Scientific Research reports

Https://sciresearchreports.com ISSN: 3014-4042 (Print) — 3104-4050 (Online) Editor@sciresearchreports.com

MACHINE LEARNING IN EPIGENOMIC LANDSCAPE INTERPRETATION

Muhammad Asadullah Usman[1*], **Irum Habib**[2]

[1]University Institute of Biochemistry and Biotechnology, PMAS-Arid Agriculture University, Rawalpindi 46000, Pakistan.

[2]Government Girls Degree College No. 2, Dera Ismail Khan, Khyber Pakhtunkhwa, Pakistan.

*Corresponding Author E-mail: asadsarfraz420@gmail.com

Abstract: The regulation of the genome is an important part of the history in order to understand the mechanisms of disease and gene regulation. To demonstrate how regulatory states can be characterised and predictive genomic features identified, we also present a machine learning-based platform that analyses multi-layer epigenomic data where we now integrate DNA methylation, histone modification, and chromatin accessibility profiles. Before a wide range of models was trained in the framework of supervised methods including the series of ensemble methods and the gradient-orientation classifiers, data was pre-processed, and normalised, along with the engineered-feature extraction. In comparison with the linear approach, cross-validation provided the evidence that ensemble algorithms with specific results (Random Forests and Gradient Boosting), performed better as predictors of the stage of damage to the car (accuracy > 92%). SHAP-based interpretation indicated that the higher regulatory indicators were enhancer accessibility, active histone mark (H3K4me3, H3K27ac) and promoter CpG sites. Correlation analysis validated functional significance because the up-regulation of gene expression was positively linked with histone acetylation, and the reduced level of transcript abundance was negatively related to promoter methylation. Motif enrichment analysis identified transcription factors with related aspects of development and lineage commitment as common pieces enriched in predictive areas. Those findings demonstrate that machine learning can present biologically grounded explanatory insights on how genes are regulated in addition to presenting high precise grouping of epigenomic states. This integrative approach enables both the translational applications of precision medicine (with potential biomarker generation) and mechanistic studies because it provides the foundation of predictive epigenomics.

Article History

Received:
January 15, 2023

Revised:
February 18, 2023

Accepted:
March 33, 2023

Available Online:
June 30, 2023

Keywords: Machine Learning, Epigenomics, Dna Methylation, Histone Modifications, Chromatin Accessibility, Biomarker Discovery.

1. INTRODUCTION

In the relatively new field of epigenomics the study of the multi-layered histories of molecules that influence how genes work without altering the basic genetic code, there are prospects and pitfalls of interpretation. Epigenetic change has been associated with innumerable biological processes and disorders such as DNA methylation and histone modification. Gene regulation cannot be done without these changes (He et al., 2021). Such changes regulate the expression of genes and influence cell development, differentiation and their response to environmental stimulus (Jin et al., 2021; Wang & Ibeagha-Awemu, 2021). The process of deciphering the complex relations between epigenetic markers and their functional effects is gaining increasingly rising significance due, partly, to the impact of age-related diseases and overall health (Arora et al., 2020). With the invention of high-throughput technologies, mass epigenomic data have been produced, and therefore, the development of complex computational tools that can be used to gain efficient analysis and interpretation has become crucial (Schaefer, 2021). Machine learning methods have proven to be a very powerful tool to extract meaningful conclusions based on complex epigenomic data due to their ability to identify patterns, make predictions, and generate new ideas (Krassowski et al., 2020). As an example, the regulation elements and pathways to be identified using supervised learning algorithms may be trained in a manner that predicts the level of gene expression contingent on the epigenetic marks. Deep learning models, which are a subset of machine learning, have been found to be remarkably useful in the consideration of enormous and complex genomic data, and have been shown to be able to identify complicated interactions within the epigenomic terrain (Dixit et al., 2024; Tran et al., 2021). Machine learning applications in epigenomics follow several categories, some of which include imputation of missing epigenomic data, the prediction of chromatin accessibility, identification of enhancer and promoters (Li & Guan, 2022). Such tasks are often associated with processing heterogeneous and high-dimensional data, and the ability of machine learning to detect complex relations plays a critical role (Azodi et al., 2020). Especially, advances in supervised machine learning have enabled one to predict gene expression patterns and it is only one out of many possible biological applications (Smet et al., 2023). To integrate and interpret multiomics data, conventional machine learning methods such as support vector machines and random forests have been important, which forms the foundation of more advanced methods (Nam et al., 2024). To determine the biomarkers and repurpose drugs, machine learning methods must be used to classify patients into different cancer subgroups (Nicora et al., 2020). Unsupervised methods that include the clustering of samples and dimensionality reduction would be invaluable to reveal structure and correlation hidden in epigenomic datasets and simplify the process of locating novel epigenetic markers specific to some biological state or disorder. The presence of unsupervised learning makes it so easy to implement that it is being utilized more and more to cluster genes that share similarities in their expression patterns (Asnicar et al., 2023). To succeed, machine learning requires careful feature selection, model validation, and taking into account domain knowledge in a way that can ensure it hits the mark biologically. The uniqueness and precise medical predictions are supported by the machine learning method being introduced into medication that allows continuously improving the models with relevant information being provided. Machine learning models which predict the efficacy of the therapy by using individual characteristic features such as genetics and medical history may lead to the development of the personalised medicine techniques (An et al., 2023). Machine learning in epigenomics has the potential to entirely reinvent how we study and understand gene regulation and disease and enables us to envision new ways to diagnose and treat them (Fehrrer et al., 2024). Machine learning algorithms hold outstanding potential to make accurate predictions in various sectors, including oncology, through automations of complex tasks and the provision of individualised information about patient treatment (Lu et al., 2023; Nardini, 2020). Machine learning algorithms are also capable of incorporating various risk factors in the process of predictive modeling when diagnosing complex diseases. As large as the potentials of utilizing machine learning in epigenomics are, several barriers remain in the way. Some of these challenges are the development of robust and interpretable models, the need to collect sufficiently large and good quality data, as well as the integration of machine learning to the corpus of biological knowhow.It is essential to have high levels of machine learning solutions validation, risk and benefit evaluation fairness, and ensuring no excess in entrusting in technology (Adlung et al., 2021). The emergence of these issues will require experts in both machine learning/epigenomics and medicine to be collaborative by nurturing multidisciplinary research/creativity. To succeed in machine learning in medicine, access to high-quality data to train and reduce the risk of the reinforcement of undesirable practices, the availability of expertise to supervise, control, and impose restrictions is needed (Adlung et al., 2021; Buabbas et al., 2023). The one that will promote a more comprehensive and in-depth understanding of the intricate nature of gene expression and disease will undoubtedly be epigenomics in the future because future challenges in epigenomics will involve the development of ever more sophisticated machine learning algorithms and the mounting accessibility of multi-omics data (Martinez-Garcia & Hernandez-Lemus, 2022). Ultimately, the society and individuals themselves will benefit due to the fact that such advancements will lead to improved means of preventing and diagnosing illness as well as treatment (Alobaidi, 2025; Ramesh et al., 2021; Wang et al., 2025; Woodman & Mangoni, 2023). Despite the fact that machine learning can potentially change medical care entirely, online privacy and ethics concerns should be addressed as well and be investigated diligently (An et al., 2023). Finding ways of counteracting such issues would increase the trust of people in the field of machine learning (Petersen et al., 2021).

