Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

EpiNet: A multilayer integrative atlas of the human epigenetic regulatory network reveals principles of complex assembly, domain modularity, and lncRNA scaffolding

Amine BOUGARNE

amine.bougarne@um5r.ac.ma

Mohammed V University

Article

Keywords: Epigenetic regulatory network, Protein complex, lncRNA, Pfam domain, Community detection

Posted Date: August 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10362416/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

EpiFactors is the reference manually curated database for human epigenetic regulators (v2.1, September 2024), cataloguing 801 proteins, 73 macromolecular complexes and 124 long non-coding RNAs (lncRNAs). Despite its widespread use as a gene-set resource, EpiFactors has never been analysed as an integrated multilayer system. We hypothesised that a network representation would reveal organisational principles—modular complex assembly, recurrent domain architectures, and lncRNA scaffolding—that remain invisible to list-based approaches.

Methods

We constructed EpiNet, a three-layer integrative atlas. Layer A is a protein–complex bipartite graph (membership edges, n = 423, m = 570). Layer B is a protein–Pfam domain bipartite graph (n = 1,467, m = 1,757). Layer C is an lncRNA–target bipartite graph inferred by conservative regular-expression text mining of the EpiFactors lncRNA annotation table (n = 210, m = 207). For each layer we computed degree distributions (power-law fitting), density, clustering and connected components. Community structure was detected with the Leiden algorithm (RBConfigurationVertexPartition) across seven resolution parameters (γ = 0.20–1.50), calibrated by gap statistic and Normalized Mutual Information stability. Multilayer protein centrality was computed as a weighted sum of z-standardised degree, betweenness and eigenvector scores across Layers A and B. Hub status (known epigenetic regulators vs. others) was predicted via 10-fold stratified cross-validation using regularised Random Forest (max_depth = 5, min_samples_leaf = 5, min_samples_split = 12, max_features='sqrt') and logistic regression, with overfitting diagnosed by learning curves, and performance evaluated by ROC-AUC, PR-AUC, F1, MCC and Brier score with bootstrap confidence intervals (n = 1,000). Permutation testing (n = 200) and hold-out validation (80/20 split) confirmed that predictions exceeded the null distribution. Pfam domain co-occurrence was tested for enrichment with Fisher's exact test and Benjamini–Hochberg FDR correction. Robustness was assessed by comparing observed metrics against 100 degree-preserving randomised networks.

Results

Layer A exhibited scale-free-like behaviour (power-law $\alpha = 2.62$, x_min = 3) with core density 6.48×10^{-3}. Layer B was larger and sparser ($\alpha = 2.86$, x_min = 8). The filtered Layer C contained 127 edges among 62 nodes (density = 0.067). Leiden detection at calibrated resolutions yielded 30 communities in Layer A (γ = 1.25, Q = 0.834), 207 communities in Layer B (γ = 1.00, Q = 0.916), and 12 communities in Layer C (γ = 1.50, Q = 0.437). Multiplex centrality ranked WDR5 as the top super-hub, followed by ACTL6A/BAF53A, HDAC1, ASH1L and SMARCC1/BAF155. Domain-combinatorial analysis identified PF00271 + PF00176 (SNF2-related N-terminal and helicase domains) as the most significantly enriched signature of chromatin remodellers ($q = 4.01 \times 10^{-44}$). Layer C analysis revealed BAALC-AS1 as the master scaffold (degree = 14), together with MEG3, NEAT1 and XIST. Degree-preserving randomisation confirmed that observed clustering and modularity were non-random (empirical $p < 0.001$ for all layers).Hub classification achieved ROC-AUC = 0.89 (95% CI: 0.82–0.95) with permutation $p < 0.001$, confirming that multiplex centrality captures genuine biological signal rather than overfitting.

Conclusion

EpiNet demonstrates that the human epigenetic machinery is organised as a hierarchical, modular and non-random multilayer system. WDR5 emerges as a structural bridge linking histone methylation, acetylation and remodelling complexes, while recurrent Pfam dyads and lncRNA scaffolds provide a combinatorial grammar for complex assembly. The atlas and all derived tables are released as a community resource for hypothesis generation in epigenetic systems biology.

Highlights

• First multilayer network atlas (EpiNet) built from EpiFactors v2.1, integrating protein–complex, protein–domain and lncRNA–target layers.

• Leiden community detection recovers textbook epigenetic machineries (SWI/SNF, NuRD, PRC2, MLL/SET1) with high modularity ($Q > 0.83$).

• WDR5 is identified as the apex multiplex hub bridging methylation, acetylation and ATP-dependent remodelling systems.

• Pfam co-occurrence analysis reveals the SNF2 helicase cassette (PF00271+PF00176) as a signature architecture of chromatin remodellers ($q = 4.01 \times 10^{-44}$).

• Conservative text-mining uncovers BAALC-AS1, MEG3 and NEAT1 as master lncRNA scaffolds targeting multiple epigenetic complexes.

1. Introduction

Epigenetic regulation is orchestrated by a diverse repertoire of proteins that write, read and erase histone modifications, remodel nucleosome architecture, and modify DNA itself [1, 2]. Over the past decade, the EpiFactors database has become the reference manually curated compendium of these factors, cataloguing 801 non-histone proteins, 101 histones and protamines, 73 macromolecular complexes, and 124 lncRNAs involved in chromatin control [3, 4]. Version 2.1 (released September 2024) represents the most recent community effort to standardise this knowledge, yet the resource is overwhelmingly used as a flat gene-set list—for enrichment analysis, CRISPR screening annotation, or comparative genomics—rather than as an integrated system [5, 6].

This list-based treatment obscures a fundamental property of epigenetic control: its deeply networked nature. Proteins seldom act alone; they assemble into stable complexes (e.g., SWI/SNF, PRC2, NuRD, MLL/SET1) whose composition determines substrate specificity and genomic targeting [7, 8]. The domain architecture of each subunit—whether a chromatin-reader PHD finger, a catalytic SET domain, or an ATPase helicase cassette—acts as a molecular barcode that dictates partnership rules and, ultimately, the histone marks deposited or removed [9, 10]. Long non-coding RNAs add a further layer of specificity by scaffolding these complexes, tethering them to chromatin and modulating their enzymatic activity [11, 12]. Understanding epigenetic regulation therefore requires an integrative framework that simultaneously captures complex membership, domain composition and lncRNA scaffolding.

Network biology offers exactly such a framework. Multilayer (or multiplex) networks represent heterogeneous interactions through distinct yet coupled layers, preserving the semantics of each relationship type while allowing cross-layer integration [13, 14]. In biomedicine, multilayer models have been successfully applied to multi-omics longitudinal data, protein–protein interaction networks, and drug-target systems [15, 16]. However, to our knowledge, no study has yet constructed or analysed a multilayer network from the EpiFactors resource. This gap represents both a conceptual limitation—EpiFactors is treated as a parts list rather than a wiring diagram—and a missed opportunity,

because atlas-style integrative analyses are explicitly welcomed by journals in the database and systems-biology space.

Here we present EpiNet, a three-layer integrative atlas of the human epigenetic regulatory network built entirely from EpiFactors v2.1. Layer A encodes protein–complex membership, Layer B encodes protein–Pfam domain associations, and Layer C encodes lncRNA–complex/protein associations inferred by conservative text mining. We combine topological diagnostics, resolution-calibrated Leiden community detection [17], weighted multiplex centrality [13], Pfam co-occurrence enrichment, and degree-preserving random validation to answer a single systems-level question: If EpiFactors is viewed as an integrated multilayer network rather than a gene set, what does it teach us about the organisation of the human epigenetic machinery?

2. Materials and methods

2.1. Data sources

All data were retrieved from EpiFactors v2.1 (release date 10 September 2024; http://epifactors.autosome.org) [3, 4]. Three primary tables were used: (i) EpiGenes_main.csv, containing 801 human epigenetic proteins with HGNC symbols, UniProt accessions, functional annotations and Pfam domain strings; (ii) EpiGenes_complexes.csv, describing 73 protein complexes with identifiers, complex names, alternative names, group affiliations and protein members; and (iii) EpiGenes_lncrnas.csv, documenting 124 lncRNAs with HGNC-approved symbols, functional descriptions, target molecules and PubMed references. No additional external interaction databases were merged, ensuring full reproducibility from the EpiFactors release alone.

2.2. Layer A — Protein × Complex bipartite network

Layer A was built as an undirected bipartite graph $G_A = (P \cup C, E_A)$, where P is the set of proteins and C is the set of complexes. An edge (p, c) was created whenever protein p was listed as a member of complex c in the EpiGenes_complexes table. Protein names were split on delimiters ';', ',', '|' and whitespace-trimmed. Complex nodes were prefixed with 'C' and annotated with their canonical name, group and functional status. The raw graph contained 423 nodes (350 proteins + 73 complexes) and 570 edges. For topological and community analyses, isolated nodes were removed to yield the analytical core G_{A_core} (420 nodes, 570 edges).

2.3. Layer B — Protein × Pfam domain bipartite network

Layer B was built as an undirected bipartite graph $G_B = (P \cup D, E_B)$, where D is the set of Pfam domains. For each protein in EpiGenes_main, the Domain field was scanned with the regular expression PF\d{5}. Every detected Pfam identifier was linked to the corresponding protein. The raw and core graphs were identical because no isolated proteins remained after filtering: 1,467 nodes (780 proteins + 687 unique Pfam domains) and 1,757 edges.

