Title

Cell type–specific functions of nucleic acid-binding proteins revealed by deep learning on co-expression networks

Naoki Osato [1*] and Kengo Sato [1]

[1] School of Life Science and Technology, Institute of Science Tokyo, Tokyo, Japan.

E-mail: naokiosato11@gmail.com

*Corresponding author.

Abstract

Nucleic acid-binding proteins (NABPs) play central roles in gene regulation, yet their functional targets and regulatory programs remain incompletely characterized due to the limited scope and context specificity of experimental binding assays. Here, we present a deep learning framework that integrates gene co-expression–derived interactions with contribution-based model interpretation to infer NABP regulatory influence across diverse cellular contexts, without relying on predefined binding motifs or direct binding evidence.

Replacing low-informative binding-based features with co-expression–derived interactions significantly improved gene expression prediction accuracy. Model-inferred regulatory targets showed strong and reproducible concordance with independent ChIP-seq and eCLIP datasets, exceeding random expectations across multiple genomic regions and threshold definitions. Functional enrichment and gene set enrichment analyses revealed coherent, cell type–specific regulatory programs, including cancer-associated pathways in K562 cells and differentiation-related processes in neural progenitor cells.

Notably, we demonstrate that DeepLIFT-derived contribution scores capture relative regulatory importance in a background-dependent but biologically robust manner, enabling systematic identification of context-dependent NABP regulatory roles. Together, this framework provides a scalable strategy for functional annotation of NABPs and highlights the utility of combining expression-driven inference with interpretable deep learning to dissect gene regulatory architectures at scale.

Keywords

Computational biology, Deep learning, Regulation of gene expression, DNA/RNA-binding proteins, Protein function prediction, Reactome, ChatGPT, GSEA

Introduction

Understanding the cell type–specific regulatory functions of nucleic acid-binding proteins (NABPs), including both DNA-binding proteins (DBPs) and RNA-binding proteins (RBPs), remains a fundamental challenge in molecular biology [1-10]. Although experimental assays such as ChIP-seq and ChIA-PET can map DBP binding sites, they require high-quality antibodies and are not easily scalable across large sets of proteins or conditions. The challenge is even greater for RBPs, many of which lack well-defined binding motifs due to their reliance on intrinsically disordered regions (IDRs) for RNA recognition, rather than sequence-specific contacts [11-13]. Interpreting low-affinity DBP binding, which contributes to enhancer activity and paralog-specific regulation, is particularly difficult—especially in the context of phase-separated transcriptional condensates [14]. Conventional methods like eQTL mapping or DNase-seq require prior motif knowledge and struggle to resolve such indirect or context-dependent interactions [15-17]. While massively parallel reporter assays (MPRAs) and large language model–based architectures such as Enformer have advanced our ability to relate sequence to function [18,19], they do not reveal which NABPs mediate the predicted regulatory effects. Thus, novel strategies are required to infer NABP–target relationships and functional roles, particularly when motif information is unavailable or direct binding evidence is missing.

Deep learning has substantially improved the prediction of gene expression by modeling regulatory inputs from NABPs, including transcription factors and microRNAs

59 (miRNAs) [20]. For example, models incorporating ChIP-seq–derived NABP binding sites
60 in promoter regions and predicted miRNA-binding sites in transcripts have achieved a
61 high correlation (r = 0.82) between predicted and observed expression levels. In a
62 previous study, we extended this approach by incorporating binding site data for DNA-
63 binding proteins within 1 Mb of transcription start sites (TSSs) in open chromatin, while
64 excluding miRNAs and RBPs. To facilitate accurate computation of importance scores
65 (also known as contribution scores) and to improve model interpretability, we unified the
66 input representation by excluding a previous input layer for RNA-binding proteins and
67 miRNAs [21]. While this modification slightly reduced predictive accuracy (r = 0.70), it
68 enabled broader coverage of distal regulatory elements and laid the foundation for
69 downstream functional analysis based on importance scores. These scores are later
70 leveraged to identify regulatory targets and infer NABP functions.

71 Gene co-expression analysis, a well-established method since the 2000s, offers a
72 complementary perspective by identifying genes with coordinated expression across
73 biological contexts [22-30]. These patterns, derived from transcriptomic datasets across
74 diverse cell types, can capture both direct and indirect regulatory associations—including
75 those involving NABPs with uncertain or difficult-to-characterize binding features. In
76 this study, we integrated gene co-expression data into a deep learning framework to assess
77 its utility for predicting gene expression and identifying context-specific regulatory
78 targets of NABPs. This combination provides a flexible means to infer functional
79 relationships even when direct binding data are lacking.

80 To test the effectiveness of this approach, we substituted the binding site features
81 of ~300 NABPs that contributed little to previous deep learning predictions with gene co-
82 expression–based interactions. The modified model successfully predicted gene

expression levels while assigning DeepLIFT-based importance scores to NABP–gene interactions. Many predicted targets overlapped with ChIP-seq or eCLIP-validated binding partners and showed higher importance scores than background interactions. In this study, we use DeepLIFT to compute attribution values for each gene. We refer to these values as DeepLIFT scores, also described as importance scores or contribution scores in the context of feature attribution. Functional enrichment analyses of twelve representative NABPs revealed consistent correspondence with known annotations in UniProt [31], and several additional functions were supported by independent publications. To enhance interpretability, we also applied a ChatGPT-based method to infer gene functions and retrieve supporting literature. The results were consistent with enrichment-based findings and highlighted biologically meaningful associations, highlighting cell type–specific processes such as neural cell differentiation, as well as cancer-related functions and pathways in neural progenitor and K562 cells.

This study provides a computational strategy to uncover the functional roles and regulatory targets of NABPs without requiring experimentally defined binding sites. By integrating gene co-expression profiles with deep learning, the framework captures both direct and indirect interactions often missed by conventional methods. This approach enables the systematic and scalable identification of NABP functions, providing a new avenue for exploring transcriptional regulation in a cell type–specific manner.

Results

Gene co-expression features enhance expression prediction accuracy

106 Replacing low-contribution DBP inputs with gene co-expression data led to a
107 substantial increase in model performance (**Fig. 1**). In human foreskin fibroblasts
108 (HFFs), substituting 302 out of 1,310 DBPs with co-expression profiles improved the
109 correlation between predicted and actual gene expression from 0.70 to 0.80 (**Fig. 2a, b**).
110 These co-expression inputs included both NABP-encoding and non-NABP mRNAs, with
111 gene symbols for DNA- and RNA-binding proteins obtained from the ENPD (Eukaryotic
112 Nucleic acid Binding Protein) database[32], and co-expression profiles sourced from
113 COXPRESdb [22]. A subsequent experiment replaced 325 low-contribution DBPs with
114 co-expression interactions involving RBPs, resulting in a further increase in prediction
115 accuracy ($r = 0.81$; **Fig. 2c**). Collectively, these findings demonstrate that integrating co-
116 expression interactions into the model enhances its ability to capture gene regulatory
117 dynamics beyond direct binding site information.

118 Contribution scores derived from co-expression–based NABP–gene interactions
119 exhibited broader distributions and higher overall magnitudes than those obtained from
120 ChIP-seq–based inputs across all four cell types examined (Wilcoxon signed-rank test, p
121 $< 2.2 \times 10^{-16}$; **Fig. 2d** and **Supplementary Fig. S2c**; **Supplementary Table S1**). Each
122 pair of violin plots represents absolute contribution scores for co-expression–based sites
123 (right plot in each pair) and the corresponding ChIP-seq sites not included in that set (left
124 plot). **A**–**P** correspond to specific cell types and NABP classes as follows: **B**, DBP co-
125 expression–based sites in neural progenitor cells (NPCs); **A**, NPC ChIP-seq sites
126 excluding **B**. **D**, RBP co-expression–based sites in NPCs; **C**, NPC ChIP-seq sites
127 excluding **D**. **F**, DBP co-expression–based sites in HepG2 cells; **E**, HepG2 ChIP-seq sites
128 excluding **F**. **H**, RBP co-expression–based sites in K562 cells; **G**, K562 ChIP-seq sites
129 excluding **H**. **J**, DBP co-expression–based sites in HMEC cells; **I,** HMEC ChIP-seq sites

excluding **J**. **L**, RBP co-expression–based sites in HMEC cells; **K**, HMEC ChIP-seq sites excluding **L**. **N**, DBP co-expression–based sites in human foreskin fibroblasts (HFFs); **M**, HFF ChIP-seq sites excluding **N**. **P**, RBP co-expression–based sites in HFFs; **O**, HFF ChIP-seq sites excluding **P**. Corresponding violin plots generated using an amplified background (input × 2) are shown in **Supplementary Figure S2 c**. However, the nature of this difference depended on the class of nucleic acid–binding protein. For DNA-binding proteins (DBPs), contribution score distributions differed significantly between ChIP-seq–derived and co-expression–derived regulatory sites, yet their mean and median values were often comparable across cell types. This indicates that, for DBPs, co-expression–based feature selection primarily reshapes the distributional structure of contribution scores without uniformly shifting their central tendency. In contrast, for RNA-binding proteins (RBPs), regulatory sites inferred from gene co-expression consistently showed substantially higher DeepLIFT contribution scores than those derived from experimental binding data alone. This pattern was observed across all examined cell types, including K562, NPCs, HMECs, and HFFs, and was reflected by marked increases in both mean and median contribution scores. These results suggest that co-expression–based inference preferentially enriches for RBP-associated regulatory interactions that contribute more directly to gene expression prediction, whereas experimental binding data for RBPs may include a larger fraction of interactions that are weakly coupled to transcriptomic output. These observations are consistent with previous reports showing that incorporating RNA-binding protein– and miRNA-associated features improves gene expression prediction accuracy, as exemplified by the DEcode framework [20]. Importantly, these trends were robust to DeepLIFT background definition: comparable differences between ChIP-seq– and co-expression–derived targets

were observed under both background scaling schemes (input × 0.5 and input × 2). Thus, while background scaling influences the sign and absolute scale of contribution scores, it does not alter the fundamental distinction between binding-based and co-expression–based regulatory inference.

To assess the influence of background selection on DeepLIFT-derived contribution scores, we compared results obtained using a reduced background (input × 0.5; **Fig. 2e**) and an amplified background (input × 2; **Fig. 2f**) for the same representative DNA- and RNA-binding proteins. Under the reduced background, contribution scores tended to show negative correlations with gene expression, whereas the amplified background produced predominantly positive correlations. These patterns appeared approximately mirrored around zero, consistent with DeepLIFT's definition as a relative contribution measure with respect to the chosen reference. Importantly, despite this inversion in sign, the relative ranking of genes and the overall cell type–specific structure of contribution score distributions remained highly consistent across background settings. In both cases, genes with larger-magnitude contribution scores were preferentially associated with higher or lower expression levels depending on the background, indicating that DeepLIFT captures relative regulatory importance rather than absolute activation or repression. A conceptual model explaining how input definition and background scaling give rise to these patterns is provided in **Supplementary Fig. S4**. Notably, background-dependent differences were more pronounced for RNA-binding proteins (RBPs). For the same representative RBPs, scatter plots generated using the reduced background (input × 0.5) exhibited more distinct, partially separated distributions across the four cell types, whereas under the amplified background (input × 2), data points from different cell types overlapped more extensively. This suggests that the reduced-background configuration

178 enhances sensitivity to cell type–specific regulatory variation for RBPs. In contrast,
179 DNA-binding proteins (DBPs) generally showed stronger correlations with gene
180 expression under both background conditions, consistent with their more direct
181 transcriptional regulatory roles. Together, these results underscore the importance of
182 background-aware and protein class–specific interpretation of DeepLIFT contribution
183 scores.
184

185 **Experimental binding data support model-inferred regulatory targets**

186 To determine whether DeepLIFT-derived regulatory contribution scores reflect
187 biologically meaningful nucleic acid–binding protein (NABP) activity, we systematically
188 compared model-inferred regulatory targets with experimentally validated ChIP-seq and
189 eCLIP binding data across multiple genomic regions, threshold definitions, and cellular
190 contexts (**Fig. 3**). Experimentally validated binding sites are consistent with deep
191 learning–derived regulatory targets. Genes with ChIP-seq or eCLIP peaks for a given
192 NABP exhibited disproportionately high or low importance ranks, indicating that
193 experimentally confirmed binding sites are associated with stronger regulatory influence
194 in gene expression prediction models. This relationship supports the biological validity
195 of DeepLIFT-based contribution scores, which quantify the importance of specific
196 NABPs in driving gene expression changes.

197 To evaluate this association, we analyzed ChIP-seq data for 32 DBPs in HepG2
198 cells and eCLIP data for 29 RBPs in K562 cells—two cell lines with the most
199 comprehensive publicly available NABP binding site datasets from ENCODE and related
200 studies. A gene was considered associated with an NABP if a ChIP-seq peak was located
201 within ±5 kb of its transcription start site (TSS) or if an eCLIP peak was mapped to its

transcribed region. As illustrated in **Fig. 3a** and **Supplementary Fig. S1**, NABPs with experimentally validated binding sites consistently showed higher absolute contribution score values across their predicted target genes, reinforcing their regulatory relevance and corresponding with prior interpretations of DeepLIFT scores [20,33].

Model-derived importance scores reliably predict DBP binding sites, as evidenced by significant overlap with ChIP-seq peaks (**Fig. 3b, Supplementary Table S2** and **S3**). To evaluate this correspondence, we analyzed ChIP-seq data for 32 DNA-binding proteins (DBPs) and quantified how frequently high-scoring gene regions—defined by DeepLIFT contribution scores—coincided with experimental binding evidence. Overlap ratios were calculated as the proportion of high-contribution regions that intersected with ChIP-seq peaks, considering both promoter-proximal (±5 kb) and distal (±100 kb) regions around transcription start sites (TSSs). To account for the functional heterogeneity and abundance of ChIP-seq peaks, we applied a precision-at-K approach, focusing on the top-ranked or top-scoring gene regions most likely to reflect regulatory relevance. Four thresholds were used to define high-contribution regions: two based on contribution score rank (top/bottom 20 or 50) and two based on absolute score magnitude (≥ 0.01 or ≤ -0.01; ≥ 0.001 or ≤ -0.001). A control analysis using randomly selected gene regions of matched size was conducted in parallel. In all tested conditions, the overlap between ChIP-seq peaks and high-contribution regions exceeded that of the random control, supporting the model's ability to prioritize functionally relevant regulatory sites.

Similarly, as illustrated in **Fig. 3c** and **Supplementary Table S4**, the model-derived importance scores consistently prioritize biologically relevant RBP–target interactions, as evidenced by the cumulative overlap between high-contribution gene regions and eCLIP peaks, which was substantially greater than that observed in the

226 random control. We assessed this by analyzing eCLIP data for 29 RNA-binding proteins
227 (RBPs) in K562 cells, comparing high-contribution regions—defined using the same
228 thresholds as for DBPs—to eCLIP peaks mapped to transcribed regions (from TSS to
229 TES). Collectively, these results demonstrate that the model effectively captures
230 functionally meaningful NABP–target associations for both DNA- and RNA-binding
231 proteins.

232 Significant differences in the predicted regulatory targets of several NABPs were
233 observed between K562 and HepG2 cells based on DeepLIFT contribution scores (**Fig.
234 3d**). To assess cell type–specific regulatory activity, we compared the DeepLIFT scores
235 of genes harboring NABP binding sites, as defined by ChIP-seq or eCLIP. Because our
236 deep learning model identifies target candidates from NABP-associated co-expressed
237 gene sets, not all predicted targets necessarily reflect direct binding; rather, they represent
238 genes that participate in shared or downstream regulatory programs. RPLP0 exhibited
239 nearly identical score distributions across the two cell types, whereas ETF1, NELFE, and
240 CAMTA2 showed marked divergence. These differences indicate that the model captures
241 and adjusts NABP-dependent regulatory relationships in a cell type–specific manner.