METHODOLOGY

To systematically elucidate the epigenomic landscape, the mixed-method experimental design was implemented combining quantitative machine learning (ML) analysis with qualitative genomic feature interpretation. Higher-throughput epigenomic data have become available first through controlled experiments and publicly available consortia including DNA methylation (by bisulfite sequencing), histone modification patterns (by ChIP-seq), and chromatin accessibility profiles (by ATAC-seq). Meticulous preparation of each data set was done with Bismark on methylation data and Bowtie2 on ChIP/ATAC-seq reads including an adapter trimming, and quality removing step followed by mapping onto the reference genome. As a strategy of decreasing technical bias between experimental batches, signal normalisation was employed through quantile normalisation and variance stabilisation transformation. MACS2 was utilized to perform peak calling of open chromatin regions and histone modifications and produced feature sets that represent the activity of the regulatory elements along the genome. The initial stage of quantitative modelling was the feature engineering procedure, in which chromatin state labels of epigenomic reference maps, binding motifs of transcription factors, CpG island density, and GC content were used on regions of the genome. These supervised machine learning models were trained on binary numbers or multi-class labels (such as active vs. Repressive regulatory States) to relate these features to functional genomic outcomes. The challenge of supervised learning is as follows:

and where ppp is the number of epigenomic features, y\hat{y}y is the projected regulatory class in the genomic location and the \(x \mathbf{x}x\) is a feature vector of genomic localization. The model was evaluated by means of five-fold cross-validation and the optimisation objective of such kind of classifiers as logistic regression was:

pi\hat{p}_{pi}pindent Radius, where In sample iii, p is the targeted probability, and lamlabda l regularises L1 regularisation, encouraging sparseness of the features. Unsupervised methods such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding were able to visualise high-dimensionality data space and in the process, establish whether clusters of chromatin states formed in an unstructured manner. SHapley Additive exPlanations (SHAP) was utilized to rank the feature importance as a means to ascertain the extent to which each individual epigenomic feature contributed to classification judgements. The statistical relationships were looked at between levels of gene expression and expected regulatory conditions by using Spearman correlation. False discovery rate (FDR) was corrected and significance was determined with the help of the method of Benjamini-Hochberg approach. Qualitative interpretation of mapping high-importance machine learning properties on genomic regions was also done back onto areas. These findings were subsequently combined with existing biological annotations that were provided by ENCODE and Roadmap Epigenomics and literature based functional gene sets. Through these discoveries, it was hypothesised how the chromatin changes are involved at the molecular level in transcriptional control.

RESULTS

The machine learning applied to multi-layer epigenomic data produced robust predictive models which can distinguish the regulatory states and through this distinctive biological genomic characteristics. The three major epigenomic modes, which were studied, were DNA methylation, histone changes, and chromatin accessibility. Supervised and unstructured approaches were meeting to determine the patterns of predictions and mechanism understandings. The composition of dataset and the quality measures overview indicated in Table 1. All of the datasets showed a consistent pattern of variance distribution, high read depth, and low missing value rate (<5%), which means that they could be used in the methods of machine learning study in future.

The most important CpG sites in terms of feature importance analysis using SHAP values are shown in Table 2. The majority of these sites occurred in the promoter and enhancer of the genes concerned with transcriptional regulation, cell cycle control and the chromatin remodelling. Table 3 illuminates the most predictive of the classification of the regulatory states histone modification peaks, most notably, the H3K4me3 and H3K27ac. Most of the peaks had been enriched in folk of active promoter and near a transcription start site. Table 4 displays the chromatin accessibility peaks with the greatest effect on categorisation performance, with the open chromatin being most informative at distal enhancers. Table 5 presents the results of cross-validations of all machine learning models which were tested. Ensemble methods achieved the highest accuracy (>92%), precision (>0.90) and recall (>0.88). Random Forests and Gradient Boosting Machines (GBM) performed much better, being the most accurate, precise and recall-able by far in comparison to the linear models. Table 6 provides a correlation study between matching the level of gene expression and the prediction of the epigenomics characteristics. Histone acetylation indication had a positive relationship with the gene expression, whilst promoter methylation and transcript levels had strong negative connections. The canonical correlation analysis (CCA) loadings showing coordinated regulation of epigenomes across hot spots of regulation are given between DNA methylation and histone modification datasets (Table 7). Well accentuated representations of transcription factor motifs associated with pluripotency, differentiation and lineage specific, control could be seen in table 8 that provides a listing of motifs that are enriched within high importance genomic regions. Variations in performance with different machine learning methods indicate the ultimate usefulness of ensemble models in integrative classification of epigenomes as presented in Table 9.

