetected as methylated and unmethylated ions were considered to be inconclusive results.

Results and discussion

MethylSight 2.0: performance and benchmarking

The predictive performances on the blind test set of the final models are tabulated in Table 4.

All models trained as part of this work perform significantly better on the blind test set than methods published previously (GPS-MSP[13], MethylSight 1.0[15], Met-Predictor[14]), although these methods have likely encountered some of the sites in our test set during training, which would confer them an unfair advantage. In fact, our worst performing model, a MLP using lysine embeddings produced by the ESM-2 pLM was associated with a 25% improvement over the state of the art (SOTA) in terms of both AUPRC and precision at 0.5 recall (Pr@0.5Re).

Among the MLP models we trained on the four representations produced by the three pLMs considered, the model trained on ESM-2 embeddings performed noticeably worse relative to the other three representations which produced similar results, though the model trained on the combined embeddings appears to have a performance modestly superior to that of the MLPs trained on ProtT5 and Ankh embeddings. The relative performance of the different representations is consistent with the sizes of the pLMs, ProtT5 being 3 times the size of Ankh, and 4.6 times the size of ESM-2 in terms of learnable parameter counts. This observation is consistent with evidence that pLM performance scales with model size following a power law[51,52]. The performance advantage gained by using ProtT5's embeddings is also partially explained by the fact that the model was pre-trained on a significantly larger dataset of ~2.1B sequences, which is orders of magnitude larger in size than the training sets used to pre-train ESM-2 and Ankh.

The use of a transformer architecture trained "from scratch" specifically for the task of predicting lysine methylation prediction did allow for an improvement in performance over the use of lysine embeddings generated by all three foundational pLMs in our MLPs. Our best single-task transformer model slightly outperformed the best MLP (*i.e.* the MLP trained on concatenated ProtT5-Ankh-ESM-2 embeddings), using AUPRC as a metric. However, it produced a more significant improvement in precision at the a 50% recall threshold of nearly 10%. This showcases the power of the self-attention mechanism, as attention layer parameters in our transformer model were learned specifically for the lysine methylation prediction task as opposed to the more general masked language modeling objective, as was case for the foundational models.

Looking at the precision-recall curves (PRCs) assessing model performance on our blind test set (Figure 4A), we see that implementing a multitask learning strategy leveraging knowledge about other PTMs is useful, as our best model (hyperparameters tabulated in Table 5) outperforms all other models over the entire range of possible recall values (or operating thresholds). However, this advantage is anticipated to dissipate at higher recall values assuming that the true class imbalance of methylation sites to non-methylation sites in the proteome is higher than that in the test set (*e.g.*, 1:36 as opposed to 1:6.5; Figure 4C). This suggests that knowledge of other PTMs of lysines can indeed transfer to lysine methylation. This is consistent with our initial hypothesis as well as with the established phenomenon of PTM competition wherein lysine modifying enzyme "compete" to modify specific lysine residues[53–55].

The predicted human lysine methylome

Using conservative settings (*i.e.* at $PCPr_{1:36}$ = 0.75), MethylSight 2.0 identified a total of 62,567 lysine methylation sites within 13,791 different proteins in the human proteome (Figure 5A). Based on our performance assessment and an estimated imbalance ratio of 1:36, we anticipate that out of these predicted sites, ~47,000 (75%) are actual methylation sites. This figure, alone, is significantly higher than the ~30,000 sites predicted with 63% precision by MethylSight 1.0. Given that the estimated recall of MethylSight 2.0 under these conditions is ~30%, we estimate the size of the lysine methylome at ~155,000.

Statistically significant enrichment in several biological process and molecular function GO terms were identified (Figure 5B). Of note, we observed overrepresentation in methylated proteins of terms related to translation, ribosomal biogenesis and structure, and cytoskeleton structure. Enrichment of related terms in methylated proteins was also observed within the MethylSight 1.0-predicted lysine methylome[15], and is also supported by the literature. For instance, methylation of several lysine residues in the human elongation factor 1A (eEF1A) is known to regulate ribosome biogenesis and actin cytoskeleton dynamics, among others[56]. The methylation of elongation factor eEF2 by the KMT FAM86A is also known to regulate translation dynamics[57]. The literature also supports the involvement of lysine methylation in cytoskeleton regulation. The role of lysine methylation in cytoskeleton regulation is well-established[58]. The α-tubulin cytoskeletal protein is known to be tri-methylated by SETD2 (and acetylated) on K40, and loss of methylation has been associated with “catastrophic microtubule defects” which impair DNA repair mechanisms[59] and cell cycle progression[60]. Recently, methylation of BCAR3 on K334 by SMYD2 in breast cancer was shown to enhance lamellipodia dynamics of breast cancer cells through the recruitment of Formin-like proteins which accelerate actin polymerization and facilitate cell proliferation and metastasis *in vivo*[61].
Interestingly, our SAFE analysis of methylated proteins mapped onto the HuRI human interactome (Figure 5C) also shows subnetworks where overrepresentation of GO terms related to RNA processing and regulation and cytoskeleton organization is observed.

In Figure 5D, we illustrate the predicted prevalence of lysine methylation events in a subset of the actin cytoskeleton pathway (KEGG[62]: hsa04810). Most proteins in this important subset of the pathway contain several lysine methylation sites.

Identification and validation of novel lysine methylation sites

The PRM-MS experiments on a HeLa cell lysate guided with an isolation list listing tryptic peptides corresponding to MethylSight 2.0-predicted methylation sites revealed a significant number of hits (Figure 6A). In fact, 68 of the 100 sites predicted to be methylated produced transitions consistent with methylated peptides with a fair or better level of confidence. In contrast, for only 6 of the sites could evidence of the unmethylated peptide and no evidence of methylation be found. The results were inconclusive for 26 peptides, *i.e.* the peptide could not be detected, neither in a unmethylated state, nor in a methylated state. In the worst case where we consider all inconclusive sites to be negative, MethylSight 2.0 would achieve a precision of 68%. The precision becomes 91.9% if we discard sites for which no transitions could be detected from the analysis. It is likely that the precision we would have observed, if all results had been conclusive, would lie somewhere within that range. This suggests that an imbalance ratio of 1:36 is a reasonable estimate.

Interestingly, we found that more methylation sites were found for lysines that were not known to be otherwise modified (39) than were found for lysines known to be acetylated, ubiquitinated, or sumoylated (29). This observation contradicts our initial hypothesis that more methylation sites would be detected among proteins which are known to be otherwise modified, because of their "modifiable" character. One plausible explanation is that one or more of these other modifications might have in fact competed with methylation, thus reducing its abundance and preventing its detection.

In order to compare the sensitivity of MethylSight 2.0 with that of competing methods assessed in this work, we applied the latter to the 68 novel sites that were predicted to be methylated by MethylSight 2.0 and confirmed by PRM-MS. Even with a liberal decision threshold of 50%, the other methods mislabeled many of those validated sites as negative. MethylSight 1.0 correctly labeled 48 sites (70%) as positive, while GPS-MSP recalled 32 sites (47%) and Met-Predictor recalled only 17 sites (25%). This finding further demonstrates the superior sensitivity of MethylSight 2.0 relative to its predecessor and other competing methods.

We choose here to highlight two sites occurring within proteins of high biological and clinical significance which produced transitions unequivocally consistent with methylation: the □ subunit of the Eukaryotic elongation factor 1 (eEF1□) and DnaJ homolog subfamily B member 11 (DNAJB11).

eEF1□ is one of the four subunits of the eEF1 complex, along with the □, □ and □ subunits[63]. Though not believed to be a catalytically active member of the complex[64], eEF1□ is believed to act as a structural scaffold for the □ subunit and to facilitate the complex's function of bringing aminoacyl tRNAs to the ribosome for translation[60]. Beyond its role in the elongation factor 1 complex, eEF1□ is predicted to have "moonlighting" roles and be involved in several other biological processes including viral ribonucleic acid (RNA) transcription, oxidative stress response, cytoskeleton-membrane linking, and cellular trafficking[65]. MethylSight 2.0 correctly predicted the methylation of K428 in eEF1□ (Figure 6B), which is located within the C-terminus domain of the protein. While the N-terminus end of eEF1□ is known to interact with eEF1□ and anchor it into the complex, the role of the C-terminus end of the protein is not a clearly understood, aside from the fact that it is highly conserved and protease resistant[66]. There is some evidence that interaction with the □ subunit of the eEF1 complex may occur at the C-terminus domain[67]. Therefore, it is plausible that the methylation status of eEF1□-K428 could modulate the interaction between these the □ and □ subunits. Given that eEF1□ has been found to interact with actin[68], it is possible that methylation of K428 could modulate this interaction and influence cytoskeleton dynamics if it indeed occurs at the C-terminus. eEF1□'s clinical significance is supported by the observation that it overexpressed in gastric carcinoma[69], colon adenocarcinoma[70], and pancreatic cancer[71], in all likelihood so that cancer cells can satisfy the higher translation load required to adapt and proliferate. Altogether, our observation that eEF1□ is methylated on K438 combined with its involvement in key biological processes and cancer warrants further investigations into the biological significance of the modification.

Clear transitions consistent with the presence of a methyllysine were also recorded for the tryptic peptide fragment from DNAJB11 containing K66 (Figure 6C). DNAJB11 is a member of the DNAJ (or HSP40) subclass of family of heat shock proteins which all share a J-domain. The role of this highly conserved domain is to stimulate the hydrolysis of ATP by chaperones in the HSP70 protein family whose main function is stabilize or

restore the native protein conformation of potentially misfolded client proteins under cellular stress[72]. Proteins in the DNAJ family have been implicated in tumor progression and metastasis[72]. DNAJB11 in particular has been overexpressed in pancreatic cancer, where exosomal DNAJB11 regulates expression of EGFR activates the MAPK pathway[73] and in liver cancer, by preventing alpha-1-antitrypsin degradation[74]. In contrast, low DNAJB11 messenger RNA (mRNA) levels appear to be correlated with worse outcomes in thyroid carcinoma[75]. In addition to its role in several cancers, research has shown that phosphorylation of T188 in DNAJB11 can reduce the aggregation of □-synuclein in Parkinson's disease[76]. As such, an association between K66 methylation and Parkinson's disease is possible, either directly via an unknown mechanism, or indirectly, through the modulation of T188 phosphorylation via cross-talk, for example. In all cases, since it is located within the J-domain, it is likely that the methylation of K66 is biologically significant, and this result provides a rationale for the characterization of K66.

An important limitation of our analysis is that the validation experiments were conducted using HeLa cells solely, *i.e.* immortalized cervical cancer cells. Given that MethylSight 2.0 is cell type-agnostic, it is probable that at least some of the 32 sites which could not be confirmed to be methylated via PRM-MS are in fact methylated in other cell types or under different conditions. Future experiments will need to be performed in order to determine how predictions made by MethylSight 2.0 generalize across cell types.

Cancer mutations associated with a predicted loss of methylation at a proximal site

MethylSight 2.0 was used to predict the impact of the 1,000 most frequent missense mutations in the COSMIC database[77] on the predicted methylation score of lysines within the mutant protein. The COSMIC database contains a curated list of somatic mutations and their impact in cancer.

We found that scores were relatively insensitive to these mutations except in select cases where the mutations were in close proximity with a lysine. Interestingly, we only observed predicted *loss* of methylation (using the conservative operating threshold). We detected 25 lysines that were no longer predicted to be methylated and associated with a decrease in methylation score ≥ 0.02 in presence of a mutation (Figure 7A).

The most striking loss of predicted methylation are associated with the $RhoA^{G17V}$ and $RhoA^{G17E}$ mutations. In these mutants, the neighbouring K18 is no longer predicted to be a methylation site. K18, like the mutated G17, is located amidst the GDP binding pocket (Figure 7B).

While methylation of RhoA-K18 has never been – to the best of our knowledge – confirmed experimentally, it is possible that methylated K18 could modulate RhoA activity. In fact, though it is believed that mutations in G17 impair GDP/GTP binding[78], it is not known exactly *how* this mutation impairs binding of GDP/GTP. Given the proximity of K18 to the GTP/GDP binding site, it is not implausible that alteration of the methylation status could alter RhoA's ability to bind and release GTP/GDP or coordinate Mg^{2+}.

Taken together, this provides an interesting avenue for further investigation so as to determine whether (1) RhoA-K18 is a true methylation site, (2) loss of methylation occurs in these mutants *in vitro*, and (3) this loss of methylation directly impacts GTP/GDP binding.

