



# Histone-Net: a multi-paradigm computational framework for histone occupancy and modification prediction

Muhammad Nabeel Asim<sup>1,2</sup> · Muhammad Ali Ibrahim<sup>1,2</sup> · Muhammad Imran Malik<sup>3</sup> · Imran Razzak<sup>4</sup> · Andreas Dengel<sup>1,2</sup> · Sheraz Ahmed<sup>1</sup>

Received: 6 November 2021 / Accepted: 27 May 2022  
© The Author(s) 2022

## Abstract

Deep exploration of histone occupancy and covalent post-translational modifications (e.g., acetylation, methylation) is essential to decode gene expression regulation, chromosome packaging, DNA damage, and transcriptional activation. Existing computational approaches are unable to precisely predict histone occupancy and modifications mainly due to the use of sub-optimal statistical representation of histone sequences. For the establishment of an improved histone occupancy and modification landscape for multiple histone markers, the paper in hand presents an end-to-end computational multi-paradigm framework “Histone-Net”. To learn local and global residue context aware sequence representation, Histone-Net generates unsupervised higher order residue embeddings (DNA2Vec) and presents a different application of language modelling, where it encapsulates histone occupancy and modification information while generating higher order residue embeddings (SuperDNA2Vec) in a supervised manner. We perform an intrinsic and extrinsic evaluation of both presented distributed representation learning schemes. A comprehensive empirical evaluation of Histone-Net over ten benchmark histone markers data sets for three different histone sequence analysis tasks indicates that SuperDNA2Vec sequence representation and softmax classifier-based approach outperforms state-of-the-art approach by an average accuracy of 7%. To eliminate the overhead of training separate binary classifiers for all ten histone markers, Histone-Net is evaluated in multi-label classification paradigm, where it produces decent performance for simultaneous prediction of histone occupancy, acetylation, and methylation.

**Keywords** Machine learning and deep learning · Supervised kmer embeddings · Multi-label and binary classification · Histone markers · Acetylation, methylation and histone occupancy

## Introduction

Cells are fundamental building blocks of living organisms. Cells constitute tissues, tissues form organs and combination of organs give birth to organ systems [1,2]. The way different living organisms grow, survive, develop and reproduce is regulated by an instruction manual called Deoxyribonucleic Acid (DNA) or Genetic code [1,2]. The genetic code is

organised into chromatin in a series of nucleosomes, where in each nucleosome, DNA is wrapped around histone octamers which are made up of four pairs of histone proteins (H2A, H2B, H3 and H4). A graphical representation of nucleosome construction with Histone Octamer and DNA binding is illustrated in Supplementary Figure 1. In the process of gene regulation, nucleosomes play an important role as gene transcription is blocked in regions, where DNA is tightly packed by nucleosomes. Nucleosome occupancy affects epigenetic silencing [3], cell replication [4], differentiation [5], and reprogramming [5]. Determining whether DNA around histone octamer is tightly wrapped or loosely wrapped, a genetic task known as histone occupancy determination has profound importance in genetic research [6,7]. Accurate determination of histone occupancy can facilitate deeper understanding of DNA accessibility to proteins [8,9], chromatin functions [10,11], and occupancy correlation with promoter strength [12].

✉ Muhammad Nabeel Asim  
muhammad\_nabeel.asim@dfki.de

<sup>1</sup> German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany  
<sup>2</sup> Department of Computer Science, Technical University of Kaiserslautern, 67663 Kaiserslautern, Germany  
<sup>3</sup> National Center for Artificial Intelligence (NCAI), National University of Sciences and Technology, Islamabad, Pakistan  
<sup>4</sup> Deakin University, Melbourne, Australia

Similar to histone occupancy, histone modifications (acetylation, methylation, phosphorylation, sumoylation, and ubiquitylation) are responsible to regulate multifarious biological processes including chromosome wrapping [13,14] transcriptional activation and de-activation [15–17], damaging and repairing of DNA [18,19]. For instance, histone amino (N)-terminal tails modifications influence internucleosomal exchanges and are capable to modify chromatin structures which ultimately affect gene expression [20] and give birth to many complex diseases, such as Cancer [13].

To acquire a deeper comprehension of epigenetic regulation at cellular level and to pave way for the development of drugs specifically targeting cancer treatment, and histone altering enzymes [21], histone modification detection is essentially required [22]. Histone modifications [23] largely affect the availability of DNA to different transcription factors and ribonucleic acid polymerases. Histone octamers repeat themselves across all nucleosomes in histone sequences, hence properties of nucleosomes primarily rely on incorporated area of histone sequences with specific acetylation and methylation sites level [24,25]. In addition, considering methylation of histone proteins H3 and H4 mainly regulate the core activity of DNA replication [26], and acetylation of different histone proteins impact chromatin structure as well as gene transcription [27,28]. A thorough analysis of histone acetylation and methylation areas in histone sequences can decipher the association of histone modification with metabolism which mediates diverse epigenetic abnormalities in multifarious pathological conditions [29].

Developing a robust computational approach for accurate histone occupancy and modification prediction has been an active area of research, since the public availability of ten benchmark data sets developed by [30]. From ten benchmark histone marker data sets, two belong to histone occupancy, three are related to histone acetylation, and five are related to histone methylation. Across 10 different benchmark data sets, histone sequences having occupancy, acetylation, or methylation level greater than 1.2 belong to positive class and lower than 0.8 belong to negative class. To perform binary classification across all ten benchmark histone markers data sets, [30] proposed the very first computational approach for histone occupancy and modification prediction for yeast genome. Their proposed approach utilized occurrence of higher order residues to generate statistical representation of histone sequences and Support Vector Machine classifier.

Using 10 different benchmark data sets facilitated by [30] related to histone occupancy, acetylation, and methylation, to date, a number of computational methodologies have been developed [31–36]. Prime focus of existing computational approaches [30,34–36] has been to generate a rich statistical representation of histone sequences. In this regard, few researchers have utilized bag of words-based

approaches [30,34], whereas others have utilized one hot encoding scheme to generate statistical representation of histone sequences [31–33,35]. While, bag of words-based statistical representation only manages to capture residue frequency and neglects rich semantic information. One-hot encoding lacks to capture comprehensive contextual information and correlations of residues. Furthermore, bag of words and one-hot encoding schemes face the curse of dimensionality issue with the induction of higher order sequence residues.

Recently, [36] proposed a deep learning approach for histone occupancy and modification prediction. For each sequence, they transformed one-hot encoded vector of higher order residues into image-like tensor through the assignment of each higher order residue to a pixel in an image by making use of Hilbert curves. Image-based representation of histone sequences was passed to a CNN model for the extraction of important residue correlations and dependencies. Although image-based representation manages to find discriminative sequence residues, however, fails to handle transnational invariance of residues mainly due to the supreme attention towards local residue context. Despite the fact that histone sequences are primarily comprised of four basic residues [adenine (A), cytosine (C), guanine (G), and thymine (T)], treating them as a simple string of repetitive letters neglects their biologically relevant and inherent spatial configuration as well as interaction between sequence residues. Complex molecular spatial composition of histone sequences indicates the relevance of a rich statistical representation which can effectively capture long-range dependencies of residues. However, due to the lack of comprehensive understanding of sequence residue patterns, a rich statistical representation scheme for histone sequences related to histone occupancy, methylation, and acetylation does not exist.

Building on these deficiencies, for the establishment of an improved and more robust histone occupancy and modification landscape, the paper in hand develops a lightweight computational multi-paradigm framework, namely, “Histone-Net”. Considering the efficacy of neural language modelling in diverse Natural Language Processing (NLP) [37] and Bioinformatics tasks [38] for capturing long-range dependencies and relatedness of sequence residues as well as improving the generalizability of predictive pipeline. Histone-Net makes use of neural language modelling to generate a rich distributed representation of histone sequences. Inspiring from the extensive usage of FastText model to generate word or higher order residue embeddings in an un-supervised manner for diverse NLP (e.g., text classification) [37,39] and Bioinformatics tasks (e.g., protein family classification [40], enhancer prediction [41], n6-methyladenine sites prediction [42]). Histone-Net generates un-supervised higher order residue embeddings (DNA2Vec) of histone sequences. Furthermore, Histone-Net presents a

different application of FastText model, where it incorporates histone occupancy and modification information while learning higher order residue embeddings (SuperDNA2Vec) of histone sequences. To investigate which distributed representation learning scheme better captures coarse-grained and fine-grained relations of higher order residues, a rigorous intrinsic evaluation of both kinds of embeddings is performed by mapping high-dimensional feature space into low-dimensional feature space using Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) schemes. Extrinsic evaluation of both types of embeddings is also performed using three different machine learning classifiers (Random Forest, AdaBoost, Support Vector Machine).

Existing approaches [31–36] have not been evaluated in cross-domain binary classification paradigm, where for each histone sequence analysis task, model is trained on one type of histone marker and tested on another type of histone marker. We explore the performance potential of proposed Histone-Net approach in cross-domain binary classification paradigm. Furthermore, a critical analysis of existing computational approaches reveals that in all existing approaches, 10 different checkpoints are obtained by rigorously training the single model separately over 10 benchmark genomic data sets to predict histone occupancy, methylation, and acetylation areas in histone sequences. In this strategy, one needs to know the target histone marker beforehand to select appropriate checkpoint amongst all model checkpoints while making prediction over unseen histone sequences. More recently, [31] developed a deep learning approach “DeepHistone” to simultaneously predict different histone markers associated with particular sequence. However, DeepHistone is only capable to detect the type of histone marker modification and unable to predict histone occupancy and modification levels. Inspiring from the work of Yin et al. [31] who treated the identification of histone markers as multi-label classification problem, we develop a multi-label classification paradigm to deal with the expensive overhead of generating separate model checkpoints for ten benchmark data sets belonging to three histone sequence analysis tasks. More specifically, in multi-label classification paradigm, performance of Histone-Net is evaluated in terms of its ability to simultaneously predict histone marker type, its occupancy, acetylation, and methylation levels. A comprehensive evaluation of proposed approach under the hood of intra-domain and cross-domain binary classification as well as multi-label classification paradigm proves the dominance of proposed approach over state-of-the-art predictor, its generalization potential across multiple histone markers as well as power to simultaneously predict histone type, its occupancy, acetylation, and methylation areas using a single deep learning model.

## Materials and methods

This section illustrates different modules of computational framework Histone-Net, benchmark binary classification data sets, the process used to develop a multi-label classification data set, and evaluation metrics used to evaluate the integrity of Histone-Net in binary and multi-label classification paradigm.

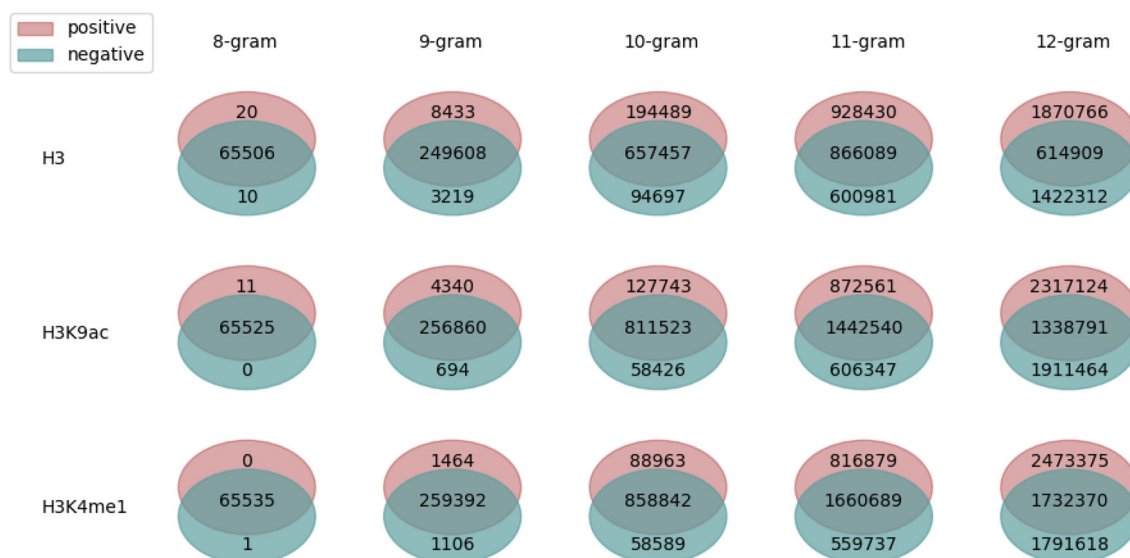
### Histone-Net framework

To better illustrate the dynamics of proposed computational framework Histone-Net, this section briefly describes four different modules of Histone-Net. First module generates overlapping higher order residues of histone sequences, details of residue generation is provided in “[Higher order residue generation of histone sequences](#)”. Second module generates unsupervised sequence embeddings (DNA2Vec) using FastText model, details of which are provided in “[Unsupervised distributed representation of histone sequences \(DNA2Vec\)](#)”. Third module presents a novel approach, namely, Histone-Net which generates histone occupancy and modification information aware residue embeddings in a supervised manner (SuperDNA2Vec) and is capable to perform histone occupancy and modification prediction in binary classification paradigm. Details of third module is provided in “[Histone-Net approach based on supervised distributed representation of histone sequences \(SuperDNA2Vec\)](#)”.