242

243 Deep learning uncovers functionally enriched NABP target genes 244 beyond co-expression networks

245 Deep learning predictions revealed that nucleic acid–binding proteins (NABPs)
246 possess regulatory target genes with coherent and biologically specific functional
247 enrichment, extending well beyond patterns detectable from co-expression alone (**Tables
248 1, 2** and **Supplementary Note**). Across twelve representative NABPs [34,35], predicted
249 target sets consistently showed significant enrichment for biologically related Gene

250 Ontology (GO) terms, whereas randomly sampled gene sets of matched size displayed no
251 such enrichment (**Fig. 4**). Notably, several enriched annotations were absent from the
252 broader pools of co-expressed genes (Pearson's *z*-score > 2), indicating that the model
253 selectively extracts biologically relevant signals from large co-expression neighborhoods.
254 Functional enrichment across GO Slim, Reactome, and PANTHER Pathways further
255 confirmed this specificity (**Table 1**) [36]. A ChatGPT-based gene set interpretation method
256 independently supported these results, providing functional summaries and literature
257 consistent with GO enrichment and UniProt annotations [31] (**Table 2** and **Supplementary**
258 **Note**) [37].
259

260 Cell type–dependent functional specificity revealed by deep learning–
261 refined NABP targets

262 Deep learning refinement further exposed cell type–specific biological functions
263 associated with regulatory targets of several RBPs. For example, the predicted targets of
264 PKM showed distinct pathway enrichments between K562 cancer cells and neural
265 progenitor cells (NPCs). Pathways such as *Pentose phosphate pathway*, *Glycolysis*, and
266 *Hypoxia response mediated by HIF activation* were significantly enriched in K562 but
267 not in NPCs, whereas *Integrin signaling pathway*, *Glycogen metabolism*, and *Glutamate*
268 *neurotransmitter release cycle* were specifically enriched in NPCs using a background
269 defined as 0.5× the original input feature values[38-47] (**Fig. 5**, **Table 3**, **Supplementary**
270 **Table S5**). ChatGPT-based interpretation produced consistent functional assignments
271 (**Supplementary Note**). Notably, these cell type–dependent associations align with
272 known biology in cancer cells and NPCs.

273 Additional RBPs exhibited similar specificity. Predicted target genes of ILF2 were
274 enriched for *G2/M DNA replication checkpoint* in K562 cells but not in NPCs[48-50] ; targets
275 of EIF2S2 were associated with *Serine and glycine biosynthesis* in NPCs but not in
276 K562[51,52] ; and targets of DNAJC2 were enriched for the *Ubiquitin proteasome pathway*
277 in K562 but not in NPCs[53,54] **(Fig. 5**, **Table 3**, **Supplementary Table S5**). ChatGPT-
278 based inference again recapitulated these cell type–dependent functional patterns
279 (**Supplementary Note**), consistent with previous reports.
280 Together, these results demonstrate that the deep learning–based refinement of
281 NABP target predictions enables systematic detection of cell type–dependent functions,
282 revealing regulatory specificity not captured by co-expression alone.
283
284

285 Functional enrichment analysis of NABP target genes using 286 contribution score–based GSEA

287 Gene set enrichment analysis (GSEA) revealed robust, cell type–dependent functional
288 differences among NABP regulatory target genes, providing clearer biological insight
289 than threshold-based enrichment approaches. Rather than ranking genes by expression
290 levels, we ranked predicted target genes according to their DeepLIFT contribution scores,
291 enabling functional interpretation based on relative regulatory importance within each
292 cell type. This framework allows pathway enrichment to be assessed by shifts in the
293 positioning of functionally related genes along the ranked list, independent of absolute
294 score magnitude or sign. Using this contribution score–based GSEA, we identified
295 pronounced differences in PKM-associated pathway enrichment between neural
296 progenitor cells (NPCs) and K562 leukemia cells (**Fig. 6**). Notably, multiple Reactome

297 pathways exhibited large differences in normalized enrichment scores ($\Delta NES = NES_{K562}$
298 $- NES_{NPC}$), indicating substantial divergence in the inferred regulatory associations of
299 PKM across these two cellular contexts. For example, pathways related to synaptic
300 organization and neural functions, such as Protein–protein interactions at synapses,
301 showed strong enrichment differences between NPCs and K562, whereas pathways
302 associated with immune signaling and cancer-related processes were preferentially
303 enriched in K562 cells. These contrasting patterns were consistently captured across
304 multiple pathways ranked by ΔNES (**Table 4**). Importantly, because DeepLIFT
305 contribution scores depend on the choice of background reference, the absolute direction
306 of enrichment (positive or negative NES) may invert under different background
307 definitions. Accordingly, we focused on relative differences in NES between cell types
308 rather than the absolute sign of enrichment. Consistent pathway-level differences between
309 NPCs and K562 were observed under both background configurations (input × 0.5, **Fig.**
310 **6**; input × 2, **Supplementary Table S8**), supporting the robustness of cell type–specific
311 functional contrasts inferred from PKM contribution scores. Notably, the amplified
312 background configuration (input × 2) tended to yield a larger number of genes with non-
313 negligible contribution scores, resulting in stronger overall enrichment scores and higher
314 NES values across pathways (**Supplementary Tables S7** and **S9**). This effect reflects
315 increased sensitivity in detecting PKM-associated regulatory targets under the input × 2
316 background, while preserving the relative enrichment differences between cell types.
317 Together, these results demonstrate that GSEA based on deep learning–derived
318 contribution scores effectively captures cell type–specific NABP functional associations
319 that are not readily detected by expression-based or threshold-dependent analyses. By
320 emphasizing relative enrichment differences across cellular contexts, this approach

321 provides a robust framework for dissecting context-dependent regulatory programs and
322 linking NABPs to biologically meaningful pathways.
323
324
325 **Discussion**
326 These findings address a long-standing challenge in molecular biology: the
327 systematic functional characterization of nucleic acid-binding proteins (NABPs) across
328 diverse cellular contexts. While traditional methods rely heavily on experimental binding
329 data and known motifs, our approach utilizes gene co-expression and machine learning
330 to circumvent these limitations. By capturing both direct and indirect regulatory signals,
331 it broadens the scope of functional prediction for NABPs that lack conventional binding
332 site information.
333 Gene expression correlations alone can serve as a powerful basis for uncovering
334 the regulatory roles of nucleic acid-binding proteins (NABPs), even in the absence of
335 experimental binding data. Leveraging transcriptomic profiles from diverse public RNA-
336 seq datasets, our deep learning framework inferred regulatory targets of NABPs based
337 solely on co-expression patterns. Unlike binding site assays such as ChIP-seq or eCLIP—
338 which are technically demanding and limited by the availability of high-quality
339 antibodies—this computational approach is scalable and applicable across NABPs with
340 poorly characterized binding motifs. Functional enrichment analysis of the predicted
341 target genes revealed biologically meaningful roles, many of which were consistent with
342 existing annotations in public databases. In addition, several previously unannotated
343 functional associations—not captured in UniProt or GO but supported by published
344 studies—were uncovered. GSEA further demonstrated that genes contributing to specific

345 cell type–associated pathways exhibited systematic shifts toward higher or lower
346 DeepLIFT scores, indicating differential regulatory influence across cellular contexts.
347 These findings were corroborated by a ChatGPT-assisted literature mining approach,
348 which independently recovered concordant biological themes and supporting references.
349 Together, these results underscore the biological relevance of NABP target predictions
350 derived from deep learning–based co-expression modeling and highlight the value of
351 integrating such predictions with AI-assisted interpretation. This strategy provides a
352 scalable framework for elucidating NABP functions and discovering previously
353 unrecognized regulatory mechanisms across diverse cell types.

354 Contribution scores derived from gene expression correlations capture both direct
355 and indirect regulatory interactions, offering insights beyond what ChIP-seq alone can
356 provide. While ChIP-seq detects thousands of DNA-binding events per protein, many of
357 these sites do not contribute directly to gene expression regulation and may reflect the
358 multifunctionality of transcription factors [55,56]. Consequently, ChIP-seq–derived binding
359 maps and DeepLIFT-based predictions show only partial overlap, underscoring both
360 methodological differences and the inherent complexity of regulatory networks. In this
361 study, gene co-expression analysis revealed fewer target genes than ChIP-seq but
362 identified subsets that are likely to be more functionally relevant. Our previous work on
363 insulator-associated factors linked such proteins to enhancer–promoter insulation and
364 chromatin organization, including phase separation and loop formation [21]. That study
365 incorporated eQTL data to anchor binding sites to expression outcomes. Here, by contrast,
366 we focused on NABPs with unknown motifs and excluded motif-based or chromatin
367 accessibility data such as DNase-seq or ATAC-seq, to assess the capacity of gene
368 expression correlations to reveal regulatory roles. Notably, the distribution of

contribution scores for co-expression–based inputs included both high and low values, implying that gene expression patterns reflect not only direct binding but also indirect regulation mediated through broader network effects. This highlights the value of contribution scores as indicators of both proximal and distal regulatory influences, even when binding motifs are unknown or experimental binding data are unavailable. The cell type–specific differences in DeepLIFT contribution score distributions observed in **Fig. 3d** suggest that the regulatory roles of individual nucleic acid-binding proteins are not fixed but are dynamically reshaped across cellular contexts. Notably, while some proteins such as RPLP0 exhibited highly conserved contribution patterns between HepG2 and K562 cells, others, including ETF1 and NELFE, displayed pronounced divergence, indicating cell type–dependent shifts in their inferred regulatory influence. These findings support the notion that NABPs participate in distinct regulatory programs depending on cellular state and lineage, and that DeepLIFT-based analyses can sensitively capture such context-dependent regulatory reorganization. This context dependence is particularly relevant for interpreting NABP function in disease versus developmental states, where subtle shifts in regulatory importance may precede overt transcriptional changes.

Functional roles of NABPs can be inferred even from small gene sets, provided that the target genes exhibit shared or functionally related patterns. In this study, twelve NABPs showed functional enrichments consistent with known annotations, supporting the validity of the deep learning–based predictions. These enrichments were identified using Fisher's exact test with false discovery rate (FDR) correction via the PANTHER database, which typically requires at least 50 genes for robust statistical significance. For proteins with fewer than 50 predicted targets, enrichment may be harder to detect depending on the functional composition of the gene set. Relaxing the ranking thresholds

increased the number of target genes and enhanced statistical power. Alternatively, more stringent thresholds could be retained if the contribution scores were sufficiently strong, offering flexibility depending on analysis goals. Additional features—such as gene expression levels or co-expression correlations—could further refine enrichment analyses, particularly for borderline cases. Notably, enrichment analysis of both high- and low-ranking targets revealed candidate novel functions for specific NABPs, suggesting that genes with large-magnitude contribution scores are biologically informative. Moreover, a complementary ChatGPT-based approach demonstrated utility in interpreting small gene sets. Unlike standard enrichment tools, this method can infer biological roles even when gene counts fall below conventional significance thresholds, provided that genes with related functions are present. This capacity helps overcome statistical limitations and broadens the interpretability of regulatory predictions in a cell type–specific context. GSEA based on DeepLIFT scores provides a complementary view of pathway-level regulation by assessing how genes associated with the same pathway are distributed along the model's importance landscape, rather than relying on gene expression levels. This approach revealed striking differences between normal and disease-associated cell types. For example, genes enriched in neural progenitor cell (NPC) pathways tended to cluster at higher or lower DeepLIFT scores in NPCs, whereas their score distributions shifted substantially in K562 cells, indicating altered regulatory programs between the two cell states. Such analyses offer a powerful means to uncover context-dependent regulatory mechanisms across diverse cell types, conditions, and developmental stages.

Strategic use of co-expression data enhances functional inference, while indiscriminate replacement can diminish biological relevance. The selective integration of gene co-expression data improved target prediction and the functional interpretability

417 of the model. In contrast, replacing all input features with co-expression-derived
418 interactions led to a decrease in enriched annotations, and known functional roles were
419 no longer recovered. Excessive replacement of input features with gene co-expression
420 data further diminished the biological relevance of model predictions. When all NABP–
421 target inputs were substituted with co-expression-based interactions, the number of
422 enriched functional annotations declined, and previously identified functions from the
423 UniProt database were no longer recovered. This suggests that indiscriminate
424 replacement may cause the model to capture non-informative or biologically spurious
425 associations. A similar trend was observed in a prior control analysis, where substituting
426 true regulatory targets with randomly selected genes led to a gradual loss of enrichment
427 as randomness increased. Based on these observations, the normalized count of enriched
428 functional terms—adjusted by the number of predicted targets—may serve as a useful
429 proxy for evaluating the accuracy and biological validity of NABP–target predictions [57].
430 These findings also raise intriguing questions about how the deep learning model
431 prioritizes informative input features during training. Future work exploring the model's
432 internal representation dynamics—such as how it distinguishes signal from noise when
433 presented with diverse regulatory interactions—may provide valuable insights into the
434 mechanisms of transcriptional regulation and model interpretability. Investigating how
435 the model converges on biologically meaningful NABP–target relationships during
436 training could deepen our understanding of both gene regulation and model
437 interpretability.

438 Cell type specificity and cooperative binding dynamics may underlie the absence
439 of detectable regulatory signals for certain NABPs. While some NABPs showed no
440 high- or low-ranking target genes in HepG2 or K562 cells, such absence does not rule out

441 regulatory relevance. These proteins may exhibit cell type–specific regulatory activity
442 detectable only in other cellular contexts, where functional enrichment analysis could
443 reveal distinct biological roles. Moreover, many transcription factors and DNA-binding
444 proteins act in concert, forming multi-protein complexes that bind cooperatively to shared
445 promoters [58-60]. In such cases, individual components may not always yield strong
446 contribution scores due to their distributed regulatory influence across the complex. This
447 highlights the potential of co-binding site analysis as a complementary strategy to uncover
448 cooperative regulatory mechanisms and to disentangle the functional interplay of NABPs
449 within transcriptional networks.

450 Beyond characterizing cell type–specific regulatory programs, our framework also
451 suggests a potential application to preventive and health-oriented research. By
452 systematically scoring "cleaning" pathways—including proteostasis, DNA damage repair,
453 RNA quality control, immune surveillance, and metabolic stress responses—based on the
454 predicted regulatory influence of nucleic acid–binding proteins, it becomes possible to
455 estimate the integrity of cellular maintenance systems before overt disease onset. Such
456 pathway-level scores may capture early functional weakening of protective mechanisms
457 that are not yet reflected by pathological phenotypes or single-gene alterations. This
458 perspective aligns with observations from cancer-resistant species and long-lived
459 individuals, in which enhanced cellular maintenance, quality control, and stress
460 adaptation are thought to contribute to disease avoidance rather than treatment. In this
461 context, deviations in cleaning pathway scores could serve as quantitative indicators of
462 declining cellular robustness, potentially enabling risk stratification or early detection of
463 disease predisposition. While the present study focuses on neural progenitor and cancer
464 cell models, the same analytical framework could, in principle, be extended to

longitudinal human data or population-scale transcriptomic profiles, providing a foundation for preventive or health-monitoring applications that complement conventional disease-centered analyses.

This study establishes a scalable computational framework for uncovering the functional roles of nucleic acid-binding proteins (NABPs) by integrating gene expression–based associations with deep learning–derived contribution scores. By leveraging co-expression–driven interactions rather than relying solely on experimentally defined binding motifs, the approach enables systematic inference of both canonical and previously uncharacterized regulatory relationships across cell types. Collectively, our results demonstrate that expression-driven contribution patterns, when coupled with pathway-level analyses and AI-assisted functional interpretation, can capture biologically meaningful differences between normal and disease-associated cellular states. This capability suggests that the framework may be particularly valuable for identifying early shifts in regulatory and cellular maintenance pathways that precede overt disease phenotypes, thereby complementing traditional binding-centric or mutation-focused analyses. Although the present study focused on a representative subset of NABPs and cell types, the framework is readily extensible to larger protein cohorts, additional cell states, and diverse transcriptomic datasets. In this way, it provides a foundation for systematic functional annotation of NABPs, including those with poorly characterized roles, and offers a path toward integrating transcriptomic signals and AI-driven predictions into large-scale functional genomics and disease-relevant research.