Performance of MethylSight 2.0 on non-human proteins

We deployed MethylSight 2.0 on the set of all known non-human lysine methylation sites catalogued in the PhosphoSitePlus database (360 sites). Interestingly, MethylSight 2.0 achieved a recall of 0.383 when applied to these sites (at $PCPr_{1:36}$ = 0.75). This is on the same order as its predicted recall (*i.e.* 0.299) on human lysines operating at the same decision threshold. This suggests that methylation sites in non-human organisms share homology with sites in human proteins.

The observation that MethylSight 2.0 achieved a better recall than predicted on this set of sites provides some evidence that the imbalance between methylated and non-methylated lysines could actually be lower than the one we estimated (*i.e.* 1:36) at the proteome scale, but further experiments would be required to confirm this.

The MethylSight 2.0 server

To make MethylSight 2.0 accessible to the broader community, we implemented a web server (Figure 8) which can be accessed at https://methylsight2.cu-bic.ca. The server is easy-to-use and allows users to select the operating threshold (precision and recall) that suits them best, depending on the application.

The web server processes *individual* protein sequences. Users interested in batch predictions may run MethylSight 2.0 as a standalone software on their own hardware. The Methylsight 2.0 source code, model weights, and instructions on how to use the software can be found on GitHub: https://github.com/ GreenCUBIC/MethylSight2.

Conclusion

Our work demonstrates that deep representations learned by pLMs trained on tens of millions of proteins are rich in information directly relevant for the task of lysine methylation site prediction and significantly improve the quality of the predictions. In fact, using these deep representations, we successfully trained a model that achieved more than double the AUPRC of previous models trained on human-engineered descriptors of protein sequences, such as those generated by ProtDCal[16] alongside the SVM-based MethylSight 1.0 predictor. We also showed that leveraging knowledge about other PTMs by means of a multitask learning strategy can further enhance the quality of the predictions. To the best of our knowledge, our model MethylSight 2.0 is the first lysine methylation prediction model to leverage pLM-generated representations and to employ a multitask learning strategy to extract knowledge from useful data that would otherwise be left unexploited. Similarly, the authors of PTM-Mamba trained a pLM capable of “zero-shot” PTM prediction. However, as it stands, PTM-Mamba[18] is unable to predict lysine methylation. We recommend that this work be extended to include lysine methylation in its language in the future.

In an effort to validate the predictions made by MethylSight 2.0 and show that it can successfully guide lysine methylation site discovery *in vivo*, we performed a validation PRM-MS experiment guided by MethylSight 2.0 predictions on a high-quality HeLa cell lysate. We uncovered 68 previously unidentified lysine methylation sites, some of which among proteins of high biological and/or therapeutic relevance, further showing the usefulness of our model as a drug target identification tool.

Applying MethylSight 2.0 to the human proteome provides insight into the extent of the lysine methylome. In fact, our analyses of the MethylSight 2.0-predicted lysine methylome suggests that the number of methyllysines in the human proteome may be even larger than previously believed (~50,000[15]), though this number remains difficult to estimate, given that our validation experiment was limited to 100 sites, and it seems unlikely that so few sites would be representative of the entire human proteome.

Additional validation experiments would be required to get a more accurate portrait of the human lysine methylome landscape.

Lysine methylation is a highly dynamic process which competes with several other PTMs, and a given lysine may be methylated to varying degrees under different conditions, *e.g.*, during development, in response to an environmental trigger, or in specific tissue types. Given that our model was trained in a tissue-blind fashion, *i.e.* positive sites in the training set were known to be methylated in *at least* one tissue type, we anticipate that validation experiments on methylation sites identified with MethylSight 2.0 may need to be performed in more than one cell type. There could be significant value in training a cell-specific lysine methylation predictor that could predict whether a lysine is methylated *in a given tissue type*, but training such a model would require tissue-specific datasets which are currently not publicly available.

Furthermore, in this study, we did not distinguish between the three possible methylation states. It is important to acknowledge that different methylation states can be associated with different – sometimes opposite – phenotypes. Several other groups have attempted to address this problem[13,14,79], but with limited success. Further research is needed to design an accurate predictor of mono-, di-, and trimethylated lysines.

Finally, MethylSight 2.0 does not attempt to associate a KMT with sites predicted to be methylated. This challenge is of prime importance, as therapeutic intervention would normally target the KMT or lysine demethylase (KDM) responsible for the modification. However, it is non-trivial given the scarcity of data for some KMTs, which in certain cases only have a few dozen known substrates or fewer. At the time of writing, we are aware of only one KMT-specific model for SET8 which could be used in conjunction with MethylSight 2.0.

Taken together, this work constitutes a significant contribution toward the elucidation of the human lysine methylome. In addition, MethylSight 2.0 can be deployed in a targeted fashion to determine whether lysines within proteins involved in a pathway of interest are probable methylation site. It therefore affords experimentalists with a tool which can help formulate rational hypotheses, guide experiments, and cut costs through prioritization of candidates for validation experiments.

Bibliography

1. Biggar, K. K. & Li, S. S.-C. Non-Histone Protein Methylation as a Regulator of Cellular Signalling and Function. *Nature Reviews Molecular Cell Biology* **16,** 5–17 (2015).
2. Carlson, S. M. & Gozani, O. Nonhistone Lysine Methylation in the Regulation of Cancer Pathways. *Cold Spring Harbor Perspectives in Medicine* **6,** a26435 (2016).
3. Han, D. *et al.* Lysine Methylation of Transcription Factors in Cancer. *Cell Death & Disease* **10,** 290 (2019).
4. Huang, M. *et al.* Methylation Modification of Non-Histone Proteins in Breast Cancer: An Emerging Targeted Therapeutic Strategy. *Pharmacological Research* **208,** 107354 (2024).
5. Straining, R. & Eighmy, W. Tazemetostat: EZH2 Inhibitor. *Journal of the Advanced Practitioner in Oncology* **13,** 158 (2022).
6. Feoli, A. *et al.* Lysine Methyltransferase Inhibitors: Where We Are Now. *RSC Chemical Biology* **3,** 359– 406 (2022).

7. Xu, K. *et al.* EZH2 Oncogenic Activity in Castration-Resistant Prostate Cancer Cells Is Polycombindependent. *Science* **338,** 1465–1469 (2012).
8. Kim, E. *et al.* Phosphorylation of EZH2 Activates STAT3 Signaling via STAT3 Methylation and Promotes Tumorigenicity of Glioblastoma Stem-like Cells. *Cancer Cell* **23,** 839–852 (2013).
9. Lanouette, S., Mongeon, V., Figeys, D. & Couture, J.-F. The Functional Diversity of Protein Lysine Methylation. *Molecular Systems Biology* **10,** 724 (2014).
10. Chen, H., Xue, Y., Huang, N., Yao, X. & Sun, Z. MeMo: A Web Tool for Prediction of Protein Methylation Modifications. *Nucleic Acids Research* **34,** W249–W253 (2006).
11. Qiu, W.-R., Xiao, X., Lin, W.-Z. & Chou, K.-C. iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach. *BioMed Research International* **2014,** 947416 (2014).
12. Kawashima, S. *et al.* AAindex: Amino Acid Index Database, Progress Report 2008. *Nucleic Acids Research* **36,** D202–205 (2008).
13. Deng, W. *et al.* Computational Prediction of Methylation Types of Covalently Modified Lysine and Arginine Residues in Proteins. *Briefings in Bioinformatics* **18,** 647–658 (2017).
14. Zheng, W., Wuyun, Q., Cheng, M., Hu, G. & Zhang, Y. Two-Level Protein Methylation Prediction Using Structure Model-Based Features. *Scientific Reports* **10,** 6008 (2020).
15. Biggar, K. K. *et al.* Proteome-Wide Prediction of Lysine Methylation Leads to Identification of H2BK43 Methylation and Outlines the Potential Methyllysine Proteome. *Cell Reports* **32,** 107896 (2020).
16. Ruiz-Blanco, Y. B., Paz, W., Green, J. & Marrero-Ponce, Y. ProtDCal: A Program to Compute GeneralPurpose-Numerical Descriptors for Sequences and 3D-structures of Proteins. *BMC Bioinformatics* **16,** 162 (2015).
17. Spadaro, A., Sharma, A. & Dehzangi, I. Predicting Lysine Methylation Sites Using a Convolutional Neural Network. *Methods (San Diego, Calif.)* **226,** 127–132 (2024).
18. Peng, F. Z. *et al.* PTM-Mamba: A PTM-aware Protein Language Model with Bidirectional Gated Mamba Blocks. *Nature Methods* **22,** 945–949 (2025).
19. Bepler, T. & Berger, B. Learning the Protein Language: Evolution, Structure, and Function. *Cell Systems* **12,** 654–669.e3 (2021).
20. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: Mutations, PTMs and Recalibrations. *Nucleic Acids Research* **43,** D512–520 (2015).
21. Li, A., Deng, Y., Tan, Y. & Chen, M. A Transfer Learning-Based Approach for Lysine Propionylation Prediction. *Frontiers in Physiology* **12,** (2021).
22. Lukinović, V., Casanova, A. G., Roth, G. S., Chuffart, F. & Reynoird, N. Lysine Methyltransferases Signaling: Histones Are Just the Tip of the Iceberg. *Current Protein and Peptide Science* **21,** 655–674 (2020).
23. Narita, T., Weinert, B. T. & Choudhary, C. Functions and Mechanisms of Non-Histone Protein Acetylation. *Nature Reviews Molecular Cell Biology* **20,** 156–174 (2019).

24. Geiss-Friedlander, R. & Melchior, F. Concepts in Sumoylation: A Decade On. *Nature Reviews Molecular Cell Biology* **8,** 947–956 (2007).

25. Damgaard, R. B. The Ubiquitin System: From Cell Signalling to Disease Biology and New Therapeutic Opportunities. *Cell Death & Differentiation* **28,** 423–426 (2021).

26. Shrestha, P., Kandel, J., Tayara, H. & Chong, K. T. DL-SPhos: Prediction of Serine Phosphorylation Sites Using Transformer Language Model. *Computers in Biology and Medicine* **169,** 107925 (2024).

27. Xue, Y. *et al.* GPS: A Comprehensive Www Server for Phosphorylation Sites Prediction. *Nucleic Acids Research* **33,** W184–W187 (2005).

28. Shi, S.-P. *et al.* PMeS: Prediction of Methylation Sites Based on Enhanced Feature Encoding Scheme. *PLOS One* **7,** e38772 (2012).

29. Shi, Y., Guo, Y., Hu, Y. & Li, M. Position-Specific Prediction of Methylation Sites from Sequence Conservation Based on Information Theory. *Scientific Reports* **5,** 12403 (2015).

30. Petersen, B., Petersen, T., Andersen, P., Nielsen, M. & Lundegaard, C. A Generic Method for Assignment of Reliability Scores Applied to Solvent Accessibility Predictions. *BMC Structural Biology* **9,** 51 (2009).

31. Høie, M. H. *et al.* NetSurfP-3.0: Accurate and Fast Prediction of Protein Structural Features by Protein Language Models and Deep Learning. *Nucleic Acids Research* **50,** W510–W515 (2022).

32. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics (Oxford, England)* **28,** 3150–3152 (2012).

33. Lin, Z. *et al.* Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **379,** 1123–1130 (2023).

34. Avraham, O., Tsaban, T., Ben-Aharon, Z., Tsaban, L. & Schueler-Furman, O. Protein Language Models Can Capture Protein Quaternary State. *BMC bioinformatics* **24,** 433 (2023).

35. Hao, X. & Fan, L. ProtT5 and Random Forests-Based Viscosity Prediction Method for Therapeutic mAbs. *European Journal of Pharmaceutical Sciences* **194,** 106705 (2024).

36. Schmirler, R., Heinzinger, M. & Rost, B. Fine-Tuning Protein Language Models Boosts Predictions across Diverse Tasks. *Nature Communications* **15,** 7407 (2024).

37. Elnaggar, A. *et al.* Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. (2023).

38. Brixi, G. *et al.* SaLT&PepPr Is an Interface-Predicting Language Model for Designing Peptide-Guided Protein Degraders. *Communications Biology* **6,** 1081 (2023).

39. Bhat, S. *et al.* De Novo Design of Peptide Binders to Conformationally Diverse Targets with Contrastive Language Modeling. *Science Advances* **11,** eadr8638 (2025).

40. Chen, L. T. *et al.* Target Sequence-Conditioned Design of Peptide Binders Using Masked Language Modeling. *Nature Biotechnology* 1–9 (2025) Preprint at https://www.nature.com/articles/s41587-02502761-2.

41. Elnaggar, A. *et al.* ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 7112–7127 (2021) Preprint at https://ieeexplore.ieee.org/document/9477085/.

42. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2623–2631 (Association for Computing Machinery, New York, NY, USA, 2019). Preprint at https://dl.acm.org/doi/10.1145/3292500.3330701.

43. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. Preprint at https://arxiv.org/abs/1912.01703 (2019).

44. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at http://arxiv.org/abs/ 1412.6980 (2017).

45. Vaswani, A. *et al.* Attention Is All You Need. (2017).

46. The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* **53,** D609–D617 (2025).

47. Berryhill, C. A. *et al.* Global Lysine Methylome Profiling Using Systematically Characterized Affinity Reagents. *Scientific Reports* **13,** 377 (2023).

48. Levitsky, L. I., Klein, J. A., Ivanov, M. V. & Gorshkov, M. V. Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. *Journal of Proteome Research* **18,** 709–714 (2018).

49. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).

50. Charih, F., Green, J. R. & Biggar, K. K. Using Machine Learning and Targeted Mass Spectrometry to Explore the Methyl-Lys Proteome. *STAR Protocols* **1**, 100135 (2020).

51. Fournier, Q. *et al.* Protein Language Models: Is Scaling Necessary? Preprint at http://biorxiv.org/ 10.1101/2024.09.23.614603v1 (2024).

52. Cheng, X. *et al.* Training Compute-Optimal Protein Language Models. Preprint at http://biorxiv.org/ lookup/doi/10.1101/2024.06.06.597716 (2024).

53. Leutert, M., Entwisle, S. W. & Villén, J. Decoding Post-Translational Modification Crosstalk With Proteomics. *Molecular & Cellular Proteomics : MCP* **20,** 100129 (2021).

54. Shukri, A. H., Lukinović, V., Charih, F. & Biggar, K. K. Unraveling the Battle for Lysine: A Review of the Competition among Post-Translational Modifications. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1866,** 194990 (2023).

55. Lee, J. M., Hammarén, H. M., Savitski, M. M. & Baek, S. H. Control of Protein Stability by Post-Translational Modifications. *Nature Communications* **14,** 201 (2023).

56. Hamey, J. J., Wienert, B., Quinlan, K. G. R. & Wilkins, M. R. METTL21B Is a Novel Human Lysine Methyltransferase of Translation Elongation Factor 1A: Discovery by CRISPR/Cas9 Knockout. *Molecular & Cellular Proteomics* **16,** 2229–2242 (2017).

57. Francis, J. W. *et al.* FAM86A Methylation of eEF2 Links mRNA Translation Elongation to Tumorigenesis. *Molecular Cell* **84,** 1753–1763.e7 (2024).

58. Michail, C., Rodrigues Lima, F., Viguier, M. & Deshayes, F. Structure and Function of the Lysine Methyltransferase SETD2 in Cancer: From Histones to Cytoskeleton. *Neoplasia (New York, N.Y.)* **59,** 101090 (2025).

59. Park, I. Y. *et al.* Dual Chromatin and Cytoskeletal Remodeling by SETD2. *Cell* **166,** 950–962 (2016).

60. Li, L. X. & Li, X. Epigenetically Mediated Ciliogenesis and Cell Cycle Regulation, and Their Translational Potential. *Cells* **10,** 1662 (2021).

61. Casanova, A. G. *et al.* Cytoskeleton Remodeling Induced by SMYD2 Methyltransferase Drives Breast Cancer Metastasis. *Cell Discovery* **10,** 1–22 (2024).

62. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Research* **53**, D672-D677 (2025).

63. Sasikumar, A. N., Perez, W. B. & Kinzy, T. G. The Many Roles of the Eukaryotic Elongation Factor 1 Complex. *WIREs RNA* **3,** 543–555 (2012).

64. Olarewaju, O., Ortiz, P. A., Chowdhury, W. Q., Chatterjee, I. & Kinzy, T. G. The Translation Elongation Factor eEF1B Plays a Role in the Oxidative Stress Response Pathway. *RNA biology* **1,** 89–94 (2004).

65. Negrutskii, B. S. *et al.* The eEF1 Family of Mammalian Translation Elongation Factors. *BBA Advances* **3,** 100067 (2023).

66. Vanwetswinkel, S. *et al.* Solution Structure of the 162 Residue C-terminal Domain of Human Elongation Factor 1Bγ. *Journal of Biological Chemistry* **278,** 43443–43451 (2003).

67. Achilonu, I. *et al.* An Update on the Biophysical Character of the Human Eukaryotic Elongation Factor 1 Beta: Perspectives from Interaction with Elongation Factor 1 Gamma. *Journal of Molecular Recognition* **31,** e2708 (2018).

68. Olatona, O. A., Choudhury, S. R., Kresman, R. & Heckman, C. A. Candidate Proteins Interacting with Cytoskeleton in Cells from the Basal Airway Epithelium in Vitro. *Frontiers in Molecular Biosciences* **11,** 1423503 (2024).

69. Mimori, K., Mori, M., Tanaka, S., Akiyoshi, T. & Sugimachi, K. The Overexpression of Elongation Factor 1 Gamma mRNA in Gastric Carcinoma. *Cancer* **75,** 1446–1449 (1995).

70. Chi, K., Jones, D. V. & Frazier, M. L. Expression of an Elongation Factor 1 Gamma-Related Sequence in Adenocarcinomas of the Colon. *Gastroenterology* **103,** 98–102 (1992).

71. Lew, Y. *et al.* Expression of Elongation Factor-1 Gamma-Related Sequence in Human Pancreatic Cancer. *Pancreas* **7,** 144–152 (1992).

72. Kim, H.-Y. & Hong, S. Multi-Faceted Roles of DNAJB Protein in Cancer Metastasis and Clinical Implications. *International Journal of Molecular Sciences* **23,** 14970 (2022).

73. Liu, P., Zu, F., Chen, H., Yin, X. & Tan, X. Exosomal DNAJB11 Promotes the Development of Pancreatic Cancer by Modulating the EGFR/MAPK Pathway. *Cellular & Molecular Biology Letters* **27,** 87 (2022).

74. Pan, J., Cao, D. & Gong, J. The Endoplasmic Reticulum Co-Chaperone ERdj3/DNAJB11 Promotes Hepatocellular Carcinoma Progression through Suppressing AATZ Degradation. *Future Oncology* **14,** 3001–3013 (2018).

75. Sun, R. *et al.* DNAJB11 Predicts a Poor Prognosis and Is Associated with Immune Infiltration in Thyroid Carcinoma: A Bioinformatics Analysis. *Journal of International Medical Research* **49,** 03000605211053722 (2021).

76. Chen, H.-Y. *et al.* ATM-mediated Co-Chaperone DNAJB11 Phosphorylation Facilitates α-Synuclein Folding upon DNA Double-Stranded Breaks. *NAR Molecular Medicine* **1,** ugae7 (2024).

77. Sondka, Z. *et al.* COSMIC: A Curated Database of Somatic Variants and Clinical Data for Cancer. *Nucleic Acids Research* **52,** D1210–D1217 (2024).

78. Sakata-Yanagimoto, M. *et al.* Somatic RHOA Mutation in Angioimmunoblastic T Cell Lymphoma. *Nature Genetics* **46,** 171–175 (2014).

79. Ju, Z., Cao, J.-Z. & Gu, H. iLM-2L: A Two-Level Predictor for Identifying Protein Lysine Methylation Sites and Their Methylation Degrees by Incorporating K-gap Amino Acid Pairs into Chou□s General PseAAC. *Journal of Theoretical Biology* **385,** 50–57 (2015).

Author contributions

François Charih: Conceptualization, Methodology, Software, Investigation, Formal analysis, Data Curation, Visualization, Writing - Original Draft, **Mullen Boulter:** Formal analysis, **Kyle K. Biggar:** Conceptualization, Resources, Formal analysis, Writing - Review & Editing, Funding acquisition, **James R. Green:** Conceptualization, Writing - Review & Editing, Funding acquisition

All authors approved of the manuscript.

Supplementary information

Supplementary Table S1: Isolation list of 100 putative methylation sites identified with MethylSight 2.0 and selected for *in vitro* validation
Supplementary Table S2: Results of mass spectrometry analysis

Data availability

The source code (models and model weights) required to run MethylSight 2.0 and the associated datasets are available on GitHub (https://github.com/GreenCUBIC/MethylSight2.git).

Funding

This research was funded by the National Science and Engineering Research Council (NSERC) Canada Discovery grant awarded to Kyle K. Biggar (RGPIN-2023-04651) and James R. Green (RGPIN-2021-04184).

Competing interests

The authors have no conflicts of interest to disclose.

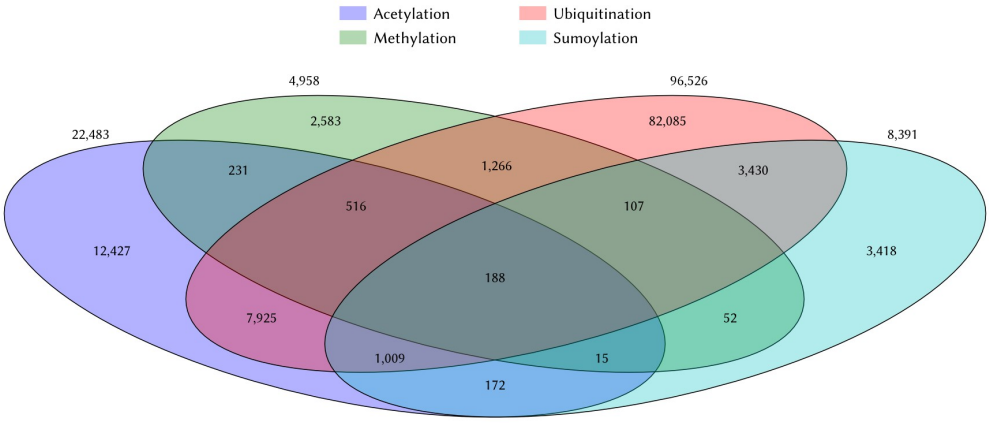


Figure 1. Co-occurrence of common post-translational modification of lysines in the PhosphoSitePlus database

The overlap of four major PTMs of lysines among human proteins are shown in this Venn diagram. These numbers were computed using the 10/17/24 update of the PhosphoSitePlus database[20]. Many yet-to-be discovered modifications remain to be deposited. Of the four major modifications of lysines, methylation is the one with the fewest annotations.

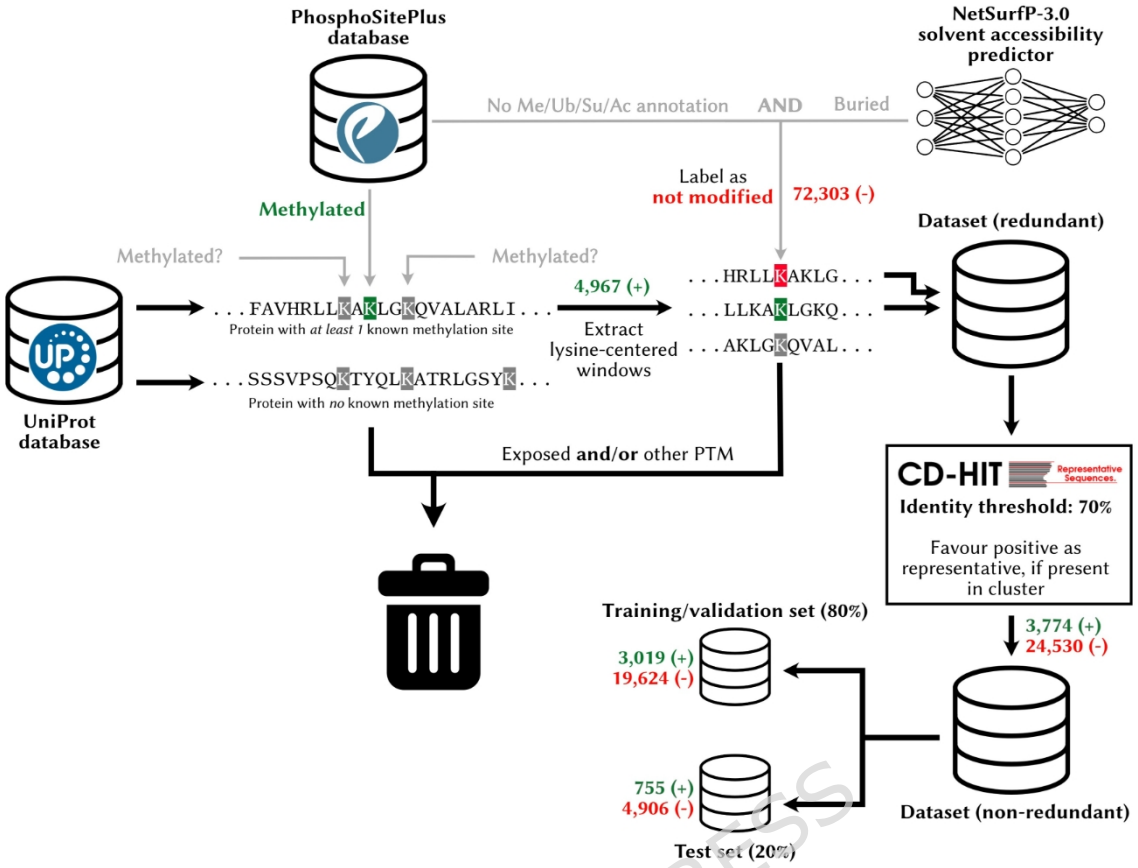


Figure 2. Preparation of a high-quality lysine modification dataset
To create the dataset used as part of this study, we sourced data from the PhosphoSitePlus database (for PTM annotations) and the UniProt database (for protein sequences). Only proteins with at least one methylated lysine residue were included in the dataset. Exposed residues of unknown status and/or having an annotation for another PTM were discarded, while the remaining lysines not known to be methylated were selected to make up the negative training data. The redundancy in the dataset was reduced with CD-HIT, using a window size of 31 for clustering and a 70% identity threshold. A blind test set was created by setting aside 20% of this data.