Fourth module presents the work of Histone-Net in multi-label classification paradigm, where it simultaneously predicts histone marker type, occupancy, acetylation, and methylation levels, a brief description of which is given in “[Histone-Net: multi-label classification paradigm](#)”.

### Higher order residue generation of histone sequences

Histone sequences are comprised of four repetitive letters adenine (A), guanine (G), cytosine (C), and thymine (T). One of the most common approach to analyze long histone sequences is segregating the sequences into higher order residues [43–48]. Mainly, two different kinds of higher order residues can be derived from the sequence including overlapping and non-overlapping [49] by rotating a fixed-size window with certain stride size (Supplementary Figure 2). Overlapping higher order residues are generated by rotating a fixed-size window over histone sequences with the stride size less than the size of the window. Whereas, non-overlapping higher-order residues are generated by rotating a fixed size window over histone sequences with stride size equal to the size of the window. Non-overlapping higher order residues lose the deep correlation information of basic residues, whereas overlapping residues retain deep correla-



**Fig. 1** Number of discriminative features of positive and negative class along with overlapping features among two classes across one data set of histone occupancy, acetylation, and methylation

tion information of residues. Following existing sequence analysis work [41,50–53], we generate overlapping higher order residues with 10 different degrees ranging from 2- to-12 to comprehensively assess the impact of higher order residues in disentangling rich contextual information for 3 distinct histone sequence analysis tasks.

Furthermore, considering, the performance of machine learning classifiers is directly proportional to the number of discriminative higher order residues, we analyze which window size of overlapping high order residues generates most discriminative residues for histone occupancy, acetylation, and methylation prediction tasks. We find that there does not exist any discriminative residues until 7-mers; however, afterward (e.g., 8-mers to 12-mers), there exist decent number of discriminative residues which occur in one class and absent in other class.

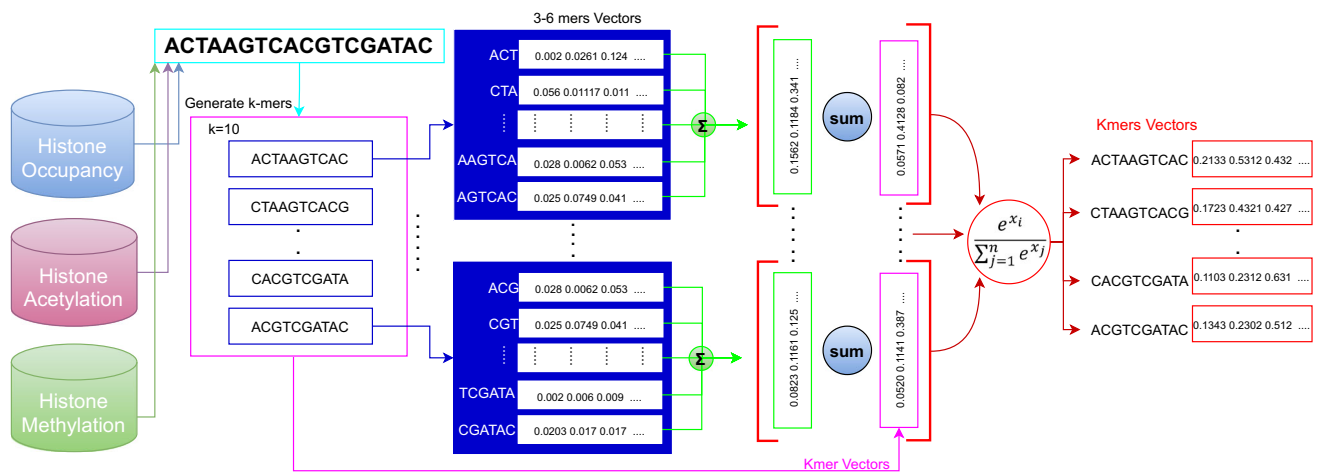
To provide a bird's eye view of discriminative potential of different overlapping higher order residues across different classes, we randomly select one data set from each histone sequence analysis task and reveal the overlapping and discriminative features for each class for 5 different higher order residues (8-mers to 12-mers) in form of venn diagrams (Fig. 1). As shown in Fig. 1, for histone occupancy H3 data set, in case of 8-mers, positive class has only 20 unique residues which are not present in the negative class and negative class has 10 unique residues which are not present in positive class. Whereas, 65,506 residues are present in both classes. With the increase of degree of  $k$ -mers, discriminative as well as overlapping residues also increase. A similar trend can also be seen with histone acetylation (H3K9ac) and methylation (H3K4me1) data sets.

### Unsupervised distributed representation of histone sequences (DNA2Vec)

In Natural Language Processing (NLP), two neural embedding generation models, namely, Continuous Bags of Words (CBOW), and Skipgram have been extensively used to learn dense word vectors based on their semantic relationships. Building on the success of these models in diverse NLP, genomics, and proteomics sequence analysis tasks, we generate rich distributed representation of histone sequences using CBOW model in an unsupervised manner (DNA2Vec), workflow of which is illustrated in Fig. 2. As is shown by the figure, for each histone sequence analysis task, higher order residues ( $k$ -mers) embedding vectors are generated in four steps. First, a corpus of overlapping high order residues ( $k$ -mers) is generated by combining all the data sets of only corresponding task. Second, we represent each higher order residue as bag of sub-mers.

Third, an element wise addition of randomly initialized vectors of different sub-mers present in high order residues is performed to generate inner residue distribution aware vectors. Finally, another vector addition is performed using these embedding vectors and complete high order residue vectors. In this manner, generated vectors of higher order residues contain the information about internal residue structure along with the properties of residues.

The training objective of un-supervised embedding generation model is to optimize higher order residue embedding matrix in such a way that model accurately estimates probability of target higher order residue (over all vocabulary of higher order residues) providing decent contextual



**Fig. 2** Workflow of unsupervisedly prepared higher order residue embeddings (DNA2Vec)

information through softmax layer. For each sequence analysis task, distributed representation of histone sequences is generated by averaging the distributed representation of higher order residues present in histone sequences.

### Histone-Net approach based on supervised distributed representation of histone sequences (SuperDNA2Vec)

The infusion of meta information (e.g., task type, task nature) in the training process of unsupervised distributed representation generation model can encapsulate a variety of inherent relationships. Such information creates a correlation of local as well as global sequence residue distribution with target task that eventually supports the accurate identification of target task. For efficient histone sequence analysis, we develop a unique distributed representation generation model (SuperDNA2Vec) which encapsulates histone occupancy, acetylation, and methylation information along with semantic information of higher order residues using feedback mechanism of supervised learning. As shown in Fig. 3, for each histone sequence analysis task including histone occupancy, acetylation, and methylation prediction, data sets of only corresponding task are combined. Using task specific data set collection, the process of generating task aware distributed representation of histone sequences starts by dividing the sequences into higher order residues (*K*-mers, e.g., 10-mers). Then, distributed representations of sequences are learned by averaging the randomly initialized vectors of higher order residues present in sequences, regularization of which is performed by applying the dropout rate of 0.5. The training objective of embedding generation model is to optimize higher order residue embedding matrix by monitoring the cross entropy loss produced while estimating the probability of target class (over all classes) through softmax layer. Through iterative learning, embedding generation

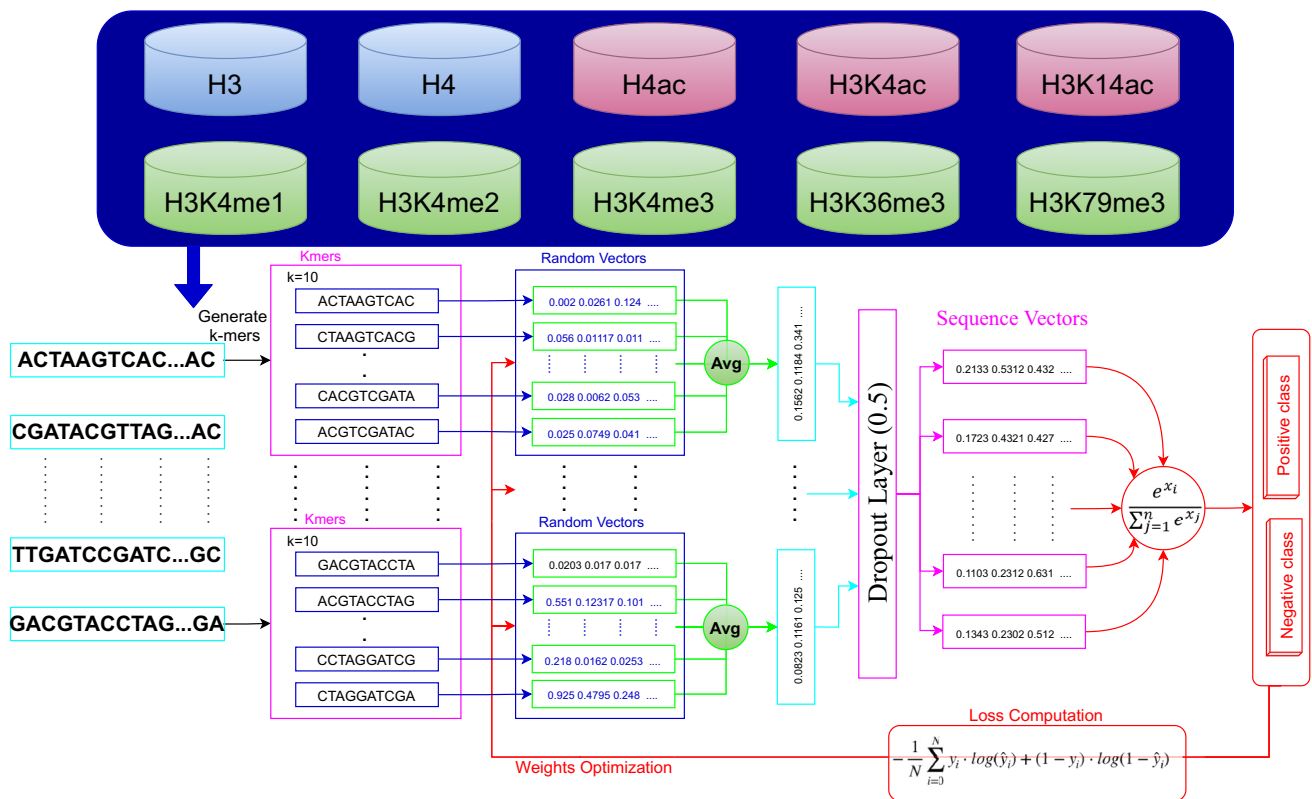
model manages to encapsulate histone occupancy, acetylation, and methylation information in embedding matrix. Precise deep learning model (Fig. 3) without softmax layer acts as an embedding generation model (SuperDNA2Vec) for test histone sequences. Whereas, entire deep learning model (Histone-Net) predicts histone occupancy, acetylation, and methylation using task aware distributed representation of histone sequences generated by SuperDNA2Vec model.

### Histone-Net: multi-label classification paradigm

Architecture of Histone-Net multi-label classification paradigm (Supplementary Figure 3) is identical to binary classification paradigm discussed earlier in “[Histone-Net approach based on supervised distributed representation of histone sequences \(SuperDNA2Vec\)](#)” except 2 major differences. Unlike Histone-Net binary classification paradigm, where SuperDNA2Vec sequence representation embeds only one type of task specific information (e.g., occupancy, acetylation, or methylation) for hand-on data set. In Histone-Net multi-label classification paradigm, SuperDNA2Vec sequence representation integrates all four histone marker type, occupancy, acetylation, and methylation information at the same time. Another difference is the use of sigmoid activation function at last layer and binary cross-entropy as a loss function instead of softmax activation and categorical cross-entropy loss function.