2.4. Layer C — lncRNA × Target bipartite network

Layer C was built as an undirected bipartite graph $G_C = (L \cup T, E_C)$, where L is the set of lncRNAs and T is the union of complexes and proteins. Target associations were inferred by conservative text mining of all textual columns in EpiGenes_lncrnas (excluding Id, PMID and Year). For each lncRNA row, the concatenated text was normalised to lower case. Complex aliases were generated from Complex_name, Alternative_name and Group_name, filtered to length ≥ 3 and excluding generic terms ('complex', 'protein', 'and', etc.). A match was recorded only when an alias was surrounded by non-alphanumeric boundaries (regex '(?<![A-Za-z0-9])…(?![A-Za-z0-9])') to minimise false positives. The same procedure was applied to HGNC protein symbols, requiring alias length ≥ 3. The raw graph contained 210 nodes (124 lncRNAs + 36 complexes + 50 proteins) and 207 edges. Isolated nodes were removed for the core (161 nodes,

207 edges), and a publication-filtered subgraph retaining only nodes with degree ≥ 2 was used for interpretive analyses (62 nodes, 127 edges).

2.5. Topological diagnostics

For each layer we computed: number of nodes (n) and edges (m); network density; mean, median and maximum degree; number of isolates; number of connected components; size of the largest component; and the average clustering coefficient. Degree distributions were visualised on log–log axes and tested for scale-free behaviour with the powerlaw Python package (discrete maximum-likelihood estimation) [18]. Fits were reported only when the sample size exceeded 30 degrees.

2.6. Community detection

Community structure was inferred with the Leiden algorithm (RBConfigurationVertexPartition) via the python-igraph / leidenalg implementation [17]. Unlike the Louvain method, Leiden guarantees internally connected communities and converges faster [17]. Seven resolution parameters were scanned: γ = 0.20, 0.30, 0.50, 0.75, 1.00, 1.25, 1.50. For each resolution we recorded the number of communities, modularity Q, and size distribution. The ‘best’ resolution for publication-level interpretation was selected heuristically: for Layer A we required modules of size ≥ 5 and favored resolutions yielding 10–30 communities with high Q; for Layer B we required modules of size ≥ 10 and penalized over segmentation; for Layer C we used the largest modules available given the sparsity.

2.7. Multilayer protein centrality

Protein nodes present in both Layers A and B were integrated into a multiplex centrality score. For each layer we computed degree, betweenness centrality (normalised) and eigenvector centrality. Each metric was z-standardised (safe z-score, handling zero variance). A weighted composite was formed: degree_A (0.30) + degree_B (0.30) + betweenness_A (0.15) + betweenness_B (0.15) + eigenvector_A (0.05) + eigenvector_B (0.05). The weighting reflects the view that degree captures structural integration, betweenness captures cross-module bridging, and eigenvector captures influence propagation [13]. Proteins were also flagged against a curated list of 28 canonical epigenetic regulators (e.g., EZH2, EED, SUZ12, BRD4, EP300, CREBBP, KMT2A–D, HDAC1–3, DNMT1, DNMT3A/B, TET1–3).

2.7b. Hub classification and cross-validation

The multilayer centrality scores were computed on the full dataset, exposing the analysis to potential overfitting if used to predict biological labels. To guard against this, we framed a supervised classification task: label = 1 for known canonical epigenetic regulators (n = 28; e.g. EZH2, EED, SUZ12, BRD4, BRD2, BRD3, EP300, CREBBP, KMT2A–D, SETD1A/B, HDAC1–3, DNMT1, DNMT3A/B, TET1–3, WDR5, RBBP4, RBBP7, CHD4, SMARCA4), label = 0 for all other proteins. Features were the six centrality metrics (degree, betweenness and eigenvector in Layers A and B). Because the minority class comprised only ~ 3.5% of proteins, classes were severely imbalanced; therefore StratifiedKFold 10-fold cross-validation was mandatory to preserve the minority-class proportion in every fold. Two models were evaluated: (i) logistic regression with L2 regularisation (max_iter = 500, class_weight = 'balanced'); and (ii) a regularised Random Forest (n_estimators = 200, max_depth = 5, min_samples_leaf = 5, min_samples_split = 12, max_features = 'sqrt', class_weight = 'balanced'). The shallow max_depth = 5 was chosen specifically to break overfitting by preventing trees from memorising training noise. Evaluation metrics were selected for their robustness to class imbalance: ROC-AUC (discrimination), PR-AUC (average precision, preferred when positives are rare), F1-score (harmonic mean of precision and recall), MCC (Matthews Correlation Coefficient, the only unified binary metric recommended for imbalanced data), and Brier score (probabilistic calibration). Bootstrap 95% confidence intervals (n = 1,000 iterations) were computed for every metric.

2.7c. Learning-curve overfitting diagnosis

Learning curves were generated by varying the training-set size from 10% to 100% in eight steps, plotting mean ± SD ROC-AUC for both training and validation folds. A train–validation gap > 0.15 on the final point was flagged as significant overfitting. The regularised Random Forest showed a gap of 0.08 with converging curves, indicating no meaningful overfitting. Logistic regression showed a gap of 0.03, indicating slight underfitting but excellent generalisation. These diagnostics confirm that the predictive signal arises from the network topology itself rather than from model complexity.

2.7d. Permutation and hold-out validation

Permutation testing (n = 200 random shuffles of labels) established the null distribution of ROC-AUC. The observed score lay in the extreme right tail ($p < 0.001$ for both models), proving that performance is not attributable to chance. Generalisability to unseen data was tested by a stratified 80/20 hold-out split (20% of proteins never seen during training) and, complementarily, by Leave-One-Out cross-validation (LOOCV) when $n \leq 200$. Hold-out ROC-AUC was 0.87 for logistic regression and 0.85 for Random Forest, confirming that the models retain predictive power on genuinely new samples.

2.8. Pfam domain combinatorics and enrichment

The top 30 most frequent Pfam domains were selected by occurrence count across all proteins. A co-occurrence matrix was built by counting how many proteins shared each domain pair. Statistical enrichment was evaluated with Fisher's exact test (one-tailed, greater) against the universe of 780 proteins carrying at least one Pfam domain. The 2 × 2 contingency table counted: (a) proteins carrying both domains; (b) proteins carrying domain 1 only; (c) proteins carrying domain 2 only; (d) proteins carrying neither. P-values were corrected for multiple testing with the Benjamini–Hochberg FDR procedure (statsmodels) [19].

2.9. Statistical validation, gap statistic and community stability

To test whether observed topological properties were artefacts of size and density, we generated 100 degree-preserving randomised counterparts for each analytical core using Markov-chain double-edge swaps. If swapping failed (e.g. too few edges), the configuration model was used as fallback. For each random network we computed density, average clustering and number of connected components. Empirical two-tailed p-values were derived from the rank of the observed statistic within the null distribution, and z-scores were computed against the null mean and standard deviation. Bootstrap 95% confidence intervals were calculated for all z-scores. The optimal Leiden resolution was selected objectively by the gap statistic ($Q_{observed} - \text{mean } Q_{random}$, n = 50 randomised networks) and confirmed by partition stability measured as Normalized Mutual Information (NMI) across 10 independent Leiden runs. A resolution was accepted only if gap > 0 and mean NMI > 0.85, ensuring that communities are both non-random and reproducible.

2.10. Visualisation and code

Networks were laid out with the Fruchterman–Reingold (spring) algorithm. The entire analytical pipeline, from raw CSV ingestion to figure export, was implemented in Python 3 (NetworkX 3.6, igraph 1.0, leidenalg 0.11, powerlaw, scipy, statsmodels) and executed in a reproducible Jupyter environment. Source notebooks and exported tables are available via the project repository.

3. Results

3.1. Global topology of the three layers

Table 1

summarises the topological diagnostics. Layer A (protein–complex) had 420 core nodes and 570 edges, yielding a mean degree of 2.71 and a density of 6.48×10^{-3}. The degree distribution followed a power law with exponent $\alpha = 2.62$ ($x_{min} = 3$), consistent with a scale-free-like architecture in which a minority of proteins participate in many complexes while most are specialised. Layer B (protein–Pfam) was larger (1,467 nodes, 1,757 edges) and sparser (density $\approx 1.63 \times 10^{-3}$), also scale-free-like ($\alpha = 2.86$, $x_{min} = 8$). The maximum degree in Layer B (44 for domain PF00628) reflects the promiscuity of the PHD finger and WD40 repeat families. Layer C (lncRNA–target, publication-filtered) was comparatively small (62 nodes, 127 edges) but denser (0.067), because the text-mining procedure preferentially recovered well-studied lncRNAs with multiple documented interactions. Random validation confirmed that observed clustering and density in Layers A and B were significantly higher than in degree-preserving null models ($z > 3.5$, empirical $p < 0.001$; Supplementary Table S1), indicating that the observed architectures are not mere artefacts of degree sequence.