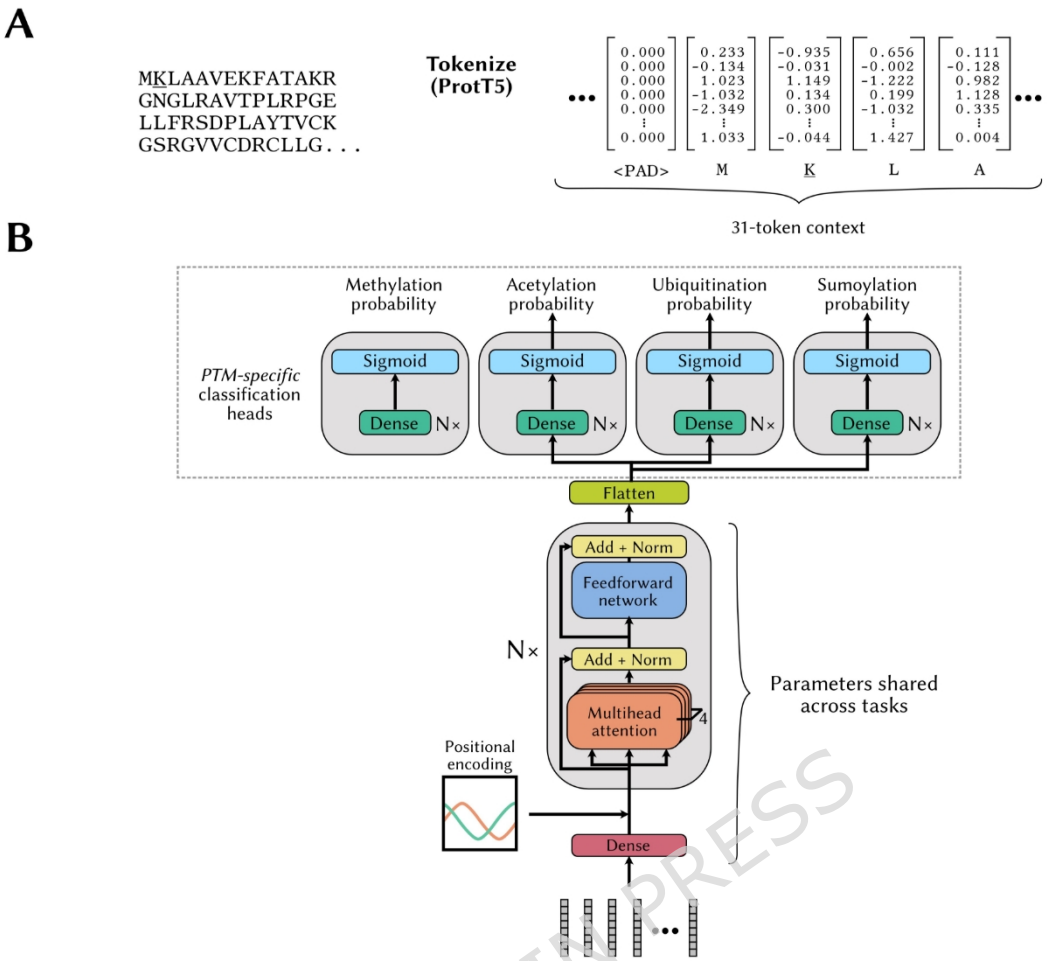


Figure 3. Transformer model architectures for methylation site prediction without and with multitask learning

(A) The inputs to our transformer-based models are the ProtT5 embeddings extracted from the full protein, with a context window of 31, centered around the lysine residue of interest. When a lysine residue is too close to an end of the protein sequence, null embeddings (<PAD>) are appended to complete the context window. **(B)** Modified transformer architecture designed to enable a multitask learning strategy. More specifically, after the flattening layer, instance representations are sent to one of four PTM-specific classification heads, depending on the task associated with the individual instances.

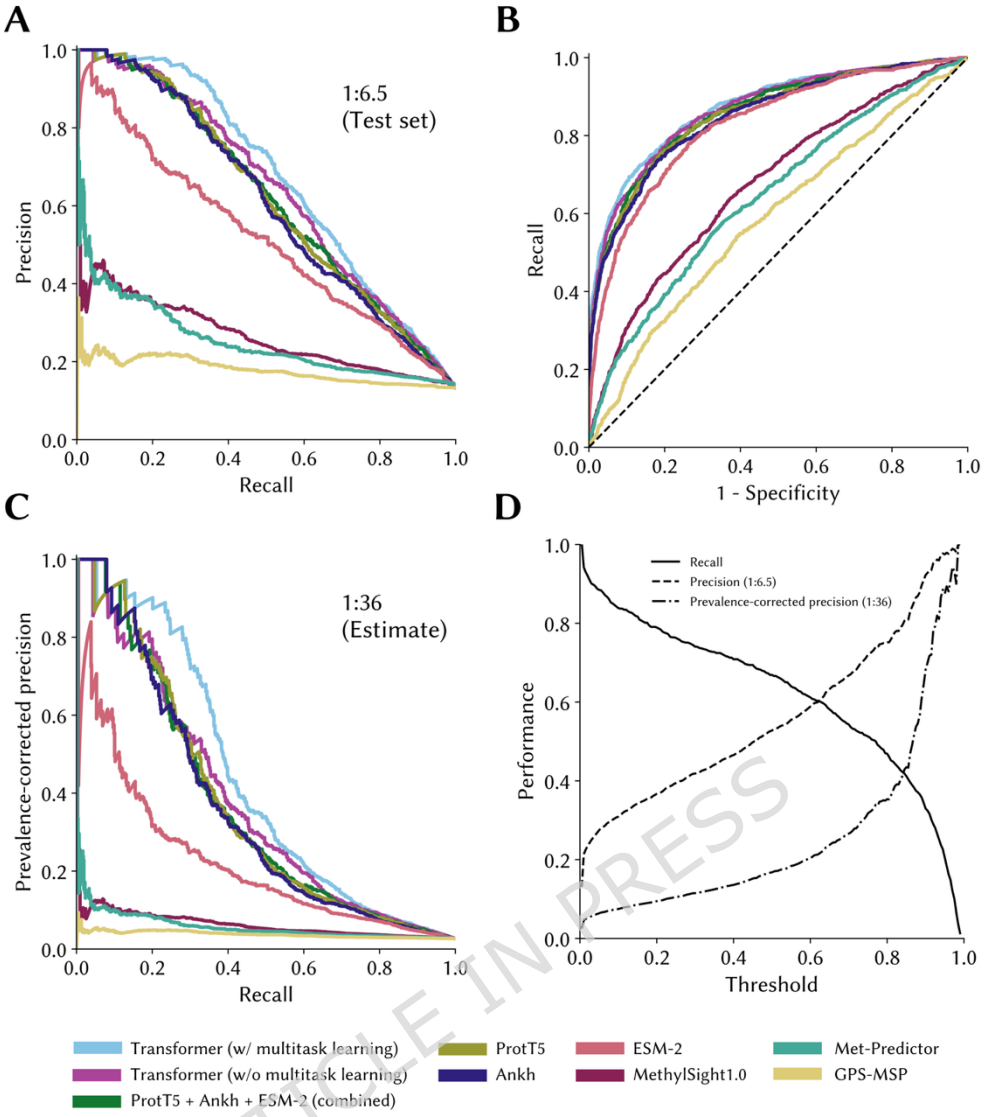


Figure 4. Precision-recall curves and receiver operating characteristic curves of the models on the independent test set

The precision-recall and receiver operating characteristic curves for the best performing models and/or training strategies were computed using a blind and independent test set of potential lysine methylation sites not seen during training. **(A)** The precision-recall curves are shown for the test set (1:6.5 imbalance). **(B)** The associated ROC curves are shown. **(C)** Prevalence-corrected precision assuming a true 1:36 imbalance between methylated and unmethylated sites provides more pessimistic estimate of performance. **(D)** The performance metrics are shown for two different imbalance ratios (1:6.5 and 1:36).

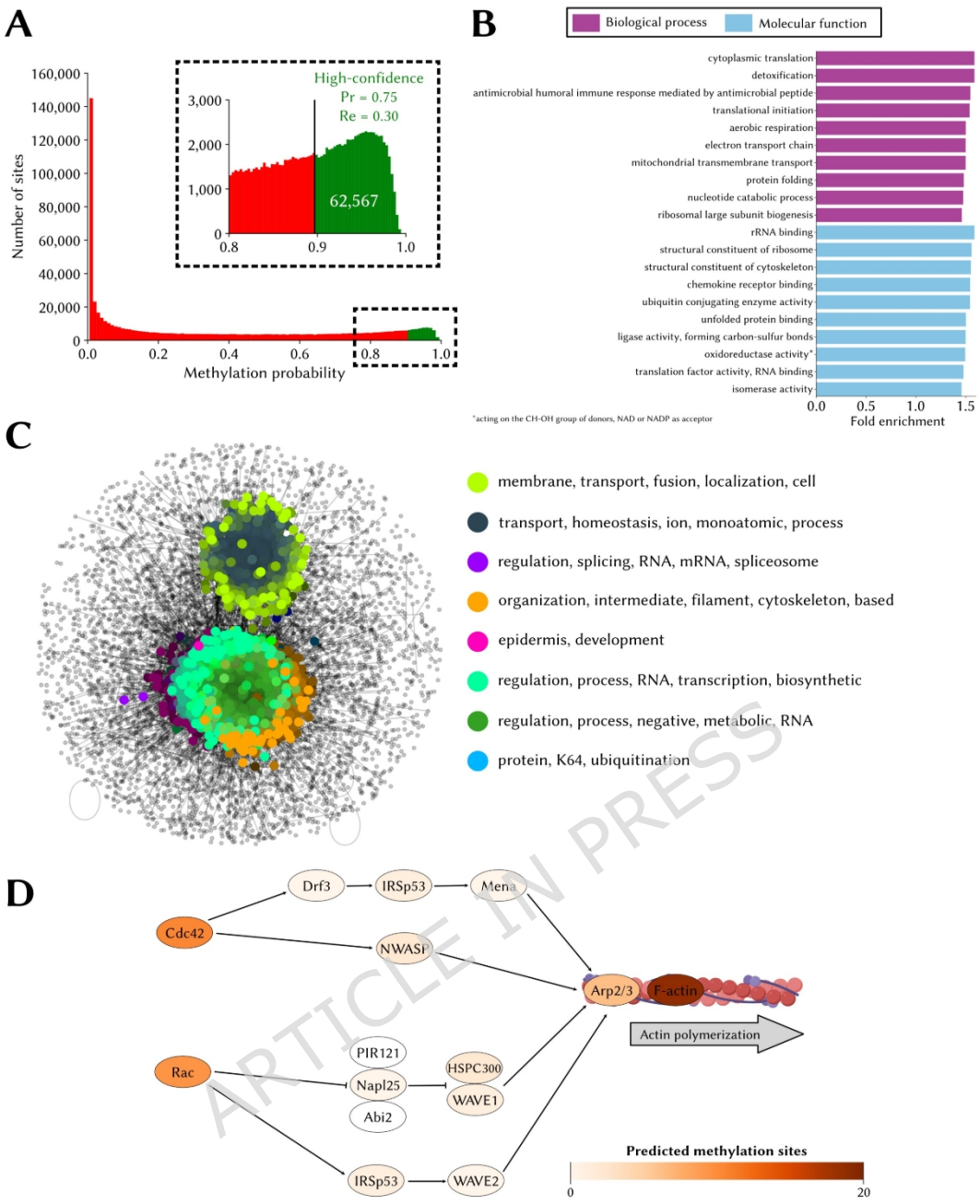


Figure 5. The lysine methylome as predicted by MethylSight 2.0
(A) Distribution of methylation probability for all 654,185 lysines in the human proteome. The sites predicted to be methylated *only* while operating under conservative settings (PCPr = 0.75) are shown in green. **(B)** Top 10 overrepresented gene ontology (GO) terms for the biological process (purple) and molecular function (blue) categories. Overrepresentation is statistically significant (α-value < 0.05; Fisher's exact test with Bonferroni correction for multiple testing). **(C)** Visual representation of the spatial analysis of functional enrichment (SAFE) analysis results; *i.e.* functional domains within the interaction network of methylated proteins and the most frequent words present in the GO terms associated with proteins in the domains. **(D)** Subset of the actin cytoskeleton regulation pathway (KEGG[62]: hsa04810). Proteins are colored on a white-to-red scale, with darker shades indicating a higher degree of methylation.

