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, 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 $< 10^{-2}$; **Fig. 2d** and **Supplementary Fig. S2c**; **Supplementary Table S1**). Each pair of
122 violin plots represents absolute contribution scores for co-expression–based sites (right
123 plot in each pair) and the corresponding ChIP-seq sites not included in that set (left plot).
124 **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

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

158 To assess the influence of background selection on DeepLIFT-derived contribution
159 scores, we compared results obtained using a reduced background (input × 0.5; **Fig. 2e**)
160 and an amplified background (input × 2; **Fig. 2f**) for the same representative DNA- and
161 RNA-binding proteins. Under the reduced background, contribution scores tended to
162 show negative correlations with gene expression, whereas the amplified background
163 produced predominantly positive correlations. These patterns appeared approximately
164 mirrored around zero, consistent with DeepLIFT's definition as a relative contribution
165 measure with respect to the chosen reference. Importantly, despite this inversion in sign,
166 the relative ranking of genes and the overall cell type–specific structure of contribution
167 score distributions remained highly consistent across background settings. In both cases,
168 genes with larger-magnitude contribution scores were preferentially associated with
169 higher or lower expression levels depending on the background, indicating that
170 DeepLIFT captures relative regulatory importance rather than absolute activation or
171 repression. A conceptual model explaining how input definition and background scaling
172 give rise to these patterns is provided in **Supplementary Fig. S4**. Notably, background-
173 dependent differences were more pronounced for RNA-binding proteins (RBPs) (**Fig.**
174 **2e,f)**. For the same representative RBPs, scatter plots generated using the reduced
175 background (input × 0.5) exhibited more distinct, partially separated distributions across
176 the four cell types, whereas under the amplified background (input × 2), data points from
177 different cell types overlapped more extensively. This suggests that the reduced-

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

provides a robust framework for dissecting context-dependent regulatory programs and linking NABPs to biologically meaningful pathways. While individual Reactome pathways detected by GSEA varied partially between background configurations, higher-level functional categories showed consistent enrichment trends across backgrounds. This suggests that differences at the individual pathway level likely reflect variations in statistical detection sensitivity associated with the number of predicted target genes, rather than qualitative changes in inferred PKM-associated functions.

Discussion

These findings address a long-standing challenge in molecular biology: the systematic functional characterization of nucleic acid-binding proteins (NABPs) across diverse cellular contexts. While traditional methods rely heavily on experimental binding data and known motifs, our approach utilizes gene co-expression and machine learning to circumvent these limitations. By capturing both direct and indirect regulatory signals, it broadens the scope of functional prediction for NABPs that lack conventional binding site information.

Gene expression correlations alone can serve as a powerful basis for uncovering the regulatory roles of nucleic acid-binding proteins (NABPs), even in the absence of experimental binding data. Leveraging transcriptomic profiles from diverse public RNA-seq datasets, our deep learning framework inferred regulatory targets of NABPs based solely on co-expression patterns. Unlike binding site assays such as ChIP-seq or eCLIP—which are technically demanding and limited by the availability of high-quality antibodies—this computational approach is scalable and applicable across NABPs with

345 poorly characterized binding motifs. Functional enrichment analysis of the predicted
346 target genes revealed biologically meaningful roles, many of which were consistent with
347 existing annotations in public databases. In addition, several previously unannotated
348 functional associations—not captured in UniProt or GO but supported by published
349 studies—were uncovered. GSEA further demonstrated that genes contributing to specific
350 cell type–associated pathways exhibited systematic shifts toward higher or lower
351 DeepLIFT scores, indicating differential regulatory influence across cellular contexts.
352 These findings were corroborated by a ChatGPT-assisted literature mining approach,
353 which independently recovered concordant biological themes and supporting references.
354 Together, these results underscore the biological relevance of NABP target predictions
355 derived from deep learning–based co-expression modeling and highlight the value of
356 integrating such predictions with AI-assisted interpretation. This strategy provides a
357 scalable framework for elucidating NABP functions and discovering previously
358 unrecognized regulatory mechanisms across diverse cell types.

359 Contribution scores derived from gene expression correlations capture both direct
360 and indirect regulatory interactions, offering insights beyond what ChIP-seq alone can
361 provide. While ChIP-seq detects thousands of DNA-binding events per protein, many of
362 these sites do not contribute directly to gene expression regulation and may reflect the
363 multifunctionality of transcription factors [55,56]. Consequently, ChIP-seq–derived binding
364 maps and DeepLIFT-based predictions show only partial overlap, underscoring both
365 methodological differences and the inherent complexity of regulatory networks. In this
366 study, gene co-expression analysis revealed fewer target genes than ChIP-seq but
367 identified subsets that are likely to be more functionally relevant. Our previous work on
368 insulator-associated factors linked such proteins to enhancer–promoter insulation and

369 chromatin organization, including phase separation and loop formation [21]. That study
370 incorporated eQTL data to anchor binding sites to expression outcomes. Here, by contrast,
371 we focused on NABPs with unknown motifs and excluded motif-based or chromatin
372 accessibility data such as DNase-seq or ATAC-seq, to assess the capacity of gene
373 expression correlations to reveal regulatory roles. Notably, the distribution of
374 contribution scores for co-expression–based inputs included both high and low values,
375 implying that gene expression patterns reflect not only direct binding but also indirect
376 regulation mediated through broader network effects. This highlights the value of
377 contribution scores as indicators of both proximal and distal regulatory influences, even
378 when binding motifs are unknown or experimental binding data are unavailable. The cell
379 type–specific differences in DeepLIFT contribution score distributions observed in **Fig.**
380 **3d** suggest that the regulatory roles of individual nucleic acid-binding proteins are not
381 fixed but are dynamically reshaped across cellular contexts. Notably, while some proteins
382 such as RPLP0 exhibited highly conserved contribution patterns between HepG2 and
383 K562 cells, others, including ETF1 and NELFE, displayed pronounced divergence,
384 indicating cell type–dependent shifts in their inferred regulatory influence. These findings
385 support the notion that NABPs participate in distinct regulatory programs depending on
386 cellular state and lineage, and that DeepLIFT-based analyses can sensitively capture such
387 context-dependent regulatory reorganization. This context dependence is particularly
388 relevant for interpreting NABP function in disease versus developmental states, where
389 subtle shifts in regulatory importance may precede overt transcriptional changes.