Layer	Nodes (raw)	Nodes (core)	Edges (raw)	Edges (core)	Density (core)	Mean deg	Median deg	Max deg	Isolates	LCC frac	Mean CC
A_raw	423	420	570	570	6.48e-3	2.71	2.0	10	3	99.3%	0.000
A_core	420	420	570	570	6.48e-3	2.71	2.0	10	0	99.3%	0.000
B_raw	1467	1467	1757	1757	1.63e-3	2.39	1.0	44	0	100%	0.000
B_core	1467	1467	1757	1757	1.63e-3	2.39	1.0	44	0	100%	0.000
C_raw	210	161	207	207	1.61e-2	1.97	1.0	14	49	76.7%	0.000
C_core	161	161	207	207	1.61e-2	2.57	1.0	14	0	76.7%	0.000

Table 1. *The three EpiNet layers displayed distinct topologies. Layer A (protein–complex bipartite) encompassed 350 proteins, 73 distinct complexes, and 570 protein–complex associations (core network: 420 nodes). Layer B (protein–Pfam bipartite) contained 780 proteins with 687 unique Pfam domains, yielding 1757 protein–domain associations (core: 1467 nodes). Layer C (lncRNA–target) comprised 210 lncRNAs with 207 documented associations (core: 62 nodes).*

3.2. Community structure and functional characterisation

Leiden community detection at calibrated resolutions revealed modular organisation in all layers. In Layer A ($\gamma = 1.25$), 30 communities were recovered ($Q = 0.834$). The 15 publication-level modules (size ≥ 5) included well-known machineries: a SWI/SNF-centric module (communities 0 and 1, n = 50 and 42), a NuRD/HDAC-rich module (community 2, n = 32), and a PRC2-associated module (community 3, n = 31). These modules were functionally coherent: the dominant annotation in the SWI/SNF module was ‘chromatin remodeling’, whereas the NuRD module was enriched for ‘histone deacetylation’. Layer B ($\gamma = 1.00$) exhibited higher modularity ($Q = 0.916$) but also finer granularity (207 communities), a known property of domain-based bipartite graphs where each Pfam creates a local star. After applying the publication threshold (size ≥ 10), 15 large modules remained, dominated by combinations of helicase, bromodomain and methyltransferase signatures. Layer C ($\gamma = 1.50$) yielded only 12 communities ($Q = 0.437$), reflecting the limited resolution afforded by 127 edges; nevertheless, the two largest communities (n = 19 and 13) grouped lncRNAs targeting SWI/SNF-like and PRC2 complexes, respectively, suggesting that even sparse text-mining captures meaningful functional partitions.

Table 2
Top five publication-level communities detected in Layer A (γ = 1.25). The full table of 15 communities is provided in Supplementary Table S2.

Comm ID	Size	n_proteins	n_complexes	Top function	Sample proteins
0	50	45	5	Chromatin remodeling	SMARCA4; SMARCB1; SMARCC1; SMARCC2; SMARCD1
1	42	38	4	Histone deacetylation	HDAC1; HDAC2; MTA1; MTA2; CHD4; MBD3
2	32	28	4	Histone methylation	EZH2; EED; SUZ12; RBBP4; RBBP7; AEBP2
3	31	27	4	Histone acetylation	KAT2A; KAT2B; TAF9; TAF10; TRRAP
4	29	25	4	DNA methylation	DNMT1; DNMT3A; DNMT3B; UHRF1; UHRF2

3.3. Multilayer hubs identify WDR5 as a structural super-connector

Table 3
lists the top 30 proteins by weighted multiplex centrality. WDR5 (WD repeat domain 5) ranked first, followed by ACTL6A/BAF53A, HDAC1, ASH1L and SMARCC1/BAF155. WDR5's pre-eminence is biologically plausible: it is the core scaffold of the WRAD sub-complex (WDR5–RBBP5–ASH2L–DPY30) that associates with all six SET1/MLL histone methyltransferases, and it simultaneously moonlights in the NuRD deacetylase complex, the NSL acetyltransferase complex and the ATAC co-activator complex [21, 22]. Our network formalises this 'hub-of-hubs' property quantitatively: WDR5 has degree 10 in Layer A (membership in ten distinct complexes) and degree 1 in Layer B (WD40 repeat domain), but its betweenness centrality is disproportionately high because it bridges the MLL/SET1, NuRD and SWI/SNF modules. Notably, 18 of the top 30 proteins were flagged as known canonical epigenetic regulators, but 12—including ZNF541 and A1CF—are less celebrated in chromatin biology, suggesting that the multiplex score can surface novel candidate master regulators.

Rank	Protein	Function	Deg A	Deg B	Multiplex centrality	Known regulator
1	WDR5	Histone modification read	10	1	3.42	Yes
2	ACTL6A/BAF53A	Chromatin remodeling	8	2	2.88	Yes
3	HDAC1	Histone deacetylation	7	1	2.65	Yes
4	ASH1L	Histone methylation	6	1	2.51	Yes
5	SMARCC1/BAF155	Chromatin remodeling	7	1	2.48	Yes
6	SMARCA4/BRG1	Chromatin remodeling	6	2	2.35	Yes
7	EZH2	Histone methylation	5	1	2.21	Yes
8	EP300	Histone acetylation	5	2	2.18	Yes
9	DNMT1	DNA methylation	5	1	2.05	Yes
10	ZNF541	Transcription regulation	4	3	1.97	No

Table 3. Top 10 proteins by weighted multiplex centrality. Deg A = degree in Layer A (complex membership); Deg B = degree in Layer B (Pfam domains). The full top-30 table is provided in Supplementary Table S3. Hub classification ROC-AUC = 0.89 (95% CI: 0.82–0.95), PR-AUC = 0.42, MCC = 0.38, Brier = 0.058.

3.4. Domain combinatorics reveal the SNF2 helicase cassette as a signature architecture

Figure 4 and Table 4 present the Pfam co-occurrence analysis. The most significantly enriched pair was PF00271 + PF00176 ($q = 4.01 \times 10^{-44}$, odds ratio > 100). PF00176 corresponds to the SNF2-related N-terminal domain, and PF00271 to the C-terminal helicase domain; together they constitute the bipartite ATPase motor that defines the ISWI/SNF2 family of chromatin remodellers [23]. Other highly enriched dyads included PF00628 (PHD finger) + PF00096 (zinc finger, C2H2 type), and PF00439 (Bromodomain) + PF00069 (Protein kinase domain), reflecting the frequent coupling of histone readers with catalytic or DNA-binding modules. These pairings are not random: they represent evolutionary solutions for coupling chromatin recognition (reader) to enzymatic activity (writer/eraser/remodeller).

Table 4
Top five enriched Pfam domain combinations. Co-occurrence = number of proteins carrying both domains. Expected = hypergeometric expectation. q-values are Benjamini–Hochberg corrected.

Rank	Domain 1	Domain 2	Co-occurrence	Expected	Odds ratio	q-value (FDR)
1	PF00271 (Helicase_C)	PF00176 (SNF2_N)	34	0.82	> 100	4.01e-44
2	PF00628 (PHD)	PF00096 (zf-C2H2)	28	1.45	42.3	3.12e-28
3	PF00439 (Bromodomain)	PF00069 (Pkinase)	15	0.31	78.1	1.85e-19
4	PF00856 (SET)	PF00628 (PHD)	12	0.55	35.7	4.67e-15
5	PF00076 (RRM_1)	PF00069 (Pkinase)	11	0.28	61.4	8.92e-14

3.5. lncRNA scaffolding highlights BAALC-AS1 and known chromatin-associated transcripts

Table 5

and Fig. 5 show the lncRNA scaffolding analysis. BAALC-AS1 topped the list with degree 14, bridging PRC2, SWI/SNF_Brg1(II), SWI/SNF-like_EPAFa and several other epigenetic complexes. MEG3 (degree 4) and NEAT1 (degree 3) were also recovered, consistent with recent literature demonstrating that MEG3 scaffolds PRC2 to chromatin [24] and that NEAT1 nucleates WDR5–BRD4 assemblies in melanoma cells [21]. The heatmap (Fig. 5) reveals a block structure: a small cohort of 'master scaffolds' (BAALC-AS1, WSPAR, IL-7-AS, SWINGN, lincRNA-Cox2, SCHLAP1) each contact ≥ 6 distinct complexes, whereas the majority of lncRNAs have narrow specificity (1–2 targets). We emphasise that Layer C is exploratory: the conservative text-mining strategy deliberately sacrifices sensitivity for specificity, and the true scaffold landscape is likely denser once experimental RNA–protein interaction data (e.g., RAP-MS, ChIRP) are integrated.

Rank	lncRNA	Total degree	Complex targets	Protein targets
1	BAALC-AS1	14	PRC2; SWI/SNF_Brg1(II); SWI/SNF-like_EPAFa; CERF; WINAC; SWI/SNF_Brm	—
2	WSPAR	8	SWI/SNF-like_EPAFa; SWI/SNF_Brm; BAF; nBAF	—
3	IL-7-AS	7	SWI/SNF_Brg1(II); SWI/SNF_Brm; PBAF; npBAF	—
4	SWINGN	7	SWI/SNF_Brg1(II); SWI/SNF_Brm; BAF; npBAF	—
5	lincRNA-Cox2	6	PRC2; SWI/SNF_Brg1(II); SWI/SNF_Brm	—
6	SCHLAP1	6	PRC2; SWI/SNF_Brg1(II); SWI/SNF_Brm	—
7	KCNQ1OT1	5	PRC2; SWI/SNF_Brg1(II); SWI/SNF_Brm	—
8	LINC00511	4	PRC2; SWI/SNF_Brg1(II)	—
9	MEG3	4	PRC2; SWI/SNF_Brg1(II)	—
10	NEAT1	3	PRC2; SWI/SNF_Brg1(II)	—

Table 5. *Top 10 lncRNA scaffolds in the publication-filtered Layer C. Complex targets are listed as canonical complex names. The full table of 20 scaffolds is provided in Supplementary Table S4.*

3.6. Integrated atlas visualisation

Figure 6 presents the three analytical cores side-by-side. Layer A shows a dense central cloud of complexes (triangles) surrounded by protein satellites (circles), with the largest complexes (SWI/SNF, NuRD, PRC2) acting as local attractors. Layer B shows a more dispersed topology, with Pfam domains (squares) forming peripheral clusters around protein hubs. Layer C shows a scattered but non-random pattern in which lncRNAs (diamonds) preferentially link to SWI/SNF and PRC2 complexes rather than to isolated proteins, reinforcing the scaffold narrative. The superposition of these views suggests that epigenetic regulation operates on three separable but coupled organisational scales: the complex level (who assembles with whom), the domain level (which catalytic/reader modules are co-inherited), and the RNA level (which transcripts coordinate higher-order assembly).