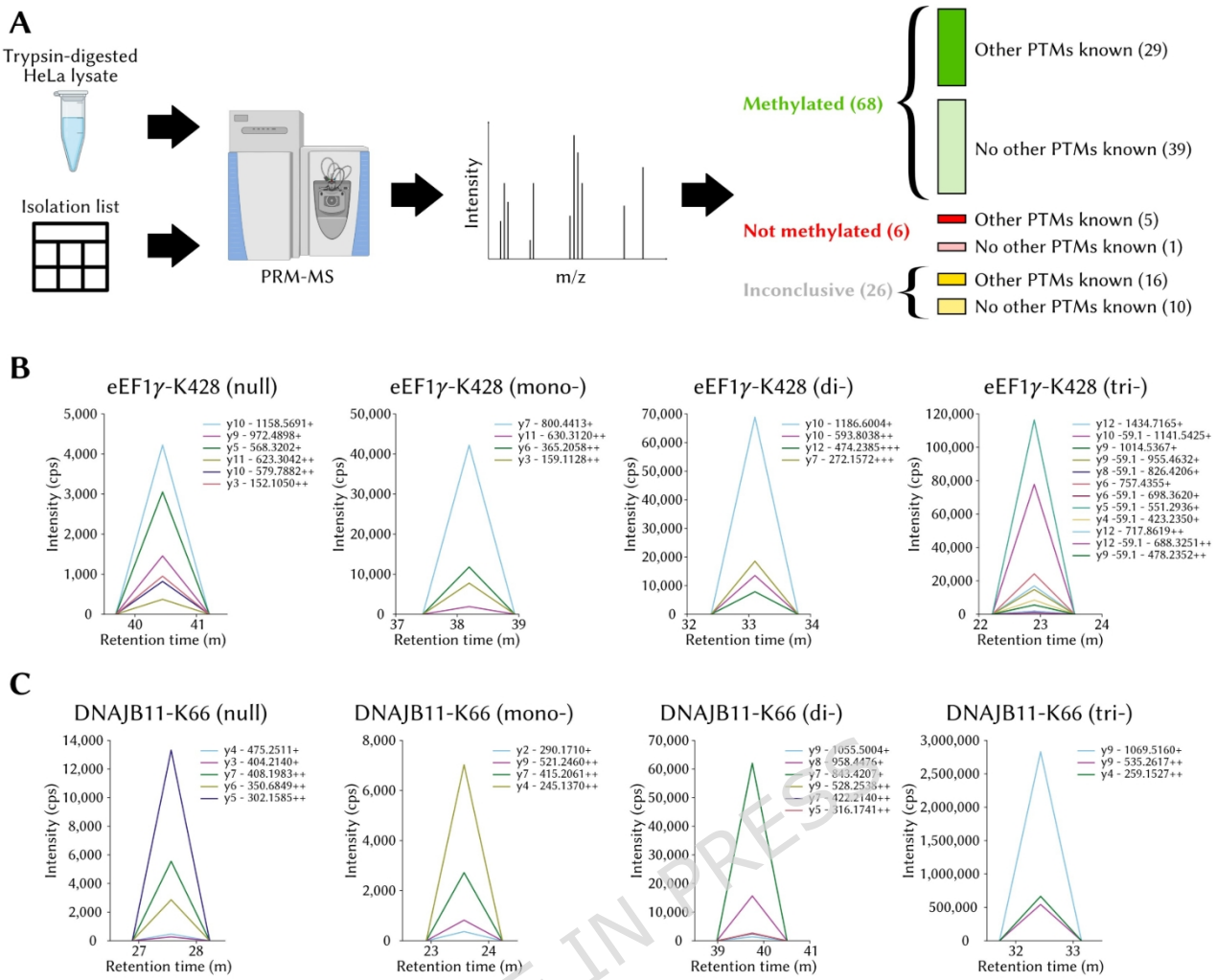


Figure 6. MethylSight 2.0-enabled discovery of novel lysine methylation sites with PRM-MS

(A) High-level overview of the methodology employed to validate 100 methylation sites identified with MethylSight 2.0 and distribution of the compiled results. **(B)** Transitions for the tryptic peptide containing eEF1□-K428 (EYFSWEGAFQHVGK). The mass-to-charge ratios of the y ions are shown. The transitions of the null, mono-, di-, and tri-methylation states are plotted separately for clarity, because the measured intensity varies in scale. **(C)** Transitions for the tryptic peptide containing DNAJB11-K66 (NPDDPQAQEK).

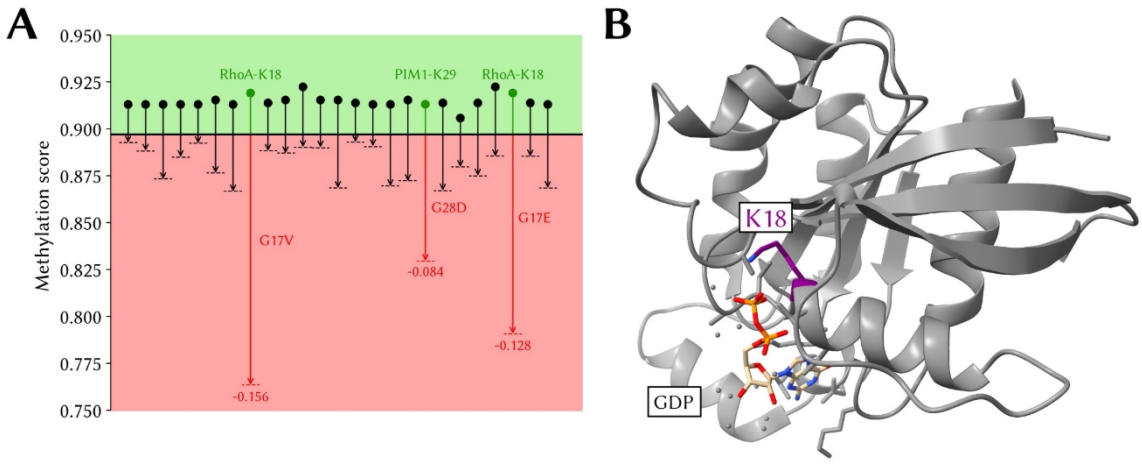


Figure 7. Predicted loss of methylation in oncogenic proteins

(A) Changes in predicted methylation score resulting in predicted loss of methylation associated with the

1,000 most commonly reported cancer-associated single amino acid mutations in the Catalog of Somatic Mutations reported in the Cancer (COSMIC) database[77]. Shown are score changes of at least 2%. **(B)** Position of K18 (in purple) in the X-ray structure of RhoA (PDB: 1DPF). The GDP co-factor is in beige with the two phosphate groups in orange.

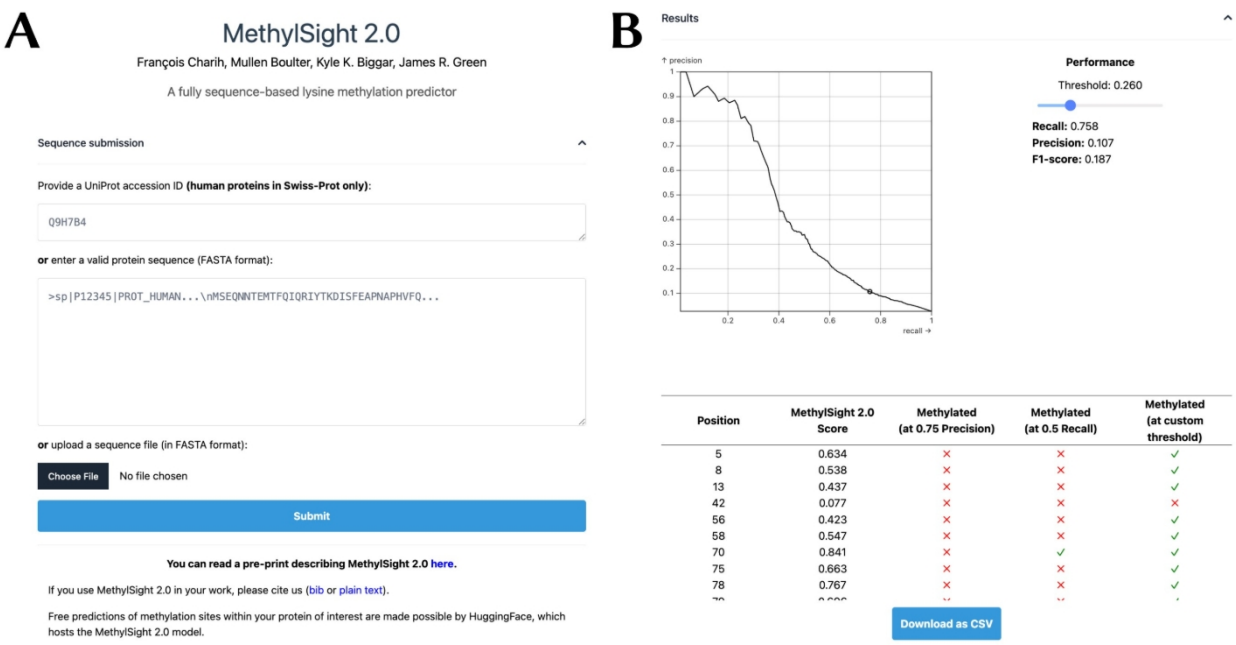


Figure 8. MethylSight 2.0 web server

(A) The user can either provide the UniProt accession ID or the sequence of the human protein of interest. If the protein is from another organism or is a non-canonical human protein (*e.g.*, isoform or mutant), the user must provide a FASTA-formatted sequence through a text box or upload the equivalent FASTA file. **(B)** Once the results are available, a precision-recall curve computed for MethylSight 2.0 on our test set with an assumed real imbalance ratio of 1:36 is presented. This provides the user with a visual interpretation of the operating threshold and the optimal threshold which can be tuned with a slider. The results are presented as a table which can be downloaded in CSV format.

Table 1. Composition of the PhosphoSitePlus dataset (human lysines)

Modification	Important functions	Unique proteins	Positive sites	Negative sites (low confidence)
Methylation	Chromatin and gene expression regulation (histones), signaling, enzyme (in)activation[22]	2,751	4,966	157,385
Acetylation	Protein stability, regulation of PPIs and protein-DNA interactions (histones)[23]	7,047	22,547	333,013
Sumoylation	Alteration of molecular interactions of substrate through addition/hiding interaction surfaces[24]	2,646	8,391	124,274
Ubiquitination	Regulation of protein degradation, autophagy, protein trafficking[25]	11,712	96,545	377,547

Table 2. Composition of the high-confidence dataset used to train and test the models

	Methylation	Ubiquitination	Acetylation	Sumoylation
Training set (pos/neg)	2,415/15,699	68,539/58,314	15,791/51,498	5,737/15,862
Validation set (pos/neg)	604/3,925			
Test set (pos/neg)	755/4,906			

Table 3. Foundational protein language models used to embed potential lysine methylation sites