Methods

Predictive modeling of gene expression across diverse cell types using deep learning and NABP binding information

To predict gene expression across multiple human cell types, we adapted a previously established deep learning framework that integrates DNA-binding site data for transcription factors and other regulatory proteins. In this implementation, regulatory elements located both proximally and distally (up to ±1 Mb) from transcription start sites (TSSs) were included as input features (**Fig. 1**) [20,21]. To minimize the effects of cellular heterogeneity in bulk tissue, RNA-seq datasets were selected at the individual cell-type level. Publicly available RNA-seq data were retrieved for four functionally distinct cell types: human foreskin fibroblasts (HFFs), mammary epithelial cells (HMECs), neural progenitor cells (NPCs), and either HepG2 hepatoma cells or K562 leukemia cells, depending on the class of regulatory protein analyzed [61,62]. Gene expression values (FPKM) were computed using the RSeQC toolkit based on GENCODE v19 annotations, and subsequently $\log_2$-transformed after adding a small offset to avoid undefined values. Following the approach described by Tasaki et al. [20,21], expression fold changes were calculated relative to the median expression level across the four cell types. For each model iteration, three common cell types (HFFs, HMECs, and NPCs) were used, and either HepG2 (for DNA-binding protein analysis) or K562 (for RNA-binding protein analysis) was included to match the availability of ChIP-seq and eCLIP validation datasets. These cell types were chosen to represent diverse biological lineages and regulatory contexts. For specific analyses requiring direct comparison of DeepLIFT contribution scores across cell types—such as the evaluation of cell type–dependent regulatory differences for the same nucleic acid-binding protein (e.g., **Fig. 3d**)—models were trained using a four-cell-type configuration comprising HFFs, NPCs, HepG2, and

513 K562. This unified configuration enabled direct comparison of contribution score
514 distributions between HepG2 and K562 while maintaining a shared background of
515 common cell types.
516

517 Extended regulatory features and DeepLIFT-compatible model design

518 To capture distal regulatory signals beyond promoter regions, we integrated DNA-
519 binding site data from putative enhancers into a deep learning framework for gene
520 expression prediction. Building upon our earlier work [20,21], this model extends the
521 architecture originally proposed by Tasaki et al. (2020) [20,21], which focused on promoter-
522 proximal regions and included RNA-binding protein and miRNA inputs. In our previous
523 study, aimed at identifying insulator-associated DNA-binding proteins, we deliberately
524 excluded RNA-binding proteins and miRNAs to focus on DNA-centered regulatory
525 elements. Instead, we combined DNA-binding sites from both promoter (-2 kb to +1 kb)
526 and enhancer-like regions (±1 Mb from the transcription start site [TSS]) into a unified
527 input matrix, enabling DeepLIFT analysis under the high-accuracy "genomics-default"
528 mode. In this study, we use DeepLIFT to compute attribution values for each gene. We
529 refer to these values as DeepLIFT scores, also described as importance scores or
530 contribution scores in the context of feature attribution.

531 Binding site coordinates were obtained from the GTRD database (v20.06) [63], and
532 regulatory associations were inferred using cis-eQTLs from GTEx v8. A ChIP-seq peak
533 that overlapped a 100-bp window centered on an expression-associated SNP was
534 considered linked to the corresponding transcript. Trans-eQTLs were excluded, and
535 coordinates were converted to the hg19 assembly using UCSC LiftOver [64]. To limit
536 feature dimensionality while maximizing genomic coverage, we computed peak lengths

537 for each DNA-binding protein (DBP) in 100-bp bins across ±1 Mb around the TSS, then
538 retained the maximum peak value within each 10-kb window, yielding up to 200 enhancer
539 bins. These were concatenated with 30 promoter-region bins to form a 230-bin unified
540 input layer. This single-input-layer structure was necessary because DeepLIFT's
541 genomics-default configuration ("RevealCancel rule on dense layers and Rescale rule on
542 conv layers") does not support models with multiple input layers. Concatenating all DBP-
543 related features into one matrix thus maintained compatibility while enhancing
544 interpretability.

545 Model training followed the protocol described by Tasaki et al. and applied in our
546 prior work [20,21]. Specifically, 2,705 genes from chromosome 1 were used for testing,
547 while 22,251 genes from other chromosomes were randomly assigned to the training set
548 and 2,472 to the validation set. Models were implemented in TensorFlow 2.6.0 and
549 trained on an NVIDIA RTX 3080 GPU, typically completing in 2 to 4 hours. Feature
550 importance was assessed using DeepLIFT, which assigned contribution scores to input
551 features indicating their relative influence on gene expression prediction.

552

553 Predicting NABP regulatory targets using gene co-expression–informed 554 substitutions

555 To uncover regulatory targets of NABPs without relying on direct binding site
556 assays, we implemented a co-expression–driven modification of the deep learning input.
557 Specifically, DNA-binding proteins with low contribution scores in the original model
558 were replaced with NABPs whose regulatory interactions were inferred from gene co-
559 expression data. Of the 1,310 DBPs initially included, 302 were substituted with other
560 DBPs, and in a separate RNA-binding protein analysis, 325 were replaced with RNA-

binding proteins (RBPs) identified from the ENPD database [32]. Putative NABP–target relationships were established using co-expression profiles provided by COXPRESdb [22], where a Pearson's *z*-score > 2 between a NABP-encoding gene and a non-NABP gene indicated a strong co-expression link. These co-expression pairs served as proxies for regulatory interactions and were used to generate new input features for the model. To integrate these proxies into the deep learning framework, hypothetical NABP binding sites were assigned at the transcription start site (TSS) and 2 kb upstream of each co-expressed gene. The model was then retrained using this modified input, and DeepLIFT was applied to calculate importance scores, evaluating the influence of each NABP–gene interaction on predicted expression levels. This approach enables the functional characterization of NABPs with unknown or uncharacterized binding motifs by leveraging gene co-expression as a proxy for regulatory targeting.

Prioritization of NABP regulatory targets based on importance score rankings

To identify the primary regulatory targets of each NABP, genes were ranked based on the importance scores assigned to NABP binding sites by the deep learning model. Genes with binding sites consistently ranked in the top five or bottom five across all NABPs were selected as putative targets, reflecting high positive or negative regulatory influence. These targets represent genes most strongly associated with a given NABP's regulatory activity. To ensure sufficient statistical power for downstream enrichment analysis, rank thresholds were adjusted for each NABP based on the number of high-confidence predictions. While a contribution score rank cutoff of five was

584 sufficient in most cases, the threshold was relaxed in specific cases—for example, up to
585 rank ten for proteins, which initially yielded a smaller number of predicted targets.

586 Functional enrichment and AI-based interpretation confirm the biological
587 relevance of NABP regulatory targets. To evaluate the biological coherence of predicted
588 NABP regulatory targets, we performed functional enrichment analysis using the
589 Statistical Overrepresentation Test from the PANTHER database alongside custom in-
590 house scripts [36]. Gene Ontology (GO) terms from biological process, molecular function,
591 and cellular component categories were analyzed for enrichment across three gene sets:
592 (1) primary regulatory target genes identified by the model, (2) all co-expressed genes
593 (Pearson's z-score > 2), and (3) randomly selected genes of matched size. Enrichment
594 was assessed using Fisher's exact test with Benjamini–Hochberg false discovery rate
595 (FDR) correction, and FDR < 0.05 was considered statistically significant. For this
596 analysis, human gene symbols were mapped to Ensembl Gene IDs using the HGNC
597 database (https://www.genenames.org/), and GO annotations were obtained from the
598 Gene Ontology resource [34,35]. Target gene IDs served as the test set, while the full
599 expression dataset used in the deep learning model formed the reference background.

600 Gene set enrichment analysis (GSEA) was performed using the pre-ranked
601 GSEA algorithm implemented in GSEApy (v1.1.11). For each DNA-binding protein
602 (DBP) and RNA-binding protein (RBP), DeepLIFT contribution scores were obtained for
603 all genes from the trained deep learning models. Because raw DeepLIFT scores vary in
604 magnitude across genes and include a large number of zeros, scores were normalized
605 within each gene by scaling the values across all 1,310 DBPs/RBPs to a continuous range
606 from −1 to 1 using min–max normalization. Genes were then ranked according to these
607 normalized contribution values specific to the DBP or RBP of interest. The ranked lists

were used as input for GSEA with the Reactome (v2025.1) and Gene Ontology (Molecular Function, Biological Process, and Cellular Component, 2021 releases) gene set collections. Analyses were performed with 10,000 permutations, weighted enrichment statistics (weight = 1), and a minimum gene-set size of 1 to retain all functional categories. Enrichment significance was assessed using normalized enrichment scores (NES), nominal *p*-values, and false-discovery rates (FDR). This procedure enabled the identification of pathways enriched among genes with high positive or negative DeepLIFT contributions, thereby revealing regulatory roles and cell type–specific pathways associated with each nucleic acid–binding protein.

To complement the statistical analyses, we applied a ChatGPT-based functional inference method [37]. This approach used curated prompts—detailed in *Extended Data Tables 1 and 5*—to predict functional roles based on gene sets, and retrieved relevant literature using a modified version of the "4.Reference search and validation.ipynb" script [37]. This script is publicly available on Code Ocean and GitHub (https://doi.org/10.24433/CO.7045777.v1 ; https://github.com/idekerlab/llm_evaluation_for_gene_set_interpretation). The ChatGPT-based results provided consistent and interpretable biological themes aligned with the enrichment findings.

Validation of predicted NABP targets using experimentally determined binding

To assess the biological relevance of the predicted regulatory targets, we compared deep learning–inferred NABP–gene associations with experimentally validated binding data. Specifically, we evaluated the spatial overlap between predicted

632 target genes and ChIP-seq peaks for 32 DNA-binding proteins (within ±5 kb of
633 transcription start sites [TSSs]) as well as eCLIP peaks for 29 RNA-binding proteins
634 (within the transcribed region from TSS to transcription end site [TES]). Gene structure
635 annotations were derived from the GENCODE v19 release [65], and binding site data were
636 obtained from the ENCODE portal and prior publications (**Supplementary Table S7**) [66].
637 This overlap analysis provides orthogonal validation for the predicted targets and
638 supports the utility of gene co-expression–based modeling for inferring NABP–target
639 relationships.
640

641 ## Statistical analysis

642 Functional enrichment was assessed using Fisher's exact test, with multiple testing
643 correction applied via the Benjamini–Hochberg false discovery rate (FDR) procedure.
644 Statistical significance was defined as FDR < 0.05. Distributions of DeepLIFT
645 contribution scores for nucleic acid-binding sites derived from gene co-expression data
646 and ChIP-seq data were compared using the Mann–Whitney U test (also known as the
647 Wilcoxon rank-sum test). A p-value < 0.05 was considered statistically significant. GSEA
648 (gene set enrichment analysis) was performed using DeepLIFT contribution scores of
649 DBPs and RBPs. The normalized enrichment score (NES) was used to quantify whether
650 regulatory target genes associated with specific biological functions were preferentially
651 located at the high or low end of the ranked gene list. Statistical significance was assessed
652 using permutation-based p-values, as implemented in the GSEA framework.
653

654 ## Data availability

RNA-seq data for human foreskin fibroblast (HFF) cells were obtained from the Gene Expression Omnibus (GEO) under accession number GSM941744. RNA-seq data for human mammary epithelial cells (HMEC) were retrieved from ENCODE via the UCSC Genome Browser (http://genome.ucsc.edu/encode/) under the file name *wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam*. RNA-seq data for human neural progenitor cells (NPC) were obtained from GEO under accession number GSM915326. RNA-seq data for HepG2 and K562 cells were obtained from the ENCODE portal under accession numbers ENCFF670LIE and ENCFF584WML, respectively. mRNA co-expression data were downloaded from COXPRESdb v7 (https://zenodo.org/record/6861444). The list of nucleic acid-binding proteins was obtained from the Eukaryotic Nucleic acid Binding Protein (ENPD) database (http://qinlab.sls.cuhk.edu.hk/ENPD/). ChIP-seq data for transcription factors were downloaded from the GTRD database (v20.06 release; https://gtrd.biouml.org). eQTL data (GTEx_Analysis_v8_eQTL.tar) were obtained from the GTEx portal (v8 release; https://gtexportal.org). eCLIP data for RNA-binding proteins were retrieved from the ENCODE portal (accession ENCSR456FVU) and from Supplementary Data 2 of Van Nostrand et al., *Nature* (2020), https://doi.org/10.1038/s41586-020-2077-3.

Code availability

The deep learning model implemented in this study is available at https://github.com/reposit2/coexgene.

Figures

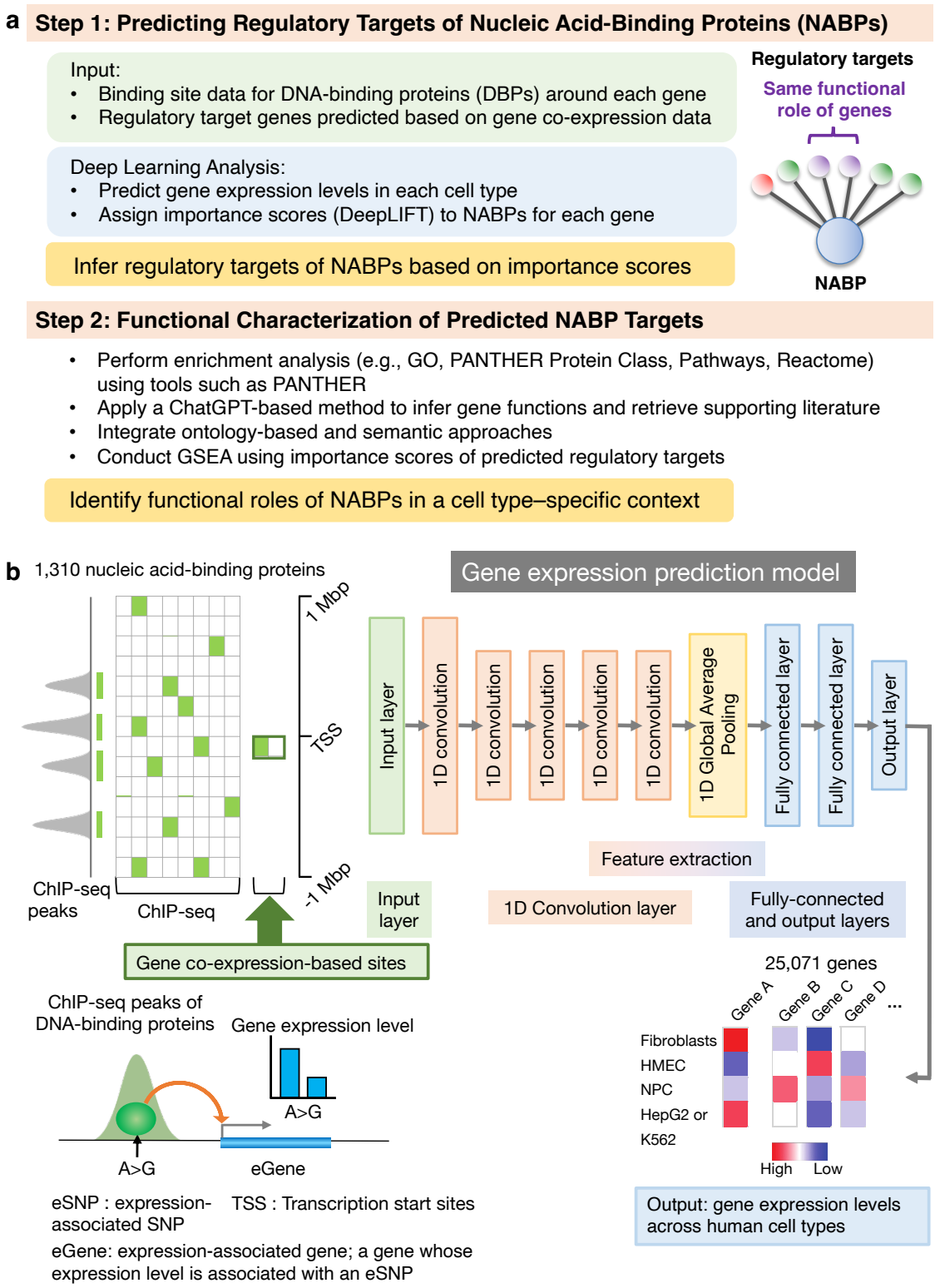

a
Step 1: Predicting Regulatory Targets of Nucleic Acid-Binding Proteins (NABPs)
Input:
• Binding site data for DNA-binding proteins (DBPs) around each gene
• Regulatory target genes predicted based on gene co-expression data
Deep Learning Analysis:
• Predict gene expression levels in each cell type
• Assign importance scores (DeepLIFT) to NABPs for each gene
Infer regulatory targets of NABPs based on importance scores
Regulatory targets
Same functional role of genes
NABP
Step 2: Functional Characterization of Predicted NABP Targets
• Perform enrichment analysis (e.g., GO, PANTHER Protein Class, Pathways, Reactome) using tools such as PANTHER
• Apply a ChatGPT-based method to infer gene functions and retrieve supporting literature
• Integrate ontology-based and semantic approaches
• Conduct GSEA using importance scores of predicted regulatory targets
Identify functional roles of NABPs in a cell type–specific context
b
1,310 nucleic acid-binding proteins
1 Mbp
TSS
-1 Mbp
ChIP-seq peaks
ChIP-seq
Gene co-expression-based sites
Gene expression prediction model
Input layer
1D convolution
1D convolution
1D convolution
1D convolution
1D convolution
1D Global Average Pooling
Fully connected layer
Fully connected layer
Output layer
Feature extraction
Input layer
1D Convolution layer
Fully-connected and output layers
ChIP-seq peaks of DNA-binding proteins
Gene expression level
A>G
A>G
eGene
25,071 genes
Gene A
Gene B
Gene C
Gene D
...
Fibroblasts
HMEC
NPC
HepG2 or K562
High
Low
Output: gene expression levels across human cell types
eSNP : expression-associated SNP
TSS : Transcription start sites
eGene: expression-associated gene; a gene whose expression level is associated with an eSNP