### Benchmark binary classification data sets for histone occupancy and modification prediction

This section illustrates the details of 10 public benchmark histone occupancy and modifications (acetylation and methylation) prediction data sets [36] used to evaluate the performance of proposed multi-paradigm computational



**Fig. 3** Workflow of supervisedly prepared higher order residues embeddings (SuperDNA2Vec) and Histone-Net approach

**Table 1** Statistics of 10 benchmark data sets including 2 data sets for histone occupancy detection, 3 data sets for acetylation and 5 data sets for methylation level prediction

Data set name	Description	Positive samples	samples
H3	H3 occupancy	7667	7298
H4	H4 occupancy	6480	8121
H3K4me1	H3K4 mono-methylation relative	17,266	14,411
H3K4me2	H3K4me2 H3K4 di-methylation relative to H3	18,143	12,540
H3K4me3	H3K4me3 H3K4 tri-methylation relative to H3	19,604	17,195
H3K36me3	H3K36me3 H3K36 tri-methylation relative to H3	18,892	15,988
H3K79me3	H3K79me3 H3K79 tri-methylation relative to H3	15,337	13,500
H3K9ac	H3K9 acetylation relative to H3	15,415	12,367
H3K14ac	H3K14 acetylation relative to H3	18,771	14,277
H4ac	H4 acetylation relative to H4	18,410	15,686

framework Histone-Net. The process used to prepare 10 benchmark data sets is described in previous studies [54], here we only summarize the statistics of 10 benchmark data sets. Table 1 describes the sample-to-label distribution of 2 histone occupancy (H3, H4), 5 methylation (H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K79me3), and 3 acetylation data sets (H3K9ac, H3K14ac, H4ac). For acetylation and methylation level prediction data sets, K with its leading number represents the Kth amino acid which has to be modified with mono, di, or tri acetyl (“ac”) and methyl

(“me”) modifications. For example, in H3K4me1 data set, 4th amino acid of H3 protein is modified with a mono methyl group.

For each benchmark data set, histone sequences having relative occupancy, methylation, and acetylation values greater than 1.2 belong to positive class and lower than 0.8 belong to negative class. In binary classification paradigm, Histone-Net is evaluated in intra-domain and cross-domain settings using 10 benchmark binary classification data sets.

### Multi-label classification data set for histone occupancy and modification prediction

This section describes the process used to develop multi-label classification data set for histone occupancy and modification prediction.

Figure 4 illustrates the complete workflow used to develop imbalanced and balanced version of multi-label histone sequence analysis data set by utilising 10 benchmark datasets given by [54]. All 10 benchmark data sets have total 2.74 million sequences, where each sequence is annotated with either 0 or 1. A closer look at sequence ids provided by [54] reveals that a significant number of sequence ids appear in multiple histone markers data sets. For instance, consider a sequence id “iTELL-Chr1\_61” which is annotated as 1 in H3 histone marker data set indicates that the sequence has histone occupancy more than 1.2, same sequence id is annotated as 0 in H3k4me1 histone marker data set indicating that the sequence methylation level is less than 0.8, and same sequence id is annotated as 1 in H4kme2 histone marker data set indicating that the sequence methylation level is more than 1.2.

This analysis serves as a basis to formulate multi-label data set, where each sequence id may have 20 labels at max instead of 1 label (0 or 1). From 20 labels, 10 labels represent the association of sequence with positive class distribution of 10 benchmark histone markers, whereas other 10 labels show the association of sequence with negative class distribution of 10 benchmark histone markers. In this manner, each sequence target label is represented as a 20-dimensional binary vector, where 1 is assigned for the association of sequence with positive class of particular histone marker and 0 is assigned for the association of sequence with negative class of particular histone marker. Considering whether multi-label data set is imbalanced or balanced largely impact the predictor perfor-

mance, two different versions of multi-label histone sequence analysis data set are prepared.

Balanced version of multi-label data set is prepared by eliminating all label cardinalities which have less than 25 sequences, whereas imbalanced version of data set is obtained by retaining all label cardinalities which have less than 25 sequences. In this manner, balanced multi-label data set of 19,088 and imbalanced data set of 22,191 sequences are obtained, sample-to-label distributions of which are illustrated in Fig. 5. In both imbalanced and balanced version of multi-label data sets, hepta-label cardinality have most number of sequences followed by octa-label cardinality. In imbalanced version which retains all label cardinalities, uni-label cardinality has least number of sequences followed by bi-label cardinality, histone markers and respective class distribution of which are depicted in Fig. 6. H3K79me3\_neg has most number of uni-label sequences, whereas H3K4me2\_pos-H3K4me2\_pos has most number of bi-label sequences. Likewise, histone markers and class distribution of tri-label and tetra-label cardinalities which have more than 25 number of sequences are shown in Fig. 7. Combination of H3k36me3\_pos, H3k4me1\_pos, and H3k4me2\_pos has most number of tri-label sequences, whereas the combination of H3k4me2\_pos, H3K79me3\_pos, H3K36me3\_pos, and H3K4me1\_pos has most number of tetra-label sequences. Beside trivial binary classification paradigm, Histone-Net is evaluated in multi-label classification paradigm using imbalanced and balanced version of multi-label histone sequence analysis data set.

### Evaluation criteria

To perform a comprehensive evaluation of Histone-Net in binary and multi-label classification paradigm under the hood of diverse settings, this paper utilizes 15 different evaluation

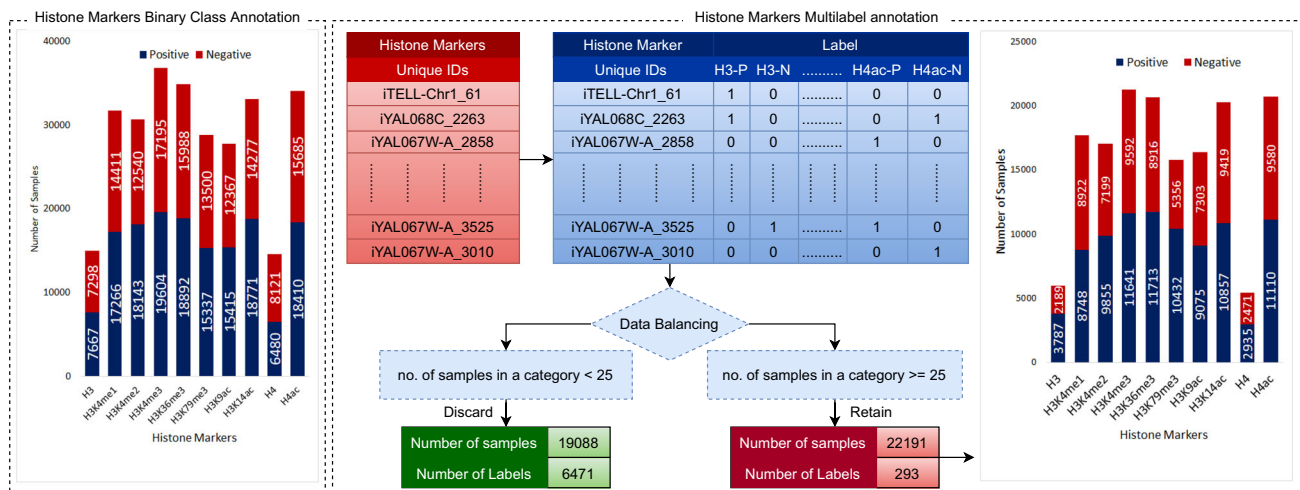


Fig. 4 Development workflow of imbalanced and balanced multi-label classification data sets for histone occupancy and modification prediction

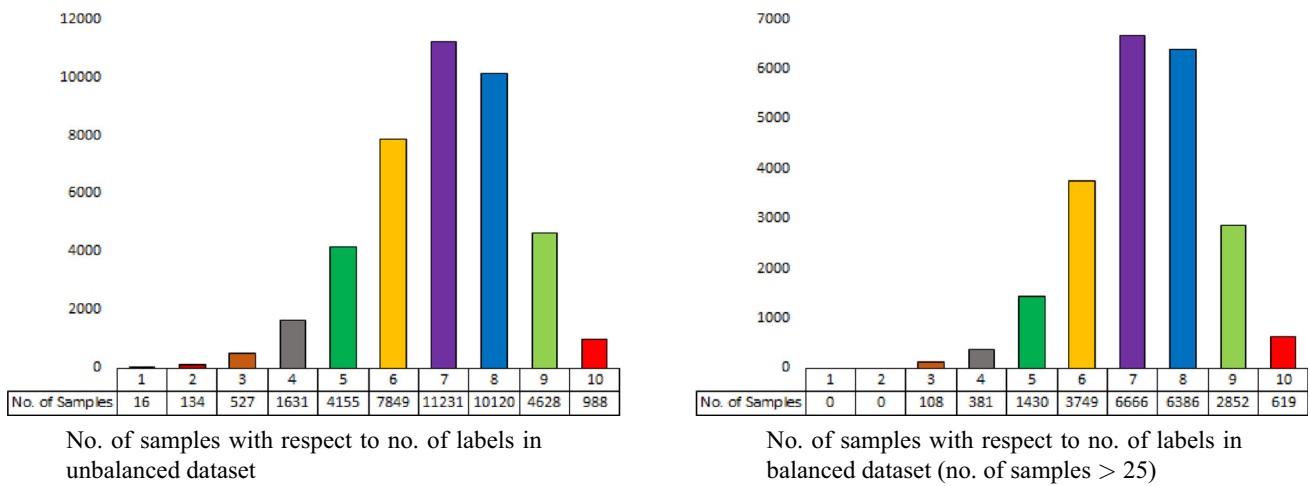


Fig. 5 Sample-to-label distribution of two versions of multi-label classification data sets for histone occupancy and modification prediction

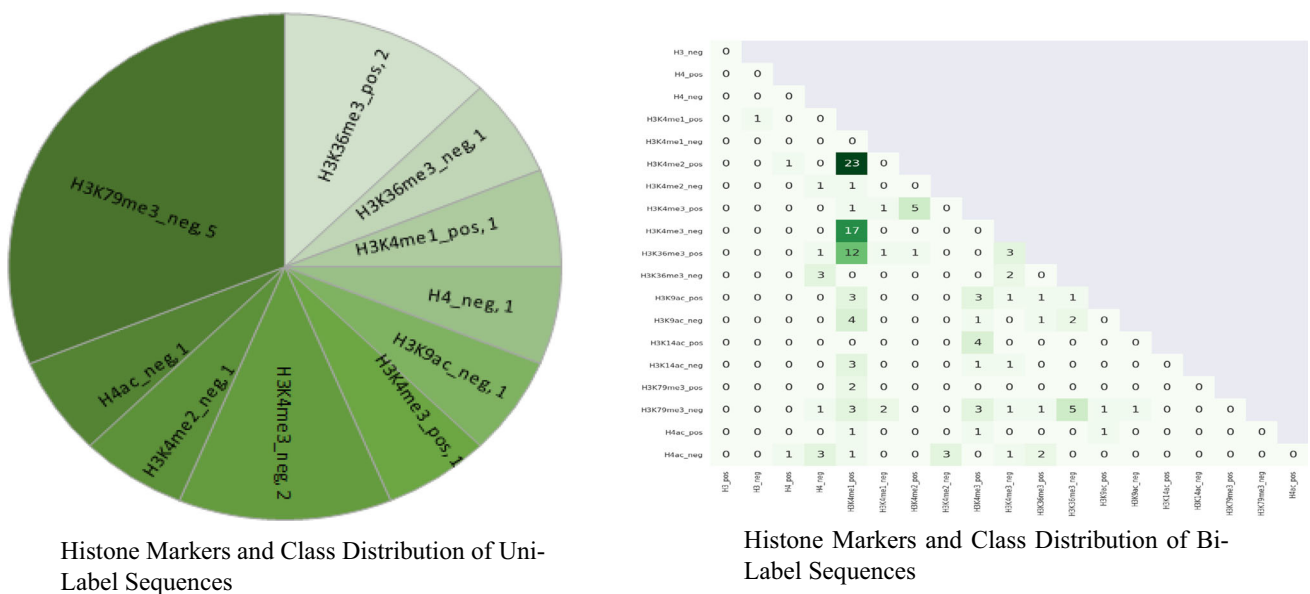


Fig. 6 Histone markers having less than 25 samples with uni and bi-label cardinalities

metrics, short descriptions of which is provided in following subsections.