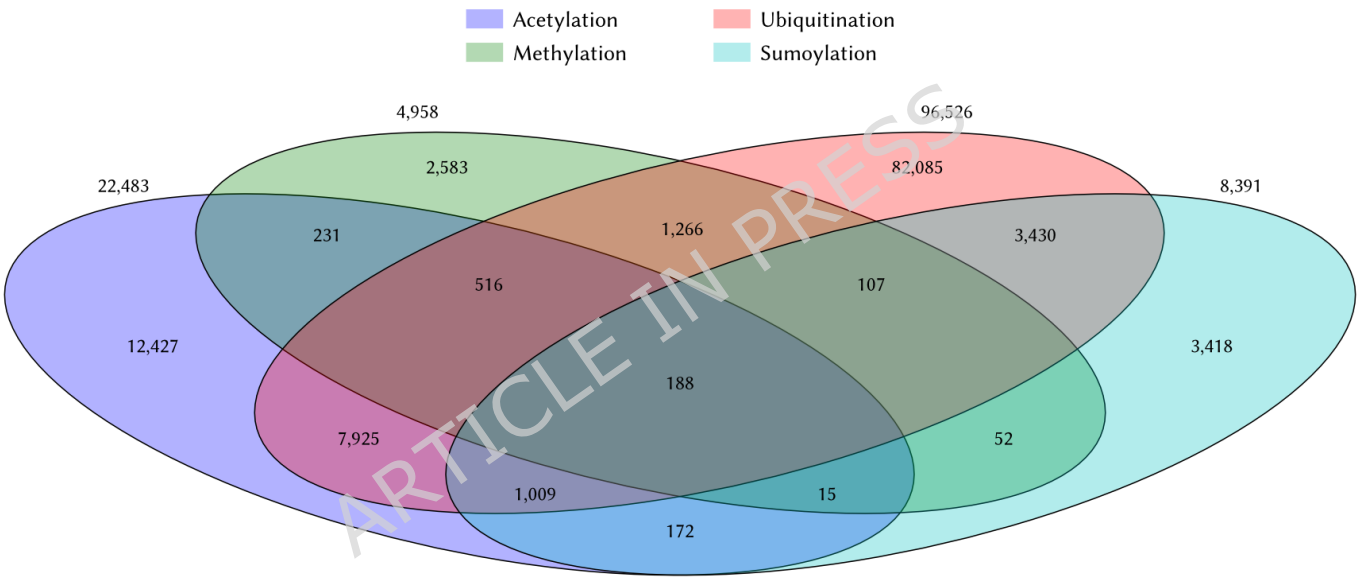
Model	Architecture	Version	Embedding dimension	Parameters (approx.)	Training strategy	Training data
ProtT5[41]	Encoder-decoder	ProtT5-XL-BFD	1,024	3B	1-gram random masking with demasking	BFD (pre-training; ~2.1B sequences) and UniRef50 (fine-tuning; ~45M sequences)
ESM-2[33]	Encoder-only	ESM-2-T33-650M-UR50D	1,280	650M	1-gram random masking with demasking	UniRef50+90 (~65M sequences)
Ankh[37]	Encoder-decoder	Ankh Large	1,536	1B	1-gram random masking with full sequence reconstruction	UniRef50 (~45M sequences)

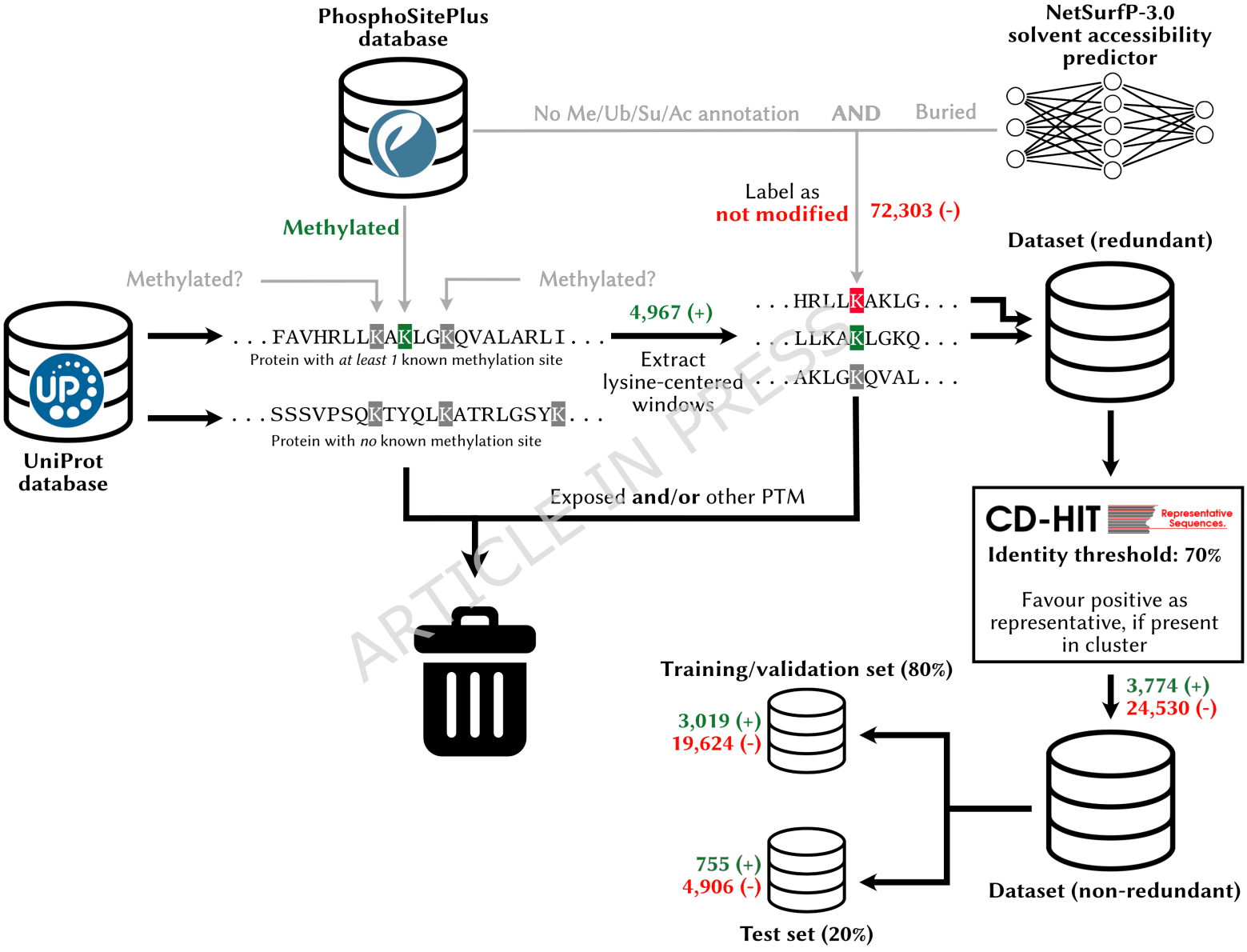
Table 4. Prediction accuracy of our models and publicly available methods on the blind test set (1:6.5 imbalance)

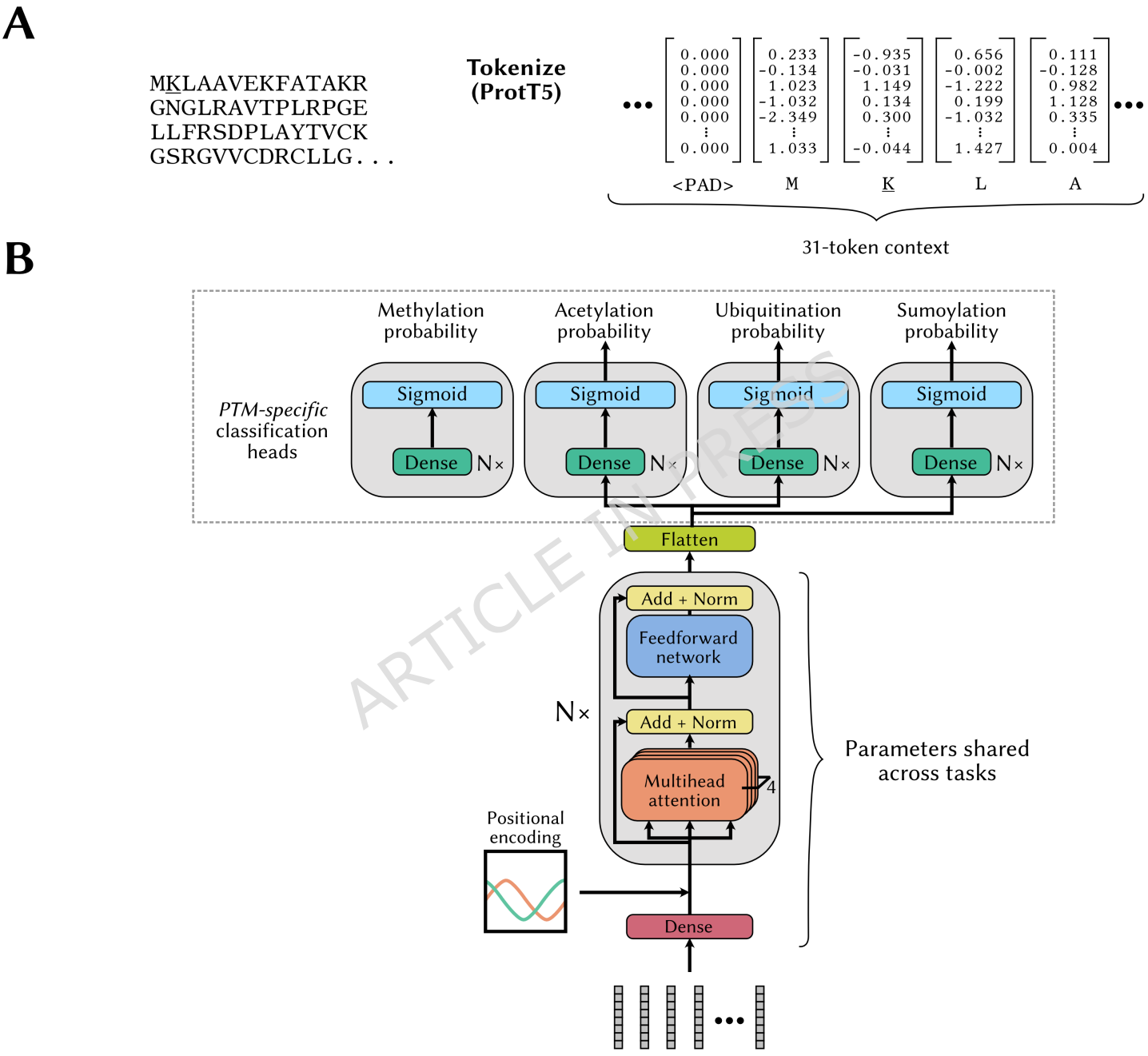
	Method	AUPRC	AUROC	Pr@0.5Re
This work	**Transformer (w/ multitask learning)**	**0.672**	**0.874**	**0.769**
	Transformer (w/o multitask learning)	0.642	0.867	0.734
	MLP (Combined embeddings)	0.624	0.859	0.637
	MLP (ProtT5 embeddings)	0.620	0.854	0.620
	MLP (Ankh embeddings)	0.611	0.848	0.609
	MLP (ESM-2 embeddings)	0.517	0.829	0.520
Previous SOTA	MethylSight 1.0[15]	0.267	0.677	0.242
	Met-Predictor[14]	0.254	0.646	0.221
	GPS-MSP[13]	0.179	0.588	0.175

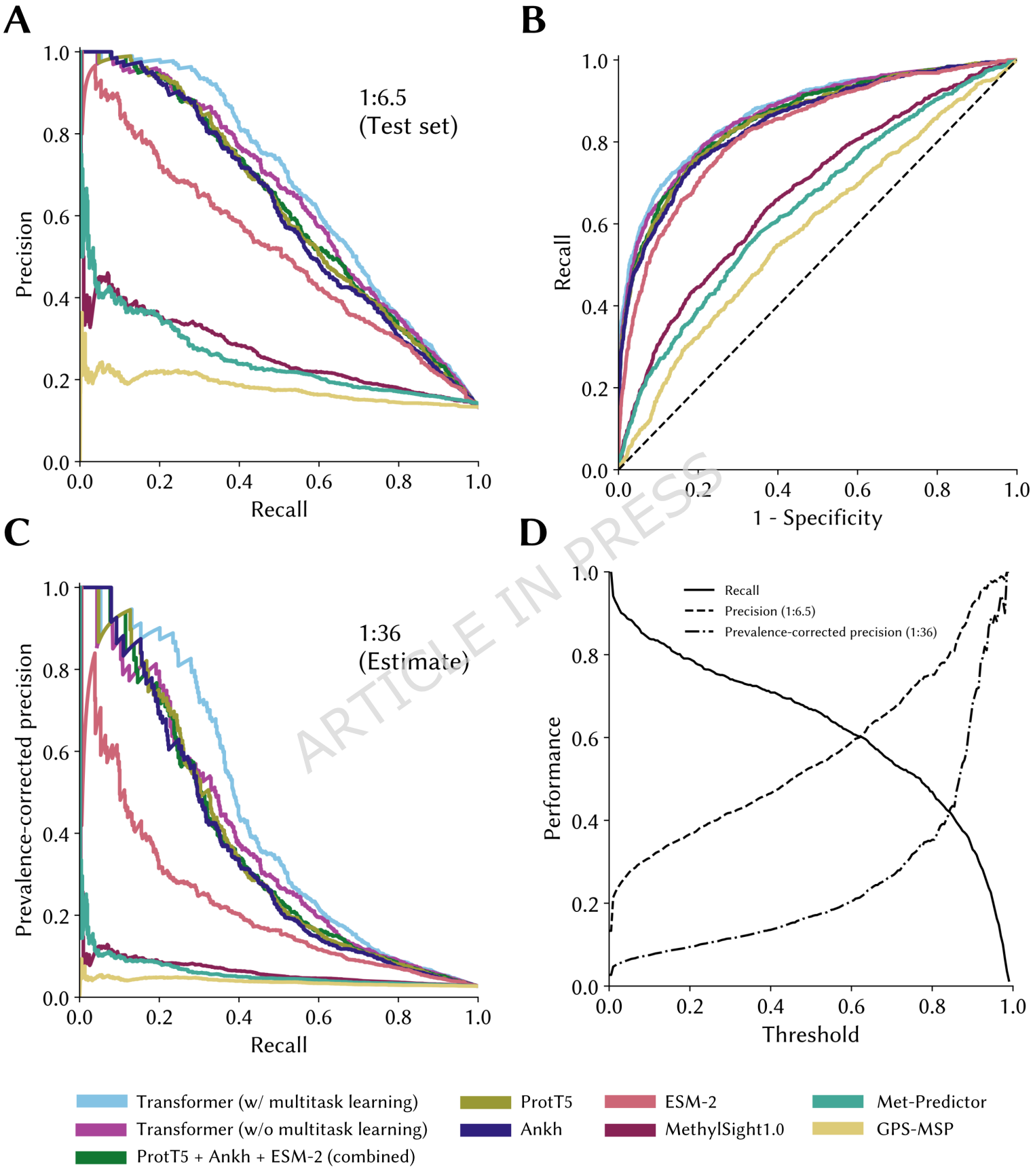
Table 5. Hyperparameters used to train the most accurate model (MethylSight 2.0)