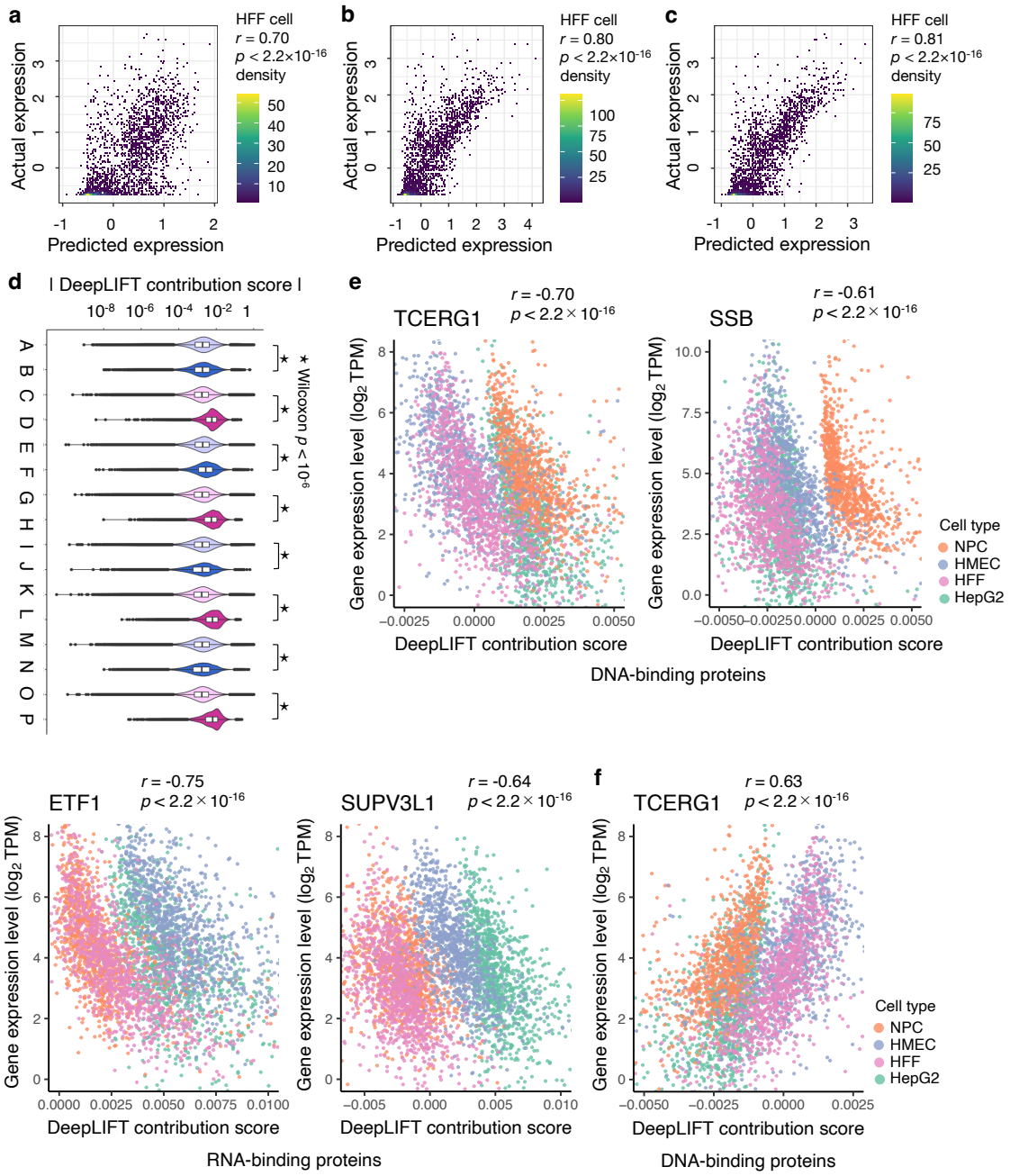
Figure 1. Overview of the deep learning framework for predicting and functionally characterizing NABP regulatory targets.

a. This diagram outlines the overall strategy used to infer the functional roles of nucleic acid-binding proteins (NABPs) from gene expression data. Gene expression profiles from various human cell types were integrated with NABP–gene co-expression relationships derived from COXPRESdb. These inferred interactions were used to replace low-contribution DNA-binding proteins in a deep learning model originally trained on ChIP-seq–based binding site data. The modified model was trained to predict gene expression levels and identify regulatory target genes using DeepLIFT-derived importance scores. Functional annotations of the predicted target genes were assessed using conventional enrichment analyses and GSEA, along with a ChatGPT-based method for literature-guided interpretation.

b. Gene expression levels in human cells were predicted using the genomic locations of DNA-binding sites at promoters and distal elements.

During model training, a DNA-binding protein bound to a distal region overlapping an expression-associated SNP (eSNP) was linked to a potentially regulated gene (eGene) based on eQTL data. For each gene–protein pair, ChIP-seq peak locations were binned (indicated by green boxes) and formatted into a matrix as input for deep learning.

In co-expression–based models, binding sites were artificially assigned near transcription start sites (TSSs). The deep learning model consisted of 1D convolutional layers, global average pooling, and fully connected layers. It predicted gene expression levels across 25,071 genes and multiple human cell types.


a
HFF cell
r = 0.70
p < 2.2×10-16
density
Actual expression
Predicted expression
b
HFF cell
r = 0.80
p < 2.2×10-16
density
Actual expression
Predicted expression
c
HFF cell
r = 0.81
p < 2.2×10-16
density
Actual expression
Predicted expression
d
| DeepLIFT contribution score |
A B C D E F G H I J K L M N O P
* Wilcoxon p < 10-6
e
TCERG1
r = -0.70
p < 2.2 × 10-16
SSB
r = -0.61
p < 2.2 × 10-16
Gene expression level (log2 TPM)
DeepLIFT contribution score
Cell type
NPC
HMEC
HFF
HepG2
DNA-binding proteins
ETF1
r = -0.75
p < 2.2 × 10-16
SUPV3L1
r = -0.64
p < 2.2 × 10-16
RNA-binding proteins
f
TCERG1
r = 0.63
p < 2.2 × 10-16
DNA-binding proteins

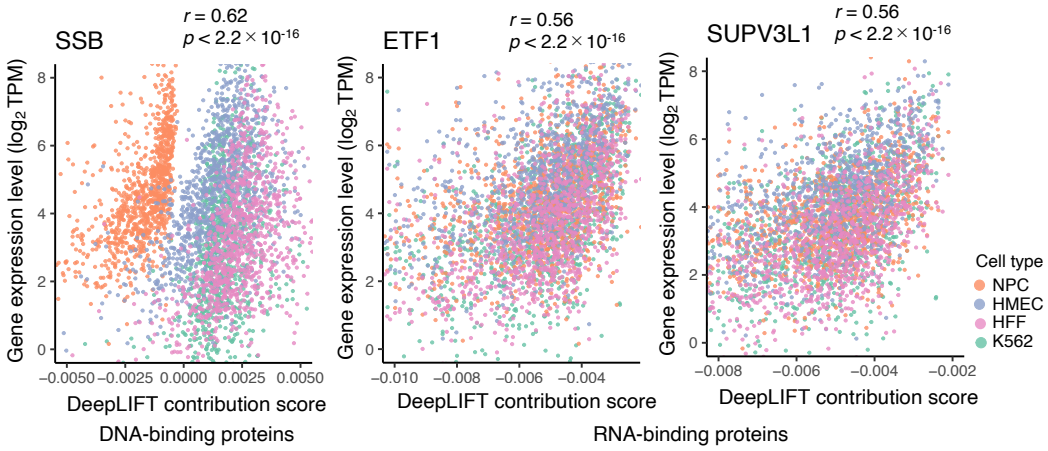


Figure 2. Relationship between gene expression levels and contribution scores.

Gene expression levels ($\log_2$ TPM) predicted by a deep learning model were compared with observed expression levels using Spearman's rank correlation.

a–c, Schematic overview of input feature configurations used for gene expression prediction.

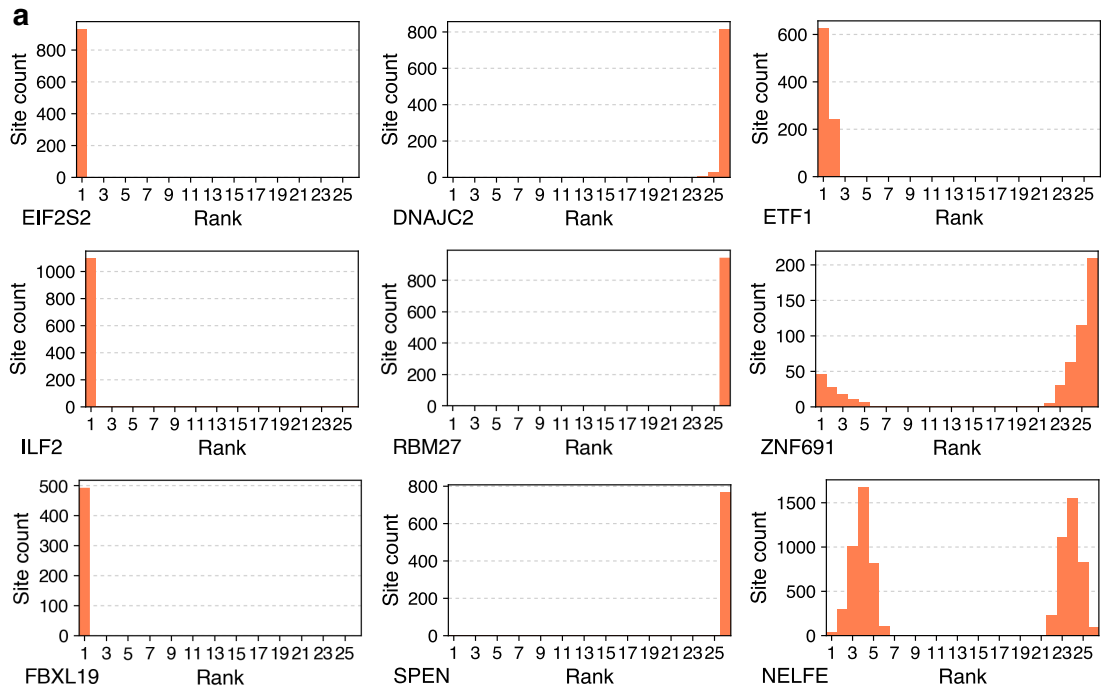
(**a**) DNA-binding sites of 1,310 proteins derived from ChIP-seq data.

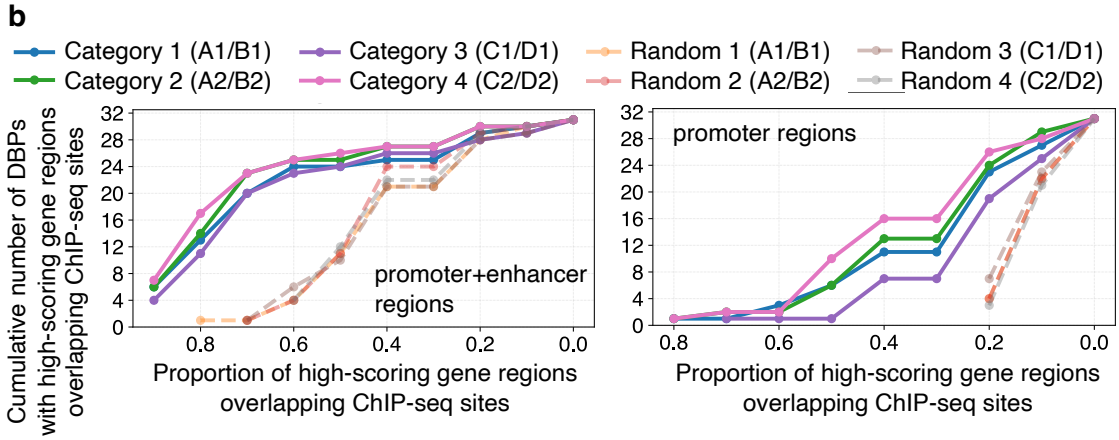
(**b**) Replacement of 302 low-contribution DNA-binding proteins with co-expression–derived DNA-binding proteins.