### Binary classification evaluation metrics

This paper utilizes 4 most widely used binary classification evaluation metrics to assess the performance of Histone-Net framework including accuracy, precision, recall, and F1-score. Accuracy is the most simplest evaluation metric which computes the ratio between correctly predicted histone sequences and total histone sequences. Recall measures the true positive rate and precision computes the percentage of correct positive predictions from all positive predictions. While recall effectively handles type II errors, precision effectively deals type I errors. F1 Score is considered a

trade-off among recall and precision that depicts orthogonal relationship. It is calculated through the weighted average of recall and precision. Among all, Mathematical expressions of four different evaluation metrics are given below:

$$f(x) = \begin{cases} \text{Accuracy} = \frac{(o^+ + o^-)}{(o^+ + o^-)} \\ (\text{Rec}) \text{ Recall} = \frac{(o^+)}{(o^+ + F^+)} \\ (\text{Pre}) \text{ Precision} = \frac{(o^+)}{(o^+ + F^+)} \\ \text{F1-score} = 2 * \frac{\text{Pre} * \text{Rec}}{\text{Pre} + \text{Rec}} \end{cases} \quad (1)$$

In Eq. 1,  $O^+$  represents false positives and true positives,  $O^-$  refers to false negatives and true negatives. Total positive class histone sequences which are accurately identified





Fig. 7 Histone markers having more than 25 samples with tri- and tetra-label cardinalities

as positive referred by  $O_+^+$ , similarly total negative class histone sequences which are corrected predicted as negative are expressed using  $O_+^-$ . Whereas, histone sequences which are wrongly predicted into positive class (False positives) are represented as  $F_+^+$  and histone sequences which are mistakenly classified into negative class (False Negatives) are referred as  $F_+^-$ .

**Multi-label classification evaluation metrics**

To evaluate the performance of Histone-Net in multi-label classification paradigm, we have utilised 11 different evaluation metrics, namely, accuracy, precision, recall, F1-score, area under the receiver operating characteristic (AU-ROC), area under precision recall curve (AU-PRC), average precision, one error, hamming loss, ranking loss, and coverage.

For multi-label classification, accuracy computes the proportion of accurately predicted label to the total number of labels, averaged over all sequences present in test set. Precision computes the proportion of correctly predicted labels to total number of predicted labels, recall calculates true positive rate, and F1-score compute harmonic mean of precision and recall, averaged over all sequences of test set. Hamming loss is the fraction of miss-classified labels, ranking loss computes the average fraction of reversely ordered label pairs for each histone sequence. Average precision computes the proportion of relevant labels in the highest ranked labels. One error calculates the fraction of histone sequences whose most confident predicted label is irrelevant. Coverage computes average number of labels which should be included to cover entire set of relevant labels. Area under precision recall curve indicates the trade-off among precision and recall. Area Under Receiver Operating Characteristic (AU-ROC) reveals

the performance of a classifier at various thresholds. It is important to mention that for most evaluation metrics, higher performance figures show better performance except hamming loss, ranking loss, one error, and coverage, where lower performance figures show better performance. Mathematical expressions of different multi-label evaluation metrics are provided below:

$$f(x) = \begin{cases} \text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \frac{|y^{(i)} \wedge \hat{y}^{(i)}|}{|y^{(i)} \vee \hat{y}^{(i)}|} \\ (\text{Rec}) \text{ Recall} = \frac{1}{N} \sum_{i=1}^N \frac{|y^{(i)} \wedge \hat{y}^{(i)}|}{|y^{(i)}|} \\ (\text{Pre}) \text{ Precision} = \frac{1}{N} \sum_{i=1}^N \frac{|y^{(i)} \wedge \hat{y}^{(i)}|}{|\hat{y}^{(i)}|} \\ \text{F1-score} = \frac{1}{N} \sum_{i=1}^N \frac{2x|Pre(x_i)*Rec(x_i)|}{(|Pre(x_i)|)+(|Rec(x_i)|)} \\ \text{Hamming Loss} = \frac{1}{NL} \sum_{i=1}^N \sum_{j=1}^L [I(y_j^{(i)} \neq \hat{y}_j^{(i)})] \\ \text{Ranking Loss}(y, \hat{f}) = \frac{1}{N} * \sum_{l=0}^{N-1} \frac{1}{\|y_l\|_0 * (m_{\text{labels}} - \|y_l\|_0)} \\ \text{Average Precision}(f, C) = \frac{1}{|N|} \sum_{i=1}^{|N|} \sum_{b \in B_i} \frac{||b'|_{\text{rank}(a_i, b') \leq \text{rank}(a_i, b)}|}{\text{rank}(a_i, b)} \\ \text{One Error}(F) = \frac{1}{N} \sum_{i=1}^N [|\arg \max F(x_i) \notin Y_i^+|] \\ \text{Coverage} = (y, \hat{f}) = \frac{1}{N} * \sum_{l=0}^{N-1} \max_j: y_{ij}=1 \text{rank}_{ij} \\ \text{rank}_{ij} = |\{k : \hat{f}_{ik} \geq \hat{f}_{ij}\}| \end{cases} \quad (2)$$

In Eq. 2,  $N$  represents total number of sequences,  $Y_i$  represents the ground truth label assignment for  $i$ th sequence,  $\hat{Y}_i$  represents the predicted label for  $i$ th sequence,  $x_i$  represents  $i$ th sequences,  $j$ th represents the class index,  $\wedge$  represents logical AND operator, and  $\vee$  represents logical OR operator.

**Experimental setup**

Proposed computational framework Histone-Net is implemented using Scikit-Learn [55] and Pytorch [56]. To perform a fair performance comparison of Histone-Net predictive

methodologies with state-of-the-art histone occupancy and modification predictor [36], following Yin et al. [36], in both adapted DeepHistone [31] and proposed Histone-Net approach, randomly chosen 90% sequences are used for training and 10% sequences are used for testing. From 90% training sequences, 10% sequences are used as a validation set. We perform a large scale experimentation to develop an optimal model for histone occupancy and modification prediction. We assess the performance of DNA2Vec and SuperDNA2Vec sequence embeddings of 8 different dimensions (25, 32, 50, 64, 75, 100, 128, 150) using three different machine learning classifiers (RFC, AdaBoost, SVM). These sequence embeddings are prepared by averaging the statistical vectors of higher order residues present in them. We find that DNA2Vec and SuperDNA2Vec 100-dimensional sequence vectors mark best performance for intra-domain and cross-domain binary classification paradigms, whereas 64-dimensional sequence vectors perform better for multi-label classification paradigm. In all settings, embedding generation model is trained for 10 epochs, where we tweak the dropout from 0.1 to 0.5 only during SuperDNA2Vec embedding generation. From different batch sizes (32, 64, 128, 256), learning rates (0.001-to-0.008), and decay rates (0.91-to-0.99), proposed deep learning approach performs better when it is trained with a batch size of 64, Adam [57,58] optimizer decay rate of 0.95, and learning rate of 0.008.

To find optimal hyperparameter values for machine learning classifier, we tweak quality of split, number of estimators, kernel type, degree, gamma, and penalty parameter using GridSearch [59]. We find that tree-based machine learning classifiers perform better with gini criteria using 50 number of estimators, discriminative classifier SVM performs better with radial basis kernel, degree of 2, penalty parameter (C) of  $2^{-5}$  and gamma value of 0.001. After finding optimal DNA2Vec and SuperDNA2Vec sequence vectors as well as hyperparameter values, we perform experimentation with 11 different higher order residues ranging from 2-to-12 to determine which higher order residue-based sequence embeddings comprehensively help the classifier to make accurate predictions. For 10 benchmark data sets of 3 different histone sequence analysis tasks, we generate 99 (tasks/unique data set groups \*  $k$ -mers \* machine learning classifiers =  $3 * 11 * 3$ ) predictive pipelines for DNA2Vec sequence embeddings and 330 (data sets \*  $k$ -mers \* machine learning classifiers =  $10 * 11 * 3$ ) predictive pipelines for SuperDNA2Vec embeddings. Proposed Histone-Net approach generates 110 (data sets \*  $k$ -mers \* deep learning classifiers =  $10 * 11 * 1$ ) predictive checkpoints. From different higher order residues, we find that 7-mers to 11-mers sequence embeddings mark best performance. More specifically, 7-mers DNA2Vec and 11-mers SuperDNA2Vec sequence embeddings mark best performance across all classifiers. To evaluate adapted DeepHistone approach [31] across 10 benchmark histone markers

data sets, we utilize the source code and parameters provided by Yin et al. [31].

## Results and discussion

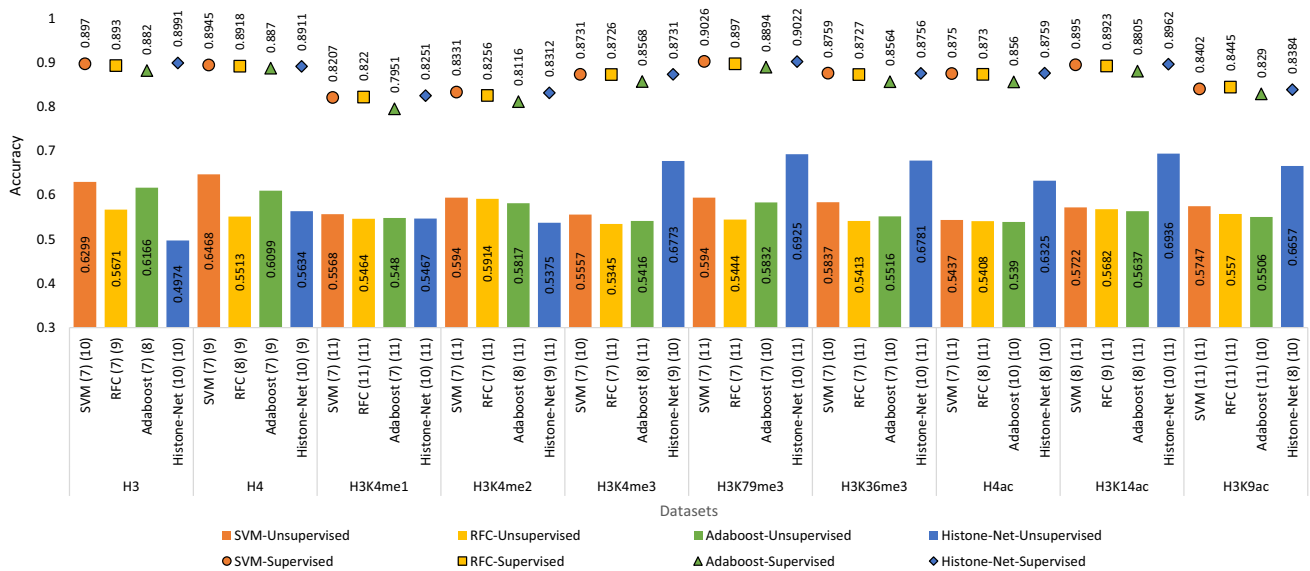
This section performs comprehensive extrinsic and intrinsic evaluation of DNA2Vec and SuperDNA2Vec sequence embeddings. It compares the performance of proposed approach with machine learning classifiers, adapted convolutional neural network-based approach DeepHistone [31], and state-of-the-art image representation-based predictor HCNN [36].

### Extrinsic evaluation of DNA2Vec and SuperDNA2Vec distributed representation in intra-domain setting using binary classification paradigm

Performance values produced by 3 machine learning classifiers with unsupervised and supervised sequence embeddings using 5 different higher order residues (7-to-11) are provided in Supplementary Tables 1 and 2 in terms of 4 distinct evaluation measures. Furthermore, to analyze which higher order residue-based supervised and unsupervised sequence embeddings perform better with what classifier across 3 different histone sequence analysis tasks, performance values are compared in terms of accuracy and F1 score, graphical illustrations of which are provided in Fig. 8 and Supplementary Figure 4, respectively.

A critical performance analysis of supplementary Table 1 indicates that among 3 machine learning classifiers, SVM classifier marks better performance with un-supervised sequence embeddings across 3 distinct histone sequence analysis tasks. Over 2 histone occupancy 5 histone methylation, and 3 histone acetylation prediction data sets, among all different higher order residues ( $K$ -mers), 7-mers sequence embeddings produce better performance across most machine learning classifiers.

Furthermore, performance analysis of Supplementary Table 2 indicates that with the induction of target histone occupancy and modification information, performance of all 3 machine learning classifier gets significantly rocketed over all histone occupancy, methylation, and acetylation data sets. Higher order residues such as 10-mers and 11-mers mark superior performance across 10 benchmark data sets belonging to histone occupancy, acetylation, and methylation prediction. Among all classifiers, SuperDNA2Vec and softmax classifier-based approach (Histone-Net) achieves best performance across 10 benchmark data sets. Second, best performance for histone occupancy, acetylation, and methylation prediction tasks across different evaluation metrics is achieved by SVM classifier.