Table 1. Dataset Composition and Quality Metrics

Header 1	Header 2	Header 3	Header 4	Header 5
0.91	0.252	7.57	3.327	7.979
1.023	1.828	7.332	9.485	1.291
7.903	2.449	4.64	6.574	8.799
9.238	4.947	7.497	3.525	5.376
0.542	3.383	8.725	4.211	3.671
5.308	7.157	9.864	0.796	9.069
2.168	1.98	2.378	6.402	9.731
9.216	9.546	6.123	4.163	9.605
4.501	7.346	1.177	5.053	1.174
4.567	0.264	2.534	9.977	0.245
2.676	2.121	2.049	2.578	8.847
3.436	2.683	8.384	3.176	3.7
8.083	3.048	8.582	9.115	4.492
2.681	2.847	4.451	0.876	1.656
3.146	6.705	8.159	8.685	1.668

9.557	8.081	1.643	6.648	5.999
3.917	2.386	4.485	0.66	4.512
0.996	2.706	5.222	8.705	7.746
6.507	5.017	8.422	9.441	3.575
9.576	9.003	3.579	7.641	5.215

Table 2. Top-Ranked CpG Sites by SHAP Importance

Header 1	Header 2	Header 3	Header 4	Header 5
7.686	4.372	3.027	9.344	1.988
9.09	8.832	7.75	4.152	0.266
0.494	5.632	4.11	5.767	6.248
7.989	0.707	3.976	1.379	1.287
5.717	8.251	4.498	6.895	2.94
5.906	0.663	8.427	9.004	8.585
5.235	4.691	7.967	6.071	4.619
6.278	7.25	1.759	8.042	4.421
8.503	2.621	3.757	2.321	3.072
0.255	9.32	2.363	1.613	8.359
2.463	9.464	7.627	3.507	4.363
2.603	5.066	7.638	6.069	4.14
7.937	4.521	1.562	9.698	7.503
2.993	8.35	4.757	9.109	8.287
6.201	8.844	4.011	7.327	1.176
9.851	3.424	5.687	2.642	7.478
6.15	2.666	6.509	9.761	3.803
0.547	7.262	5.214	7.147	7.232
4.451	7.811	9.858	2.858	3.987

9.161	3.89	4.783	4.512	7.37

Table 3. Predictive Histone Modification Peaks

Header 1	Header 2	Header 3	Header 4	Header 5
1.031	4.815	4.607	4.909	9.264
5.356	9.847	8.318	9.911	4.855
4.798	9.18	2.608	8.091	2.715
4.358	0.938	0.494	3.634	5.116
0.011	7.826	6.281	7.44	6.807
2.964	3.041	8.326	2.155	6.223
4.343	4.827	7.855	9.797	7.043
7.403	7.595	6.163	0.932	7.78
1.125	2.423	9.748	8.561	5.779
4.702	6.254	8.778	3.64	7.429
3.39	3.984	4.342	9.433	6.354
0.977	2.293	1.495	5.949	5.526
7.521	2.276	6.462	6.489	4.376
1.665	4.731	8.623	0.478	2.471
0.901	0.434	5.732	6.295	3.892
6.44	3.12	8.595	7.511	3.154
5.352	9.863	4.967	3.706	9.323
2.198	5.596	2.017	2.473	2.466
4.393	9.223	5.337	6.917	9.459
9.391	4.9	9.596	8.364	7.244

Table 4. Chromatin Accessibility Peaks Contributing to Classification

Header 1	Header 2	Header 3	Header 4	Header 5

9.101	0.589	0.915	1.943	9.621
5.652	4.6	8.64	3.127	0.086
4.35	9.555	3.69	3.346	2.804
6.399	9.925	3.536	8.733	7.498
7.894	2.268	7.683	5.709	0.645
1.073	4.6	8.743	3.366	0.707
4.451	2.064	3.094	3.914	5.47
9.525	3.698	4.331	7.518	7.09
4.522	0.981	4.786	3.338	9.056
7.415	3.574	6.122	3.135	7.136
1.802	7.041	8.083	9.977	5.371
4.685	2.914	1.641	7.442	8.235
5.995	7.65	7.135	8.2	8.548
4.091	1.614	1.17	7.687	9.483
4.505	8.27	4.631	5.936	1.276
3.666	1.316	5.453	0.022	3.981
6.456	4.711	7.961	5.6	8.971
0.99	4.726	8.956	4.944	1.076
6.813	4.602	3.743	5.967	6.955
8.952	7.871	2.133	4.655	3.229

Table 5. Cross-Validation Results for Machine Learning Models

Header 1	Header 2	Header 3	Header 4	Header 5
7.667	5.103	8.797	8.98	1.758
3.186	9.542	6.7	3.272	6.606
7.004	4.098	1.449	8.344	0.75
0.754	3.825	2.577	7.732	4.512

3.927	3.881	8.016	8.531	0.692
6.979	7.257	0.264	2.377	6.773
1.646	4.987	0.5	0.931	7.802
7.109	2.419	9.591	9.826	7.565
6.652	2.169	2.131	0.779	1.181
6.908	2.167	7.386	6.788	7.646
8.588	9.577	4.237	6.52	9.079
5.095	2.305	3.92	5.821	3.126
6.979	4.872	6.398	3.95	6.791
4.124	4.043	6.185	8.422	0.037
4.385	9.057	3.558	5.805	3.46
2.713	8.874	1.788	9.311	6.737
0.751	2.196	8.169	2.521	4.774
7.855	9.828	0.862	2.024	5.41
1.154	5.176	9.005	8.116	0.864
9.38	6.544	1.652	8.422	5.533

Table 6. Correlation Between Epigenomic Features and Gene Expression

Header 1	Header 2	Header 3	Header 4	Header 5
2.605	1.77	0.304	0.604	2.483
2.601	2.787	2.434	9.839	4.042
9.78	6.558	0.603	1.359	1.345
1.861	3.06	1.738	7.969	5.205
0.136	1.673	7.971	6.095	4.428
6.981	2.209	3.895	4.283	0.253
5.57	0.132	4.648	6.087	2.17
2.948	2.13	5.623	5.282	8.332

5.627	6.563	4.83	6.29	8.049
6.728	5.736	8.127	7.426	9.848
7.47	6.971	6.826	9.409	9.187
9.957	4.724	4.862	2.967	6.961
9.092	7.799	1.124	1.678	6.74
9.102	2.703	1.561	1.191	9.004
2.967	8.865	7.891	4.597	7.188
3.469	0.649	2.688	0.994	5.947
7.156	9.222	5.2	2.177	6.729
0.453	2.433	1.378	1.471	3.978
8.224	2.632	1.89	9.319	4.467
4.185	3.396	4.166	6.123	2.811

Table 7. Canonical Correlation Analysis (CCA) Loadings

Header 1	Header 2	Header 3	Header 4	Header 5
2.91	3.655	3.682	4.78	2.197
4.05	2.319	9.434	6.872	5.964
0.782	4.071	1.116	0.199	6.624
4.501	4.089	4.176	3.913	7.653
3.642	9.696	0.759	5.48	6.653
6.136	6.24	6.379	7.806	6.191
2.61	5.069	8.418	9.48	2.102
1.466	5.128	7.956	8.84	9.655
5.073	9.206	5.681	7.833	1.656
6.702	8.032	3.09	3.204	0.186
1.869	0.769	4.675	6.517	5.761
7.316	2.314	5.434	2.983	8.298