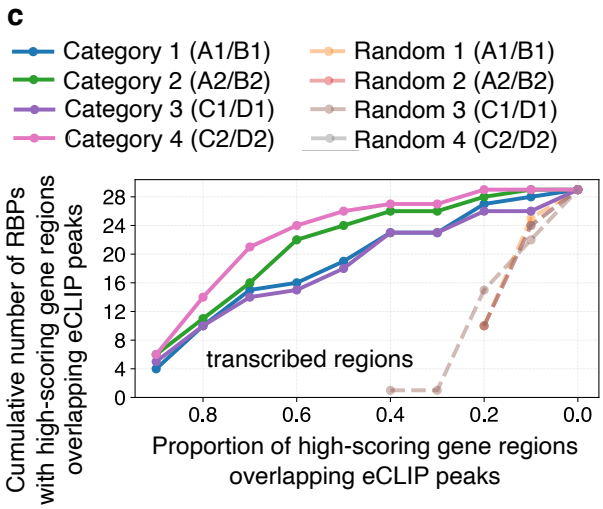
(**c**) Replacement of 325 DNA-binding proteins with RNA-binding proteins identified from co-expression relationships.

d, Comparison of absolute DeepLIFT contribution scores for predicted regulatory sites derived from gene co-expression versus ChIP-seq data, using a background defined as 0.5× the original input feature values. Each pair of violin plots represents co-expression–based sites (right) and corresponding ChIP-seq sites not included in that set (left). Contribution scores were significantly higher for co-expression–based sites (Wilcoxon signed-rank test, $p < 2.2 \times 10^{-16}$). Results obtained using an amplified background (input × 2) are shown in **Supplementary Fig. S2c**.

e,f, Relationship between gene expression levels and DeepLIFT contribution scores for representative DNA- and RNA-binding proteins across four cell types. Contribution scores were computed using a background defined as input × 0.5 (**e**) or input × 2 (**f**). The same set of proteins is shown in both panels, enabling direct comparison of background effects. While background scaling alters the sign of the DeepLIFT–expression correlation, the relative ordering of contribution scores and cell type–specific patterns are preserved.




b
Category 1 (A1/B1)
Category 2 (A2/B2)
Category 3 (C1/D1)
Category 4 (C2/D2)
Random 1 (A1/B1)
Random 2 (A2/B2)
Random 3 (C1/D1)
Random 4 (C2/D2)
Cumulative number of DBPs with high-scoring gene regions overlapping ChIP-seq sites
promoter+enhancer regions
promoter regions
Proportion of high-scoring gene regions overlapping ChIP-seq sites


c
Category 1 (A1/B1)
Category 2 (A2/B2)
Category 3 (C1/D1)
Category 4 (C2/D2)
Random 1 (A1/B1)
Random 2 (A2/B2)
Random 3 (C1/D1)
Random 4 (C2/D2)
Cumulative number of RBPs with high-scoring gene regions overlapping eCLIP peaks
transcribed regions
Proportion of high-scoring gene regions overlapping eCLIP peaks

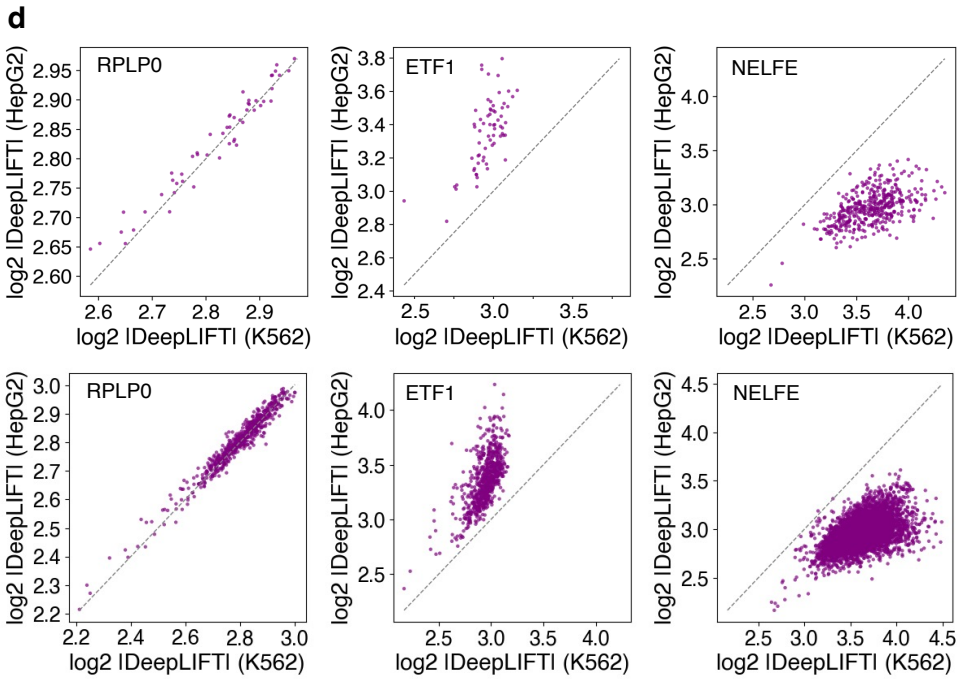


Figure 3. Comparison of predicted regulatory target genes with experimental ChIP-seq and eCLIP binding data. This figure illustrates the correspondence between gene-level regulatory contribution scores inferred by deep learning and experimentally identified nucleic acid-binding protein (NABP) binding sites, using a DeepLIFT background defined as 0.5× the original input feature values.

a. Rank distribution of DeepLIFT contribution scores for nucleic acid-binding sites validated by ChIP-seq (DNA-binding proteins) or eCLIP (RNA-binding proteins). Contribution scores were ranked from highest to lowest and grouped into equal-sized rank bins. Binding sites located within promoters or transcribed regions of experimentally bound genes showed non-random enrichment toward higher or lower contribution ranks compared with all binding sites.

b. Proportion of high-contribution genomic regions overlapping ChIP-seq binding sites in promoter-associated regions. The fractions of high-contribution regions

751 overlapping ChIP-seq peaks are shown for promoter + enhancer regions (±100 kb from
752 the transcription start site) and promoter-only regions (±5 kb). Multiple rank- and score-
753 based thresholds were applied to define high- and low-contribution regions (see
754 **Supplementary Note and Tables S2 and S3**).

755 **c. Overlap between high-contribution regions and eCLIP peaks for RNA-binding**
756 **proteins.** Cumulative distributions show the proportion of high-contribution gene regions
757 overlapping experimentally determined eCLIP peaks within transcribed regions (from
758 transcription start site to transcription end site), compared with random controls.
759 Corresponding analyses using an amplified DeepLIFT background (input × 2) produced
760 similar results and are shown in **Supplementary Figure S3**.

761 **d. Comparison of DeepLIFT contribution scores between HepG2 and K562 cells**
762 **under different background definitions.** Scatter plots display contribution scores for
763 predicted regulatory target genes of three representative NABPs (RPLP0, ETF1, and
764 NELFE), comparing K562 (x-axis) and HepG2 (y-axis). Absolute contribution scores
765 were $\log_2$-transformed for visualization, and only genes with experimentally validated
766 binding sites were included. For each NABP, results are shown for two background
767 configurations (input × 0.5 and input × 2).

768

769

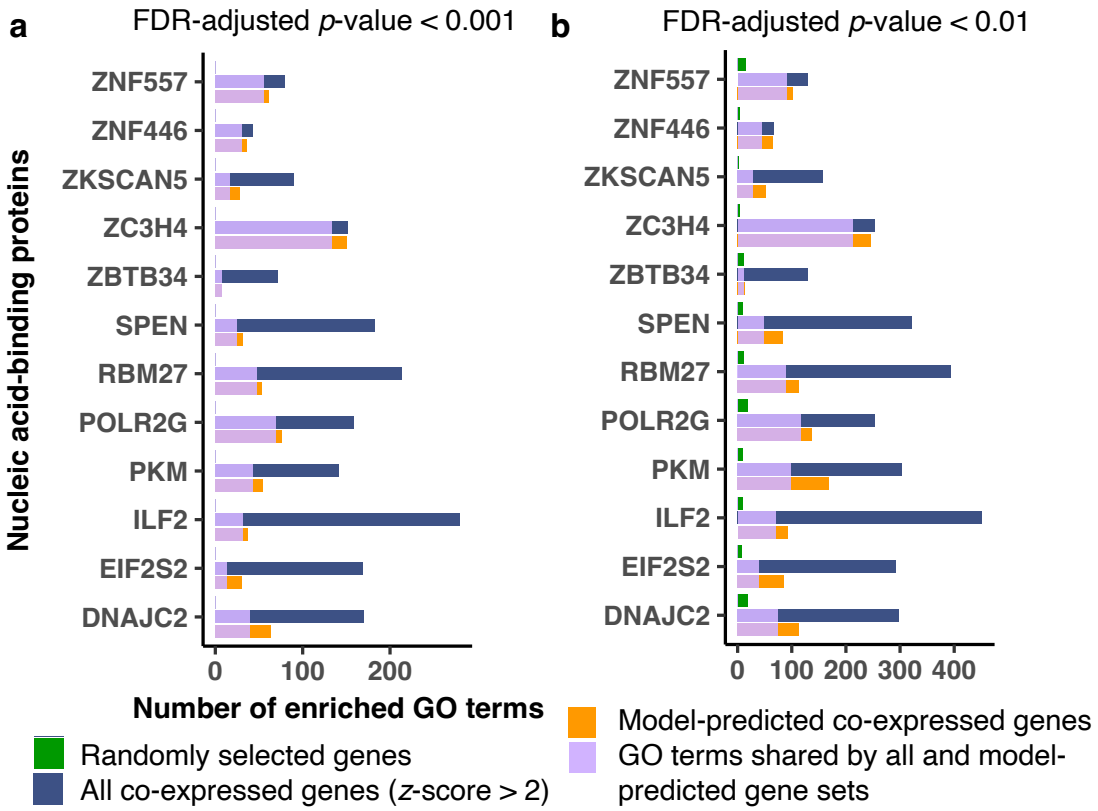


Figure 4. Overrepresented functional annotation terms among target genes of nucleic acid-binding proteins.

Panels **a** and **b** show results filtered at false discovery rate (FDR)–adjusted *p*-values of < 0.001 and < 0.01, respectively. Functional enrichment analysis was conducted for three gene sets: (1) model-predicted co-expressed genes, (2) all co-expressed genes with a Pearson correlation *z*-score > 2, and (3) randomly selected genes of equal size. Gene Ontology (GO) terms from the categories of biological process, molecular function, and cellular component were combined. Fisher's exact test *p*-values were adjusted using the Benjamini–Hochberg method to control the false discovery rate (FDR).

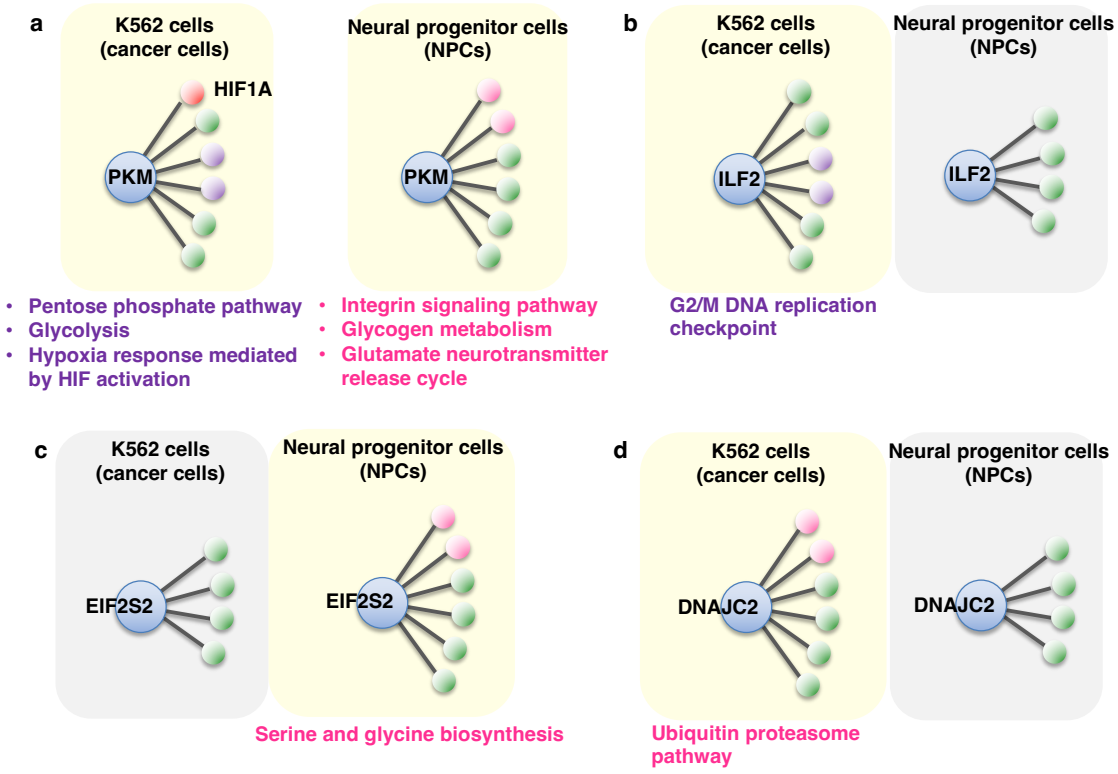
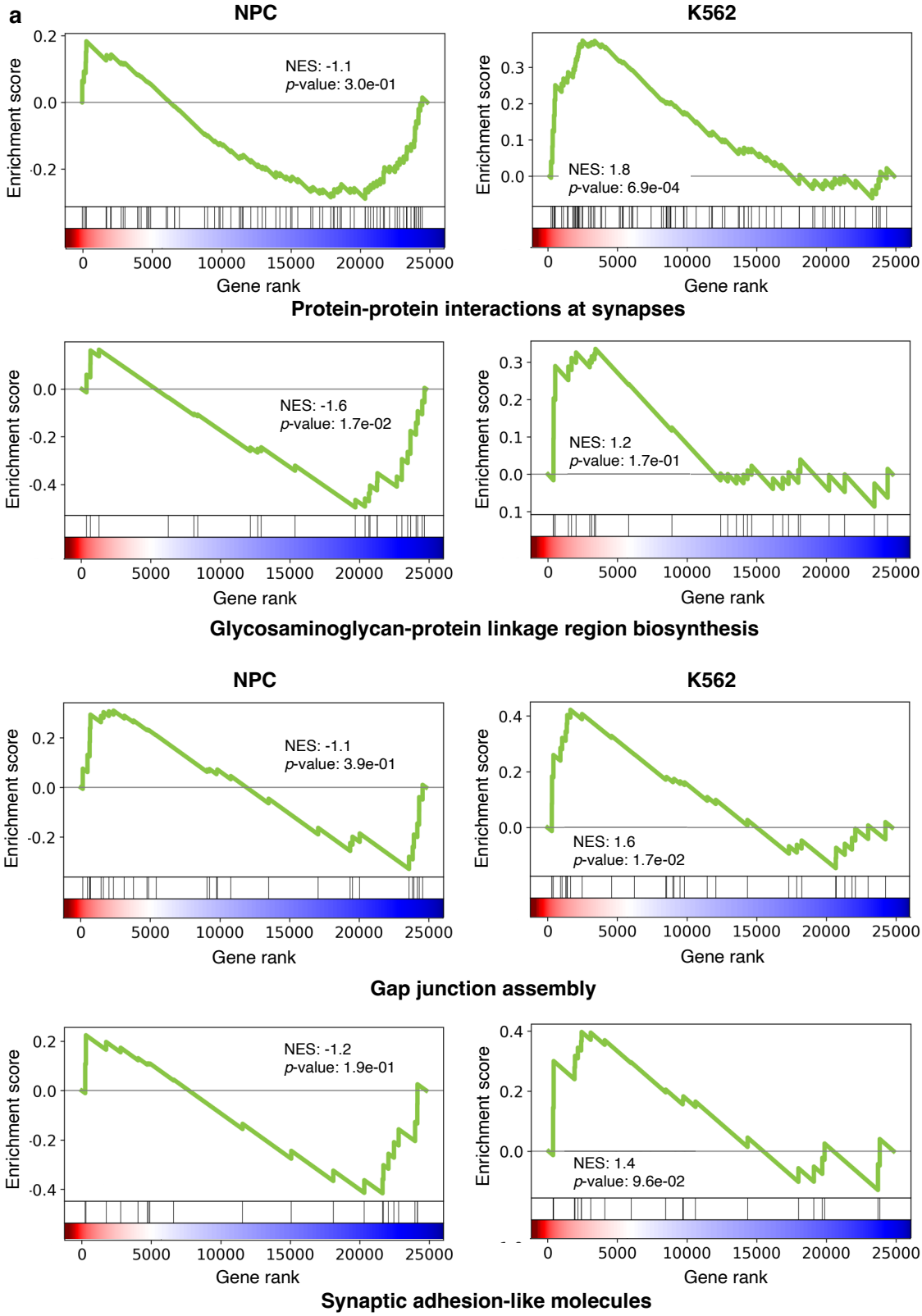
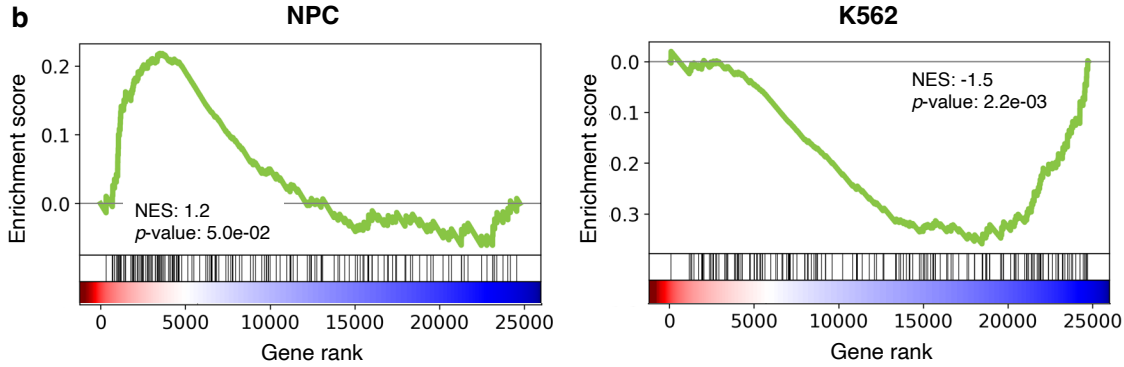


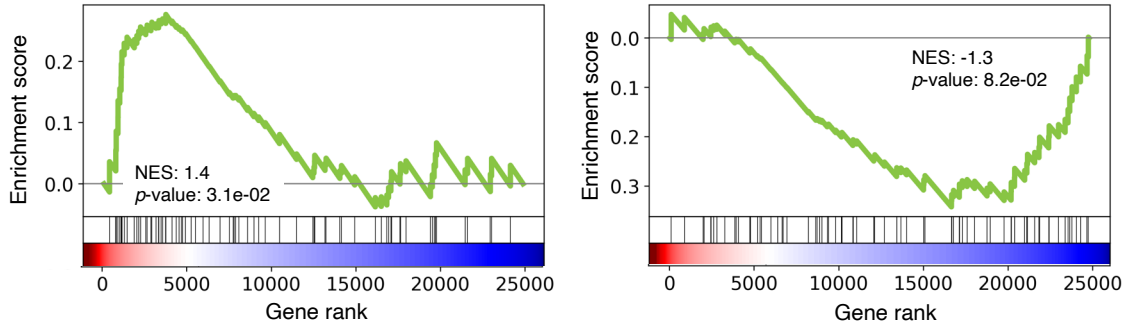
Figure 5. Functional mapping of nucleic acid-binding proteins reveals context-dependent regulatory roles in cancer and neural progenitor cells.