**Fig. 8** Difference in accuracy of classifiers over 10 different histone occupancy, acetylation and methylation prediction data sets produced using unsupervised and supervised sequence embeddings

In both bar graphs (Fig. 8, Supplementary Figure 4), it can be seen that SuperDNA2Vec sequence embeddings significantly raises the performance of all classifiers. A bird's eye view of the accuracy graph (Fig. 8) reveals that, using SuperDNA2Vec sequence representation, over both histone occupancy prediction data sets (H4, H3), among all 4 predictors, Histone-Net performance takes the major leap of almost 37% followed by 33.5% of RFC, 27%, AdaBoost, and 26% of SVM when compared against the performance produced using unsupervised higher order residues-based sequence embeddings. For most methylation prediction data sets, RFC performance reach the increment figure of 31%, SVM, and Adaboost performance increases with an average figure of 28%, and Histone-Net incurs 26% increment in performance. Likewise, across all acetylation prediction data sets (H4ac, H3K14ac, H3K9ac), SVM and Adaboost performance increment is identical (31%), RFC marks highest accretion of 32%, whereas Histone-Net performance increases by the figure of 21% in terms of accuracy.

Using supervised sequence embeddings (SuperDNA2Vec), assessing the performance rise in terms of F1-score (Supplementary Figure 4), among all, Histone-Net performance gets rocketed by the average figure of 46% on histone occupancy prediction data sets (H4, H3). Whereas, RFC marks the average increment of 47% over acetylation prediction data sets and Adaboost manages to achieve the average performance increment of 53% for most methylation prediction data sets.

To summarize, higher order residues assist to disentangle rich contextual information and contribute to learn better distributed representation. From the comparison of DNA2Vec and SAuperDNA2Vec, it is evident that SuperDNA2Vec

sequence representation effectively exploits the correlation of higher order residues with target histone occupancy and modification information which eventually greatly influences the performance of all classifiers. Although, all machine learning classifiers performance produced using unsupervised sequence embeddings gets improved significantly with the utilization of supervised sequence embeddings (SuperDNA2Vec); however, among all, softmax classifier outperforms the top performance of all baseline classifiers with decent margin across most histone markers. It is important to mention that machine learning classifiers (e.g., SVM) marginally performs better than proposed Histone-Net approach on few histone markers data sets, because these classifiers make use of SuperDNA2Vec sequence embeddings. Here, the aim of analyzing the performance of different machine learning classifier is to show that even simple classifiers can produce a performance similar to proposed deep learning approach using better statistical sequence representation.

### Intrinsic evaluation of DNA2Vec AND SuperDNA2Vec distributed representation

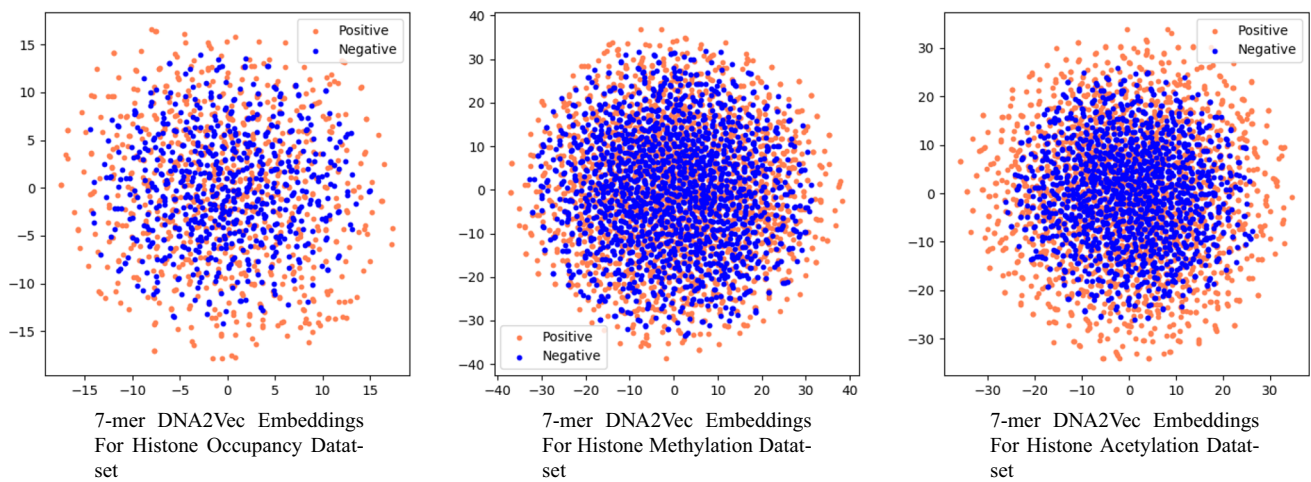
To analyze whether DNA2Vec or SuperDNA2Vec distributed representation learning scheme generates comprehensive occurrence and positional information aware vectors for higher order residues. We perform intrinsic evaluation of both presented distributed representation generation schemes on randomly selected three data sets, one from each histone sequence analysis task. It can be concluded from Supplementary Tables 1 and 2, among different higher order residues-based sequence embeddings, across 10 different his-

tone marker data sets, overall, most classifiers perform better with 7-mer unsupervised and 11-mer supervised sequence embeddings. Hence, we perform intrinsic evaluation using only 7-mer unsupervised and 11-mer supervised sequence embeddings. To effectively visualize the intrinsic patterns on each selected histone occupancy, acetylation, and methylation data set, we randomly select 10% sequences from positive as well as negative class, 100-dimensional DNA2Vec representations of which are passed to PCA approach that reduces the dimensions from 100 to 25. These 25 dimensional sequence vectors are finally passed to T-SNE visualizer that further reduces the dimensions and create mappings in two-dimensional space, where histone sequences of each class are represented with unique color.

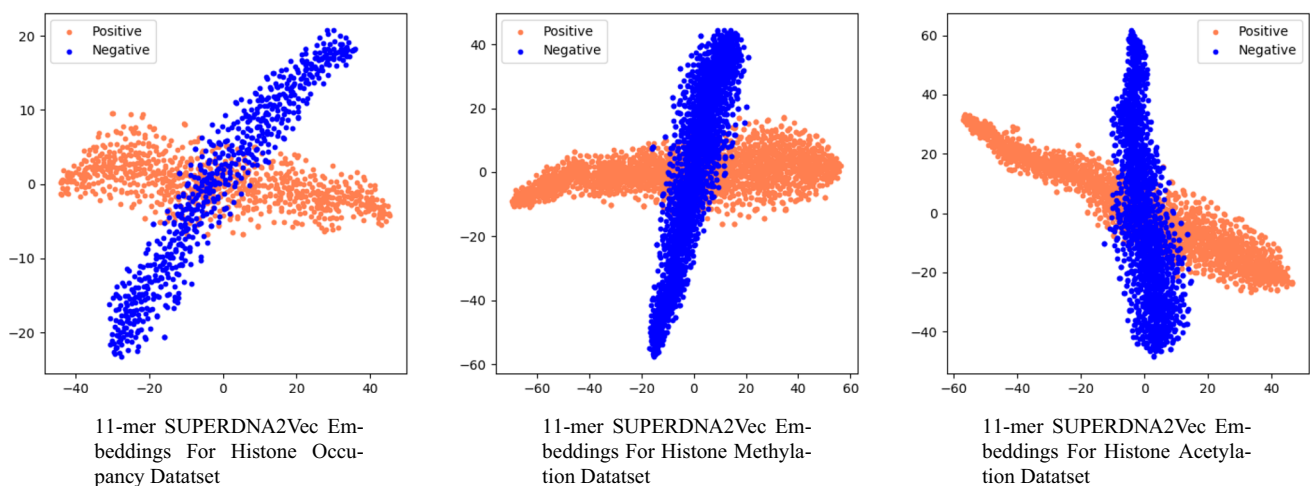
Likewise, to prepare supervised higher order sequence embeddings SuperDNA2Vec, each histone occupancy and modification data set is splitted into train, validations and

test sets containing 80%, 10%, and 10% histone sequences, respectively. For each histone sequence analysis task, we utilize the training and validation data to effectively train SuperDNA2Vec model which is later used to generate embeddings for histone sequences present in test set. Visualization of histone sequences present in test set is performed in a similar manner using PCA and T-SNE.

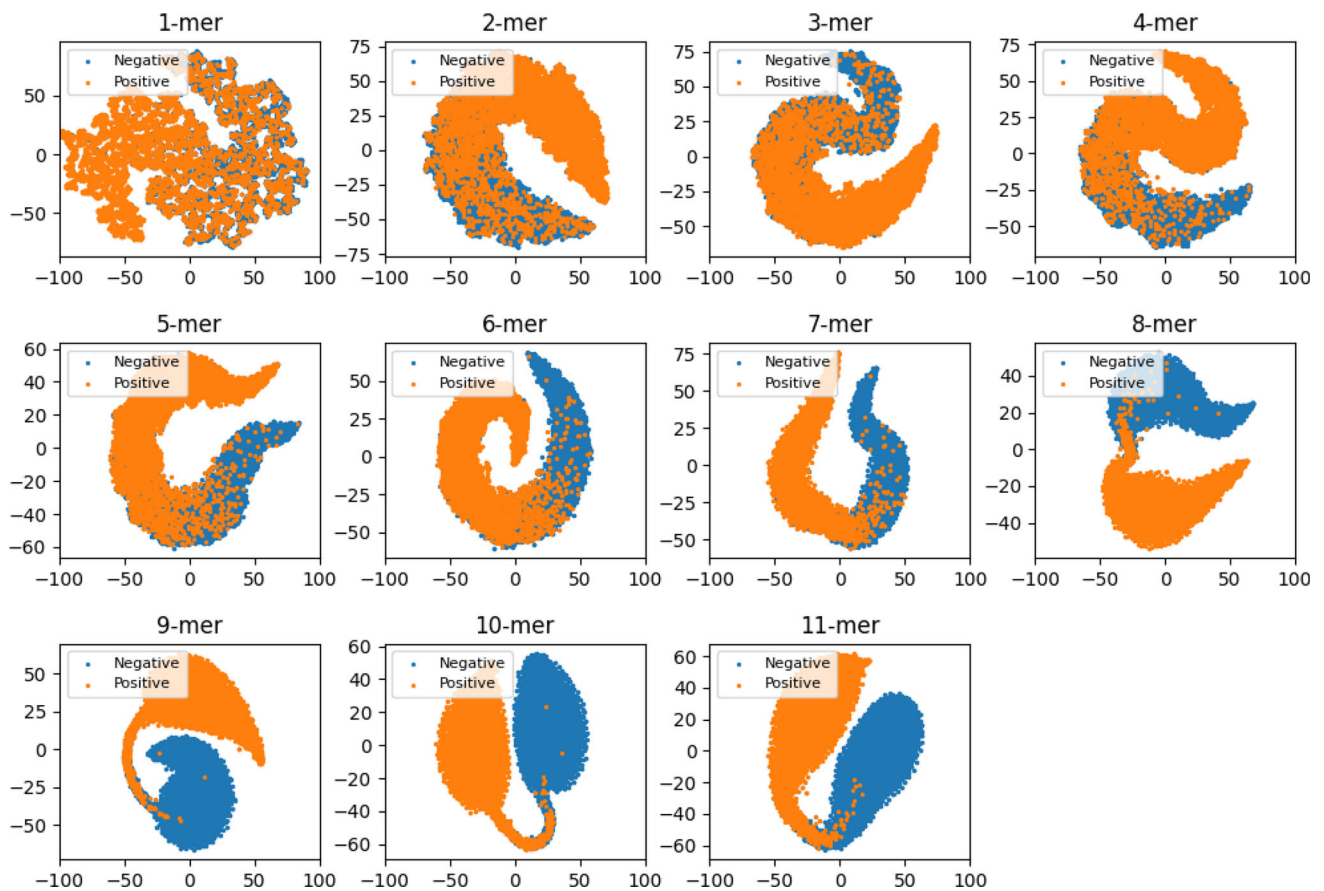
As is depicted by the embedding charts Figs. 9 and 10, clusters for both positive and negative class produced by supervised sequence embeddings (SuperDNA2Vec) of all three selected data sets are far less overlapping as compared to the charts produced by unsupervised sequence embeddings (DNA2Vec), where the clusters are highly overlapping. In addition, it is quite evident that supervised sequence embeddings better captures the local and global semantic composition of higher order residues which eventually assists to develop optimal sequence clusters. Therefore, building on



**Fig. 9** Unsupervisedly learned higher order residue-based embeddings of 10% histone sequences for histone occupancy, acetylation, and methylation tasks



**Fig. 10** Supervisedly learned higher order residue-based embeddings of 10% histone sequences for histone occupancy, acetylation, and methylation tasks



**Fig. 11** Intrinsic evaluation of different higher order residues-based supervisedly learned sequence embeddings (superdna2vec) over histone occupancy (H3) data set

these clusters, for new instances, there are strong chances that un-supervised sequence representation is more likely to cause mis-classification as compared to supervised sequence representation, where the chances of correctly classifying the given instance are far high.