390 Functional roles of NABPs can be inferred even from small gene sets, provided
391 that the target genes exhibit shared or functionally related patterns. In this study, twelve
392 NABPs showed functional enrichments consistent with known annotations, supporting

393 the validity of the deep learning–based predictions. These enrichments were identified
394 using Fisher's exact test with false discovery rate (FDR) correction via the PANTHER
395 database, which typically requires at least 50 genes for robust statistical significance. For
396 proteins with fewer than 50 predicted targets, enrichment may be harder to detect
397 depending on the functional composition of the gene set. Relaxing the ranking thresholds
398 increased the number of target genes and enhanced statistical power. Alternatively, more
399 stringent thresholds could be retained if the contribution scores were sufficiently strong,
400 offering flexibility depending on analysis goals. Additional features—such as gene
401 expression levels or co-expression correlations—could further refine enrichment analyses,
402 particularly for borderline cases. Notably, enrichment analysis of both high- and low-
403 ranking targets revealed candidate novel functions for specific NABPs, suggesting that
404 genes with large-magnitude contribution scores are biologically informative. Moreover,
405 a complementary ChatGPT-based approach demonstrated utility in interpreting small
406 gene sets. Unlike standard enrichment tools, this method can infer biological roles even
407 when gene counts fall below conventional significance thresholds, provided that genes
408 with related functions are present. This capacity helps overcome statistical limitations and
409 broadens the interpretability of regulatory predictions in a cell type–specific context.
410 GSEA based on DeepLIFT scores provides a complementary view of pathway-level
411 regulation by assessing how genes associated with the same pathway are distributed along
412 the model's importance landscape, rather than relying on gene expression levels. This
413 approach revealed striking differences between normal and disease-associated cell types.
414 For example, genes enriched in neural progenitor cell (NPC) pathways tended to cluster
415 at higher or lower DeepLIFT scores in NPCs, whereas their score distributions shifted
416 substantially in K562 cells, indicating altered regulatory programs between the two cell

417 states. Such analyses offer a powerful means to uncover context-dependent regulatory
418 mechanisms across diverse cell types, conditions, and developmental stages. The
419 increased enrichment strength observed under the amplified background configuration is
420 consistent with gene-level analyses (**Fig. 3d**), which indicate that this setting identifies a
421 broader set of PKM-associated regulatory targets. Thus, background scaling primarily
422 affects sensitivity, while preserving relative cell type–specific enrichment patterns.

423 Strategic use of co-expression data enhances functional inference, while
424 indiscriminate replacement can diminish biological relevance. The selective integration
425 of gene co-expression data improved target prediction and the functional interpretability
426 of the model. In contrast, replacing all input features with co-expression-derived
427 interactions led to a decrease in enriched annotations, and known functional roles were
428 no longer recovered. Excessive replacement of input features with gene co-expression
429 data further diminished the biological relevance of model predictions. When all NABP–
430 target inputs were substituted with co-expression-based interactions, the number of
431 enriched functional annotations declined, and previously identified functions from the
432 UniProt database were no longer recovered. This suggests that indiscriminate
433 replacement may cause the model to capture non-informative or biologically spurious
434 associations. A similar trend was observed in a prior control analysis, where substituting
435 true regulatory targets with randomly selected genes led to a gradual loss of enrichment
436 as randomness increased. Based on these observations, the normalized count of enriched
437 functional terms—adjusted by the number of predicted targets—may serve as a useful
438 proxy for evaluating the accuracy and biological validity of NABP–target predictions [57].
439 These findings also raise intriguing questions about how the deep learning model
440 prioritizes informative input features during training. Future work exploring the model's

441 internal representation dynamics—such as how it distinguishes signal from noise when
442 presented with diverse regulatory interactions—may provide valuable insights into the
443 mechanisms of transcriptional regulation and model interpretability. Investigating how
444 the model converges on biologically meaningful NABP–target relationships during
445 training could deepen our understanding of both gene regulation and model
446 interpretability.

447 Cell type specificity and cooperative binding dynamics may underlie the absence
448 of detectable regulatory signals for certain NABPs. While some NABPs showed no
449 high- or low-ranking target genes in HepG2 or K562 cells, such absence does not rule out
450 regulatory relevance. These proteins may exhibit cell type–specific regulatory activity
451 detectable only in other cellular contexts, where functional enrichment analysis could
452 reveal distinct biological roles. Moreover, many transcription factors and DNA-binding
453 proteins act in concert, forming multi-protein complexes that bind cooperatively to shared
454 promoters [58-60]. In such cases, individual components may not always yield strong
455 contribution scores due to their distributed regulatory influence across the complex. This
456 highlights the potential of co-binding site analysis as a complementary strategy to uncover
457 cooperative regulatory mechanisms and to disentangle the functional interplay of NABPs
458 within transcriptional networks.