A deep learning model trained on transcriptomic co-expression data was used to predict the regulatory influence of nucleic acid-binding proteins (NABPs) on gene expression. Contribution scores were computed to quantify the impact of each NABP across target genes, and the highest-ranking genes were designated as predicted regulatory targets. Functional interpretation was performed using a large language model–based inference method, which identified key biological pathways associated with each NABP.

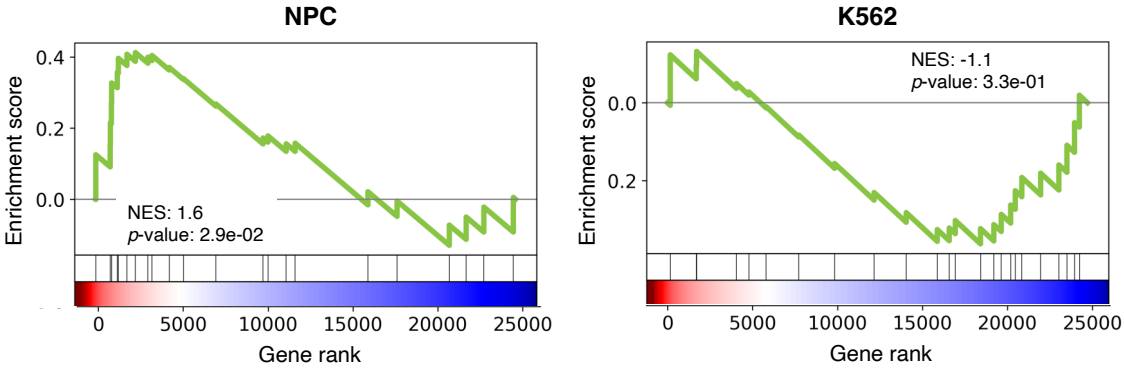

a
NPC
K562
Enrichment score
Gene rank
NES: -1.1
p-value: 3.0e-01
NES: 1.8
p-value: 6.9e-04
Protein-protein interactions at synapses
NES: -1.6
p-value: 1.7e-02
NES: 1.2
p-value: 1.7e-01
Glycosaminoglycan-protein linkage region biosynthesis
NES: -1.1
p-value: 3.9e-01
NES: 1.6
p-value: 1.7e-02
Gap junction assembly
NES: -1.2
p-value: 1.9e-01
NES: 1.4
p-value: 9.6e-02
Synaptic adhesion-like molecules



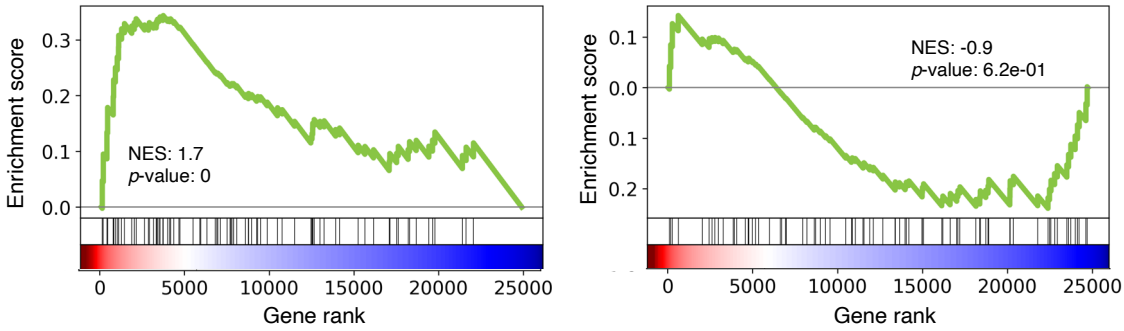
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell



Role of phospholipids in phagocytosis



WNT ligand biogenesis and trafficking



FCERI mediated MAPK activation

Figure 6. Cell type–specific pathway associations inferred from contribution score–based GSEA.

GSEA (gene set enrichment analysis) was performed using DeepLIFT contribution scores of nucleic acid-binding proteins (NABPs). This analysis revealed that genes with high or low contribution scores for a given NABP are preferentially associated with biological pathways relevant to each cell type.

The x-axis represents genes ranked by their DeepLIFT contribution scores for each NABP, with higher-scoring genes placed toward the left. PKM is shown here as a representative RNA-binding protein to illustrate contribution score–based pathway analysis. The y-axis indicates the running enrichment score, which increases when pathway-associated genes are concentrated among genes with higher contribution scores and decreases when they are enriched among genes with lower contribution scores. The normalized enrichment score (NES) summarizes the strength and direction of this enrichment across pathways.

For example, genes annotated to "Protein–protein interactions at synapses" were preferentially enriched among genes with lower PKM contribution scores in NPCs but among genes with higher contribution scores in K562 cells, indicating a marked cell type–specific difference in the inferred association between PKM and this pathway. Conversely, genes associated with "Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell" showed contrasting enrichment patterns between K562 and NPCs, highlighting differential pathway associations inferred from PKM contribution scores across the two cell types.

Tables

822 **Table 1. Functional roles inferred for nucleic acid-binding proteins: integration of**
823 **UniProt annotations, ChatGPT-based predictions, and confidence scores.**
824 Functional annotations obtained from UniProt entries were systematically compared with
825 significantly enriched biological processes among the model-predicted regulatory targets,
826 as identified by the PANTHER Overrepresentation Test (see **Supplementary Table S5**).
827 Shared or overlapping functions between each nucleic acid-binding protein (NABP) and
828 its predicted targets are indicated. Functional roles inferred using a ChatGPT-based
829 interpretation method are highlighted in yellow, with confidence scores—defined as the
830 proportion of annotated genes within the gene set—provided in parentheses. Confidence
831 levels are categorized as high (0.87–1.00), medium (0.82–0.86), low (0.01–0.81), or none
832 (0). For methodological reference, see Hu M. et al., *Nat. Methods*, 2025
833 (https://doi.org/10.1038/s41592-024-02525-x) [37].
834

ZBTB34 (HepG2 cells)
C2H2 zinc finger transcription factor
RNA polymerase II transcription regulatory region sequence-specific DNA binding
regulation of transcription by RNA polymerase II
C2H2 zinc-finger–directed regulation of RNA polymerase II transcription (0.88)

ZC3H4 (HepG2 cells)
chromatin/chromatin-binding, or -regulatory protein
RNA processing factor
basal RNA polymerase II transcription machinery binding
Chromatin-regulated RNA polymerase II transcription and RNA processing (0.89)

ZKSCAN5 (HepG2 cells)
C2H2 zinc finger transcription factor
RNA polymerase II cis-regulatory region sequence-specific DNA binding
autophagy
macroautophagy
Transcription-coupled autophagy–lysosome remodeling and membrane trafficking (0.78)

ZNF446 (HepG2 cells)
DNA-binding transcription activator activity, RNA polymerase II-specific
protein import into peroxisome matrix
non-motile cilium assembly
Transcription factor–directed ciliogenesis with peroxisome import and genome maintenance (0.86)

ZNF557 (HepG2 cells)
DNA-binding transcription activator activity, RNA polymerase II-specific
mRNA polyadenylation factor
RNA methyltransferase
RNA processing factor
RNA metabolism protein
Transcription-coupled RNA processing and RNP biogenesis (0.90)

POLR2G (NPC cells)
regulation of transcription by RNA polymerase II
RNA processing factor
translation initiation factor
Mitochondrial translation
RNA processing–coupled cytosolic and mitochondrial translation (0.90)

RBM27 (K562 cells)
RNA processing factor
RNA splicing factor
RNA metabolism protein
chromatin/chromatin-binding, or -regulatory protein
SUMO is proteolytically processed
Ub-specific processing proteases
RNAPII-coupled RNA processing and proteostasis axis (0.86)

SPEN
chromatin/chromatin-binding, or -regulatory protein
non-receptor serine/threonine protein kinase
ubiquitin-protein ligase
Signaling by Rho GTPases
Chromatin remodeling–anchored RNAPII transcription coordinated by ubiquitin and Rho-kinase signaling (0.78)

PKM (K562 cells)
Hypoxia response via HIF activation
glucose 6-phosphate metabolic process
Disorders of transmembrane transporters
Membrane Trafficking
Hypoxia/HIF-1–orchestrated metabolic reprogramming with cytoskeletal–membrane remodeling (0.80)

PKM (NPC cells)
Cellular response to hypoxia
KEAP1-NFE2L2 (NRF2) pathway
actin or actin-binding cytoskeletal protein
Hypoxia-driven metabolic rewiring with KEAP1–NRF2 signaling and actin–endosomal remodeling (0.86)

DNAJC2 (K562 cells)
RNA processing factor
chaperone
proteasome-mediated ubiquitin-dependent protein catabolic process
proteasomal protein catabolic process
ribosome biogenesis
RNA processing–coupled ribosome biogenesis and translation with proteostasis surveillance (0.88)

EIF2S2 (NPC cells)
proteinogenic amino acid biosynthetic process
Serine glycine biosynthesis
Unfolded Protein Response (UPR)
NFE2L2 regulating anti-oxidant/detoxification enzymes
ATF4 activates genes in response to endoplasmic reticulum stress
Integrated stress response coordinating UPR, antioxidant defense, and amino-acid/secretory metabolism (0.86)

ILF2 (K562 cells)
RNA processing factor
nucleic acid binding
G2/M DNA replication checkpoint
MASTL Facilitates Mitotic Progression
Coupled RNA processing–nuclear export aligned with mitotic cell-cycle control (0.86)

Table 2. Functional characterization of predicted regulatory targets of nucleic acid-binding proteins using a ChatGPT-based inference framework. Representative biological functions were inferred for each NABP by applying a large language model–based gene set interpretation method. Functional themes are

842 summarized in bold for each NABP, while numbered subcategories denote distinct

843 functional roles associated with individual gene subsets.

ZBTB34 (HepG2 cells)
C2H2 zinc-finger–directed regulation of RNA polymerase II transcription (0.88)
1. Sequence-specific transcriptional control (core signal)
2. Transcriptional modulators and repressors (adjacent layer)
3. Signaling inputs that gate transcriptional outputs
4. RNA- and chromatin-adjacent regulators
5. Contextual effectors shaped by the transcriptional program

ZC3H4 (HepG2 cells)
Chromatin-regulated RNA polymerase II transcription and RNA processing (0.89)
1. Sequence-specific control + chromatin remodeling
2. RNAPII initiation and elongation
3. Co-transcriptional RNA processing
4. RNP assembly and nuclear–cytoplasmic transport
5. Ribosome biogenesis and rRNA metabolism
6. Post-translational regulation of the transcriptional apparatus
7. Genome maintenance coupling
8. Growth-signal gating
9. Peripheral modules

ZKSCAN5 (HepG2 cells)
Transcription-coupled autophagy–lysosome remodeling and membrane trafficking (0.78)
1. Transcriptional control as the top-level driver
2. Phosphoinositide signaling primes autophagosome initiation
3. Core autophagy execution and cargo selection
4. Endosome–lysosome tethering and autophagosome fusion
5. Proteasome and ubiquitylation complement lysosomal turnover
6. Organelle interfaces that feed the degradative axis
7. Ciliary/centrosomal trafficking as a specialized downstream client
8. Stress-responsive transcriptional tuning

ZNF446 (HepG2 cells)
Transcription factor–directed ciliogenesis with peroxisome import and genome maintenance (0.86)
1. Transcriptional control and RNA maturation
2. Ciliogenesis and membrane/organelle logistics
3. Peroxisome matrix import and organization
4. Genome maintenance and telomere control
5. Metabolic and cofactor support
6. Growth and sensory context
7. Network integration

ZNF557 (HepG2 cells)
Transcription-coupled RNA processing and RNP biogenesis (0.90)
1. Core transcriptional control
2. Co-transcriptional RNA processing
3. RNA modification and quality control
4. RNP assembly and ribosome/tRNA/rRNA biogenesis
5. Broad TF superstructure
6. Coupled auxiliary programs (contextual, not dominant)
7. Network integration

PKM (K562 cells)
Hypoxia/HIF-1–orchestrated metabolic reprogramming with cytoskeletal–membrane remodeling (0.80)
1. Transcriptional hypoxia module setting the program
2. Metabolic rewiring toward glycolysis and redox support
3. Proteostasis and catabolic throughput to sustain the shift
4. Actin–cytoskeleton remodeling and focal adhesion dynamics
5. Vesicle trafficking and membrane remodeling to deliver the phenotype
6. mRNA processing and translation control coupled to stress
7. KEAP1–NFE2L2 (NRF2)–linked oxidative stress buffering (adjacent)

PKM (NPC cells)
Hypoxia-driven metabolic rewiring with KEAP1–NRF2 signaling and actin–endosomal remodeling (0.86)
1. Metabolic rewiring and hypoxia response
2. KEAP1–NRF2 cytoprotective axis
3. Actin/adhesion remodeling and endocytic–lysosomal traffic
4. Proteostasis, autophagy, and apoptosis gating
5. Integrated model

ILF2 (K562 cells)
Coupled RNA processing–nuclear export aligned with mitotic cell-cycle control (0.86)
1. Transcription initiation and elongation across Pol I/II/III
2. Co-transcriptional RNA maturation: splicing, 3′-end formation, RNA modification
3. mRNP assembly, nuclear export, and nucleocytoplasmic transport
4. Nuclear envelope/pore remodeling synchronized with mitosis
5. Post-transcriptional regulators and chromatin/epigenetic control
6. Secretory and membrane-traffic support + stress responses
7. Gene-set composition is concordant with these functions

ILF2 (NPC cells)
RNAP II/III–coupled RNA biogenesis and ribosome assembly (0.88)
1. Transcription initiation and elongation
2. Co-transcriptional RNA processing and RNP assembly
3. Nuclear export, surveillance, and decay
4. rRNA processing and ribosome biogenesis
5. tRNA metabolism and charging
6. Chromatin and repair context

DNAJC2 (K562 cells)
RNA processing–coupled ribosome biogenesis and translation with proteostasis surveillance (0.88)
1. RNA processing and export
2. Ribosome biogenesis and rRNA metabolism
3. Translation and initiation control
4. Proteostasis and stress-response coupling
5. Genome maintenance and cell-cycle gating
6. Organelle and mitochondrial support

844

845

846

847 **Table 3. Functional enrichments among predicted regulatory target genes of nucleic**

848 **acid-binding proteins.**

849 Functional enrichment analysis was performed using the PANTHER database. This table
850 provides selected examples of annotations that were statistically enriched among
851 regulatory target genes predicted by the deep learning model trained on gene co-
852 expression data. '**Reference**' indicates the total number of human genes; '**Analyzed**'
853 denotes the number of regulatory target genes examined; '**Expected**' represents the
854 number of genes expected to have the annotation by chance; '**Fold Enrichment**' reflects
855 the observed-to-expected ratio; '**Raw *p*-value**' reports the result of Fisher's exact test; and
856 '**FDR**' refers to the Benjamini–Hochberg false discovery rate.