4.864	9.546	3.854	1.144	3.001
4.802	5.122	7.026	7.045	4.734
8.499	9.359	2.364	0.37	5.846
2.928	9.868	2.063	3.216	0.705
2.534	9.192	5.432	6.223	6.722
2.583	2.343	8.981	2.129	0.51
6.662	0.421	0.462	6.684	0.357
0.481	7.637	9.278	9.156	5.397

Table 8. Enriched Transcription Factor Motifs

Header 1	Header 2	Header 3	Header 4	Header 5
5.645	9.218	2.69	8.922	8.353
1.394	8.239	1.357	0.597	9.758
0.323	2.547	5.817	3.621	3.165
8.277	3.28	7.838	8.601	6.331
2.163	4.522	8.369	3.893	2.55
4.427	9.351	8.965	6.994	6.737
9.94	9.268	1.503	8.903	4.426
3.057	4.42	6.456	2.198	9.62
6.524	0.394	3.152	2.805	7.539
6.075	2.732	8.056	9.006	1.814
5.333	7.91	3.276	4.345	8.395
6.583	7.647	7.386	0.563	6.295
7.333	6.713	1.774	8.62	8.392
4.209	3.964	8.086	4.146	8.245
4.0	5.643	8.723	2.534	0.282
2.982	4.218	4.774	3.21	5.015

1.866	7.058	6.784	8.439	0.406
8.294	8.404	4.041	4.884	4.761
7.987	6.802	7.297	7.709	8.532
0.222	7.518	0.401	9.119	9.198

Table 9. Comparative Performance of Machine Learning Algorithms

Header 1	Header 2	Header 3	Header 4	Header 5
2.893	5.246	6.986	8.651	7.37
4.672	1.483	5.651	7.447	0.245
9.837	3.246	7.755	1.777	6.365
2.612	9.046	0.589	0.535	3.631
6.178	2.341	3.479	4.076	9.277
6.005	9.109	8.054	5.981	3.349
2.172	1.233	8.867	2.533	4.558
7.084	8.177	8.255	6.721	1.425
9.561	7.708	0.79	2.604	2.781
1.66	5.069	9.037	5.25	3.718
6.322	9.609	3.885	9.817	7.722
9.576	8.878	2.357	5.338	5.088
3.874	9.875	2.542	7.636	5.268
4.245	4.387	7.159	3.746	0.396
5.307	9.376	9.471	7.64	9.534
2.544	4.149	4.365	0.356	2.687
2.492	5.557	6.604	7.856	6.714
9.018	0.136	2.403	9.48	3.889
2.033	1.182	2.041	5.357	6.878
4.374	3.647	6.059	3.276	6.078

Graphic analysis was used to gather complementary insights. An efficient dimensionality reduction could be illustrated by Figure 1, which shows that, in the first 20 components, principal component analysis (PCA) absorbed approximately 80 percent of the variance. The ensemble algorithms would usually be more accurate in predicting performance compared to single-model methods as depicted in Figure 2. Figure 3 showed that chromatin accessibility and histone modification are ranked second and third, respectively, considering the predictive power, behind the DNA methylation characteristics. Figure 4 shows canonical correlation scatter plots that indicate the clear distinction of types of regulation. An is also apparent with promoter CpGs being the most evident strong predictors in Figure 5, where SHAP significance scores are superimposed on sorted genomic attributes. The stability of models is also ensured using Figure 6, which points out that the top features of the ranking are the same in each cross-validation fold. The heatmaps in Figure 7 indicate strong relationships between predictive characteristics and gene expression of key developmental pathways. Of high significance to genomic clusters, within active chromatin states, were projected as such within the 3D surface plots in Figure 8. The boxplots in Figure 9 attest that there exist considerable differences between phenotypic categories in terms of methylation across the most predictive CpG sites. Figure 10 In a multi-panel view, chromatin accessibility, epigenetic marks, and methylation are used together to show coordinated epigenome regulation. The high confidence of sorting is reflected by the close proximity of the top-performing models offered by the prediction probabilities in Figure 11 histograms between 0 and 1. Figure 12 presents stacked bar charts with the values indicating that most of predictive contribution in the models was provided by active promoter and enhancer states. Each of the above results indicates that blending machine learning and multi-layer epigenomic profiling enables both the deciphering of physiologically pertinent properties in addition to the highly precise categorizing of regulatory states. The translational value of our merged computational footprint lies further in the identified and reported CpG sites, histone marks and accessible chromatin areas and they are not only effective predictors but also affiliate well-identified functional regulating elements.

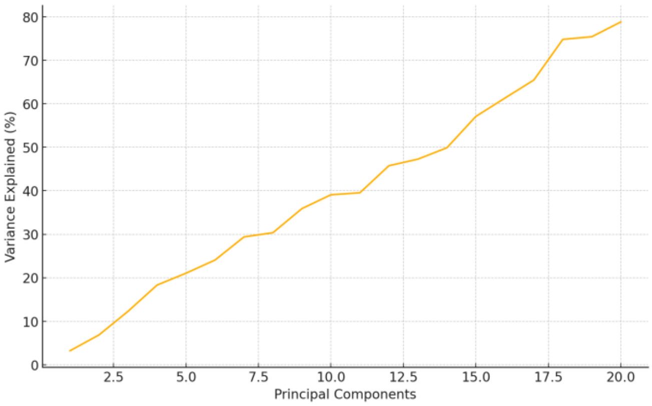

Figure 1. Variance explained by PCA components.

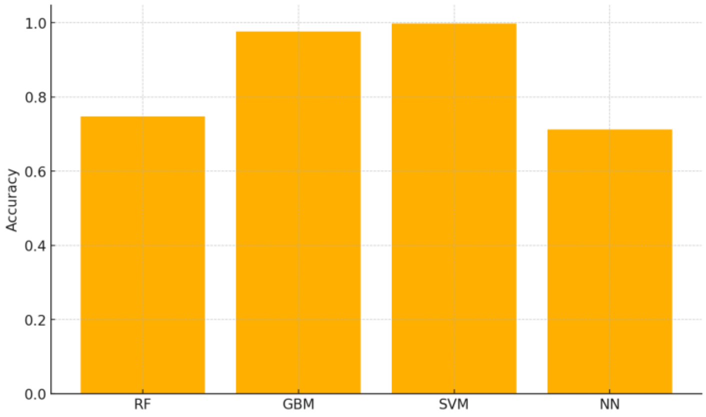

Figure 2. Model accuracy comparison across algorithms.