459 Beyond characterizing cell type–specific regulatory programs, our framework also
460 suggests a potential application to preventive and health-oriented research. By
461 systematically scoring "cleaning" pathways—including proteostasis, DNA damage repair,
462 RNA quality control, immune surveillance, and metabolic stress responses—based on the
463 predicted regulatory influence of nucleic acid–binding proteins, it becomes possible to
464 estimate the integrity of cellular maintenance systems before overt disease onset. Such

465 pathway-level scores may capture early functional weakening of protective mechanisms
466 that are not yet reflected by pathological phenotypes or single-gene alterations. This
467 perspective aligns with observations from cancer-resistant species and long-lived
468 individuals, in which enhanced cellular maintenance, quality control, and stress
469 adaptation are thought to contribute to disease avoidance rather than treatment. In this
470 context, deviations in cleaning pathway scores could serve as quantitative indicators of
471 declining cellular robustness, potentially enabling risk stratification or early detection of
472 disease predisposition. While the present study focuses on neural progenitor and cancer
473 cell models, the same analytical framework could, in principle, be extended to
474 longitudinal human data or population-scale transcriptomic profiles, providing a
475 foundation for preventive or health-monitoring applications that complement
476 conventional disease-centered analyses.

477 This study establishes a scalable computational framework for uncovering the
478 functional roles of nucleic acid-binding proteins (NABPs) by integrating gene
479 expression–based associations with deep learning–derived contribution scores. By
480 leveraging co-expression–driven interactions rather than relying solely on experimentally
481 defined binding motifs, the approach enables systematic inference of both canonical and
482 previously uncharacterized regulatory relationships across cell types. Collectively, our
483 results demonstrate that expression-driven contribution patterns, when coupled with
484 pathway-level analyses and AI-assisted functional interpretation, can capture biologically
485 meaningful differences between normal and disease-associated cellular states. This
486 capability suggests that the framework may be particularly valuable for identifying early
487 shifts in regulatory and cellular maintenance pathways that precede overt disease
488 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.

Beyond delineating cell type–specific regulatory programs, the present framework highlights a general strategy for interpreting the functional roles of nucleic acid–binding proteins through expression-informed, contribution score–based analyses. By integrating gene co-expression–derived regulatory associations with deep learning–based attribution, this approach enables systematic inference of NABP functions without relying exclusively on experimentally defined binding events or motifs. Importantly, pathway-level analyses revealed coherent differences between normal and disease-associated cellular states, underscoring the capacity of contribution score–based GSEA to capture context-dependent regulatory logic that is not readily apparent from expression levels or binding data alone. More broadly, this framework suggests a potential extension toward health-oriented and preventive applications. Pathway-level scoring of cellular maintenance processes—such as proteostasis, DNA damage repair, RNA quality control, immune surveillance, and metabolic stress responses—may provide quantitative indicators of cellular robustness prior to overt disease manifestation. Such signals could reflect early functional weakening of protective regulatory programs rather than established pathological states. While this study focused on a limited set of cell types and NABPs, the approach is readily scalable to larger transcriptomic datasets and diverse biological contexts. Together, these findings establish a foundation for integrating deep

513 learning–derived regulatory contribution patterns with pathway-level interpretation,
514 offering a scalable route to functional annotation of poorly characterized NABPs and a
515 complementary perspective to binding-centric or mutation-focused analyses in functional
516 genomics and disease-related research.
517
518

519 Methods

520 Predictive modeling of gene expression across diverse cell types using 521 deep learning and NABP binding information

522 To predict gene expression across multiple human cell types, we adapted a
523 previously established deep learning framework that integrates DNA-binding site data for
524 transcription factors and other regulatory proteins. In this implementation, regulatory
525 elements located both proximally and distally (up to ±1 Mb) from transcription start sites
526 (TSSs) were included as input features (**Fig. 1**) [20,21]. To minimize the effects of cellular
527 heterogeneity in bulk tissue, RNA-seq datasets were selected at the individual cell-type
528 level. Publicly available RNA-seq data were retrieved for four functionally distinct cell
529 types: human foreskin fibroblasts (HFFs), mammary epithelial cells (HMECs), neural
530 progenitor cells (NPCs), and either HepG2 hepatoma cells or K562 leukemia cells,
531 depending on the class of regulatory protein analyzed [61,62]. Gene expression values
532 (FPKM) were computed using the RSeQC toolkit based on GENCODE v19 annotations,
533 and subsequently $\log_2$-transformed after adding a small offset to avoid undefined values.
534 Following the approach described by Tasaki et al. [20,21], expression fold changes were
535 calculated relative to the median expression level across the four cell types. For each
536 model iteration, three common cell types (HFFs, HMECs, and NPCs) were used, and

537 either HepG2 (for DNA-binding protein analysis) or K562 (for RNA-binding protein
538 analysis) was included to match the availability of ChIP-seq and eCLIP validation
539 datasets. These cell types were chosen to represent diverse biological lineages and
540 regulatory contexts. For specific analyses requiring direct comparison of DeepLIFT
541 contribution scores across cell types—such as the evaluation of cell type–dependent
542 regulatory differences for the same nucleic acid-binding protein (e.g., **Fig. 3d**)—models
543 were trained using a four-cell-type configuration comprising HFFs, NPCs, HepG2, and
544 K562. This unified configuration enabled direct comparison of contribution score
545 distributions between HepG2 and K562 while maintaining a shared background of
546 common cell types.
547

548 **Extended regulatory features and DeepLIFT-compatible model design**

549 To capture distal regulatory signals beyond promoter regions, we integrated DNA-
550 binding site data from putative enhancers into a deep learning framework for gene
551 expression prediction. Building upon our earlier work [20,21], this model extends the
552 architecture originally proposed by Tasaki et al. (2020) [20,21], which focused on promoter-
553 proximal regions and included RNA-binding protein and miRNA inputs. In our previous
554 study, aimed at identifying insulator-associated DNA-binding proteins, we deliberately
555 excluded RNA-binding proteins and miRNAs to focus on DNA-centered regulatory
556 elements. Instead, we combined DNA-binding sites from both promoter (-2 kb to +1 kb)
557 and enhancer-like regions (±1 Mb from the transcription start site [TSS]) into a unified
558 input matrix, enabling DeepLIFT analysis under the high-accuracy "genomics-default"
559 mode. In this study, we use DeepLIFT to compute attribution values for each gene. We

560 refer to these values as DeepLIFT scores, also described as importance scores or
561 contribution scores in the context of feature attribution.