DBP, RBP	Functional annotation	Enriched	Not Enriched	Analyzed / Reference	Expected	Fold Enrichment	Raw *p*-value	FDR
PKM	Pentose phosphate pathway (PC0762)	K562	NPC	3 / 9	0.15	20.24	3.E-04	1.E-02
PKM	Pentose phosphate pathway (R-HSA-71336)	K562	NPC	3 / 15	0.25	12.14	2.E-03	2.E-02
PKM	Glycolysis (PC00024)	K562	NPC	4 / 20	0.33	12.14	3.E-04	1.E-02
PKM	Hypoxia response via HIF activation (P00030)	K562	NPC	4 / 30	0.49	8.09	1.E-03	4.E-02
PKM	response to hypoxia (GO:0001666)	K562	NPC	5 / 18	0.30	16.86	8.E-06	4.E-04
ILF2	G2/M DNA replication checkpoint (R-HSA-69478)	K562	NPC	2 / 5	0.06	30.95	2.E-03	4.E-02
POLR2G	ALKBH3 mediated reversal of alkylation damage (R-HAS-112126)	K562	NPC	2 / 4	0.11	17.80	5.E-03	5.E-02
DNAJC2	Ubiquitin proteasome pathway (P00060)	K562	NPC	8 / 58	0.60	13.26	1.E-07	2.E-05
PKM	Integrin signalling pathway (P00034)	NPC	K562	12 / 192	2.70	4.45	2.E-05	1.E-03
PKM	Glycogen metabolism (R-HSA-8982491)	NPC	K562	4 / 24	0.34	11.87	3.E-04	6.E-03
PKM	Glutamate Neurotransmitter Release Cycle (R-HSA-210500)	NPC	K562	4 / 23	0.32	12.38	3.E-04	5.E-03
EIF2S2	Serine glycine biosynthesis (PC02776)	NPC	K562	3 / 6	0.09	33.63	6.E-05	1.E-02

857

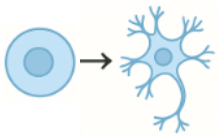
858

859

860 **Table 4. Cell type−dependent enrichment of functional pathways identified by**
861 **GSEA**

862 NES denotes the normalized enrichment score obtained from gene set enrichment
863 analysis (GSEA). ΔNES represents the difference in NES between NPC and K562. This
864 table summarizes the pathway enrichment results for PKM, presented as an example
865 RBP consistent with **Figure 6**. The seventeen Reactome pathways with the highest and
866 lowest ΔNES values are listed in the upper and lower sections, respectively. Pathways

with high ΔNES values were associated with neural cell differentiation, whereas those with low ΔNES values were associated with cancer-related functions.



Pathways associated with neural cell differentiation

Pathways associated with cancer cells

Reactome pathway description	Reactome Pathway ID	NES K562	NES NPC	Δ NES K562 - NPC
Protein-protein interactions at synapses	R-HSA-6794362	1.79	-1.11	2.89
Glycosaminoglycan-protein linkage region biosynthesis	R-HSA-1971475	1.23	-1.56	2.79
Golgi-to-ER retrograde transport	R-HSA-8856688	1.62	-1.15	2.77
Attachment and Entry	R-HSA-9694614	1.39	-1.37	2.76
Defective EXT2 causes exostoses 2	R-HSA-3656237	1.39	-1.36	2.75
Defective B4GALT7 causes EDS, progeroid type	R-HSA-3560783	1.13	-1.60	2.73
Phase 0 - rapid depolarisation	R-HSA-5576892	1.45	-1.24	2.69
Gap junction assembly	R-HSA-190861	1.62	-1.06	2.68
Diseases associated with glycosaminoglycan metabolism	R-HSA-3560782	1.07	-1.61	2.68
Extracellular matrix organization	R-HSA-1474244	1.46	-1.21	2.67
Synaptic adhesion-like molecules	R-HSA-8849932	1.37	-1.24	2.61
Neuronal System	R-HSA-112316	1.63	-0.96	2.59
Degradation of the extracellular matrix	R-HSA-1474228	1.50	-1.08	2.59
IRE1alpha activates chaperones	R-HSA-381070	1.25	-1.31	2.56
Smooth Muscle Contraction	R-HSA-445355	1.51	-1.04	2.54
FGFR2c ligand binding and activation	R-HSA-190375	1.29	-1.23	2.52
Receptor-type tyrosine-protein phosphatases	R-HSA-388844	1.62	-0.88	2.50
Muscle contraction	R-HSA-397014	1.57	-0.92	2.49
LGI-ADAM interactions	R-HSA-5682910	1.53	-0.97	2.49
Respiratory syncytial virus (RSV) attachment and entry	R-HSA-9820960	1.39	-1.1	2.49
Activation of anterior HOX genes in hindbrain development during early embryogenesis	R-HSA-5617472	-1.11	1.50	-2.62
Role of LAT2/NTAL/LAB on calcium mobilization	R-HSA-2730905	-1.12	1.50	-2.62
FCERI mediated MAPK activatio	R-HSA-2871796	-0.92	1.72	-2.64
WNT ligand biogenesis and trafficking	R-HSA-3238698	-1.10	1.56	-2.66
Role of phospholipids in phagocytosis	R-HSA-2029485	-1.31	1.38	-2.69
Nucleotide Excision Repair	R-HSA-5696398	-1.30	1.41	-2.71
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	R-HSA-198933	-1.52	1.24	-2.75
Transcription-Coupled Nucleotide Excision Repair (TC-NER)	R-HSA-6781827	-1.47	1.33	-2.79
Creation of C4 and C2 activators	R-HSA-166786	-1.13	1.70	-2.83
CD22 mediated BCR regulation	R-HSA-5690714	-1.11	1.72	-2.83
Initial triggering of complement	R-HSA-166663	-1.23	1.64	-2.87
FCGR activation	R-HSA-2029481	-1.17	1.79	-2.96
Scavenging of heme from plasma	R-HSA-2168880	-1.26	1.74	-3.00
Mitotic G1 phase and G1/S transition	R-HSA-453279	-1.27	1.73	-3.01
Response to metal ions	R-HSA-5660526	-1.29	1.77	-3.06
Mitotic Metaphase and Anaphase	R-HSA-2555396	-1.27	1.79	-3.06
M Phase	R-HSA-68886	-1.27	1.84	-3.11
DNA Replication	R-HSA-69306	-1.30	1.94	-3.23
Synthesis of DNA	R-HSA-69239	-1.29	1.98	-3.28
Metallothioneins bind metals	R-HSA-5661231	-1.04	2.61	-3.65

Acknowledgements

Supercomputing resources were provided by the Human Genome Center, the Institute of Medical Science, the University of Tokyo, and the ROIS National Institute of Genetics.

Author contributions

N.O. designed and conducted the research, and wrote and revised the manuscript. K.S. revised the manuscript.

Supplementary data

Supplementary data are available at

Conflict of interests

None declared.

Funding

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI [Grant Number 16K00387], a start-up research grant from Osaka University, the KIOXIA Young Scientist Research Grant at Waseda University, the Grant for Basic Science Research Projects from the Sumitomo Foundation, the Research Grant from the CASIO Science Promotion Foundation, and the Research Grant from the Okawa Foundation for Information and Telecommunications awarded to N.O., JST CREST [Grant Number JPMJCR23N1], and the GteX Program Japan [Grant Number JPMJGX23B4] awarded to K.S.

897 # Supplementary Note and Tables

898

899 ## Supplementary Note for Figure 3. Definition of contribution score–
900 based regions, thresholds, and enrichment analyses

901 **Contribution score ranking and binning**

902 For each nucleic acid-binding protein (NABP), DeepLIFT contribution scores were
903 computed for all predicted gene-level regulatory regions and ranked from highest positive
904 to lowest negative values. Ranked scores were divided into consecutive bins of 50 ranks,
905 yielding 26 bins per NABP (rank 1: 1–50; rank 2: 51–100; …; rank 26: 1,251–1,310),
906 unless otherwise specified.

907

908 **Definition of high- and low-contribution regions**

909 High- and low-contribution regions were defined using both rank-based and score-
910 based thresholds to assess the robustness of enrichment results. Rank-based thresholds
911 included upper and lower rank cutoffs (e.g., top or bottom 20 and 50 ranks), while score-
912 based thresholds were defined using absolute contribution score cutoffs (e.g., ≥ 0.01 or ≥
913 0.001). All threshold combinations used in the analyses are summarized in **Tables S2** and
914 **S3.**

915

916 **Genomic regions used for overlap analyses**

917 For DNA-binding proteins, ChIP-seq overlap analyses were conducted within two
918 genomic contexts: promoter-only regions defined as ±5 kb from the transcription start site
919 (TSS), and extended promoter–enhancer regions defined as ±100 kb from the TSS. For
920 RNA-binding proteins, eCLIP overlap analyses were restricted to transcribed regions

921 spanning from the TSS to the transcription end site (TES), including both exonic and
922 intronic intervals.
923
924 **Random controls**
925 Random control regions were generated by sampling genomic intervals matched
926 for size and genomic context, excluding annotated binding sites, to estimate the expected
927 overlap by chance. All enrichment and overlap analyses were repeated using these
928 controls for comparison.
929
930 **Background configuration for DeepLIFT analyses**
931 Unless otherwise noted, DeepLIFT contribution scores were computed using a
932 background defined as 0.5× the original input feature values. Analyses using an
933 alternative background configuration (input × 2) yielded qualitatively similar results and
934 are presented in **Supplementary Figures S2** and **S3.**
935
936 **Cell type configurations used for model training and comparative analyses**
937 To accommodate differences in experimental data availability and analytical
938 objectives, multiple cell type configurations were used across analyses. For the primary
939 analyses of DNA-binding proteins (DBPs), models were trained using expression data
940 from HFFs, HMECs, NPCs, and HepG2, enabling validation against ChIP-seq datasets
941 available for HepG2 cells. For RNA-binding protein (RBP) analyses, models were trained
942 using HFFs, HMECs, NPCs, and K562 to facilitate validation using eCLIP datasets
943 available for K562 cells. For analyses requiring direct comparison of regulatory
944 contribution patterns between HepG2 and K562 for the same nucleic acid-binding

945 proteins (e.g., **Fig. 3d**), a unified four–cell-type configuration comprising HFFs, NPCs,
946 HepG2, and K562 was employed. This configuration enabled direct comparison of
947 DeepLIFT contribution score distributions across these two cell types while maintaining
948 a shared transcriptomic background. In all cases, cell type selection was driven by
949 analytical requirements and the availability of independent experimental validation data,
950 and no validation datasets were used as direct inputs during model training.
951

952 Gene co-expression–based predictions align with ChIP-seq–defined 953 DNA-binding targets

954 High-ranking predicted binding regions frequently coincide with ChIP-seq peaks,
955 particularly in promoter–enhancer intervals. Deep learning–derived DNA-binding site
956 predictions based on gene co-expression data show substantial concordance with ChIP-
957 seq results, especially when broader genomic regions encompassing promoters and
958 putative enhancers are considered. When overlap was assessed within ±100 kb of
959 transcription start sites (TSSs)—a window supported by prior enhancer–gene mapping
960 studies [19,67]—the mean and median overlap ratios reached 0.70 and 0.77, respectively
961 (Ratio A2 / B2; **Supplementary Tables S2 and S3**). In contrast, overlap was lower for
962 promoter regions alone (±5 kb), with corresponding mean and median values of 0.35 and
963 0.34. Several DNA-binding proteins, including AKAP8, POGK, SPEN, THAP7, ZNF142,
964 and ZNF691, showed notably high promoter-region overlap ratios (>0.5), suggesting
965 potential direct regulatory activity. Nevertheless, not all ChIP-seq peaks necessarily
966 reflect functionally relevant transcriptional binding events. For RNA-binding proteins,
967 overlap was examined within the transcribed regions of RNA molecules, and the mean
968 and median overlap ratios were 0.70 and 0.76, respectively (Ratio A2 / B2;

969 **Supplementary Table S4**). These findings indicate that model-inferred high-scoring
970 sites do not universally coincide with experimentally determined binding regions,
971 underscoring the complexity and context dependence of regulatory interactions.
972
973

974 Supplementary Table S1. Co-expression–based selection enhances the 975 contribution scores of NABP binding sites.

976 Binding sites identified through gene co-expression analysis exhibited higher
977 absolute DeepLIFT contribution scores than those selected solely from ChIP-seq data
978 without co-expression information. This trend was consistently observed for DNA-
979 binding proteins and was significantly stronger for RNA-binding proteins across the
980 analyzed cell types. For each group, the table presents the mean and median values of
981 both positive and negative contribution scores. These results indicate that gene co-
982 expression–based selection preferentially highlights functionally relevant regulatory
983 interactions. The upper and lower sections of the table correspond to analyses performed
984 using DeepLIFT background references scaled to input × 0.5 and input × 2, respectively.

DBP / RBP	Cell type	**Co-expression–based predictions** Upper row: mean values Lower row: median values	**Non-co-expression–based predictions** Upper row: mean values Lower row: median values
DBP	**HepG2**	5.3.E-03 2.8.E-03	3.7.E-03 1.9.E-03
RBP	**K562**	8.0.E-03 5.6.E-03	3.6.E-03 1.8.E-03
DBP	**NPC**	3.9.E-03 1.9.E-03	3.8.E-03 1.9.E-03
RBP	**NPC**	8.2.E-03 5.9.E-03	3.6.E-03 1.7.E-03
DBP	**HMEC**	4.0.E-03 1.8.E-03	3.8.E-03 1.8.E-03
RBP	**HMEC**	8.9.E-03 6.4.E-03	3.5.E-03 1.7.E-03
DBP	**HFF**	3.9.E-03 1.7.E-03	3.8.E-03 1.8.E-03
RBP	**HFF**	8.8.E-03 6.5.E-03	3.5.E-03 1.7.E-03

DBP / RBP	Cell type	**Co-expression–based:** mean (upper) median (lower)	**Non-co-expression–based:** mean (upper) median (lower)
DBP	**HepG2**	5.28.E-03 2.78.E-03	3.71.E-03 1.90.E-03
RBP	**K562**	7.95.E-03 5.60.E-03	3.56.E-03 1.78.E-03
DBP	**NPC**	3.88.E-03 1.87.E-03	3.82.E-03 1.87.E-03
RBP	**NPC**	8.20.E-03 5.90.E-03	3.56.E-03 1.73.E-03
DBP	**HMEC**	4.01.E-03 1.80.E-03	3.81.E-03 1.80.E-03
RBP	**HMEC**	8.92.E-03 6.37.E-03	3.55.E-03 1.68.E-03
DBP	**HFF**	3.89.E-03 1.70.E-03	3.82.E-03 1.79.E-03
RBP	**HFF**	8.76.E-03 6.49.E-03	3.55.E-03 1.71.E-03

988 **Supplementary Table S2. Enrichment of high-contribution promoter**
989 **regions with experimentally validated DBP binding sites.** (See
990 Supplementary Excel file)

991 This table summarizes the proportion of promoter regions (±5 kb from transcription
992 start sites [TSSs]) that were predicted to be functionally important by DeepLIFT
993 contribution scores and validated by ChIP-seq data for the corresponding DNA-binding
994 proteins (DBPs). Overlap ratios were calculated across four thresholds—two based on
995 rank and two based on absolute score—to assess the robustness of the predictions.