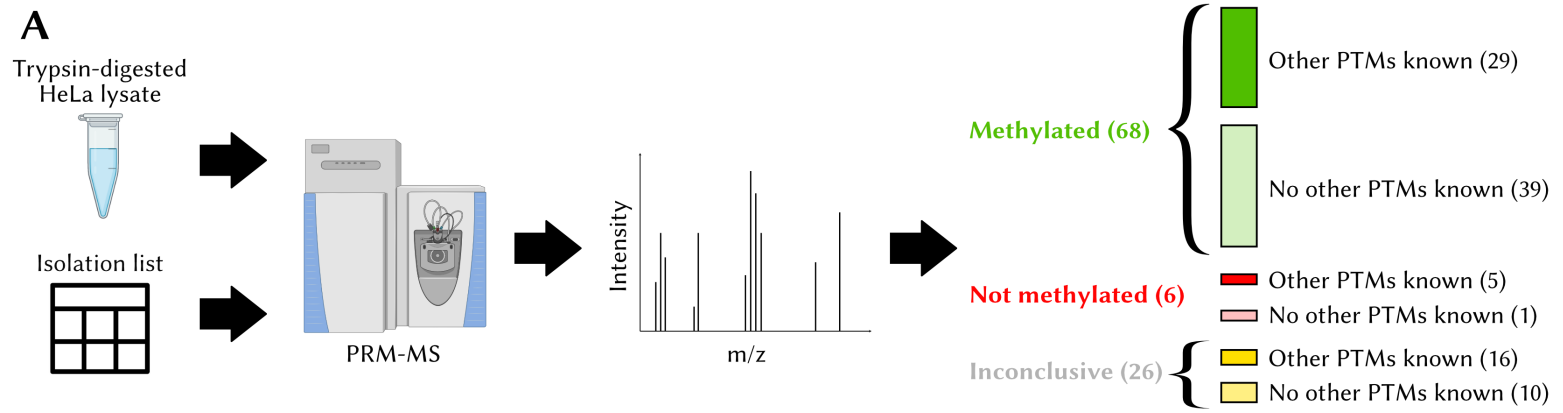
Parameter	Value
Learning rate	8×10^{-7}
Number of epochs	100
Weight factor of methylation loss (□)	20
Batch size	128
Number of transformer blocks	2
Heads per self-attention layer	4
Width of the first (pre-attention) dense layer	1,600
Widths of the dense layers (classification heads)	1,797; 1,803; 338; 493
Embedding dimension	1,024
Dropout rate	0.15