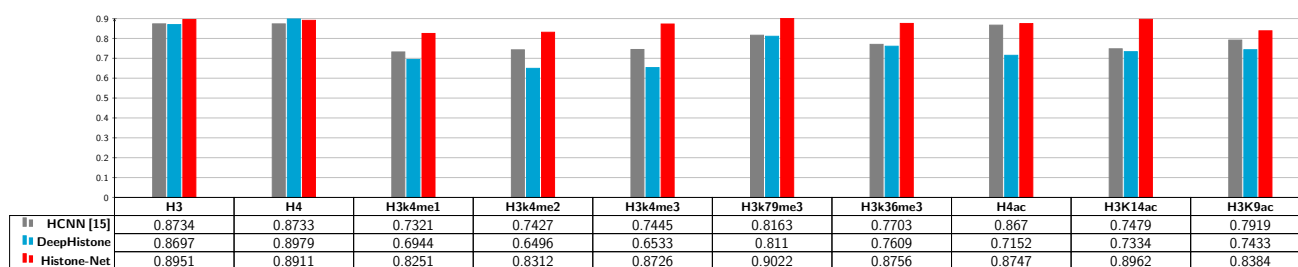
Furthermore, to intrinsically illustrate the discriminative impact of different higher residues-based SuperDNA2Vec embeddings, Fig. 11 shows the positive and negative histone sequence vectors clusters produced using 11 types of higher order residues for histone occupancy data set H3. It is evident from Fig. 11 that clusters are highly overlapping at lower higher order residues (1-mer to 6-mer); however, they are less overlapping at upper degree higher order residues (7-mer to 11-mer).

### Performance comparison of Histone-Net approach with adapted and state-of-the-art histone occupancy and modification predictors

We perform a fair performance comparison of SuperDNA2Vec and softmax classifier-based Histone-Net approach with image representation- based state-of-the-art histone occu-

pancy and modification predictor, namely, “HCNN” [36] and adapted convolutional neural network-based approach DeepHistone [31].

Figure 12 reports the performance of Histone-Net approach, state-of-the-art HCNN [36] approach, and adapted DeepHistone [31] approach over 10 different histone occupancy, methylation and acetylation prediction data sets in terms of accuracy. Performance of proposed Histone-Net, state-of-the-art HCNN [36], and adapted DeepHistone [31] approach in terms of other evaluation metrics (e.g., precision, recall) is given in Supplementary Table 3. Considering bag-of-words and one-hot encoding approaches face the issue of data sparsity and fail to consider deep contextual information of higher order residues, HCNN transformed histone sequences into 2-dimensional image-like arrays to magnify the strengths of convolutional neural network by making use of space-filling curves (Hilbert curve). HCNN [36] managed to outshine SVM, [34] CNN, and LSTM [35,36]-based approaches with a decent margin. As illustrated in Fig. 12, for both histone occupancy prediction data sets (H3,H4), on average, HCNN achieves the performance figures around 87%. For most methylation prediction data sets, HCNN average



**Fig. 12** Accuracy comparison of proposed Histone-Net approach with state-of-the-art HCNN [36] and adapted DeepHistone [31] approach. Accuracy values of HCNN are taken from Table 3 of Yin et al. [36]

performance falls around 75% except H3k79me3 data set, where it crosses the figure of 80% when evaluated in terms of accuracy, precision, and recall. Similarly, for acetylation prediction data set, it manages to mark the performance of nearly 80%.

To perform a rich performance assessment of proposed Histone-Net approach, we adapt a convolutional neural network-based approach DeepHistone proposed by Yin et al. [31]. Adapted DeepHistone process raw histone sequences, statistical representation of which is generated using one-hot encoding scheme. As shown in Fig. 12, DeepHistone only manages to achieve over 85% accuracy on 2 histone occupancy prediction data sets, over 80% accuracy on only 1 histone methylation prediction data set (H3K79me3) from 5 histone methylation prediction data sets, and over 70% accuracy on 3 histone acetylation prediction data sets. The reasons behind the limited performance of DeepHistone [31] in comparison to state-of-the-art HCNN [36] is the use of sub-optimal statistical representation learning scheme which lacks to capture translational invariance of residues.

Although image-based sequence representation approach successfully extracts discriminative higher order residues-based features, however, it still fails to acquire the comprehensive semantic information of diverse higher order residues. In addition, image-based sequence representation approach fails to extract correlations and diverse interactions of higher order residues important for concerned histone sequence analysis task. Due to these factors, the idea of incorporating target histone occupancy and modification information while generating distributed representation of higher order residues through language modelling (SuperDNA2Vec) proves extremely effective.

SuperDNA2Vec sequence embedding scheme largely assists the softmax classifier in achieving superior performance than HCNN [36] across 10 different benchmark data sets. While for histone occupancy and acetylation prediction data sets (H3,H4), on average, Histone-Net performance falls around 90% and 86%, respectively. Whereas, for most methylation prediction data sets, its performance crosses the

work and accuracy values of DeepHistone are obtained by processing raw histone sequences of various histone markers using convolutional neural network model presented by the authors [31]

landmark of 85% and on H3k79me3 data set it manages to achieve the top performance of 90% when assessed in terms of accuracy, precision, and recall. Likewise, Histone-Net significantly outperforms adapted DeepHistone [31] approach across all 10 benchmark histone markers data sets for 3 different histone sequence analysis tasks. For histone occupancy prediction, on average, Histone-Net achieves an increment of 3%, for histone methylation prediction, it attains an increment of 15% and for histone acetylation prediction, it achieves an increment of 14%. A similar performance trend in terms of other evaluation metrics can be seen in Supplementary Table 3.

In a nutshell, a different application of language modelling, where sequence embeddings are learned by taking the association of higher order residues with target histone occupancy and histone modification into account outperforms bag-of-words, one-hot encoding, image-based, and un-supervised higher order residues-based sequence representation schemes. SuperDNA2Vec effectively captures comprehensive long-range dependencies of higher order residues, their unique interactions, and occurrence in the distribution of diverse histone proteins. Using SuperDNA2Vec, simple softmax classifier manages to outperform state-of-the-art image representation and CNN-based histone occupancy, acetylation, and methylation predictor with significant margin.

### Evaluation of Histone-Net approach in cross-domain binary classification paradigm

In biomedical sequence analysis, generally, cross-domain evaluation is used to examine the effectiveness of computational predictors for practical applications. Considering existing histone occupancy and modification predictors are not evaluated in cross-domain setting, we validate the generalization potential of Histone-Net approach by performing cross-domain evaluation. In cross-domain setting, for histone occupancy sequence analysis task, Histone-Net approach is trained over the sequences of different histone markers

belonging to Histone occupancy and tested on one of the test set of particular histone marker. This process is repeated to ensure that Histone-Net approach is evaluated on the test set of each histone marker belonging to histone occupancy. Similar process is repeated for histone acetylation and methylation prediction tasks to ensure that Histone-Net is not biased towards specific histone marker data. In this manner, cross-domain performance of Histone-Net approach over test sets of 10 benchmark data sets belonging to 3 distinct histone sequence analysis task is computed.

Like intra-domain setting, in cross-domain setting, performance of Histone-Net approach is assessed using 5 different higher order residues (7-to-11). Table 2 summarizes the peak performance achieved by Histone-Net approach under different higher order residues over the test sets of 10 different benchmark data sets belonging to histone occupancy, acetylation, and methylation prediction. As indicated in Table 2, just like intra-domain setting, Histone-Net achieves top performance of around 86% in terms of 4 different evaluation metrics using upper degree higher order residues (11-mers)-based sequence representation in cross-domain setting for the task of histone acetylation prediction. Whereas, for 2 other histone sequence analysis tasks including Histone Occupancy and Histone Methylation prediction, unlike intra-domain setting, here Histone-Net marks better performance with medium degree higher order residue (7-mers)-based sequence representation. For histone occupancy, Histone-Net achieves best performance of 71% on test set of H3 histone marker as compared to H4 across all 4 evaluation metrics. For histone acetylation prediction, Histone-Net achieves better performance of 86% on test set of H4ac followed by H3K14ac and H3K9ac. Whereas, for histone methylation prediction, Histone-Net attains best performance of 72% on the test of H3K79me3 data set.

Empirical evaluation on the test sets of 10 benchmark data sets belonging to 3 distinct histone sequence analysis tasks indicates that Histone-Net manages to attain the average performance of more than 80% for histone acetylation

(H3K14ac, H3K9ac, H4ac), 70% for histone occupancy, and 60% for histone methylation prediction. Across 10 benchmark data sets, compared to average performance of 87% and peak performance of 90% achieved by Histone-Net in intra-domain setting, Histone-Net manages to attain an average performance of 70% with the peak performance of 86% in cross-domain setting. Usually, the performance of computational approaches drops up to great extent when evaluated using cross-domain paradigm; however, SuperDNA2Vec and softmax classifier-based Histone-Net methodology shows decent generalization potential across a variety of data sets belonging to 3 distinct histone sequence analysis tasks.

### Evaluation of Histone-Net approach in multi-label classification paradigm

This sections briefly describes the performance of Histone-Net multi-label classification paradigm using balanced and imbalanced version of multi-label data set in terms of 11 different evaluation metrics. To better quantify the effectiveness of Histone-Net multi-label classification paradigm, it compares the performance of Histone-Net with an adapted convolutional neural network-based methodology DeepHistone [31].

Supplementary Table 4 reports the performance produced by Histone-Net over im-balanced and balanced version of multi-label histone occupancy and modification prediction data set using different higher order residues-based SuperDNA2Vec sequence embeddings. As is evident by Supplementary Table 4, across both versions of multi-label data set, Histone-Net performance almost gradually improves by increasing the degree of higher order residues, indicating Histone-Net achieves best performance with upper degree higher order residues (10-mers, 11-mers)-based supervised sequence embeddings across all evaluation metrics. Furthermore, across all different higher order residue thresholds, Histone-Net achieves slightly better performance on balanced version of multi-label data set as compared to

**Table 2** Performance produced by Histone-Net using different degree higher order residue-based sequence representation in cross-domain setting

Histone marker test set	<i>K</i> -mers	Accuracy	Precision	Recall	F1 score
H3	7	0.7112	0.7214	0.7112	0.7069
H4	7	0.6753	0.6894	0.6753	0.6742
H3K14ac	10	0.8493	0.8461	0.8493	0.8495
H3K9ac	10	0.8297	0.8321	0.8297	0.8285
H4ac	11	0.8618	0.8625	0.8618	0.8615
H3K4me1	7	0.5792	0.5724	0.5792	0.565
H3K4me2	7	0.5913	0.5717	0.5913	0.5869
H3K4me3	7	0.4511	0.4433	0.4511	0.4471
H3K79me3	7	0.7153	0.7146	0.7153	0.7146
H3K36me3	7	0.6127	0.6087	0.6127	0.6103

imbalanced version in terms of most evaluation metrics. On balanced version of multi-label data set, Histone-Net achieves the F1-score of 72% and hamming loss of 0.20 which surpasses the Histone-Net performance achieved on imbalanced version by the figure of 3% and 1%.

Furthermore, Table 3 compares the performance of Histone-Net approach with adapted DeepHistone approach using imbalanced and balanced versions of multi-label histone occupancy and modification prediction data set. It is evident from Table 3 that Histone-Net significantly outperforms adapted DeepHistone approach across all 11 evaluation metrics. On im-balanced version of multi-label histone occupancy and modification prediction data set, Histone-Net achieves the accuracy increment of 14%, precision increment of 8%, recall increment of 21%, F1-score increment of 15%, average precision increment of 13%, AU-PRC increment of 18%, AU-ROC increment of 12%, hamming loss improvement of 7%, and coverage improvement of 2%. On the balanced version of multi-label histone occupancy and modification prediction data set, Histone-Net achieves the increment of 10%, 4%, 15%, 10%, 15%, 17%, 10%, 5%, and 2% in terms of aforementioned distinct evaluation metrics. On average, Histone-Net supersedes the performance of adapted DeepHistone by the figure 10% and 8% on im-balanced and balanced version of multi-label histone occupancy and modification prediction data set, respectively.