996 High- and low-contribution regions were classified using the following thresholds
997 for the 1,310 NABPs analyzed:

998 **Rank-based thresholds**:

999 (A1 and B1) proteins ranked in the top or bottom 20 (i.e., rank ≤ 20 or ≥ 1,290)

1000 (A2 and B2) proteins ranked in the top or bottom 50 (i.e., rank ≤ 50 or ≥ 1,260)

1001 **Score-based thresholds**:

1002 (C1 and D1) contribution score ≥ 0.01 or ≤ –0.01

1003 (C2 and D2) contribution score ≥ 0.001 or ≤ –0.001

1004

1005

1006 **Supplementary Table S3. Enrichment of high-contribution promoter**
1007 **and enhancer regions with ChIP-seq–validated DBP binding sites.**

1008 This table summarizes the proportion of predicted high-contribution regions—
1009 defined within ±100 kb of transcription start sites (TSSs), encompassing both promoter
1010 and distal enhancer regions—that overlap with ChIP-seq binding sites for the
1011 corresponding DNA-binding proteins (DBPs). Strong concordance between model-

1012 derived predictions and ChIP-seq peaks across multiple thresholds highlights the
1013 reliability of the contribution scoring method for identifying functionally relevant
1014 regulatory sites.

1015 High- and low-contribution regions were classified using the following thresholds
1016 for the 1,310 NABPs analyzed:

1017 **Rank-based thresholds**:

1018 (A1 and B1) proteins ranked in the top or bottom 20 (i.e., rank ≤ 20 or ≥ 1,290)

1019 (A2 and B2) proteins ranked in the top or bottom 50 (i.e., rank ≤ 50 or ≥ 1,260)

1020 **Score-based thresholds**:

1021 (C1 and D1) contribution score ≥ 0.01 or ≤ –0.01

1022 (C2 and D2) contribution score ≥ 0.001 or ≤ –0.001

1023

1024

1025 Supplementary Table S4. Enrichment of high-contribution transcribed 1026 regions with eCLIP–validated RBP binding sites.

1027 This table summarizes the proportion of predicted high-contribution regions—
1028 defined within the transcribed regions of genes (from transcription start site [TSS] to
1029 transcription end site [TES])—that overlap with eCLIP peaks for the corresponding
1030 RNA-binding proteins (RBPs). Strong concordance between model-based predictions
1031 and eCLIP peaks across multiple thresholds underscores the reliability of the contribution
1032 scoring method for identifying functionally relevant regulatory sites.

1033 High- and low-contribution regions were classified using the following
1034 thresholds applied to the 1,310 NABPs analyzed:

1035 **Rank-based thresholds**:

1036 (A1 and B1) proteins ranked in the top or bottom 20 (i.e., rank ≤ 20 or ≥ 1,290)

1037 (A2 and B2) proteins ranked in the top or bottom 50 (i.e., rank ≤ 50 or ≥ 1,260)

1038 **Score-based thresholds**:

1039 (C1 and D1) contribution score ≥ 0.01 or ≤ –0.01

1040 (C2 and D2) contribution score ≥ 0.001 or ≤ –0.001

1041

1042

1043 **References**

1044

1045 1 Belz, G. T. & Nutt, S. L. Transcriptional programming of the dendritic cell network.
1046 *Nature reviews. Immunology* **12**, 101-113 (2012). https://doi.org/10.1038/nri3149
1047 2 Miller, J. C. *et al.* Deciphering the transcriptional network of the dendritic cell lineage.
1048 *Nature immunology* **13**, 888-899 (2012). https://doi.org/10.1038/ni.2370
1049 3 Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man.
1050 *Cell* **140**, 744-752 (2010). https://doi.org/10.1016/j.cell.2010.01.044
1051 4 Gertz, J. *et al.* Distinct properties of cell-type-specific and shared transcription factor
1052 binding sites. *Molecular cell* **52**, 25-36 (2013).
1053 https://doi.org/10.1016/j.molcel.2013.08.037
1054 5 Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665 (2018).
1055 https://doi.org/10.1016/j.cell.2018.01.029
1056 6 Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of
1057 human transcription factors: function, expression and evolution. *Nature reviews. Genetics*
1058 **10**, 252-263 (2009). https://doi.org/10.1038/nrg2538
1059 7 Pierson, E. *et al.* Sharing and Specificity of Co-expression Networks across 35 Human
1060 Tissues. *PLoS Comput Biol* **11**, e1004220 (2015).
1061 https://doi.org/10.1371/journal.pcbi.1004220
1062 8 Sonawane, A. R. *et al.* Understanding Tissue-Specific Gene Regulation. *Cell Rep* **21**,
1063 1077-1088 (2017). https://doi.org/10.1016/j.celrep.2017.10.001
1064 9 Saha, A. *et al.* Co-expression networks reveal the tissue-specific regulation of transcription
1065 and splicing. *Genome research* **27**, 1843-1858 (2017).
1066 https://doi.org/10.1101/gr.216721.116
1067 10 Boyle, A. P. *et al.* Comparative analysis of regulatory information and circuits across

1068 distant species. *Nature* **512**, 453-456 (2014). https://doi.org/10.1038/nature13668
1069 11 Ray, D. *et al.* RNA-binding proteins that lack canonical RNA-binding domains are rarely
1070 sequence-specific. *Sci Rep* **13**, 5238 (2023). https://doi.org/10.1038/s41598-023-32245-
1071 9
1072 12 Zeke, A. *et al.* Deep structural insights into RNA-binding disordered protein regions.
1073 *Wiley Interdiscip Rev RNA* **13**, e1714 (2022). https://doi.org/10.1002/wrna.1714
1074 13 Zaharias, S. *et al.* Intrinsically disordered electronegative clusters improve stability and
1075 binding specificity of RNA-binding proteins. *J Biol Chem* **297**, 100945 (2021).
1076 https://doi.org/10.1016/j.jbc.2021.100945
1077 14 Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J. & Mann, R. S. Low-Affinity Binding Sites
1078 and the Transcription Factor Specificity Paradox in Eukaryotes. *Annu Rev Cell Dev Biol*
1079 **35**, 357-379 (2019). https://doi.org/10.1146/annurev-cellbio-100617-062719
1080 15 Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks.
1081 *Cell* **150**, 1274-1286 (2012). https://doi.org/10.1016/j.cell.2012.04.040
1082 16 Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular
1083 perturbations across complex diseases. *Nature methods* **13**, 366-370 (2016).
1084 https://doi.org/10.1038/nmeth.3799
1085 17 Fagny, M. *et al.* Exploring regulation in tissues with eQTL networks. *Proc Natl Acad Sci*
1086 *U S A* **114**, E7841-E7850 (2017). https://doi.org/10.1073/pnas.1707375114
1087 18 Agarwal, V. *et al.* Massively parallel characterization of transcriptional regulatory elements.
1088 *Nature* (2025). https://doi.org/10.1038/s41586-024-08430-9
1089 19 Avsec, Z. *et al.* Effective gene expression prediction from sequence by integrating long-
1090 range interactions. *Nature methods* **18**, 1196-1203 (2021).
1091 https://doi.org/10.1038/s41592-021-01252-x
1092 20 Tasaki, S., Gaiteri, C., Mostafavi, S. & Wang, Y. Deep learning decodes the principles of
1093 differential gene expression. *Nat Mach Intell* **2**, 376-386 (2020).
1094 https://doi.org/10.1038/s42256-020-0201-6
1095 21 Osato, N. & Hamada, M. Systematic discovery of directional regulatory motifs associated
1096 with human insulator sites. *bioRxiv* (2025).
1097 https://doi.org/https://doi.org/10.1101/2024.01.20.573595
1098 22 Obayashi, T., Kodate, S., Hibara, H., Kagaya, Y. & Kinoshita, K. COXPRESdb v8: an
1099 animal gene coexpression database navigating from a global view to detailed investigations.
1100 *Nucleic acids research* **51**, D80-D87 (2023). https://doi.org/10.1093/nar/gkac983
1101 23 Yang, S. *et al.* COEXPEDIA: exploring biomedical hypotheses via co-expressions
1102 associated with medical subject headings (MeSH). *Nucleic acids research* **45**, D389-D396
1103 (2017). https://doi.org/10.1093/nar/gkw868

24 Zhu, Q. *et al.* Targeted exploration and analysis of large cross-platform human transcriptomic compendia. *Nature methods* **12**, 211-214, 213 p following 214 (2015). https://doi.org/10.1038/nmeth.3249

25 Miller, H. E. & Bishop, A. J. R. Correlation AnalyzeR: functional predictions from gene co-expression correlations. *BMC Bioinformatics* **22**, 206 (2021). https://doi.org/10.1186/s12859-021-04130-7

26 Raina, P. *et al.* GeneFriends: gene co-expression databases and tools for humans and model organisms. *Nucleic acids research* **51**, D145-D158 (2023). https://doi.org/10.1093/nar/gkac1031

27 Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)* **302**, 249-255 (2003). https://doi.org/10.1126/science.1087447

28 Lemoine, G. G., Scott-Boyer, M. P., Ambroise, B., Perin, O. & Droit, A. GWENA: gene co-expression networks analysis and extended modules characterization in a single Bioconductor package. *BMC Bioinformatics* **22**, 267 (2021). https://doi.org/10.1186/s12859-021-04179-4

29 Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research* **38**, W214-220 (2010). https://doi.org/10.1093/nar/gkq537

30 Wong, A. K., Krishnan, A. & Troyanskaya, O. G. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic acids research* **46**, W65-W70 (2018). https://doi.org/10.1093/nar/gky408

31 UniProt, C. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic acids research* **53**, D609-D617 (2025). https://doi.org/10.1093/nar/gkae1010

32 Tak Leung, R. W., Jiang, X., Chu, K. H. & Qin, J. ENPD - A Database of Eukaryotic Nucleic Acid Binding Proteins: Linking Gene Regulations to Proteins. *Nucleic acids research* **47**, D322-D329 (2019). https://doi.org/10.1093/nar/gky1112

33 Shrikumar, A., Greenside, P. & Kundaje, A. in *Proceedings of the 34th International Conference on Machine Learning* Vol. 70 (eds Precup Doina & Teh Yee Whye) 3145--3153 (PMLR, Proceedings of Machine Learning Research, 2017).

34 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29 (2000). https://doi.org/10.1038/75556

35 Gene Ontology, C. *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224** (2023). https://doi.org/10.1093/genetics/iyad031

36 Thomas, P. D. *et al.* PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci* **31**, 8-22 (2022). https://doi.org/10.1002/pro.4218

37 Hu, M. *et al.* Evaluation of large language models for discovery of gene set function. *Nature methods* **22**, 82-91 (2025). https://doi.org/10.1038/s41592-024-02525-x

38 Zahra, K., Dey, T., Ashish, Mishra, S. P. & Pandey, U. Pyruvate Kinase M2 and Cancer: The Role of PKM2 in Promoting Tumorigenesis. *Front Oncol* **10**, 159 (2020). https://doi.org/10.3389/fonc.2020.00159

39 Zhang, Z. *et al.* PKM2, function and expression and regulation. *Cell Biosci* **9**, 52 (2019). https://doi.org/10.1186/s13578-019-0317-8

40 Alquraishi, M. *et al.* Pyruvate kinase M2: A simple molecule with complex functions. *Free Radic Biol Med* **143**, 176-192 (2019). https://doi.org/10.1016/j.freeradbiomed.2019.08.007

41 Azoitei, N. *et al.* PKM2 promotes tumor angiogenesis by regulating HIF-1alpha through NF-kappaB activation. *Mol Cancer* **15**, 3 (2016). https://doi.org/10.1186/s12943-015-0490-2

42 He, P. *et al.* PKM2 is a key factor to regulate neurogenesis and cognition by controlling lactate homeostasis. *Stem Cell Reports* **20**, 102381 (2025). https://doi.org/10.1016/j.stemcr.2024.11.011

43 Bonar, N. A. & Petersen, C. P. Integrin suppresses neurogenesis and regulates brain tissue assembly in planarian regeneration. *Development* **144**, 784-794 (2017). https://doi.org/10.1242/dev.139964

44 Porcheri, C., Suter, U. & Jessberger, S. Dissecting integrin-dependent regulation of neural stem cell proliferation in the adult brain. *J Neurosci* **34**, 5222-5232 (2014). https://doi.org/10.1523/JNEUROSCI.4928-13.2014

45 Zheng, X. *et al.* Metabolic reprogramming during neuronal differentiation from aerobic glycolysis to neuronal oxidative phosphorylation. *eLife* **5** (2016). https://doi.org/10.7554/eLife.13374

46 Agostini, M. *et al.* Metabolic reprogramming during neuronal differentiation. *Cell Death Differ* **23**, 1502-1514 (2016). https://doi.org/10.1038/cdd.2016.36

47 Saez, I. *et al.* Neurons have an active glycogen metabolism that contributes to tolerance to hypoxia. *J Cereb Blood Flow Metab* **34**, 945-955 (2014). https://doi.org/10.1038/jcbfm.2014.33

48 Matthews, H. K., Bertoli, C. & de Bruin, R. A. M. Cell cycle control in cancer. *Nat Rev Mol Cell Biol* **23**, 74-88 (2022). https://doi.org/10.1038/s41580-021-00404-3

49 Dominguez-Brauer, C. *et al.* Targeting Mitosis in Cancer: Emerging Strategies. *Molecular cell* **60**, 524-536 (2015). https://doi.org/10.1016/j.molcel.2015.11.006

50 Levine, M. S. & Holland, A. J. The impact of mitotic errors on cell proliferation and tumorigenesis. *Genes Dev* **32**, 620-638 (2018). https://doi.org/10.1101/gad.314351.118

51 Huang, X. *et al.* D-Serine regulates proliferation and neuronal differentiation of neural stem cells from postnatal mouse forebrain. *CNS Neurosci Ther* **18**, 4-13 (2012). https://doi.org/10.1111/j.1755-5949.2011.00276.x

52 El-Hattab, A. W. Serine biosynthesis and transport defects. *Mol Genet Metab* **118**, 153-159 (2016). https://doi.org/10.1016/j.ymgme.2016.04.010

53 Kaida, A. & Iwakuma, T. Regulation of p53 and Cancer Signaling by Heat Shock Protein 40/J-Domain Protein Family Members. *Int J Mol Sci* **22** (2021). https://doi.org/10.3390/ijms222413527

54 Sterrenberg, J. N., Blatch, G. L. & Edkins, A. L. Human DNAJ in cancer and stem cells. *Cancer Lett* **312**, 129-142 (2011). https://doi.org/10.1016/j.canlet.2011.08.019

55 Ladomery, M. Multifunctional proteins suggest connections between transcriptional and post-transcriptional processes. *Bioessays* **19**, 903-909 (1997). https://doi.org/10.1002/bies.950191010

56 Kim, S., Yu, N. K. & Kaang, B. K. CTCF as a multifunctional protein in genome regulation and gene expression. *Exp Mol Med* **47**, e166 (2015). https://doi.org/10.1038/emm.2015.33

57 Osato, N. Characteristics of functional enrichment and gene expression level of human putative transcriptional target genes. *BMC Genomics* **19**, 957 (2018). https://doi.org/10.1186/s12864-017-4339-5

58 Burley, S. K. & Kamada, K. Transcription factor complexes. *Curr Opin Struct Biol* **12**, 225-230 (2002). https://doi.org/10.1016/s0959-440x(02)00314-7

59 Goos, H. *et al.* Human transcription factor protein interaction networks. *Nat Commun* **13**, 766 (2022). https://doi.org/10.1038/s41467-022-28341-5

60 Berenson, A. *et al.* Paired yeast one-hybrid assays to detect DNA-binding cooperativity and antagonism across transcription factors. *Nat Commun* **14**, 6570 (2023). https://doi.org/10.1038/s41467-023-42445-6

61 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research* **41**, D991-995 (2013). https://doi.org/10.1093/nar/gks1193

62 Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**, 207-210 (2002). https://doi.org/10.1093/nar/30.1.207

63 Kolmykov, S. *et al.* GTRD: an integrated view of transcription regulation. *Nucleic acids research* **49**, D104-D111 (2021). https://doi.org/10.1093/nar/gkaa1057

64 Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic acids research* **34**, D590-598 (2006). https://doi.org/10.1093/nar/gkj144

65 Mudge, J. M. *et al.* GENCODE 2025: reference gene annotation for human and mouse.

Nucleic acids research **53**, D966-D975 (2025). https://doi.org/10.1093/nar/gkae1078

66 Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711-719 (2020). https://doi.org/10.1038/s41586-020-2077-3

67 Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 1516 (2019). https://doi.org/10.1016/j.cell.2019.02.027