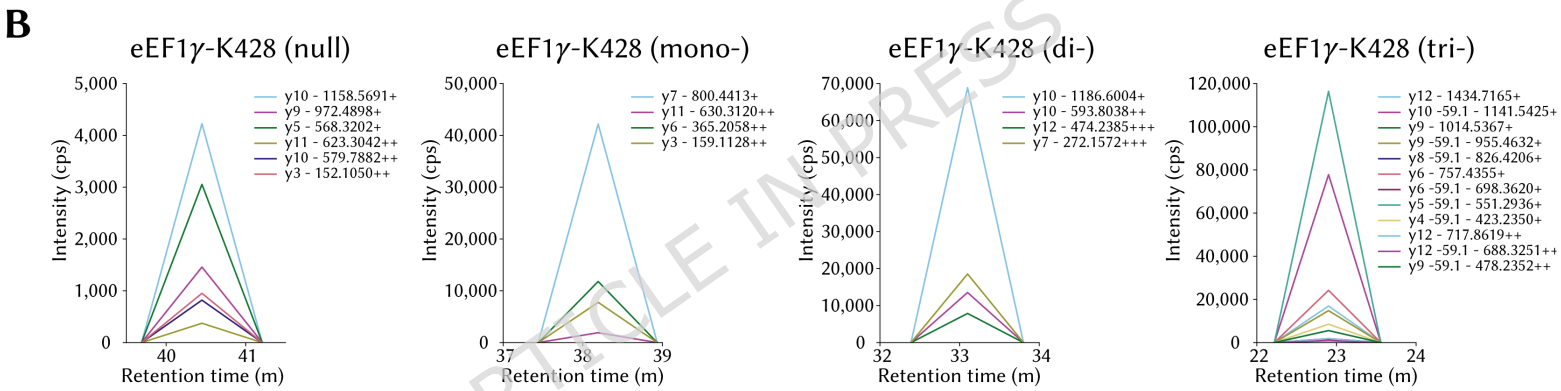


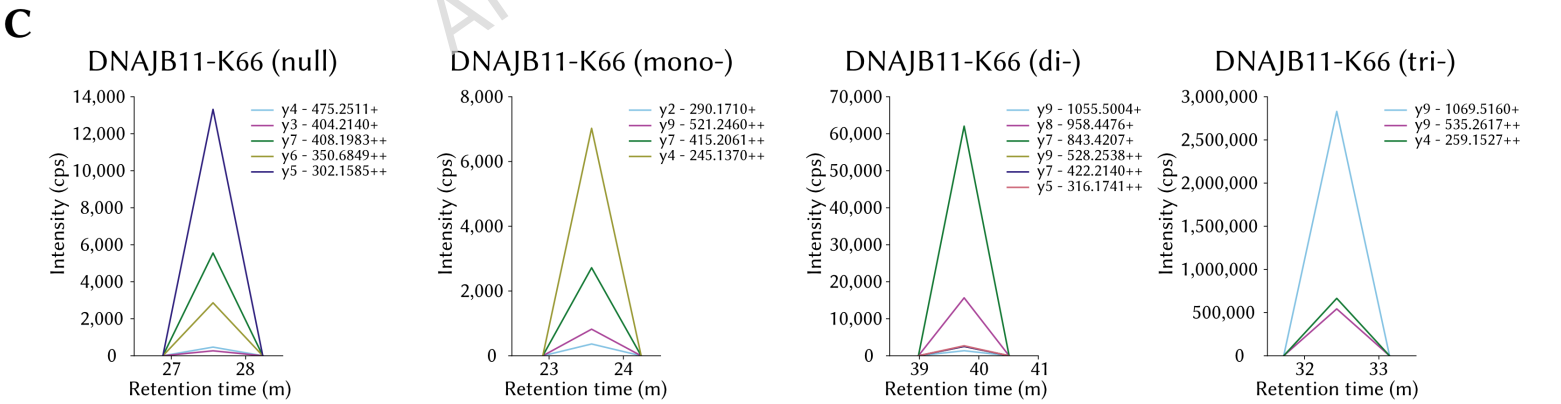


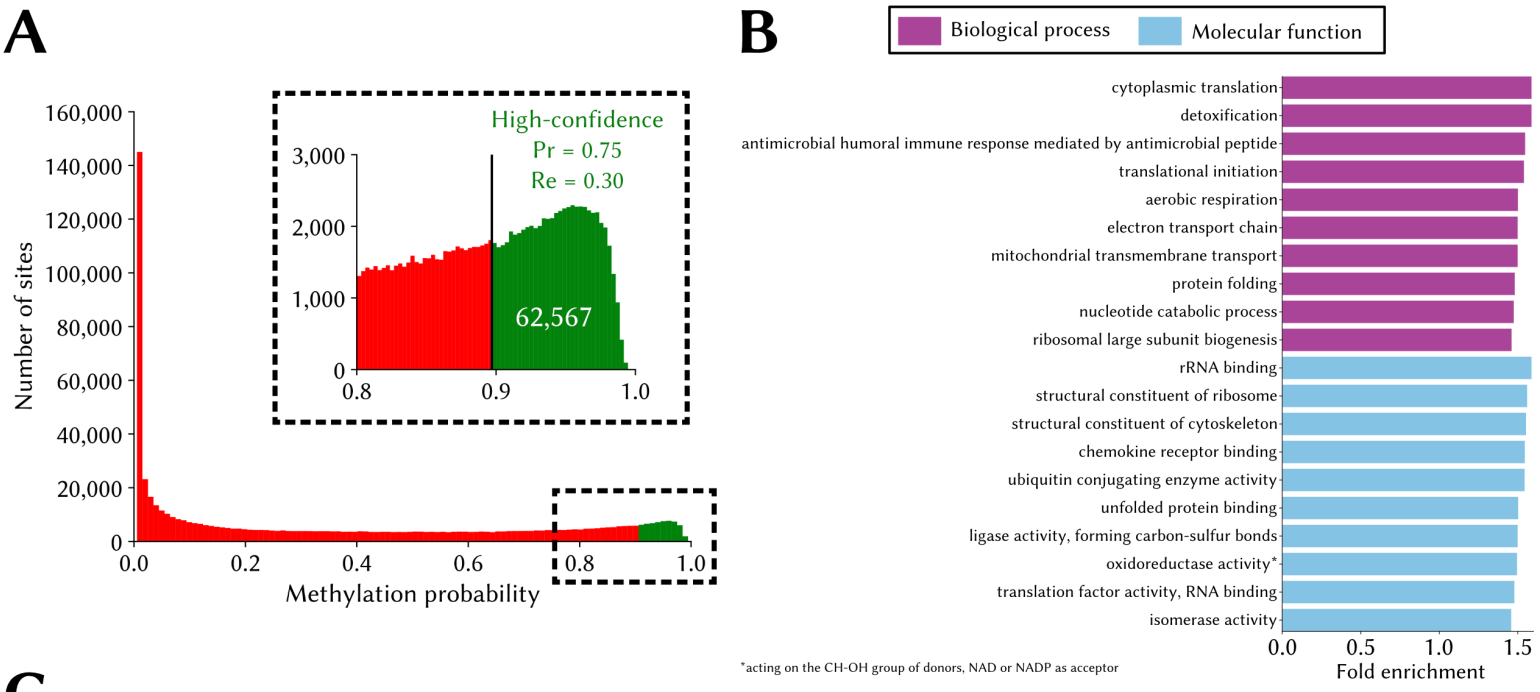


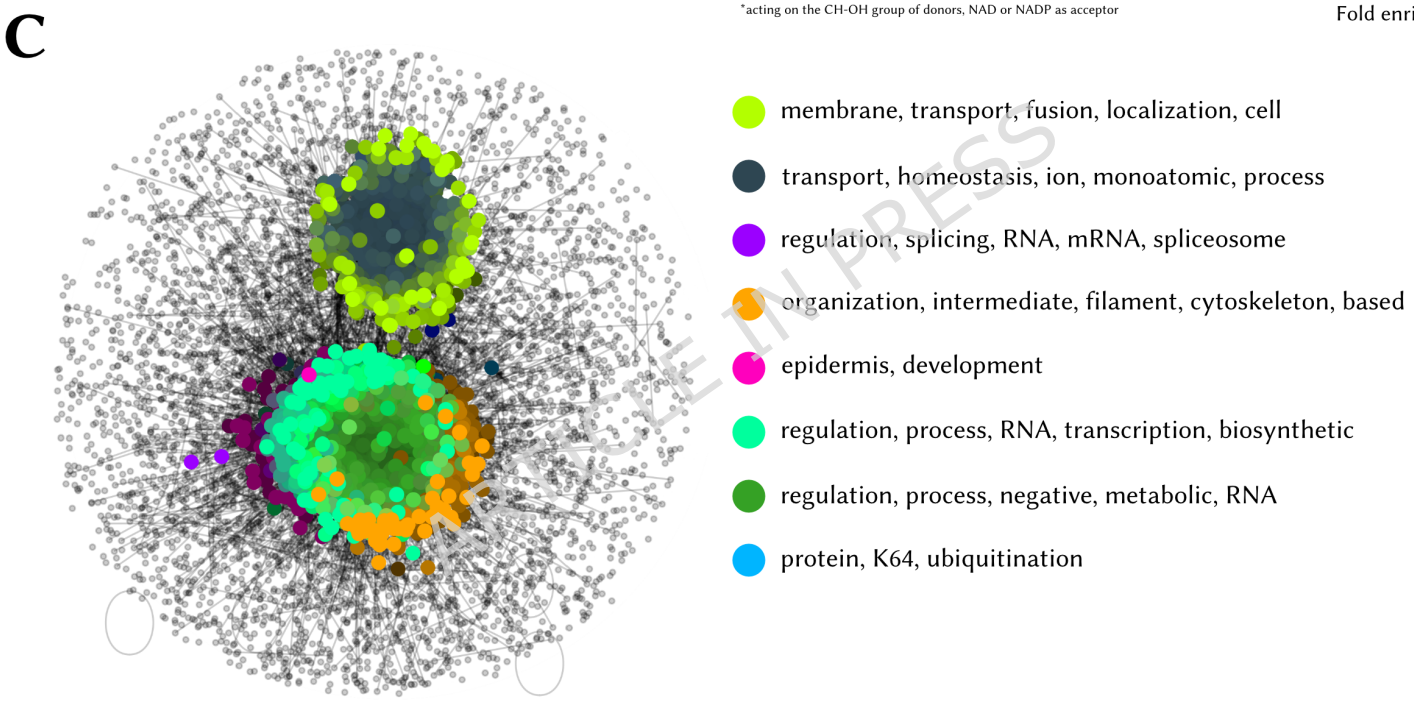


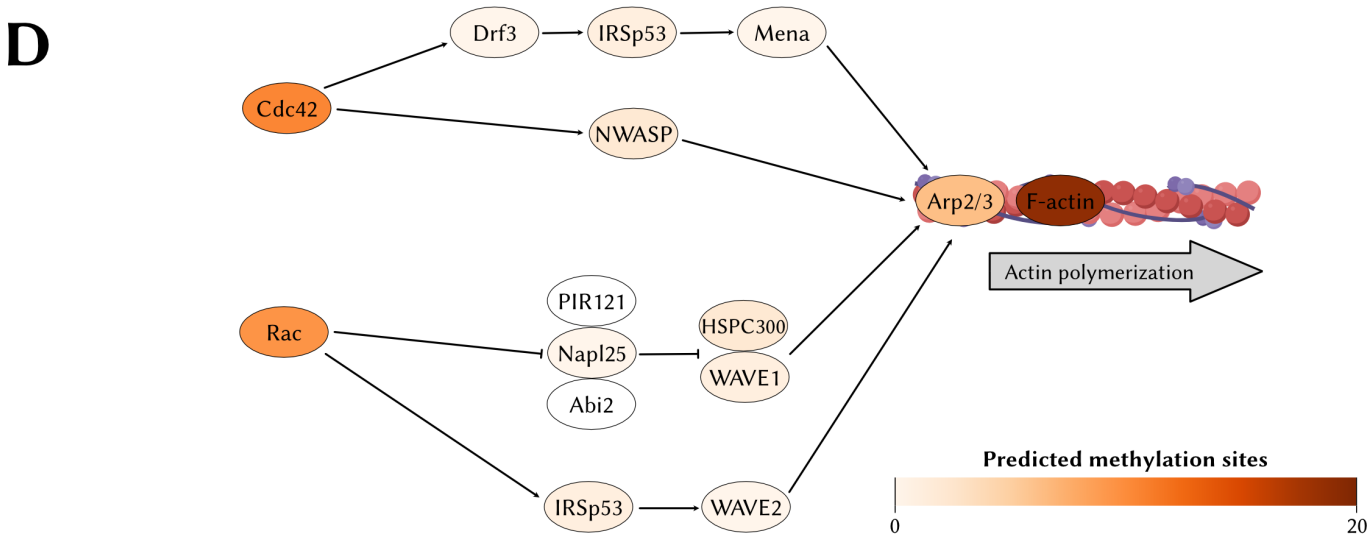


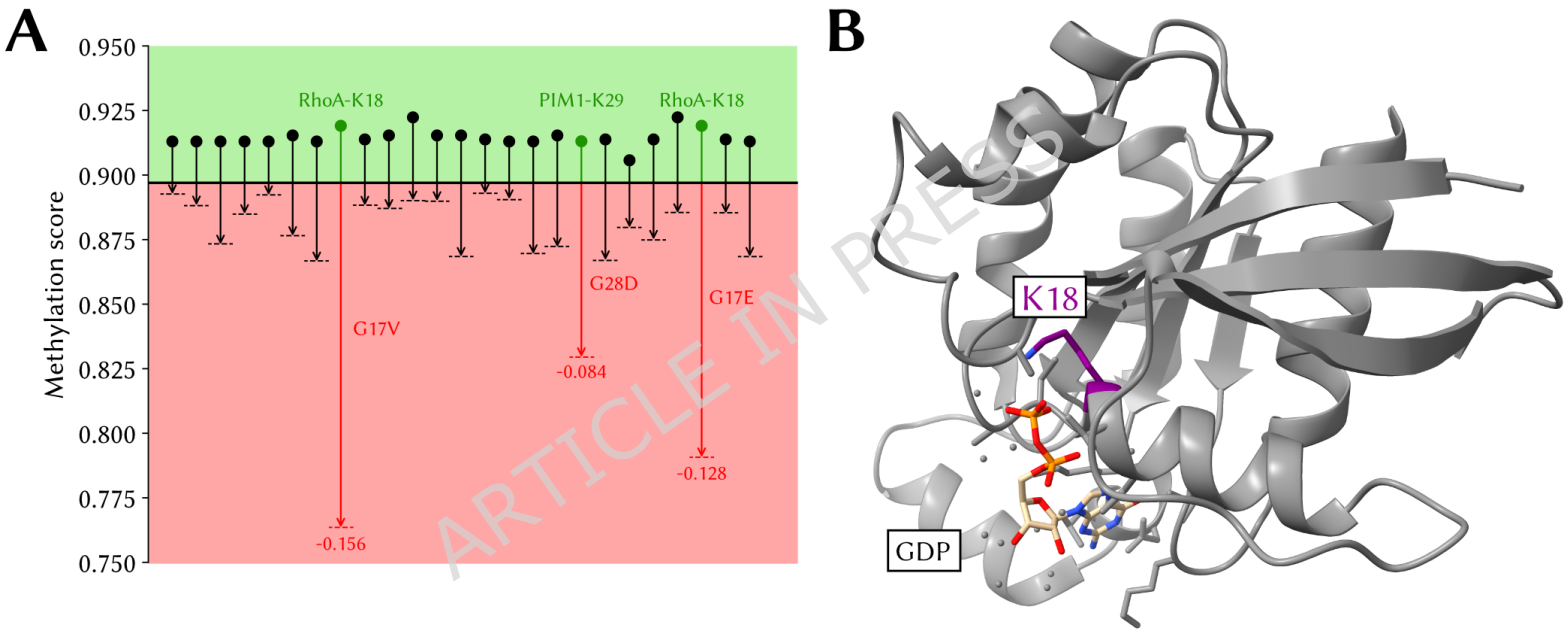












A

MethylSight 2.0

François Charih, Mullen Boulter, Kyle K. Biggar, James R. Green

A fully sequence-based lysine methylation predictor

Sequence submission

Provide a UniProt accession ID **(human proteins in Swiss-Prot only)**:

```
Q9H7B4
```

or enter a valid protein sequence (FASTA format):

```
>sp|P12345|PROT_HUMAN...\nMSEQNNTEMTFQIQRIYTKDISFEAPNAPHVFQ...
```

or upload a sequence file (in FASTA format):

Choose File No file chosen

Submit

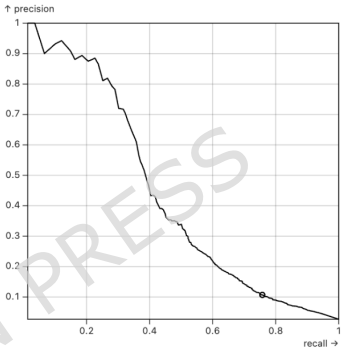
You can read a pre-print describing MethylSight 2.0 here.

If you use MethylSight 2.0 in your work, please cite us (bib or plain text).

Free predictions of methylation sites within your protein of interest are made possible by HuggingFace, which hosts the MethylSight 2.0 model.

B

Results



Performance

Threshold: 0.260

Recall: 0.758
Precision: 0.107
F1-score: 0.187

Position	MethylSight 2.0 Score	Methylated (at 0.75 Precision)	Methylated (at 0.5 Recall)	Methylated (at custom threshold)
5	0.634	✗	✗	✓
8	0.538	✗	✗	✓
13	0.437	✗	✗	✓
42	0.077	✗	✗	✗
56	0.423	✗	✗	✓
58	0.547	✗	✗	✓
70	0.841	✗	✓	✓
75	0.663	✗	✗	✓
78	0.767	✗	✗	✓

Download as CSV

Table 1. Composition of the PhosphoSitePlus dataset (human lysines)

Modification	Important functions	Unique proteins	Positive sites	Negative sites (low confidence)
Methylation	Chromatin and gene expression regulation (histones), signaling, enzyme (in)activation[22]	2,751	4,966	157,385
Acetylation	Protein stability, regulation of PPIs and protein-DNA interactions (histones)[23]	7,047	22,547	333,013
Sumoylation	Alteration of molecular interactions of substrate through addition/hiding interaction surfaces[24]	2,646	8,391	124,274
Ubiquitination	Regulation of protein degradation, autophagy, protein trafficking[25]	11,712	96,545	377,547

Table 2. Composition of the high-confidence dataset used to train and test the models

	Methylation	Ubiquitination	Acetylation	Sumoylation
Training set (pos/neg)	2,415/15,699	68,539/58,314	15,791/51,498	5,737/15,862
Validation set (pos/neg)	604/3,925			
Test set (pos/neg)	755/4,906			

Table 3. Foundational protein language models used to embed potential lysine methylation sites

Model	Architecture	Version	Embedding dimension	Parameters (approx.)	Training strategy	Training data
ProtT5[41]	Encoder-decoder	ProtT5-XL-BFD	1,024	3B	1-gram random masking with demasking	BFD (pre-training; ~2.1B sequences) and UniRef50 (fine-tuning; ~45M sequences)
ESM-2[33]	Encoder-only	ESM-2-T33-650M-UR50D	1,280	650M	1-gram random masking with demasking	UniRef50+90 (~65M sequences)
Ankh[37]	Encoder-decoder	Ankh Large	1,536	1B	1-gram random masking with full sequence reconstruction	UniRef50 (~45M sequences)

Table 4. Prediction accuracy of our models and publicly available methods on the blind test set (1:6.5 imbalance)

	Method	AUPRC	AUROC	Pr@0.5Re
This work	**Transformer (w/ multitask learning)**	**0.672**	**0.874**	**0.769**
	Transformer (w/o multitask learning)	0.642	0.867	0.734
	MLP (Combined embeddings)	0.624	0.859	0.637
	MLP (ProtT5 embeddings)	0.620	0.854	0.620
	MLP (Ankh embeddings)	0.611	0.848	0.609
	MLP (ESM-2 embeddings)	0.517	0.829	0.520
Previous SOTA	MethylSight 1.0[15]	0.267	0.677	0.242
	Met-Predictor[14]	0.254	0.646	0.221
	GPS-MSP[13]	0.179	0.588	0.175

Table 5. Hyperparameters used to train the most accurate model (MethylSight 2.0)

Parameter	Value
Learning rate	8×10^{-7}
Number of epochs	100
Weight factor of methylation loss (□)	20
Batch size	128
Number of transformer blocks	2
Heads per self-attention layer	4
Width of the first (pre-attention) dense layer	1,600
Widths of the dense layers (classification heads)	1,797; 1,803; 338; 493
Embedding dimension	1,024
Dropout rate	0.15