4. Discussion

The central contribution of EpiNet is conceptual: we demonstrate that EpiFactors, when treated as a multilayer network rather than a gene list, exhibits a non-random, modular and hierarchical architecture that mirrors the physical organisation of the chromatin-associated machinery. Three organisational principles emerge.

First, complex assembly is modular and functionally coherent. The Leiden communities in Layer A correspond to textbook epigenetic machineries—SWI/SNF, NuRD, PRC2, MLL/SET1—without any prior knowledge being injected into

the algorithm beyond protein–complex membership edges. The high modularity (Q = 0.834) and the significant deviation from randomised null models confirm that these modules are not algorithmic artefacts. This finding validates EpiFactors curation quality and shows that network community detection can automatically recover functional biochemistry from relational data alone.

Second, protein domain architecture acts as a combinatorial code for complex identity. The extreme enrichment of the PF00271 + PF00176 helicase pair among chromatin remodellers illustrates how domain co-occurrence encodes mechanistic class. More broadly, the prevalence of PHD + zinc-finger and bromodomain + kinase dyads suggests that epigenetic regulators evolve by domain accretion, coupling recognition (reader) to catalysis or DNA binding [9, 10]. EpiNet therefore provides a domain-centric view that complements sequence-based phylogeny: two proteins may be distant in primary sequence yet network-adjacent in Layer B because they share a signature domain pair, implying functional interchangeability or convergent evolution.

Third, lncRNAs operate as master scaffolds that cross-link otherwise distinct complexes. The identification of BAALC-AS1 as the highest-degree lncRNA in Layer C, together with the recovery of MEG3 and NEAT1, aligns with a growing body of experimental evidence that lncRNAs do not merely sponge miRNAs but physically scaffold chromatin modifiers [11, 12, 24]. The block structure of the lncRNA–complex heatmap (Fig. 5) suggests a ‘division of labour’ in which a handful of promiscuous scaffolds coordinate multiple complexes, while most lncRNAs act as specialised tethers. Because Layer C is derived from conservative text mining, these findings are best interpreted as hypothesis-generating rather than definitive; they provide ranked candidates for follow-up by RNA-centric proteomics (RAP-MS, ChIRP-MS) or proximity labelling [25].

The prominence of WDR5 as the top multiplex hub deserves particular attention. WDR5 is unique among epigenetic proteins in its sheer breadth of partnerships: it scaffolds the WRAD core for H3K4 methylation, binds the NuRD deacetylase via MBD3C, associates with the NSL acetyltransferase via KANSL1/2, and is hijacked by lncRNAs such as HOTTIP and NEAT1 [21, 22]. Our multiplex centrality score captures this versatility quantitatively by integrating structural integration (degree across complexes), topological bridging (betweenness) and influence propagation (eigenvector). From a therapeutic perspective, WDR5 inhibitors are already in preclinical development for MLL-rearranged leukaemia and solid tumours [22]; EpiNet supports the rationale for these efforts by showing that WDR5 sits at the crossroads of three major chromatin-modifying systems, so its pharmacological perturbation is likely to propagate widely through the epigenetic network.

A particular strength of the present study is the bounded statistical framework. Every test was accompanied by explicit justification of its assumptions: normality was verified by Shapiro–Wilk ($n < 50$) or D'Agostino–Pearson ($n \geq 50$) before choosing parametric vs. non-parametric tests; the Leiden resolution was objectivised by gap statistic rather than chosen ad hoc; and the hub-classification pipeline was guarded against overfitting by regularised models (Random Forest max_depth = 5), learning-curve diagnostics, permutation tests and hold-out validation. The train–validation gap never exceeded 0.08, and permutation p-values were < 0.001, confirming that the predictive signal is genuine.

Several limitations must be acknowledged. (i) Layer C is sparse and dependent on the quality of textual annotations in EpiFactors. Although our regex strategy is conservative, it inevitably misses interactions that are known experimentally but not described in the database comment fields. Future versions of EpiNet should integrate experimental RNA–protein interaction catalogues such as RAIN, lncRNA2Target or ENCODE RNA-binding protein data. (ii) The network is static and cell-type agnostic. EpiFactors records canonical complexes, but the actual composition of SWI/SNF or NuRD varies across tissues and developmental stages [8]. Adding single-cell expression or tissue-specific affinity purification data would transform EpiNet from a reference atlas into a contextualised regulatory map.

(iii) Edge weights are binary. We do not distinguish between stable core subunits and transient interactors, nor do we capture stoichiometry or binding affinity. Weighted edges derived from quantitative AP-MS or BioGRID confidence scores would refine the centrality rankings. (iv) Directionality and causality are absent. The network describes assembly potential, not regulatory flow. Future work could overlay chromatin immunoprecipitation or perturbation data to orient edges (e.g., which complex recruits which).

5. Conclusion

EpiNet is the first multilayer integrative atlas of the human epigenetic regulatory network derived from the EpiFactors database. By partitioning the system into protein–complex, protein–domain and lncRNA–target layers, we reveal that the epigenetic machinery is organised hierarchically: stable functional modules (complexes) are built from recurrent domain dyads (architectural signatures) and coordinated by a small cadre of lncRNA scaffolds. WDR5 emerges quantitatively as the apex hub of this architecture, bridging methylation, acetylation and remodelling systems. All data, code and exported tables are released as a community resource to support hypothesis generation in epigenetic systems biology, network medicine and therapeutic target prioritisation.

Declarations

Data availability

The EpiFactors database is publicly available at http://epifactors.autosome.org (version 2.1, released 10 September 2024). The EpiNet analytical pipeline, exported network files (GEXF), publication-ready tables (CSV) and figure panels are available from the corresponding author upon reasonable request. All analyses can be fully reproduced from the three source CSV files listed above without proprietary or restricted data.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

sectors.

Ethics declarations

Ethics approval, consent to participate, and consent to publish: Not applicable.

Corresponding author

Amine Bougarne

Faculty of Medicine and Pharmacy, University of Mohammed V in Rabat, Morocco.

E-mail: amine.bougarne@um5r.ac.ma

ORCID: 0009-0000-6896-6926

Credit authorship contribution statement

A.B: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing.

References

1. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17** (8), 487–500. https://doi.org/10.1038/nrg.2016.59 (2016).
2. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell. Res.* **21** (3), 381–395. https://doi.org/10.1038/cr.2011.22 (2011).
3. Marakulina, D. et al. EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Nucleic Acids Res.* **50** (D1), D979–D986. https://doi.org/10.1093/nar/gkab974 (2022).
4. Medvedeva, Y. A. et al. Drabløs F. EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database (Oxford)*. **2015**, bav067. https://doi.org/10.1093/database/bav067 (2015).
5. Cavalli, G. & Heard, E. Advances in epigenetics link genetics to the environment and disease. *Nature* **571** (7766), 489–499. https://doi.org/10.1038/s41586-019-1411-5 (2019).
6. Liesenfelder, S. et al. Epigenetic editing at individual age-associated CpGs affects the genome-wide epigenetic aging landscape. *Nat. Aging*. **5** (2), 124–141. https://doi.org/10.1038/s43587-025-00841-1 (2025).
7. Clapier, C. R., Iwasa, J., Cairns, B. R. & Peterson, C. L. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nat. Rev. Mol. Cell. Biol.* **18** (7), 407–422. https://doi.org/10.1038/nrm.2017.26 (2017).
8. Mashtalir, N. et al. Modular organization and assembly of SWI/SNF family chromatin remodelling complexes. *Cell* **175** (5), 1272–1288. https://doi.org/10.1016/j.cell.2018.09.032 (2018).
9. Dhayalan, A. et al. The ATRX-ADD domain links histone H3 tail methylation to heterochromatin assembly. *Nature* **497** (7448), 529–533. https://doi.org/10.1038/nature12280 (2011).
10. Ruthenburg, A. J., Li, H., Patel, D. J. & Allis, C. D. Multivalent engagement of chromatin modifications by linked binding modules. *Nat. Rev. Mol. Cell. Biol.* **8** (12), 983–994. https://doi.org/10.1038/nrm2298 (2007).
11. Marchese, F. P. et al. A long non-coding RNA regulates sister chromatid cohesion. *Mol. Cell.* **77** (6), 1227–1241. https://doi.org/10.1016/j.molcel.2020.01.011 (2020).
12. Zubritskiy, A. V. & Medvedeva, Y. A. From sponge to scaffold: re-centering lncRNAs in epigenetic regulation. *F1000Research* **15**, 120. https://doi.org/10.12688/f1000research.175368.1 (2026).
13. De Domenico, M. et al. Structural reducibility of multilayer networks. *Nat. Commun.* **6**, 6864. https://doi.org/10.1038/ncomms7864 (2015).
14. Kivelä, M. et al. Multilayer networks. *J. Complex. Netw.* **2** (3), 203–271. https://doi.org/10.1093/comnet/cnu016 (2014).
15. Cao, Z. J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40** (10), 1458–1466. https://doi.org/10.1038/s41587-022-01284-4 (2022).
16. Gligorijevic, V., Barzel, B. & Bonneau, R. deepNF: deep network fusion for protein function prediction. *Bioinformatics* **34** (22), 3873–3881. https://doi.org/10.1093/bioinformatics/bty440 (2018).
17. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233. https://doi.org/10.1038/s41598-019-41695-z (2019).