Furthermore, to analyze the effectiveness of proposed Histone-Net approach for accurately predicting most number of samples of each histone marker, we utilize one-versus-all strategy to generate 20 binary confusion matrices for 10 histone markers for imbalanced (Supplementary Figure 5) and balanced version of multi-label data set (Supplementary Figure 6). In one-versus-all strategy, false positives, false negatives, true negatives, and true positives, are computed by treating one particular histone marker class as positive and all other histone markers classes belonging to same histone sequence analysis task as negative irrespective of the multi-label problem. More specifically, we want to evaluate the behaviour of Histone-Net when there is a decent gap between the total number of positive and negative sequences.

A critical analysis of 20 confusion matrices (Supplementary Figure 5) produced by Histone-Net over imbalanced data set shows that overall 64% positive histone marker appearances and 82% negative histone marker appearances (represented as rest) are correctly predicted by Histone-Net. Top true positive figure of 89% is achieved on H3 histone marker, whereas top true negative figure of 94% is achieved on H3ac histone marker. Among histone markers related to occupancy, higher number of positive and negative sequences are correctly classified on H3 makers as compared to H4. Among histone markers related to acetylation, most number of positive class sequences are accurately classified in H3K14ac histone marker with the performance around

**Table 3** Performance produced by proposed Histone-Net and adapted DeepHistone using optimal higher order residues-based SuperDNA2Vec sequence representation over imbalanced and balanced version of multi-label data set in terms of 11 distinct evaluation metrics

Data sets	Data set	Accuracy	Precision	Recall	F1	Average precision	AU-PRC	AU-ROC	Ranking loss	OneError	Hamming loss	Coverage
Imbalanced data set	Histone-net	<b>0.5714</b>	<b>0.6882</b>	<b>0.7257</b>	<b>0.693</b>	<b>0.529</b>	<b>0.788</b>	<b>0.868</b>	0.3812	0.8556	<b>0.2049</b>	<b>15.8996</b>
	DeepHistone [31]	0.4269	0.6127	0.5245	0.5467	0.4014	0.6114	0.7502	<b>0.2295</b>	<b>0.8143</b>	0.2717	17.7378
Balanced data set	Histone-net	<b>0.6171</b>	<b>0.7113</b>	<b>0.7652</b>	<b>0.7299</b>	<b>0.5712</b>	<b>0.8244</b>	<b>0.8821</b>	0.3412	0.8324	<b>0.1977</b>	<b>15.7452</b>
	DeepHistone [31]	0.5235	0.6693	0.6164	0.6296	0.4088	0.6464	0.7842	<b>0.1958</b>	<b>0.8297</b>	0.2439	18.082

Higher performance figures are highlighted in bold



82%, whereas most number of negative class sequences are correctly predicted in H34ac histone marker with the performance around 95%. Turning towards the performance of methylation related histone markers, 80% positive class sequences are correctly classified in 3 histone markers (H3K36me3, H3K4me1, H3K4me3), whereas 88% negative class sequences are correctly predicted by Histone-Net on H3K79me3 histone marker.

In one-versus-all setting as negative class gets more number of samples which is why usually there exist a huge gap between the performance of positive and negative class; however, here, the gap is not large at all due to the robustness of Histone-Net approach towards imbalance class distribution.

In imbalanced version of multi-label data set, there exist only 5 uni-label and 24 bi-label sequences which are too little for effective model training. Considering uni-label and bi-label sequences act as a noise and derail the generalizability of classifier, we perform experimentation on balanced version of multi-label data set prepared after eliminating all uni-label and bi-label sequences. Performance analysis on 20 confusion matrices produced by Histone-Net on balanced version of multi-label data set (Supplementary Figure 6) reveals that overall 66% positive and 80% negative histone marker appearances are predicted accurately.

Across different histone markers, overall Histone-Net marks better performance on balanced version of multi-label data set as compared to imbalanced version of multi-label data set (Supplementary Figures 5 and 6). Highest true positive figure of 97% and true negative figure of 99% is achieved on H3 and H4ac histone markers, respectively, achieving an increment of 8% and 5% as compared to the peak performance achieved by Histone-Net on imbalanced version of data set. From histone markers related to occupancy, while most number of positive sequences are correctly predicted in H3 histone marker, higher number of negative sequences are correctly classified in H4 histone marker. Among histone markers related to acetylation, greater number of positive class sequences are accurately classified in H3K14ac histone marker with the performance around 86%, whereas most number of negative class sequences are correctly predicted in H34ac histone marker with the performance around 99%, outperforming the performance attained on imbalanced version by 4%. Concerning the performance of methylation related histone markers, 93% positive class sequences are correctly classified in H3K36me3 histone marker, whereas 89% negative class sequences are correctly predicted by Histone-Net on H3K4me2 histone marker, achieving an increment of 13% and 1%, respectively, when compared with top performance attained by Histone-Net on methylation histone marker of imbalanced data set.

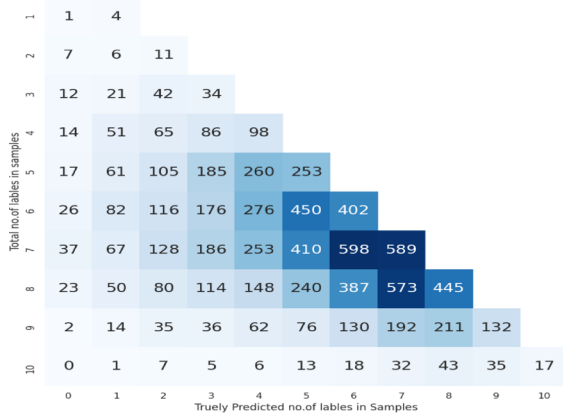
In a nutshell, across different evaluation metrics, although average performance figures attained by Histone-Net on imbalanced and balanced version of multi-label data set are

comparable. However, a close look indicates that across most histone markers, Histone-Net achieves better performance on balanced version of multi-label data set as compared to imbalanced version.

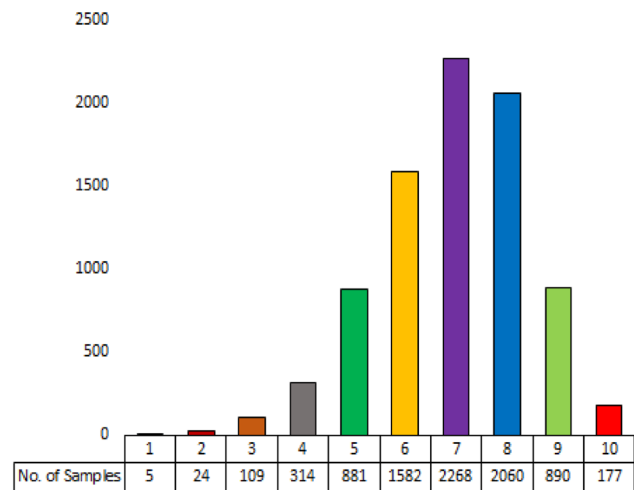
To identify up to what degree Histone-Net manages to simultaneously predict histone-occupancy, acetylation, and methylation areas in novel histone sequences, performance of Histone-Net is analyzed over imbalanced and balanced version of multi-label data set in terms of multi-label confusion matrices corresponding to unique sample-label distributions. In both versions multi-label data set, number of correctly predicted histone markers out of all actual histone markers are highlighted in confusion matrices (Fig. 13).

For imbalanced version of multi-label data set, a closer look at sample-label distribution (Fig. 13) and confusion matrix (Fig. 13) reveals that, Histone-Net manages to make accurate prediction for 90% uni-label sequences as only 1 sequence is miss-classified out of 5 sequences. For bi-label sequences, it correctly predicts 46% sequences, because 11 bi-label sequences are correctly classified out of 24 sequences. For tri-label sequences, Histone-Net performance drops further, where it only manages to identify the target histone markers of 31% sequences. For tetra-label sequences, Histone-Net achieves best performance around 83% as it makes correct predictions for 260 sequences out of 314 sequences. However, afterward, with the increase of histone marker combinations, Histone-Net best performance of 83% keeps on declining with great margin, dropping to 51%, 25%, 26%, and 22%, for penta, hexa, hepta, and octa-label sequences, respectively, achieving lowest performance 15%, 9% on highest label cardinalities including nona-label and deca-label sequences.

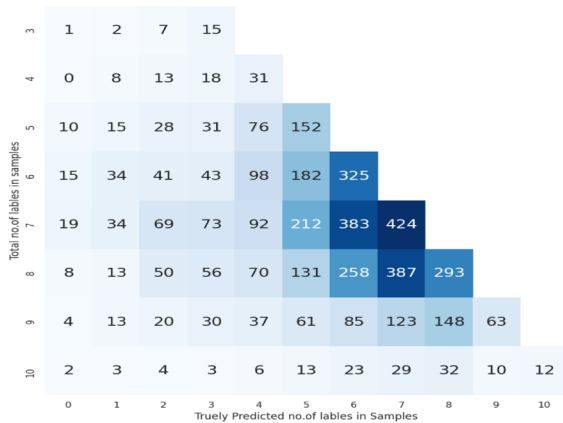
On the other hand, for balanced version of multi-label data set, uni and bi-label samples are removed from the data set and label cardinalities which have 25 or more samples are kept. For tri-label and tetra-label sequences, Histone-Net manages to correctly predict 60% and 26% sequences. Afterward, with the increase of label cardinality, unlike imbalanced data set, here Histone-Net performance decreases with less margin, it manages to correctly predict penta, hexa, hepta, and octa-label sequences with 49%, 44%, 32%, and 23% accuracy. However, for highest label cardinalities such as nona-label and deca-label sequences, Histone-Net once again only manages to correctly predict 11% and 9% sequences. Overall, Histone-Net achieves better performance on balanced version of multi-label data set. For balanced version of data set, Histone-Net achieves better performance with medium to higher level of histone marker combinations. Whereas, for imbalanced version, Histone-Net achieves better performance with low to medium level of histone marker combinations. However, for highest histone marker combinations (9 and 10), Histone-Net makes correct prediction for only 10% of sequences.



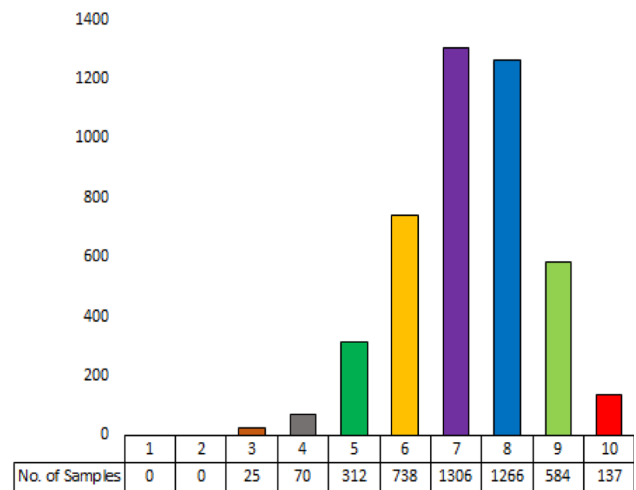
A. Multi-Label Confusion Matrix Imbalanced Dataset



B. Sample-Label Distribution of Imbalanced Dataset



C. Multi-Label Confusion Matrix of Balanced Dataset



D. Sample-Label Distribution of Balanced Dataset

**Fig. 13** Performance figures produced by Histone-Net over imbalanced and balanced version of multi-label data sets corresponding to unique sample-label distributions

To summarize, a comprehensive evaluation of Histone-Net in multi-label sequence classification paradigm using imbalanced and balanced data sets proves the capability of Histone-Net for simultaneously predicting histone type, occupancy, acetylation, and methylation levels in histone sequences. Furthermore, Histone-Net achieves decent performance on both imbalanced and balanced version of multi-label data set, showing its robustness to handle diverse data and sample-to-label distributions.