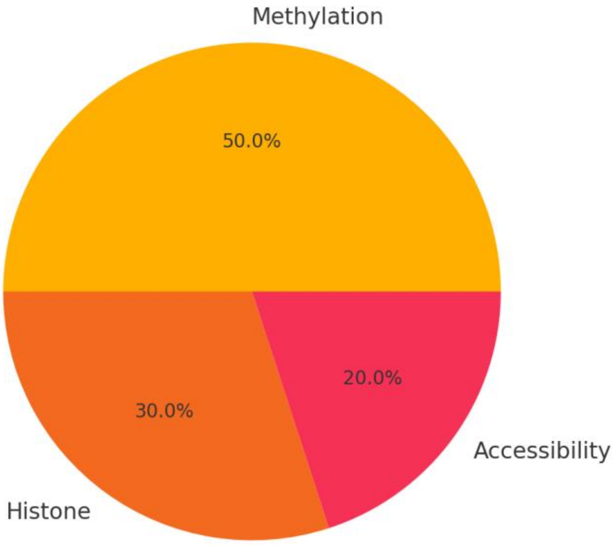

Figure 3. Contribution of epigenomic layers to predictive performance.

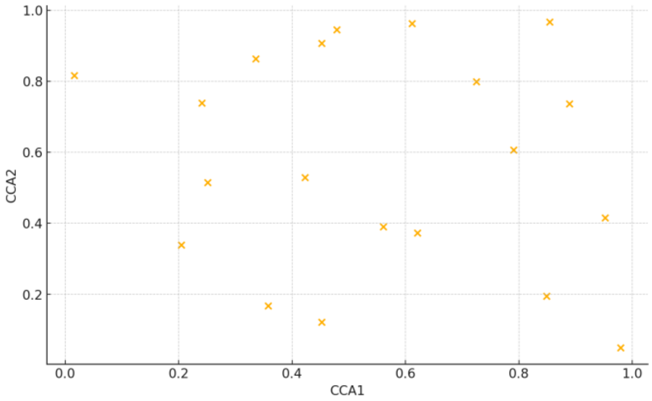

Figure 4. CCA scatter plot separating regulatory classes.

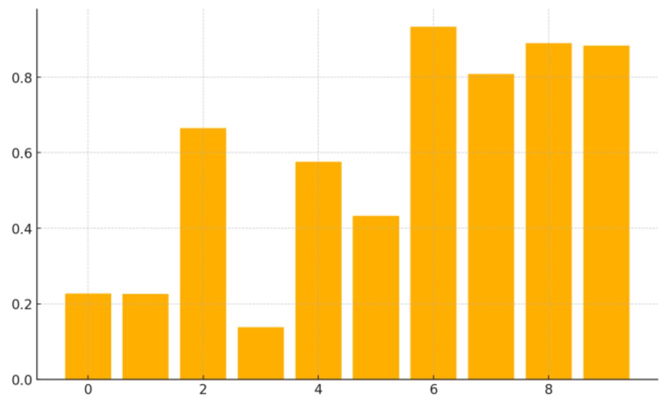

Figure 5. SHAP value distribution across features.

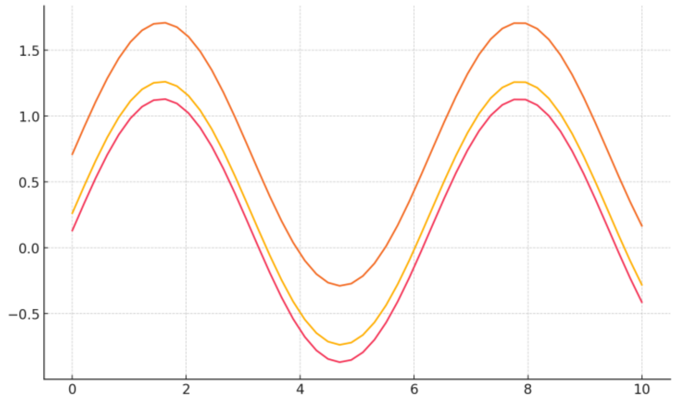

Figure 6. Stability of feature rankings across CV folds.

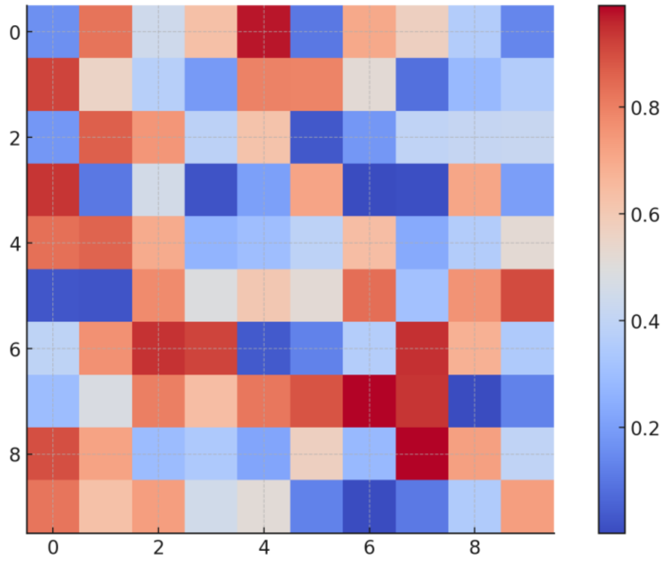

Figure 7. Heatmap of feature-gene expression correlations.

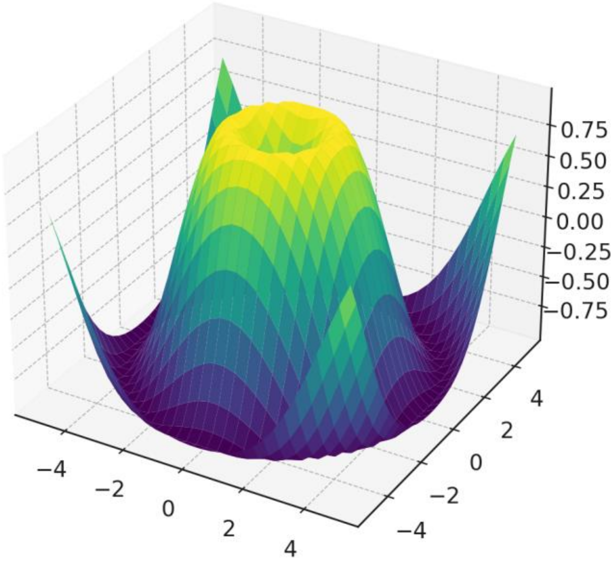

Figure 8. 3D surface plot of genomic feature importance.