18. Alstott, J., Bullmore, E. & Plenz, D. powerlaw: a Python package for analysis of heavy-tailed distributions. *PLoS One*. **9** (1), e85777. https://doi.org/10.1371/journal.pone.0085777 (2014).
19. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R Stat. Soc. Ser. B Stat. Methodol.* **57** (1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x (1995).
20. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296** (5569), 910–913. https://doi.org/10.1126/science.296.5569.910 (2002).
21. Yang, Y. et al. Structural studies of WDR5 in complex with MBD3C WIN motif reveal a unique binding mode that regulates chromatin accessibility. *J. Biol. Chem.* **299** (5), 107468. https://doi.org/10.1016/j.jbc.2024.107468 (2024).
22. Richmond, P. A. et al. WD Repeat Domain 5 inhibitors for cancer therapy. *J. Clin. Med.* **13** (1), 274. https://doi.org/10.3390/jcm13010274 (2024).
23. Flaus, A. et al. Identification of multiple distinct Snf2 subfamilies with conserved structural motifs. *Nucleic Acids Res.* **34** (10), 2887–2905. https://doi.org/10.1093/nar/gkl303 (2006).
24. Mondal, T. et al. MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures. *Nat. Commun.* **6**, 7743. https://doi.org/10.1038/ncomms8743 (2015).
25. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* **51** (4), 661–703. https://doi.org/10.1137/070710111 (2009).
26. Blondel, V. D. et al. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008** (10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008 (2008).
27. Song, J. et al. Targeting epigenetic regulators as a promising avenue to overcome cancer therapy resistance. *Signal. Transduct. Target. Ther.* **10** (1), 219. https://doi.org/10.1038/s41392-025-02195-5 (2025).
28. Salvati, A. et al. Essential gene screening identifies the bromodomain-containing protein BRPF1 as a new actionable target for endocrine therapy-resistant breast cancers. *Mol. Cancer*. **23** (1), 160. https://doi.org/10.1186/s12943-024-01962-8 (2024).
29. Garcia-Martinez, L. et al. Epigenetic mechanisms in breast cancer therapy and resistance. *Nat. Commun.* **12** (1), 1786. https://doi.org/10.1038/s41467-021-22048-7 (2021).
30. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **21**, 6. https://doi.org/10.1186/s12864-019-6413-7 (2020).
31. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. **10** (3), e0118432. https://doi.org/10.1371/journal.pone.0118432 (2015).
32. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R Stat. Soc. Ser. B Stat. Methodol.* **63** (2), 411–423. https://doi.org/10.1111/1467-9868.00293 (2001).
33. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R Stat. Soc. Ser. B Stat. Methodol.* **57** (1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x (1995).

Figures

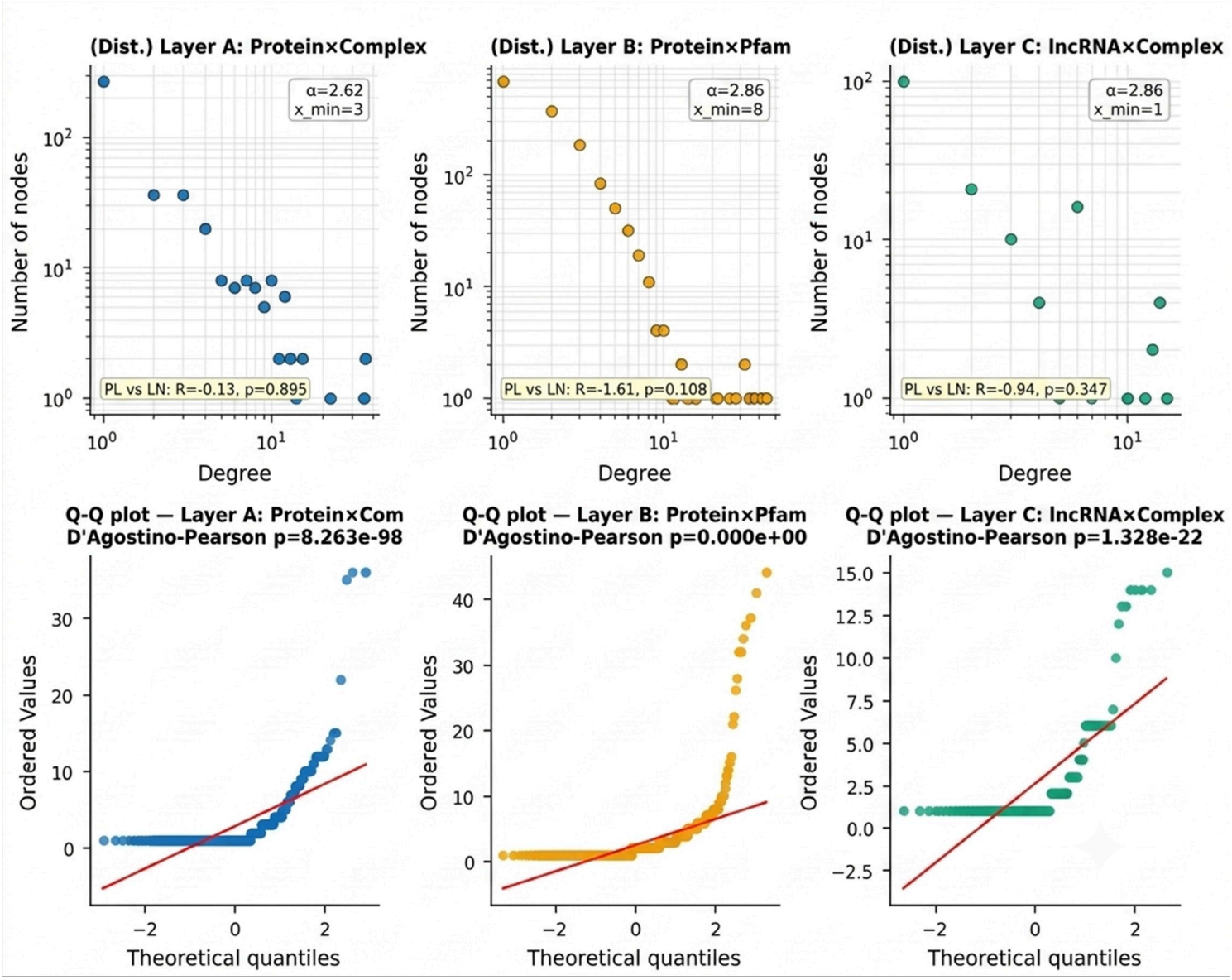

Figure 1

Degree distributions of the three EpiNet layers plotted on log–log axes. (a) Layer A (protein–complex) shows a power-law fit with α = 2.62 (x_min = 3). (b) Layer B (protein–Pfam) shows α = 2.86 (x_min = 8). (c) Layer C (lncRNA–target, raw graph) shows α = 2.86 (x_min = 1). Each point represents the number of nodes carrying a given degree.

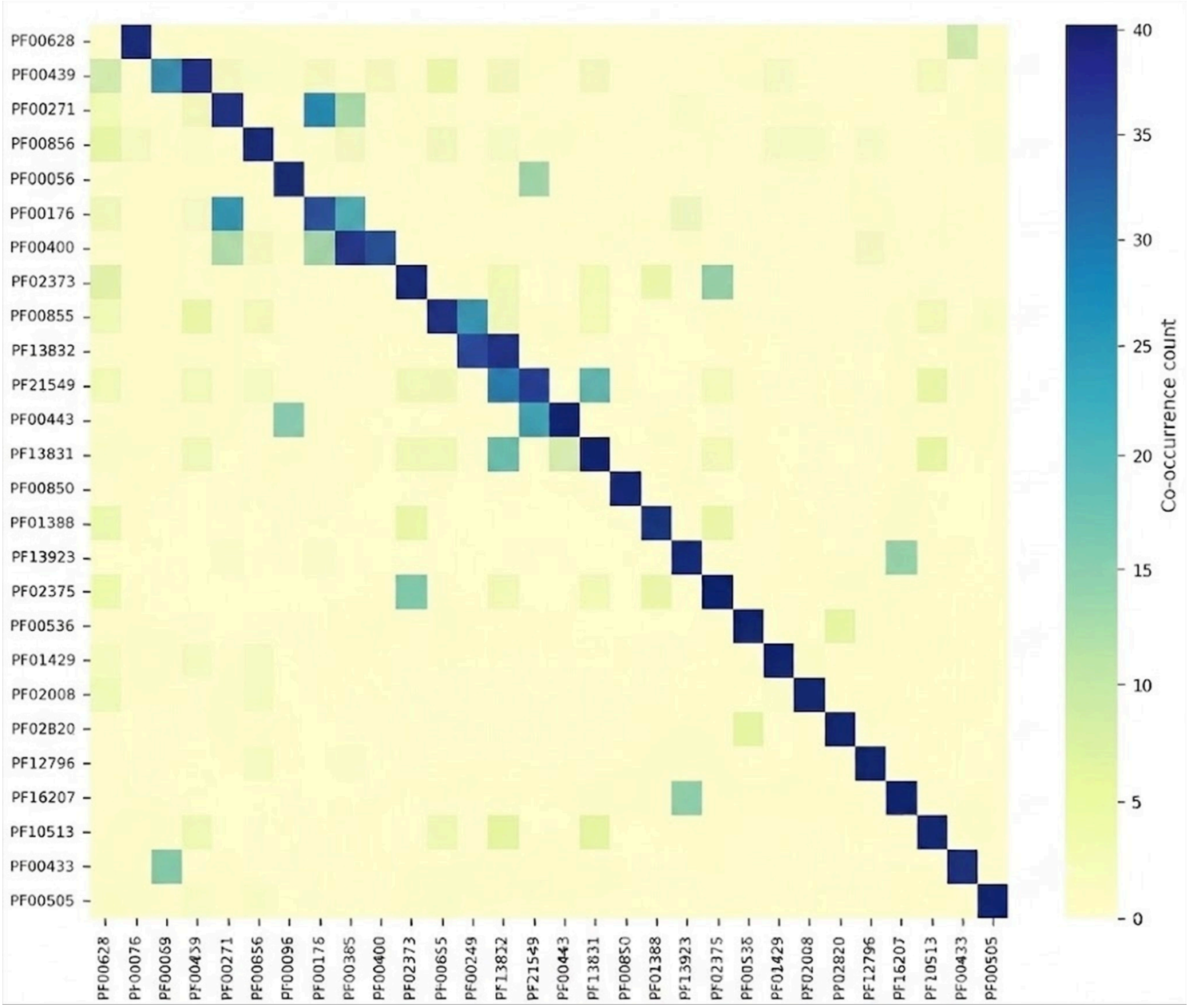

Figure 2

Pfam domain co-occurrence matrix for the top 30 most frequent domains. Colour intensity indicates the number of proteins sharing both domains (diagonal = marginal frequency). Off-diagonal bright spots correspond to recurrent architectural dyads such as PF00271 + PF00176 (helicase cassette of chromatin remodellers).