## An interactive and user-friendly Histone-Net web server

Histone-Net web server makes the lives of genomics researchers and practitioners easier by facilitating an interactive and user-friendly web interface capable to perform robust histone sequence analysis. Unlike other web server developed for biomedical sequence analysis which only supports inference on new sequences and even that for one

particular task. Histone-Net web server can be used to perform and visualize a multi-dimensional exploratory analysis of histone sequences. In addition, it can be used to train diverse predictive pipelines from scratch, tweak most crucial hyper-parameters, inference on new histone sequences for a variety of histone sequence analysis tasks including histone occupancy, acetylation, or methylation level prediction under binary and multi-label classification paradigm. Different modules of Histone-Net web server provide interactive session artifacts which can be downloaded and used for various purposes.

## Conclusions

Researchers have experimented with a variety of statistical representation learning approaches and strategies (from distributed representation to attention mechanism) to capture relatedness of residues, their diver interactions, and distribution among different classes. This paper develops unsupervised higher order residues embeddings of histone sequences using FastText model and explores a different application of FastText model to develop SuperDNA2Vec which encapsulates histone occupancy and modification information while learning higher order residues embeddings in a supervised manner. It presents a computational multi-paradigm framework Histone-Net to perform a comprehensive intrinsic and extrinsic evaluation of 2 differently learned embeddings using 3 machine learning classifiers. In addition, it develops a precisely deep neural network Histone-Net for robust histone occupancy, acetylation, and methylation prediction. A comprehensive empirical evaluation of Histone-Net in intra-domain and cross-domain settings under the hood of binary and multi-label classification paradigms proves its effectiveness over state-of-the-art, generalization potential across multiple histone markers, and aptitude to simultaneously predict histone type, its occupancy, acetylation, and methylation levels.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s40747-022-00802-w>.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Data Availability** To facilitate researchers and practitioners, Histone-Net multi-paradigm framework is deployed as an interactive and user-friendly web server at <https://histone.opendfki.de/>.

## Declarations

**Conflict of interest** Corresponding author, on the behalf of all authors declares that no conflict of interest is present.

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389(6648):251–260
- Wilson HJ (2005) Cells and tissues. In: Hench L, Jones J (eds) *Biomaterials, artificial organs and tissue engineering*. Elsevier, Amsterdam, pp 59–70
- Lin JC, Jeong S, Liang G, Takai D, Fatemi M, Tsai YC, Egger G, Gal-Yam EN, Jones PA (2007) Role of nucleosomal occupancy in the epigenetic silencing of the MLH1 CpG island. *Cancer Cell* 12(5):432–444
- Rodriguez J, Lee L, Lynch B, Tsukiyama T (2017) Nucleosome occupancy as a novel chromatin parameter for replication origin functions. *Genome Res* 27(2):269–277
- West JA, Cook A, Alver BH, Stadtfeld M, Deaton AM, Hochedlinger K, Park PJ, Tolstorukov MY, Kingston RE (2014) Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming. *Nat Commun* 5(1):1–12
- Lai WKM, Pugh BF (2017) Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nat Rev Mol Cell Biol* 18(9):548–562
- Larochelle S (2013) Dictating histone occupancy. *Nat Struct Mol Biol* 20(10):1145
- Zhou J, Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 12(10):931–934
- Liu Q, Xia F, Yin Q, Jiang R (2018) Chromatin accessibility prediction via a hybrid deep convolutional neural network. *Bioinformatics* 34(5):732–738
- Kelley DR, Snoek J, Rinn JL (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* 26(7):990–999
- Liu Q, Gan M, Jiang R (2017) A sequence-based method to predict the impact of regulatory variants using random forest. *BMC Syst Biol* 11(2):1–9
- Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL (2004) Global nucleosome occupancy in yeast. *Genome Biol* 5(9):1–11
- Bannister AJ, Kouzarides T (2011) Regulation of chromatin by histone modifications. *Cell Res* 21(3):381–395
- Brehove M, Wang T, North J, Luo Y, Dreher SJ, Shimko JC, Ottesen JJ, Luger K, Poirier MG (2015) Histone core phosphorylation regulates DNA accessibility. *J Biol Chem* 290(37):22612–22621

15. Cheung P, Allis CD, Sassone-Corsi P (2000) Signaling to chromatin through histone modifications. *Cell* 103(2):263–271
16. Binder H, Steiner L, Przybilla J, Rohlf T, Prohaska S, Galle J (2013) Transcriptional regulation by histone modifications: towards a theory of chromatin re-organization during stem cell differentiation. *Phys Biol* 10(2):026006
17. Kouzarides T (2007) Chromatin modifications and their function. *Cell* 128(4):693–705
18. Narlikar GJ, Fan H-Y, Kingston RE (2002) Cooperation between complexes that regulate chromatin structure and transcription. *Cell* 108(4):475–487
19. Kristeleit R, Stimson L, Workman P, Aherne W (2004) Histone modification enzymes: novel targets for cancer drugs. *Expert Opin Emerg Drugs* 9(1):135–154
20. Peterson CL, Laniel M-A (2004) Histones and histone modifications. *Curr Biol* 14(14):R546–R551
21. O'Geen H, Echipare L, Farnham PJ (2011) Using ChIP-seq technology to generate high-resolution profiles of histone modifications. In: Tollefsbol T (ed) *Epigenetics protocols*. Springer, Berlin, pp 265–286
22. Alhamwe BA, Khalaila R, Wolf J, von Bülow V, Harb H, Alhamdan F, Hii CS, Prescott SL, Ferrante A, Renz H et al (2018) Histone modifications and their role in epigenetics of atopy and allergic diseases. *Allergy Asthma Clin Immunol* 14(1):1–16
23. Chen ZJ, Pikaard CS (1997) Epigenetic silencing of RNA polymerase i transcription: a role for DNA methylation and histone modification in nucleolar dominance. *Genes Dev* 11(16):2124–2136
24. Kurdistani SK, Tavazoie S, Grunstein M (2004) Mapping global histone acetylation patterns to gene expression. *Cell* 117(6):721–733
25. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Alex Rolfe P, Herbolsheimer E et al (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* 122(4):517–527
26. Dorn ES, Cook JG (2011) Nucleosomes in the neighborhood: new roles for chromatin modifications in replication origin control. *Epigenetics* 6(5):552–559
27. Jayani RS, Ramanujam PL, Galande S (2010) Studying histone modifications and their genomic functions by employing chromatin immunoprecipitation and immunoblotting. *Methods Cell Biol* 98:35–56
28. Cusack M, King HW, Spingardi P, Kessler BM, Klose RJ, Kriacucionis S (2020) Distinct contributions of DNA methylation and histone acetylation to the genomic occupancy of transcription factors. *Genome Res* 30(10):1393–1406
29. Ye C, Tu BP (2018) Sink into the epigenome: histones as repositories that influence cellular metabolism. *Trends Endocrinol Metab* 29(9):626–637
30. Phaml TH, Tran DH, Ho TB, Satou K, Valiente G (2005) Qualitatively predicting acetylation and methylation areas in DNA sequences. *Genome Inform* 16(2):3–11
31. Yin Q, Wu M, Liu Q, Lv H, Jiang R (2019) DeepHistone: a deep learning approach to predicting histone modifications. *BMC Genom* 20(2):193
32. Baisya DR, Lonardi S (2020) Prediction of histone post-translational modifications using deep learning. *Bioinformatics* 36(24):5610–5617
33. Li J, Jie Z, Zuo L, Chang D (2018) Reveal the cognitive process of deep learning during identifying nucleosome occupancy and histone modification. In: 2018 Chinese automation congress (CAC). IEEE, pp 1856–1860. <https://tinyurl.com/2jmymxnm>
34. Higashihara M, Rebolledo-Mendez JD, Yamada Y, Satou K (2008) Application of a feature selection method to nucleosome data: accuracy improvement and comparison with other methods. *WSEAS Trans Biol Biomed* 5(5):95–104
35. Nguyen NG, Tran VA, Ngo DL, Phan D, Lumbanraja FR, Faisal MR, Abapihi B, Kubo M, Satou K et al (2016) DNA sequence classification by convolutional neural network. *J Biomed Sci Eng* 9(05):280
36. Yin B, Balvert M, Zambrano D, Schönhuth A, Bohte S (2018) An image representation based convolutional network for DNA classification. *arXiv preprint*. [arXiv:1806.04931](https://arxiv.org/abs/1806.04931)
37. Chiu B, Baker S (2020) Word embeddings for biomedical natural language processing: a survey. *Lang Linguist Compass* 14(12):e12402
38. Almeida AMG, Cerri R, Paraiso EC, Mantovani RG, Junior SB (2018) Applying multi-label techniques in emotion identification of short texts. *Neurocomputing* 320:35–46
39. Joulin A, Grave E, Bojanowski P, Mikolov T (2016) Bag of tricks for efficient text classification. *arXiv preprint*. [arXiv:1607.01759](https://arxiv.org/abs/1607.01759)
40. Asgari E, Mofrad MRK (2015) Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS one*. <https://doi.org/10.1371/journal.pone.0141287>
41. Le NQK, Yapp EKY, Ho Q-T, Nagasundaram N, Ou Y-Y, Yeh H-Y (2019) iEnhancer-5step: identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. *Anal Biochem* 571:53–61
42. Le NQK (2019) iN6-methylat (5-step): identifying DNA N6-methyladenine sites in rice genome using continuous bag of nucleobases via Chou's 5-step rule. *Mol Genet Genom* 294(5):1173–1182
43. Islam SMA, Heil BJ, Kearney CM, Baker EJ (2018) Protein classification using modified n-grams and skip-grams. *Bioinformatics* 34(9):1481–1487
44. Ding J, Zhou S, Guan J (2011) miRFam: an effective automatic miRNA classification method based on n-grams and a multiclass SVM. *BMC Bioinform* 12(1):1–11
45. Elattar M (2017) Classification of thermophilic and mesophilic proteins using N-grams. PhD thesis
46. El Boujnoui M (2020) A study and identification of COVID-19 viruses using n-grams with Naïve Bayes, k-nearest neighbors, artificial neural networks, decision tree and support vector machine. <https://tinyurl.com/ya83nv9y>
47. Le NQK, Yapp EKY, Nagasundaram N, Yeh H-Y (2019) Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous fast-text n-grams. *Front Bioeng Biotechnol* 7:305
48. Oberti M, Vaisman II (2017) Identification and prediction of intrinsically disordered regions in proteins using n-grams. In: Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics, pp 67–72. <https://tinyurl.com/ca4aknbv>
49. Chen Z, Zhao P, Li C, Li F, Xiang D, Chen Y-Z, Akutsu T, Daly RJ, Webb GI, Zhao Q et al (2021) iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Res* 49(10):e60–e60
50. Ng Patrick (2017) dna2vec: Consistent vector representations of variable-length k-mers. *arXiv preprint*. [arXiv:1701.06279](https://arxiv.org/abs/1701.06279)
51. Asim MN, Ibrahim MA, Malik MI, Dengel A, Ahmed S (2020) Enhancer-dsnet: a supervisedly prepared enriched sequence representation for the identification of enhancers and their strength. In: Yang H, Pasupa K, Leung AC, Kwok JT, Chan JH, King I (eds) *International conference on neural information processing*. Springer, Berlin, pp 38–48
52. Asim MN, Malik MI, Zehe C, Trygg J, Dengel A, Ahmed S (2020) MirLocPredictor: a ConvNet-based multi-label microRNA sub-cellular localization predictor by incorporating k-mer positional information. *Genes* 11(12):1475

53. Asim MN, Malik MI, Dengel A, Ahmed S (2020) K-mer neural embedding performance analysis using amino acid codons. In: IJCNN, pp 1–8. <https://tinyurl.com/2t63ht4y>
54. Pham TH, Ho TB, Tran DH, Satou K (2007) Prediction of histone modifications in DNA sequences. In: 2007 IEEE 7th international symposium on bioinformatics and bioengineering. IEEE, pp 959–966. <https://tinyurl.com/2pcsjw8e>
55. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
56. Chen KM, Cofer EM, Zhou J, Troyanskaya OG (2019) Selene: a PyTorch-based deep learning library for sequence data. *Nat Methods* 16(4):315–318
57. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y (eds) 3rd International conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference track proceedings. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980)
58. Linder J, Seelig G (2020) Fast differentiable DNA and protein sequence optimization for molecular design. arXiv preprint. [arXiv:2005.11275](https://arxiv.org/abs/2005.11275)
59. Shekar BH, Dagnev G (2019) Grid search-based hyperparameter tuning and classification of microarray cancer data. In: 2019 Second international conference on advanced computational and communication paradigms (ICACCP). IEEE, pp 1–8. <https://tinyurl.com/5y9bdthy>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.