562 Binding site coordinates were obtained from the GTRD database (v20.06) [63], and
563 regulatory associations were inferred using cis-eQTLs from GTEx v8. A ChIP-seq peak
564 that overlapped a 100-bp window centered on an expression-associated SNP was
565 considered linked to the corresponding transcript. Trans-eQTLs were excluded, and
566 coordinates were converted to the hg19 assembly using UCSC LiftOver [64]. To limit
567 feature dimensionality while maximizing genomic coverage, we computed peak lengths
568 for each DNA-binding protein (DBP) in 100-bp bins across ±1 Mb around the TSS, then
569 retained the maximum peak value within each 10-kb window, yielding up to 200 enhancer
570 bins. These were concatenated with 30 promoter-region bins to form a 230-bin unified
571 input layer. This single-input-layer structure was necessary because DeepLIFT's
572 genomics-default configuration ("RevealCancel rule on dense layers and Rescale rule on
573 conv layers") does not support models with multiple input layers. Concatenating all DBP-
574 related features into one matrix thus maintained compatibility while enhancing
575 interpretability.

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

583

584 **Predicting NABP regulatory targets using gene co-expression–informed**
585 **substitutions**

586 To uncover regulatory targets of NABPs without relying on direct binding site
587 assays, we implemented a co-expression–driven modification of the deep learning input.
588 Specifically, DNA-binding proteins with low contribution scores in the original model
589 were replaced with NABPs whose regulatory interactions were inferred from gene co-
590 expression data. Of the 1,310 DBPs initially included, 302 were substituted with other
591 DBPs, and in a separate RNA-binding protein analysis, 325 were replaced with RNA-
592 binding proteins (RBPs) identified from the ENPD database [32]. Putative NABP–target
593 relationships were established using co-expression profiles provided by COXPRESdb [22],
594 where a Pearson's *z*-score > 2 between a NABP-encoding gene and a non-NABP gene
595 indicated a strong co-expression link. These co-expression pairs served as proxies for
596 regulatory interactions and were used to generate new input features for the model. To
597 integrate these proxies into the deep learning framework, hypothetical NABP binding
598 sites were assigned at the transcription start site (TSS) and 2 kb upstream of each co-
599 expressed gene. The model was then retrained using this modified input, and DeepLIFT
600 was applied to calculate importance scores, evaluating the influence of each NABP–gene
601 interaction on predicted expression levels. This approach enables the functional
602 characterization of NABPs with unknown or uncharacterized binding motifs by
603 leveraging gene co-expression as a proxy for regulatory targeting.

604

605 **Prioritization of NABP regulatory targets based on importance score**
606 **rankings**

607 To identify the primary regulatory targets of each NABP, genes were ranked
608 based on the importance scores assigned to NABP binding sites by the deep learning
609 model. Genes with binding sites consistently ranked in the top five or bottom five across
610 all NABPs were selected as putative targets, reflecting high positive or negative
611 regulatory influence. These targets represent genes most strongly associated with a given
612 NABP's regulatory activity. To ensure sufficient statistical power for downstream
613 enrichment analysis, rank thresholds were adjusted for each NABP based on the number
614 of high-confidence predictions. While a contribution score rank cutoff of five was
615 sufficient in most cases, the threshold was relaxed in specific cases—for example, up to
616 rank ten for proteins, which initially yielded a smaller number of predicted targets.

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

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

655 ChatGPT-based results provided consistent and interpretable biological themes aligned
656 with the enrichment findings.
657

658 Validation of predicted NABP targets using experimentally determined 659 binding

660 To assess the biological relevance of the predicted regulatory targets, we
661 compared deep learning–inferred NABP–gene associations with experimentally
662 validated binding data. Specifically, we evaluated the spatial overlap between predicted
663 target genes and ChIP-seq peaks for 32 DNA-binding proteins (within ±5 kb of
664 transcription start sites [TSSs]) as well as eCLIP peaks for 29 RNA-binding proteins
665 (within the transcribed region from TSS to transcription end site [TES]). Gene structure
666 annotations were derived from the GENCODE v19 release [65], and binding site data were
667 obtained from the ENCODE portal and prior publications (**Supplementary Table S7**) [66].
668 This overlap analysis provides orthogonal validation for the predicted targets and
669 supports the utility of gene co-expression–based modeling for inferring NABP–target
670 relationships.
671

672 Statistical analysis

673 Functional enrichment was assessed using Fisher's exact test, with multiple testing
674 correction applied via the Benjamini–Hochberg false discovery rate (FDR) procedure.
675 Statistical significance was defined as FDR < 0.05. Distributions of DeepLIFT
676 contribution scores for nucleic acid-binding sites derived from gene co-expression data
677 and ChIP-seq data were compared using the Mann–Whitney U test (also known as the
678 Wilcoxon rank-sum test). A p-value < 0.05 was considered statistically significant. GSEA