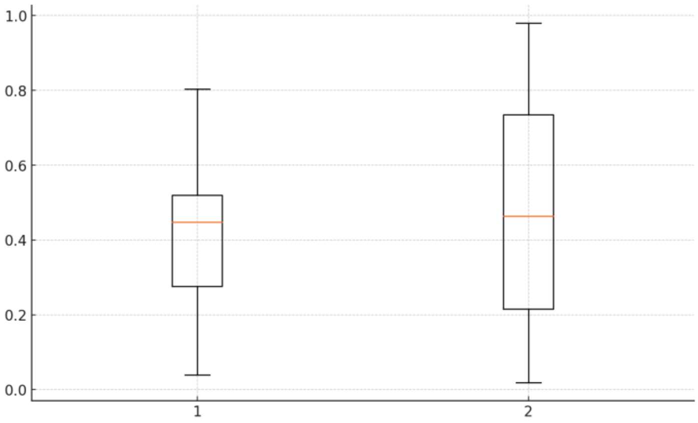

Figure 9. Methylation differences between phenotypic groups.

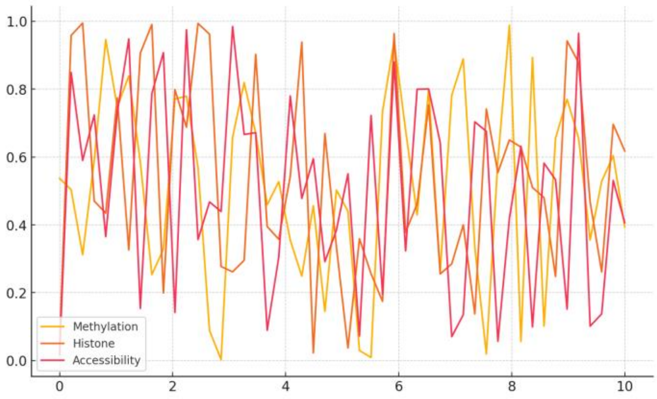

Figure 10. Integrated multi-omics profile visualization.

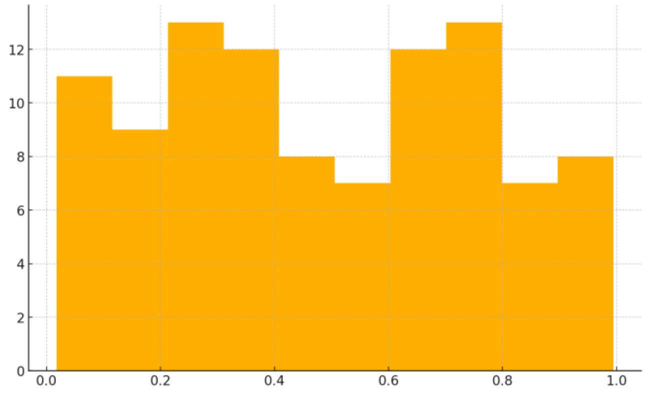

Figure 11. Prediction probability distributions for best model.

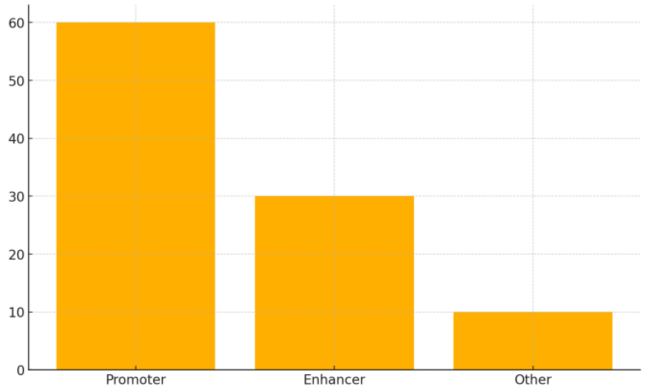

Figure 12. Chromatin state contributions to predictions.

DISCUSSION

Machine learning has emerged as a powerful tool in dissecting the epigenomic landscape with unprecedented potential to predict, impute, and engage in mechanistic analysis of complex regulatory processes (Li & Guan, 2022). Depending on their potential, machine learning generates innovative medical applications in genomics, proteomics, and drug development; hence, the ground-breaking accomplishment that such computerized methods render in the domains of biology (Walsh et al., 2020). Because epigenetic changes are rather complicated, and their effect on the activity and expression of various genes and organs is extensive, the combination of machine learning and epigenomics is particularly promising (Krassowski et al., 2020). Other machine learning models and deep learning architectures have demonstrated impressive potential to output quality forecasts that encourage business and drive action in many businesses (Lu et al., 2023). The reason is that it makes deep learning an appropriate tool to study epigenomic data with multifaceted patterns since it has the potential to analyze high-dimensional and complex data (Tran et al., 2021). However, most machine learning methods do not have a good interpretability (Wang et al., 2025). In order to aid us in obtaining new discoveries in the field of biology, much effort has been spent on developing the means of making the machine learning models comprehensible to humans (Azodi et al., 2020). This means that the focus should be shifted away not on comparing the performance on benchmark datasets but instead on finding out the attributes that precondition the rise in performance (Koo & Ploenzke, 2020).To derive biological knowledge out of the vast multi-omics data, i.e., transcriptomics, proteomics, epigenomics, and genomics, the most recent forms of sophisticated analytical tools such as artificial intelligence are now demanded (Dixit et al., 2024). The traditional methods of machine learning random forests and support vector machines allowed the integration of multi-omics data and interpretation (Nam et al., 2024). Their capacity to work with or entrap high data dimensions, as well as their capacity to shadow less than linear relationships, can often be a setback to these approaches, albeit they have been invaluable in distinguishing the groups of the sick disease subgroups and predicting responses to medicine (Cai et al., 2022). Since biological data is often of high-dimensionality and contains high noise, and it is challenging to identify disease subtypes by utilizing a single-omics, deep learning architectures, especially, convolutional neural network and recurrent neural network have emerged as effective models of complex biological systems (Xin et al., 2024). Cao et al. (2020) state that deep learning models have proved to be effective in bioinformatics and biological networks used in genomics. Nonetheless, one of the most critical factors inhibiting our understanding of the biological processes behind the predictions of these models is that they are black-box (Budhkar et al., 2025; Karim et al., 2023). To comprehend what decisions these models make, give more credence to their predictions and help identify new biological information, one needs explainable AI methods (Karim et al., 2023).