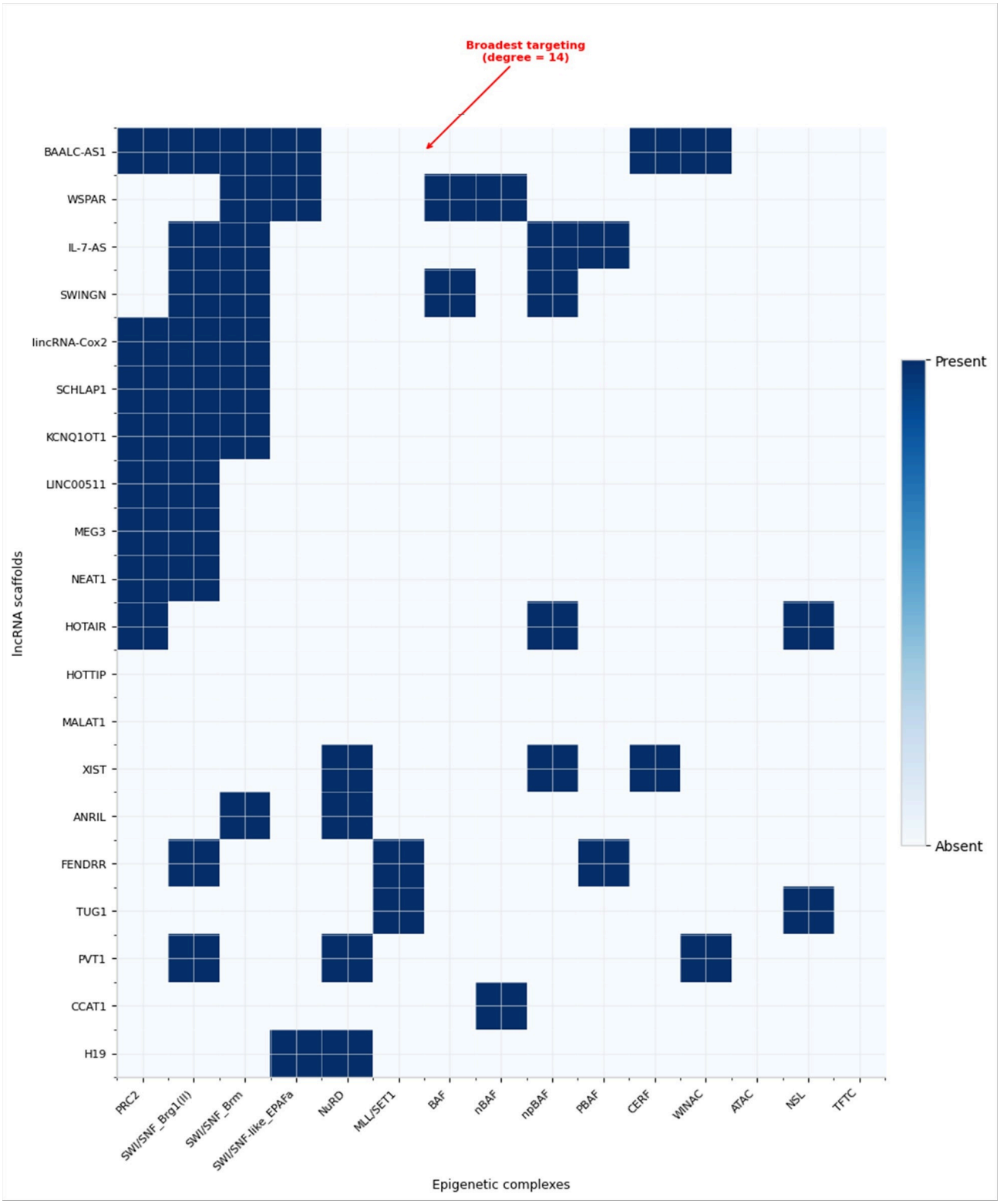

Figure 3

lncRNA–target interaction heatmap (publication-filtered Layer C). Rows: top 20 lncRNA scaffolds. Columns: top 15 epigenetic complexes ranked by degree. Blue cells indicate a documented association inferred by conservative text mining. BAALC-AS1, WSPAR and IL-7-AS show the broadest complex targeting profiles

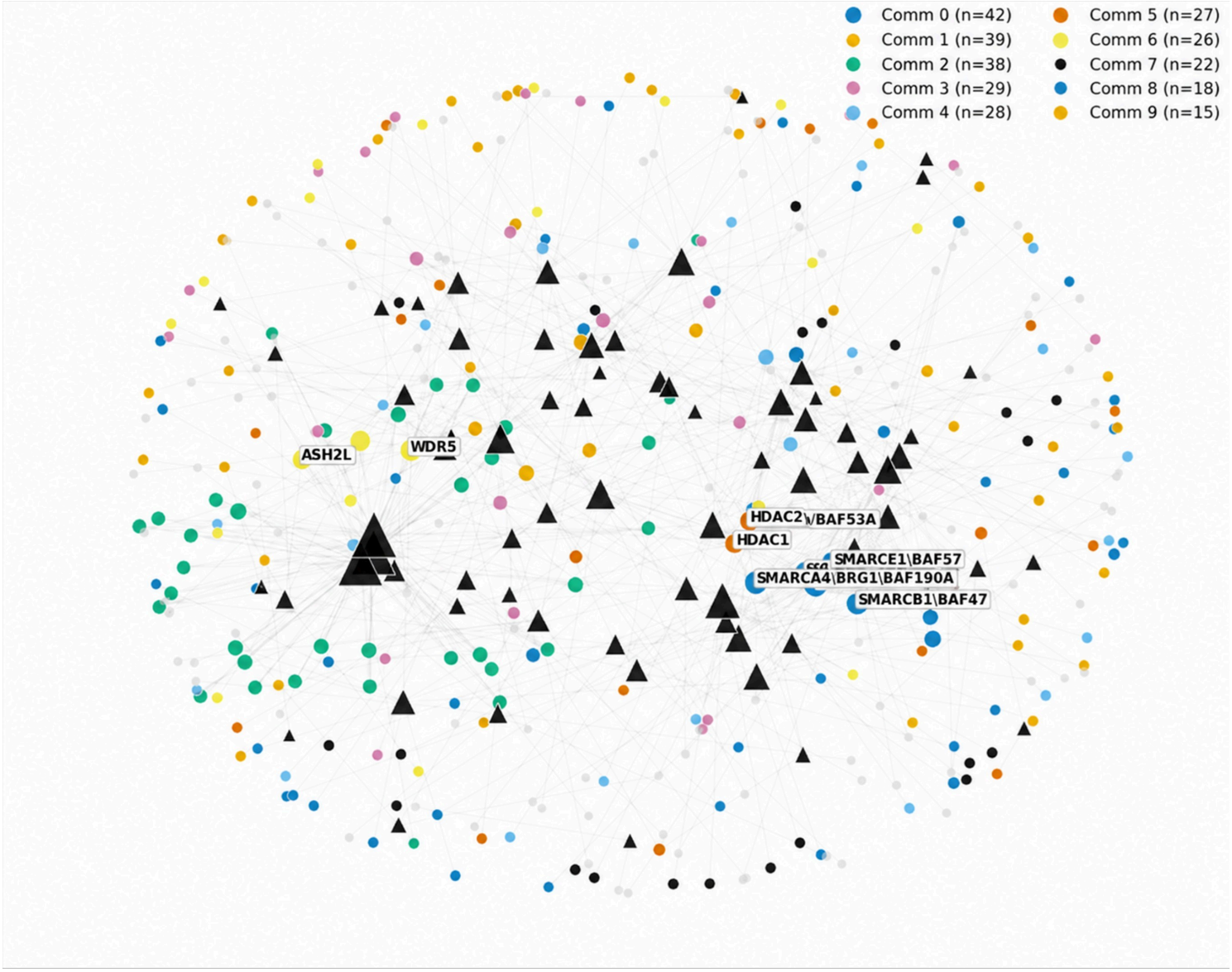

Figure 4

Force-directed layout of Layer A core (protein–complex) coloured by Leiden community (γ = 1.25). Proteins are circles (size ∝ degree), complexes are triangles (size ∝ degree). The ten largest communities are shown in distinct Okabe–Ito colours; remaining proteins are grey. Annotated labels highlight multiplex hubs: WDR5, ASH2L, HDAC1/2, SMARCA4/BRG1 and SMARCB1/BAF47