679 (gene set enrichment analysis) was performed using DeepLIFT contribution scores of
680 DBPs and RBPs. The normalized enrichment score (NES) was used to quantify whether
681 regulatory target genes associated with specific biological functions were preferentially
682 located at the high or low end of the ranked gene list. Statistical significance was assessed
683 using permutation-based *p*-values, as implemented in the GSEA framework.
684

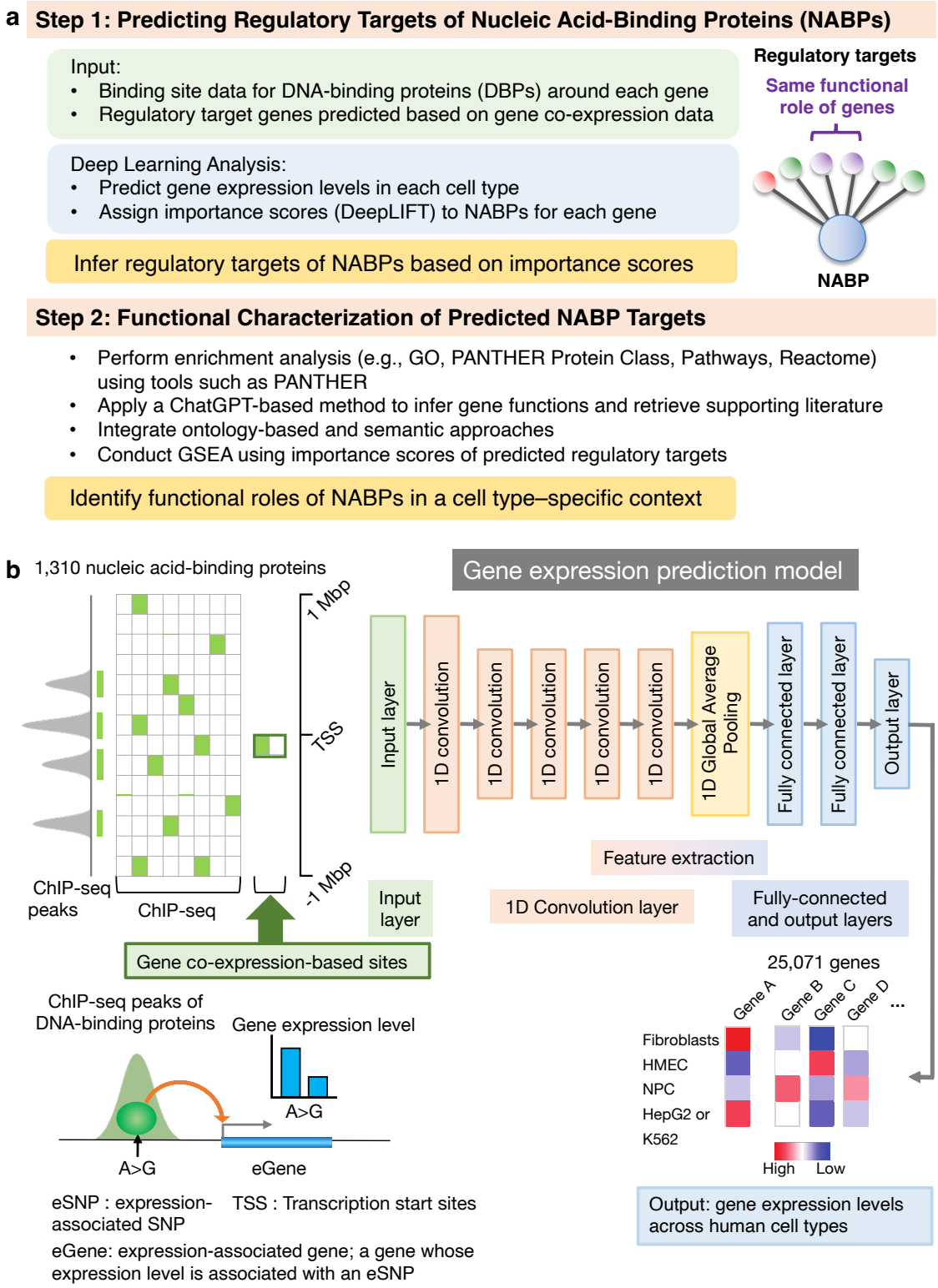
685 **Data availability**

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

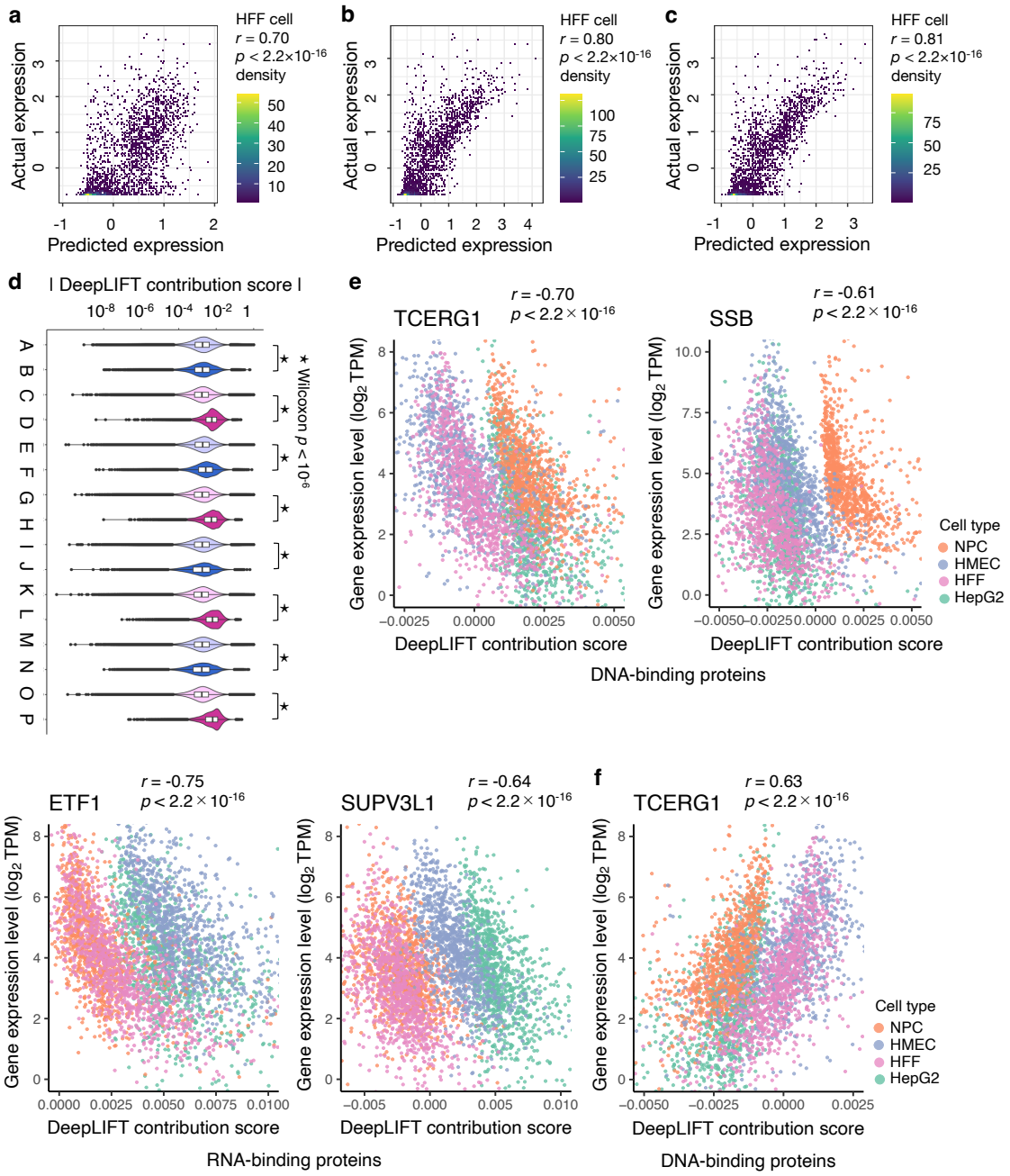
Code availability

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

711 **Figure 1. Overview of the deep learning framework for predicting and functionally**
712 **characterizing NABP regulatory targets.**
713 **a.** This diagram outlines the overall strategy used to infer the functional roles of nucleic
714 acid-binding proteins (NABPs) from gene expression data. Gene expression profiles from
715 various human cell types were integrated with NABP–gene co-expression relationships
716 derived from COXPRESdb. These inferred interactions were used to replace low-
717 contribution DNA-binding proteins in a deep learning model originally trained on ChIP-
718 seq–based binding site data. The modified model was trained to predict gene expression