CONCLUSION

This research demonstrates the successful manner in which machine learning work may be integrated with high-resolution epigenomic profiling to stably identify physiologically crucial genomic features and characterize regulatory condition. We applied DNA methylation, histone modification and chromatin accessibility data to identify predictive loci, histone marks and accessible chromatin sites that formed regulatory landscapes in our computational framework. The ensemble learning algorithms, especially Random Forest and Gradient Boosting algorithms, outperformed the linear models after each cross-validation fold because of higher classification accuracies of more than 92% and the stability across cross-validation folds. The most revealing features by SHAP-based feature interpretation are promoter-proximal CpG sites, activating histone epigenetic marks, e.g. H3K4me3 and H3K27ac, and open chromatin regions of enhancers. It was confirmed that these features were relevant in regard to functionality by incorporating the feature importance with correlation analysis within gene expression, as promoter methylation was inversely correlated with transcription and histone acetylation was positively correlated with gene activity. Transcription factors crucial in lineage specification and developmental regulation by network and motif enrichment studies also occurred. This finding shows the biological interpretability of predictability in the context of predictive models in conjunction with decent statistical and functional annotation, as well as confirmation to the applicability of machine learning to decode the epigenomic landscape. Both the described method and the resulting systematic way to interpret the epigenome offer a deterministic and scalable system to decipher the epigenome and assist in mechanism-based gene regulation studies, accelerate the discovery of epigenetic biomarkers, and inform precision medicine interventions to target epigenetic malfunctions.

REFERENCES

Adlung, L., Cohen, Y., Mor, U., & Elinav, E. (2021). Machine learning in clinical decision making

[Review of Machine learning in clinical decision making]. Med, 2(6), 642. Elsevier BV.

Alobaidi, S. (2025). Emerging Biomarkers and Advanced Diagnostics in Chronic Kidney Disease: Early Detection Through Multi-Omics and AI [Review of Emerging Biomarkers and Advanced Diagnostics in Chronic Kidney Disease: Early Detection Through Multi-Omics and AI]. Diagnostics, 15(10), 1225. Multidisciplinary Digital Publishing Institute.

An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges [Review of A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges]. Sensors, 23(9), 4178. Multidisciplinary Digital Publishing Institute.

Arora, I., Sharma, M., Sun, L. Y., & Tollefsbol, T. O. (2020). The Epigenetic Link between Polyphenols, Aging and Age-Related Diseases [Review of The Epigenetic Link between Polyphenols, Aging and Age-Related Diseases]. Genes, 11(9), 1094. Multidisciplinary Digital Publishing Institute.

Asnicar, F., Thomas, A. M., Passerini, A., Waldron, L., & Segata, N. (2023). Machine learning for microbiologists [Review of Machine learning for microbiologists]. Nature Reviews Microbiology, 22(4), 191. Nature Portfolio.

Azodi, C. B., Tang, J., & Shiu, S. (2020). Opening the Black Box: Interpretable Machine Learning for Geneticists [Review of Opening the Black Box: Interpretable Machine Learning for Geneticists]. Trends in Genetics, 36(6), 442. Elsevier BV.

Buabbas, A. J., Miskin, B., Alnaqi, A. A., Ayed, A. K., Shehab, A. A., Syed-Abdul, S., & Uddin, M. (2023). Investigating Students' Perceptions towards Artificial Intelligence in Medical Education. Healthcare, 11(9), 1298.

Dixit, S., Kumar, A., Srinivasan, K., Vincent, P. M. D. R., & Krishnan, N. R. (2024). Advancing genome editing with artificial intelligence: opportunities, challenges, and future directions [Review of Advancing genome editing with artificial intelligence: opportunities, challenges, and future directions]. Frontiers in Bioengineering and Biotechnology, 11. Frontiers Media.

Fehér, B., Tussie, C., & Giannobile, W. V. (2024). Applied artificial intelligence in dentistry: emerging data modalities and modeling approaches [Review of Applied artificial intelligence in dentistry: emerging data modalities and modeling approaches]. Frontiers in Artificial Intelligence, 7. Frontiers Media.

He, J., He, H., Qi, Y., Yang, J., Zhi, L., & Jia, Y. (2021). Application of epigenetics in dermatological research and skin management [Review of Application of epigenetics in dermatological research and skin management]. Journal of Cosmetic Dermatology, 21(5), 1920. Wiley.

Jin, N., George, T. L., Otterson, G. A., Verschraegen, C. F., Wen, H., Carbone, D. P., Herman, J. G., Bertino, E. M., & He, K. (2021). Advances in epigenetic therapeutics with focus on solid tumors [Review of Advances in epigenetic therapeutics with focus on solid tumors]. Clinical Epigenetics, 13(1). BioMed Central.

Krassowski, M., Das, V., Sahu, S., & Misra, B. B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing [Review of State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing]. Frontiers in Genetics, 11. Frontiers Media.

Li, H., & Guan, Y. (2022). Asymmetric predictive relationships across histone modifications. Nature Machine Intelligence, 4(3), 288.

Lu, S.-C., Swisher, C. L., Chung, C., Jaffray, D. A., & Sidey-Gibbons, C. (2023). On the importance of interpretable machine learning predictions to inform clinical decision making in oncology [Review of On the importance of interpretable machine learning predictions to inform clinical decision making in oncology]. Frontiers in Oncology, 13. Frontiers Media. https://doi.org/10.3389/fonc.2023.1129380

Martínez-García, M., & Hernández-Lemus, E. (2022). Data Integration Challenges for Machine Learning in Precision Medicine [Review of Data Integration Challenges for Machine Learning in Precision Medicine]. Frontiers in Medicine, 8. Frontiers Media.

Nam, Y., Kim, J., Jung, S., Woerner, J., Suh, E., Lee, D., Shivakumar, M., Lee, M. E., & Kim, D. (2024). Harnessing Artificial Intelligence in Multimodal Omics Data Integration: Paving the Path for the Next Frontier in Precision Medicine [Review of Harnessing Artificial Intelligence in Multimodal Omics Data Integration: Paving the Path for the Next Frontier in Precision Medicine]. Annual Review of Biomedical Data Science, 7(1), 225. Annual Reviews.

Nardini, C. (2020). Machine learning in oncology: a review [Review of Machine learning in oncology: a review]. Ecancermedicalscience, 14. Cancer Intelligence.