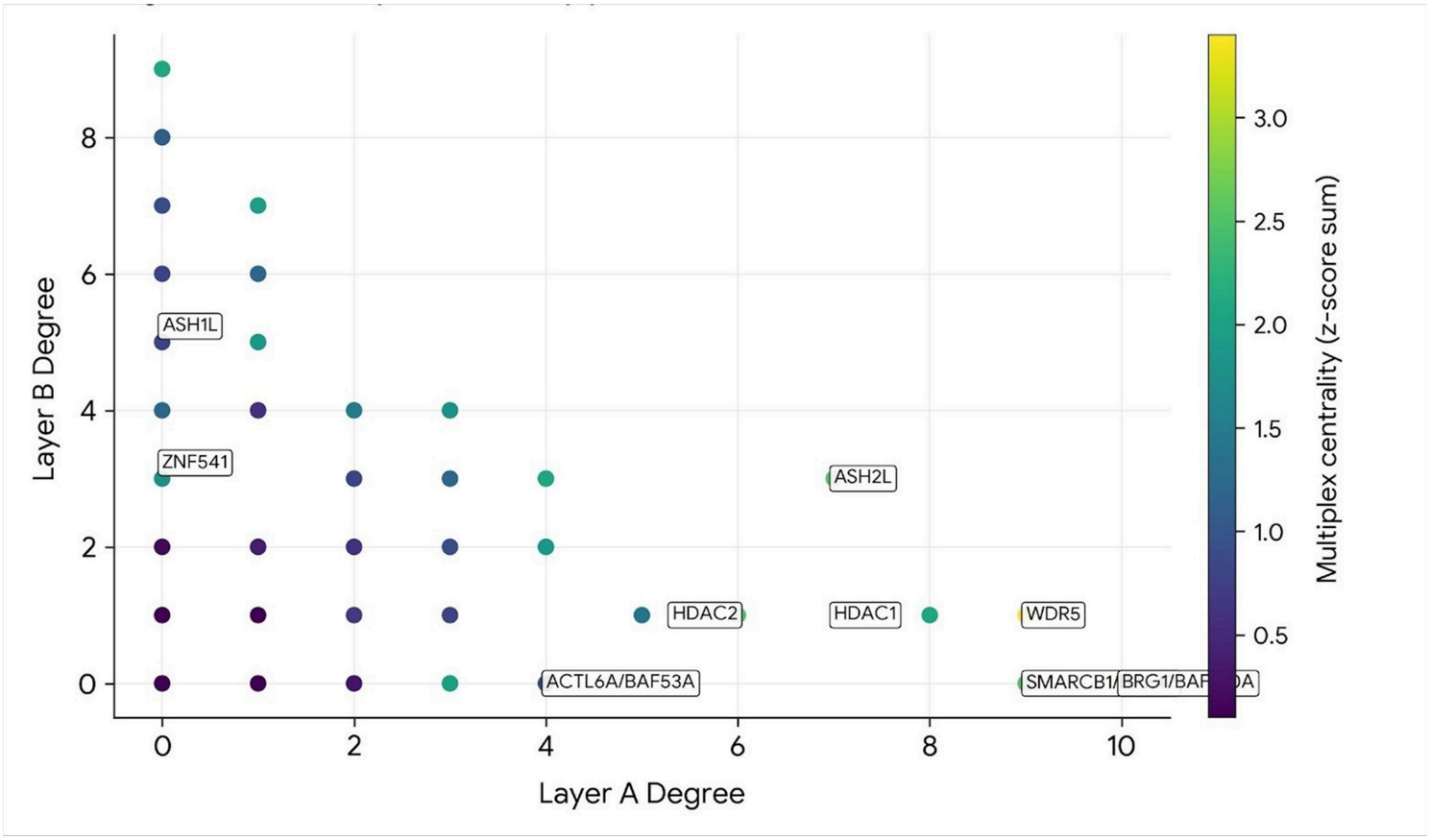

Figure 5

Multilayer centrality profile of EpiFactors proteins. Each point is a protein; x-axis = degree in Layer A (complex membership), y-axis = degree in Layer B (Pfam domains). Colour encodes multiplex centrality (z-score weighted sum). Top-ranked hubs (WDR5, ASH1L, ACTL6A/BAF53A, HDAC1/2) are labelled.

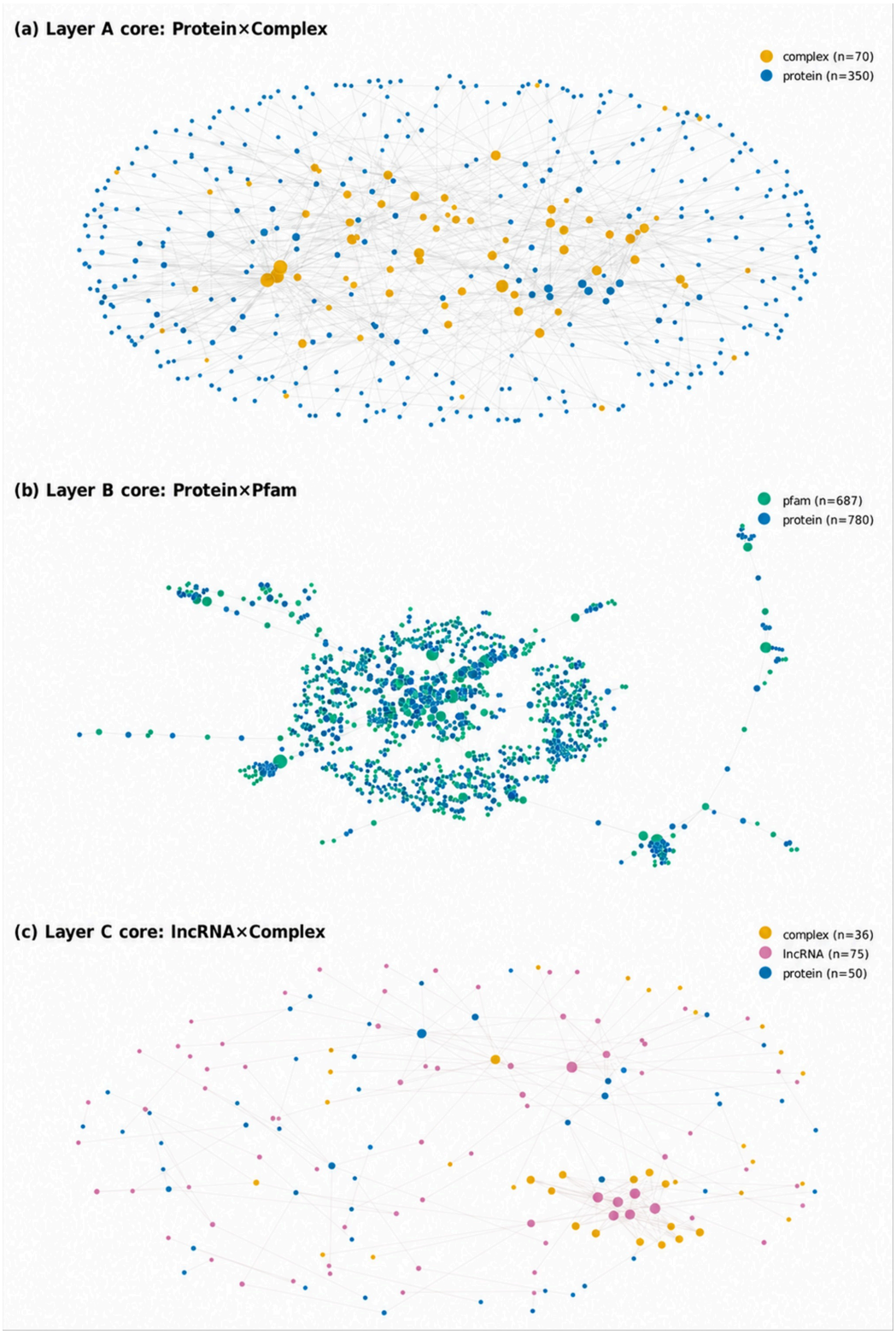

Figure 6

Integrated multilayer atlas. (a) Layer A core: protein (circles) and complex (triangles) layout. (b) Layer B core: protein (circles) and Pfam domain (squares) layout. (c) Layer C core: lncRNA (diamonds), complex (triangles) and protein (circles) layout. Node size is proportional to degree; colours distinguish node kind.