719 levels and identify regulatory target genes using DeepLIFT-derived importance scores.
720 Functional annotations of the predicted target genes were assessed using conventional
721 enrichment analyses and GSEA, along with a ChatGPT-based method for literature-
722 guided interpretation.
723 **b.** Gene expression levels in human cells were predicted using the genomic locations of
724 DNA-binding sites at promoters and distal elements.
725 During model training, a DNA-binding protein bound to a distal region overlapping an
726 expression-associated SNP (eSNP) was linked to a potentially regulated gene (eGene)
727 based on eQTL data. For each gene–protein pair, ChIP-seq peak locations were binned
728 (indicated by green boxes) and formatted into a matrix as input for deep learning.
729 In co-expression–based models, binding sites were artificially assigned near transcription
730 start sites (TSSs). The deep learning model consisted of 1D convolutional layers, global
731 average pooling, and fully connected layers. It predicted gene expression levels across
732 25,071 genes and multiple human cell types.
733
734


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
f
TCERG1
r = 0.63
p < 2.2 × 10-16
RNA-binding proteins
DNA-binding proteins

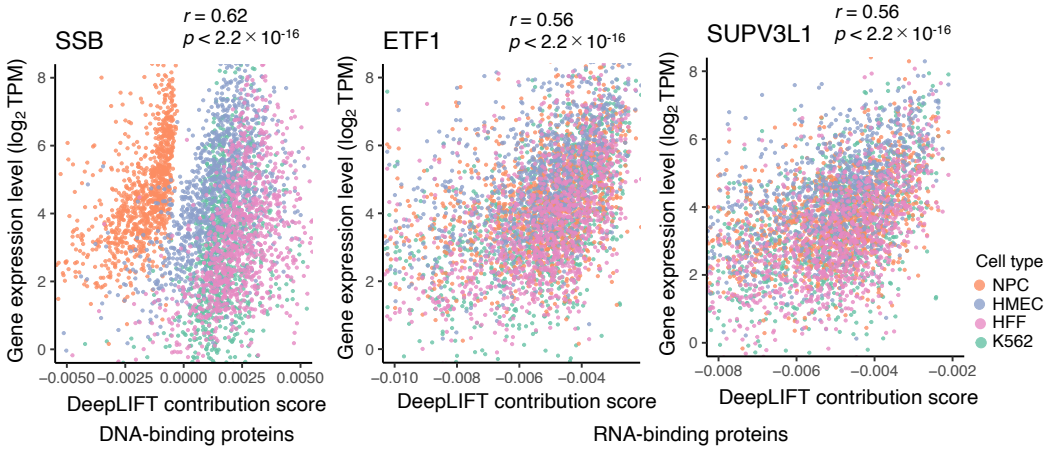


Figure 2. Relationship between gene expression levels and contribution scores.