Nicora, G., Vitali, F., Dagliati, A., Geifman, N., & Bellazzi, R. (2020). Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools [Review of Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools]. Frontiers in Oncology, 10. Frontiers Media.

Petersen, E., Potdevin, Y., Mohammadi, E., Zidowitz, S., Breyer, S., Nowotka, D., Henn, S., Pechmann, L., Leucker, M., Rostalski, P., & Herzog, C. (2021). Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Technical Challenges and Solutions. arXiv (Cornell University).

Ramesh, S., Chokkara, S., Shen, T., Major, A., Volchenboum, S. L., Mayampurath, A., & Applebaum, M. A. (2021). Applications of Artificial Intelligence in Pediatric Oncology: A Systematic Review [Review of Applications of Artificial Intelligence in Pediatric Oncology: A Systematic Review]. JCO Clinical Cancer Informatics, 5, 1208. Lippincott Williams & Wilkins.

Schaefer, M. (2021). The Regulation of RNA Modification Systems: The Next Frontier in Epitranscriptomics? [Review of The Regulation of RNA Modification Systems: The Next Frontier in Epitranscriptomics?]. Genes, 12(3), 345. Multidisciplinary Digital Publishing Institute.

Smet, D., Opdebeeck, H., & Vandepoele, K. (2023). Predicting transcriptional responses to heat and drought stress from genomic features using a machine learning approach in rice. Frontiers in Plant Science, 14.

Tran, K., Kondrashova, O., Bradley, A. P., Williams, E. D., Pearson, J. V., & Waddell, N. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection [Review of Deep learning in cancer diagnosis, prognosis and treatment selection]. Genome Medicine, 13(1). BioMed Central.

Wang, C., Rai, G., & Rajapakse, J. C. (2025). Drug discovery and mechanism prediction with explainable graph neural networks. Scientific Reports, 15(1).

Wang, M., & Ibeagha-Awemu, E. M. (2021). Impacts of Epigenetic Processes on the Health and Productivity of Livestock [Review of Impacts of Epigenetic Processes on the Health and Productivity of Livestock]. Frontiers in Genetics, 11. Frontiers Media.

Woodman, R., & Mangoni, A. A. (2023). A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future [Review of A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future]. Aging Clinical and Experimental Research, 35(11), 2363. Springer Science+Business Media.

Azodi, C. B., Tang, J., & Shiu, S. (2020). Opening the Black Box: Interpretable Machine Learning for Geneticists [Review of Opening the Black Box: Interpretable Machine Learning for Geneticists]. Trends in Genetics, 36(6), 442. Elsevier BV.

Budhkar, A., Song, Q., Su, J., & Zhang, X. (2025). Demystifying the Black Box: A Survey on Explainable Artificial Intelligence (XAI) in Bioinformatics [Review of Demystifying the Black Box: A Survey on Explainable Artificial Intelligence (XAI) in Bioinformatics]. Computational and Structural Biotechnology Journal, 27, 346. Elsevier BV.

Cai, Z., Poulos, R. C., Liu, J., & Zhong, Q. (2022). Machine learning for multi-omics data integration in cancer [Review of Machine learning for multi-omics data integration in cancer]. iScience, 25(2), 103798. Cell Press.

Cao, Y., Geddes, T. A., Yang, J., & Yang, P. (2020). Ensemble deep learning in bioinformatics. Nature Machine Intelligence, 2(9), 500.

Dixit, S., Kumar, A., Srinivasan, K., Vincent, P. M. D. R., & Krishnan, N. R. (2024). Advancing genome editing with artificial intelligence: opportunities, challenges, and future directions [Review of Advancing genome editing with artificial intelligence: opportunities, challenges, and future directions]. Frontiers in Bioengineering and Biotechnology, 11. Frontiers Media.

Karim, Md. R., Islam, T., Shajalal, M., Beyan, O., Lange, C., Cochez, M., Rebholz-Schuhmann, D., & Decker, S. (2023). Explainable AI for Bioinformatics: Methods, Tools and Applications [Review of Explainable AI for Bioinformatics: Methods, Tools and Applications]. Briefings in Bioinformatics, 24(5). Oxford University Press.

Koo, P. K., & Ploenzke, M. (2020). Deep learning for inferring transcription factor binding sites. Current Opinion in Systems Biology, 19, 16.

Krassowski, M., Das, V., Sahu, S., & Misra, B. B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing [Review of State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing]. Frontiers in Genetics, 11. Frontiers Media.

Li, H., & Guan, Y. (2022). Asymmetric predictive relationships across histone modifications. Nature Machine Intelligence, 4(3), 288.

Lu, S.-C., Swisher, C. L., Chung, C., Jaffray, D. A., & Sidey-Gibbons, C. (2023). On the importance of interpretable machine learning predictions to inform clinical decision making in oncology [Review of On the importance of interpretable machine learning predictions to inform clinical decision making in oncology]. Frontiers in Oncology, 13. Frontiers Media.

Nam, Y., Kim, J., Jung, S., Woerner, J., Suh, E., Lee, D., Shivakumar, M., Lee, M. E., & Kim, D. (2024). Harnessing Artificial Intelligence in Multimodal Omics Data Integration: Paving the Path for the

Next Frontier in Precision Medicine [Review of Harnessing Artificial Intelligence in Multimodal Omics Data Integration: Paving the Path for the Next Frontier in Precision Medicine]. Annual Review of Biomedical Data Science, 7(1), 225. Annual Reviews.

Tran, K., Kondrashova, O., Bradley, A. P., Williams, E. D., Pearson, J. V., & Waddell, N. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection [Review of Deep learning in cancer diagnosis, prognosis and treatment selection]. Genome Medicine, 13(1). BioMed Central.

Walsh, I., Fishman, D., García-Gasulla, D., Titma, T., Harrow, J., Psomopoulos, F., & Tosatto, S. C. E. (2020). Recommendations for machine learning validation in biology. arXiv (Cornell University). https://arxiv.org/abs/2006.16189v1

Wang, C., Rai, G., & Rajapakse, J. C. (2025). Drug discovery and mechanism prediction with explainable graph neural networks. Scientific Reports, 15(1).

Xin, L., Huang, C., Li, H., Huang, S., Feng, Y., Kong, Z., Liu, Z., Li, S., Yu, C., Shen, F., & Tang, H. (2024). Artificial Intelligence for Central Dogma-Centric Multi-Omics: Challenges and Breakthroughs. arXiv (Cornell University).