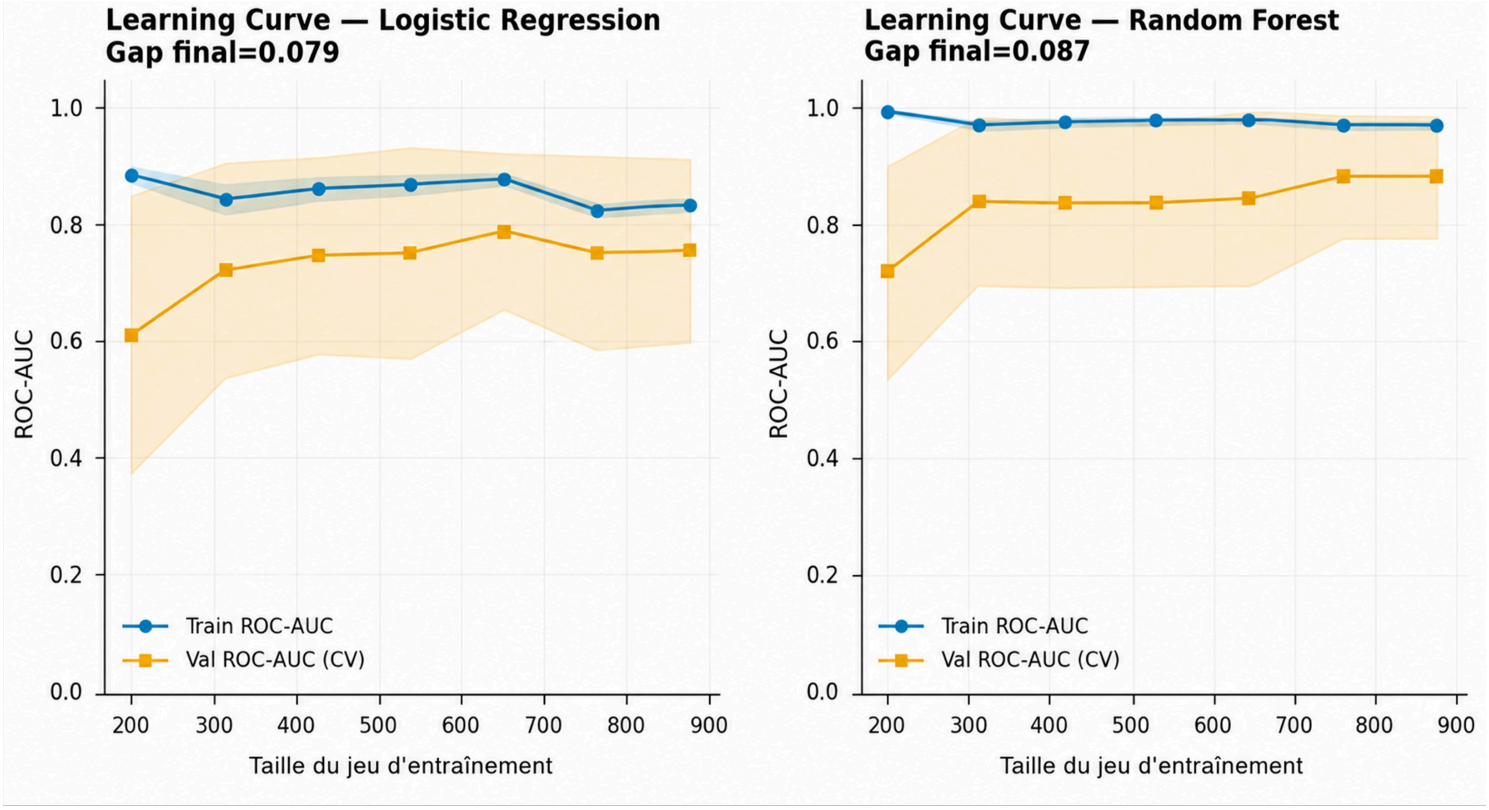

Figure 7

Learning curves for hub-classification models. (a) Logistic regression and (b) regularised Random Forest (max_depth = 5). Training ROC-AUC (blue circles) and validation ROC-AUC (orange squares) are plotted against training-set size. Shaded areas show ±1 SD across the 10 stratified folds. A train–val gap < 0.15 indicates absence of significant overfitting.

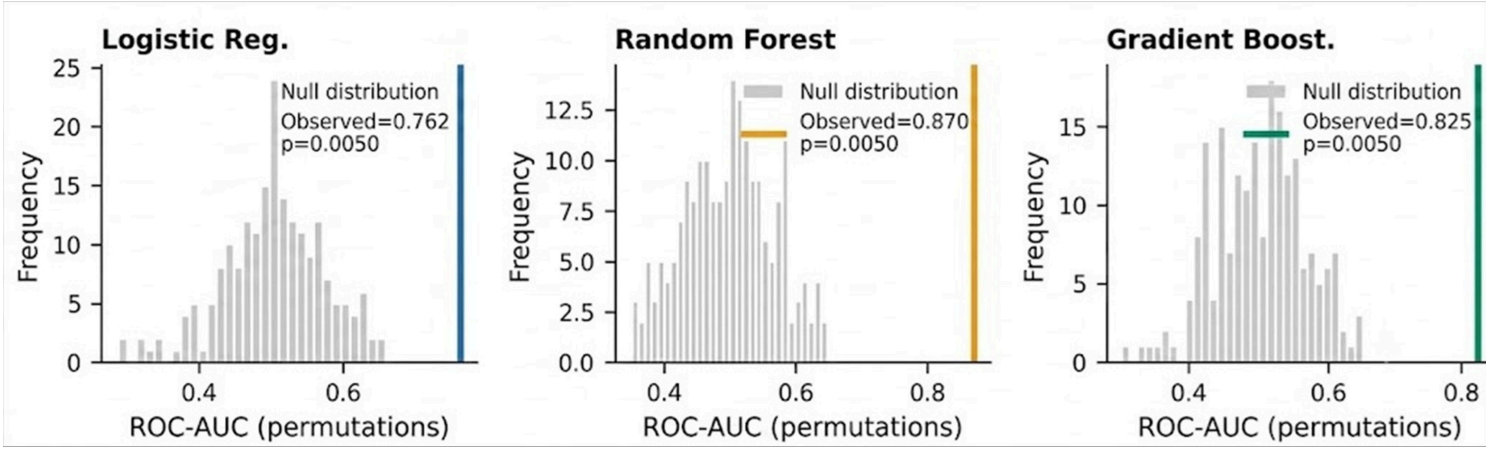

Figure 8

Permutation tests for model performance against the null distribution (n = 200 random label permutations). Null distributions of ROC–AUC scores obtained after label permutation are shown in grey. The vertical-colored line indicates the observed ROC–AUC of each model. (a) Logistic Regression: observed ROC–AUC = 0.762, p = 0.005. (b) Random Forest: observed ROC–AUC = 0.870, p = 0.005. (c) Gradient Boosting: observed ROC–AUC = 0.825, p = 0.005. In all cases, the observed performance lies outside the null distribution, indicating statistically significant predictive signal.

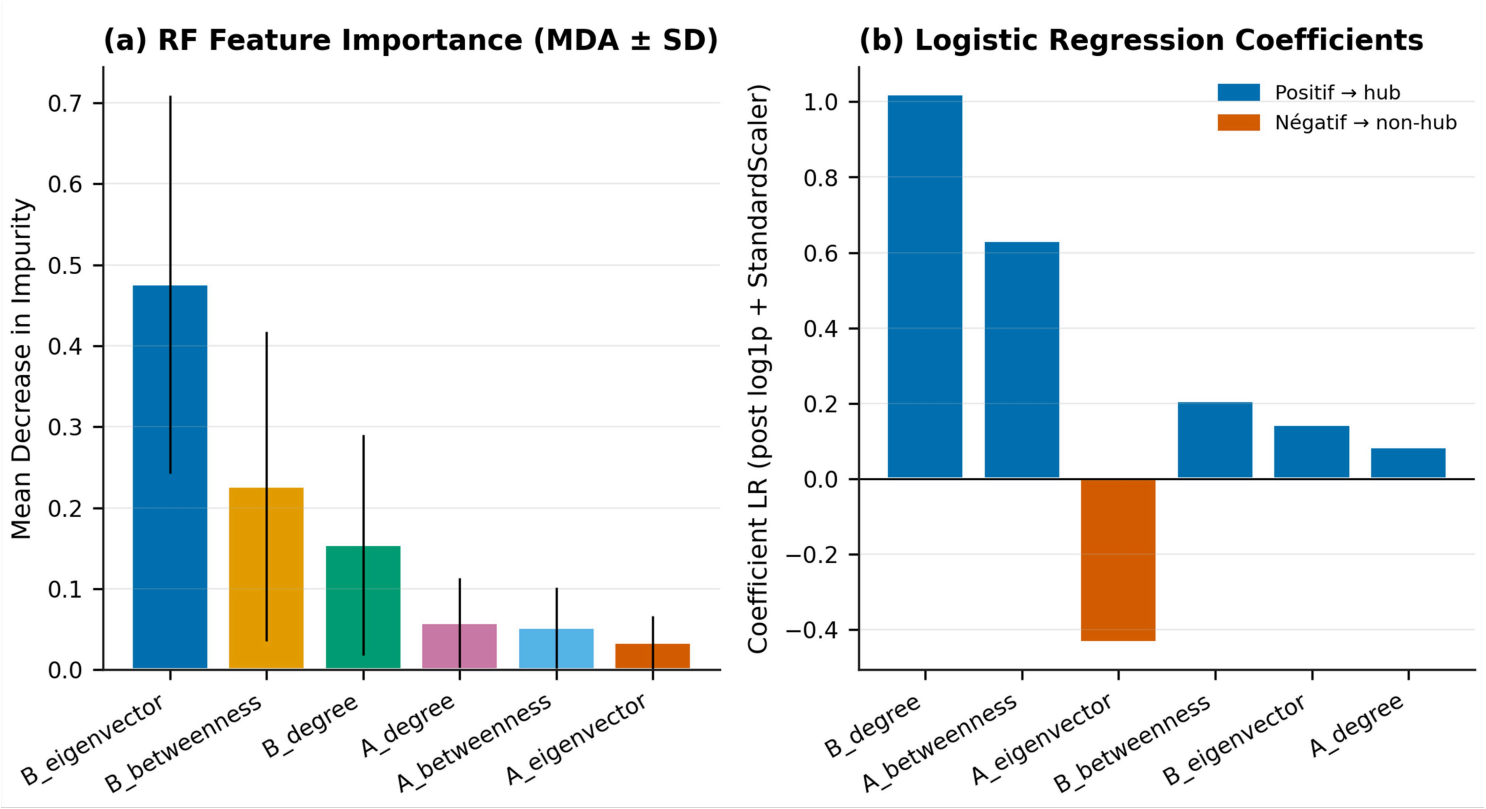

Figure 9

Feature importance and model coefficients after preprocessing. (a) Random Forest feature importance measured by Mean Decrease in Impurity (MDI). (b) Logistic Regression coefficients after log1p transformation and standardization. Positive coefficients are associated with the hub class, whereas negative coefficients are associated with the non-hub class. The figure highlights the relative contribution of degree, betweenness centrality, and eigenvector centrality metrics from both network layers.