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.

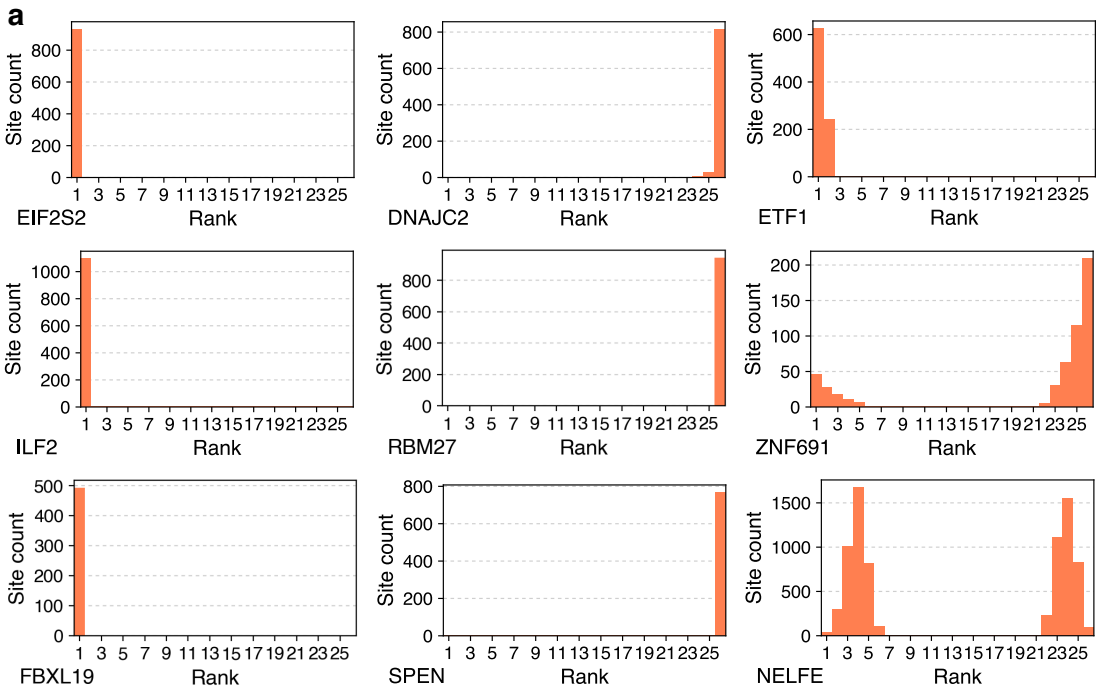
a. DNA-binding sites of 1,310 proteins derived from ChIP-seq data.

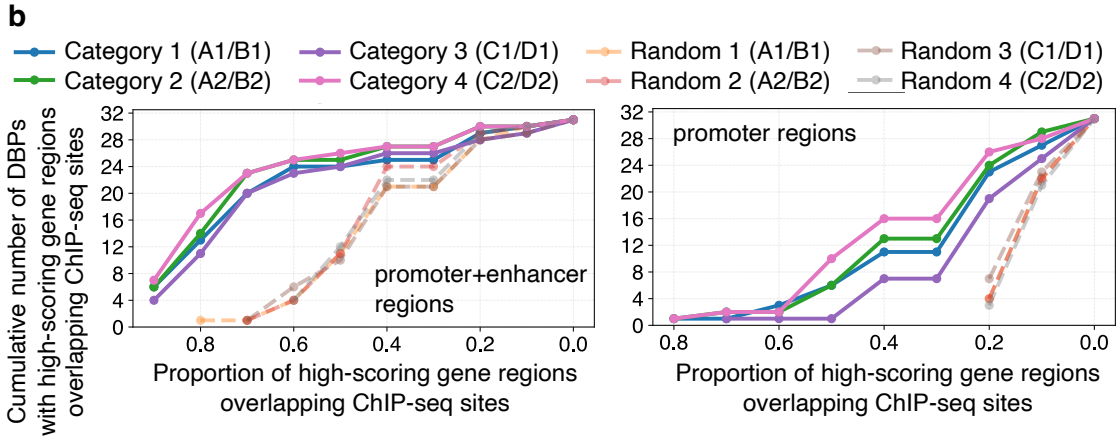
b. Replacement of 302 low-contribution DNA-binding proteins with co-expression–derived DNA-binding proteins.

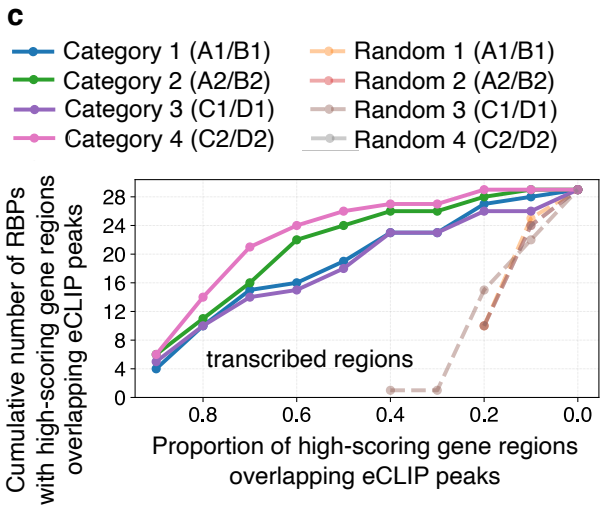
c. Replacement of 325 DNA-binding proteins with RNA-binding proteins identified from co-expression relationships.

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-values $< 10^{-6}$). 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. For each protein, the Spearman correlation coefficient and corresponding *p*-value displayed above the plot correspond to the cell type showing the strongest correlation among the four analyzed cell types. Data points are color-coded by cell type, revealing consistent distributional trends across all four cell types.




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

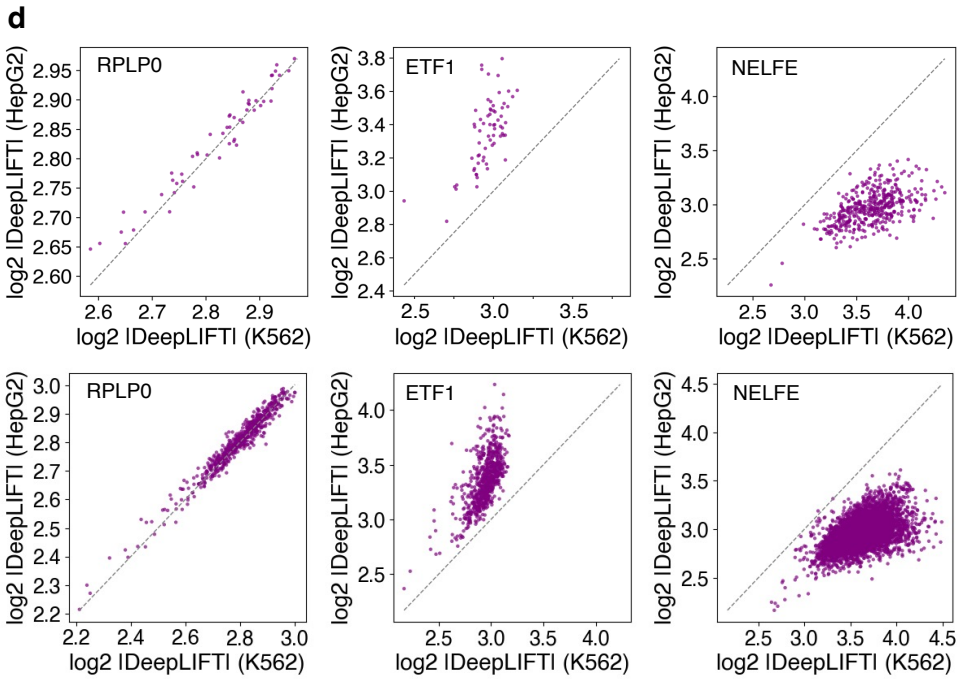


Figure 3. Comparison of predicted regulatory target genes with experimental ChIP-seq and eCLIP binding data. This figure illustrates the correspondence between gene-level regulatory contribution scores inferred by deep learning and experimentally identified nucleic acid-binding protein (NABP) binding 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

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

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

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

802

803

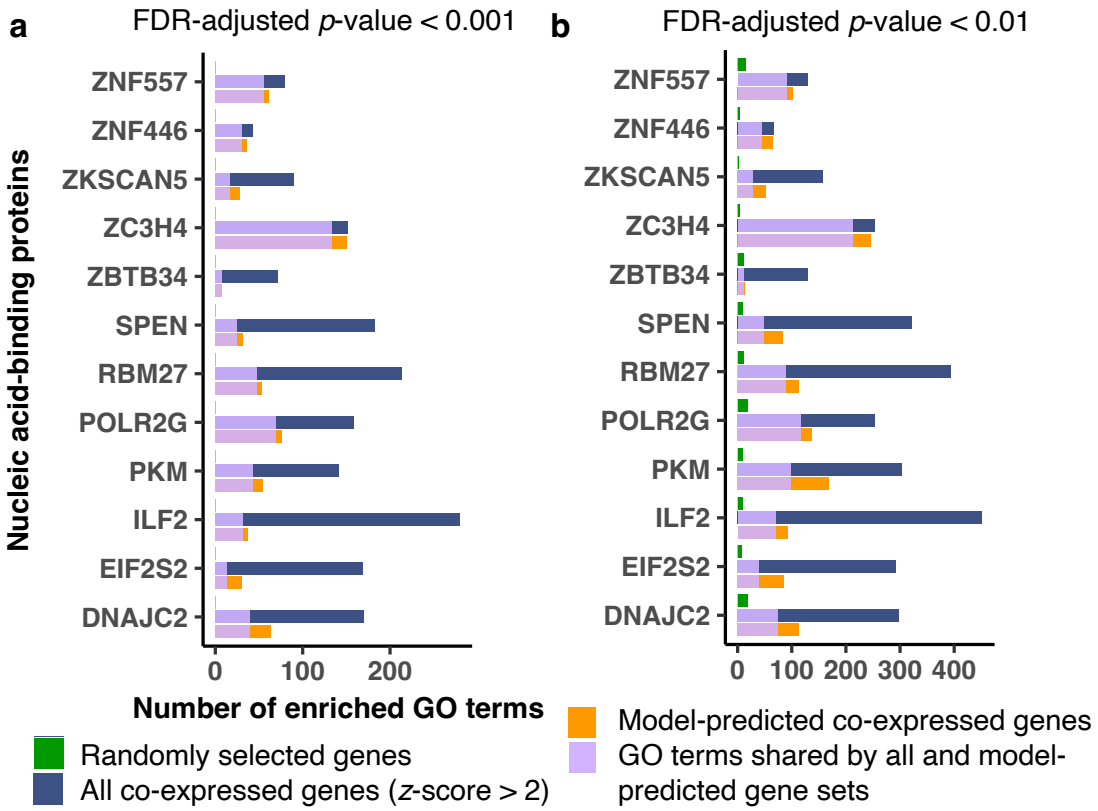


Figure 4. Overrepresented functional annotation terms among target genes of nucleic acid-binding proteins.

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).

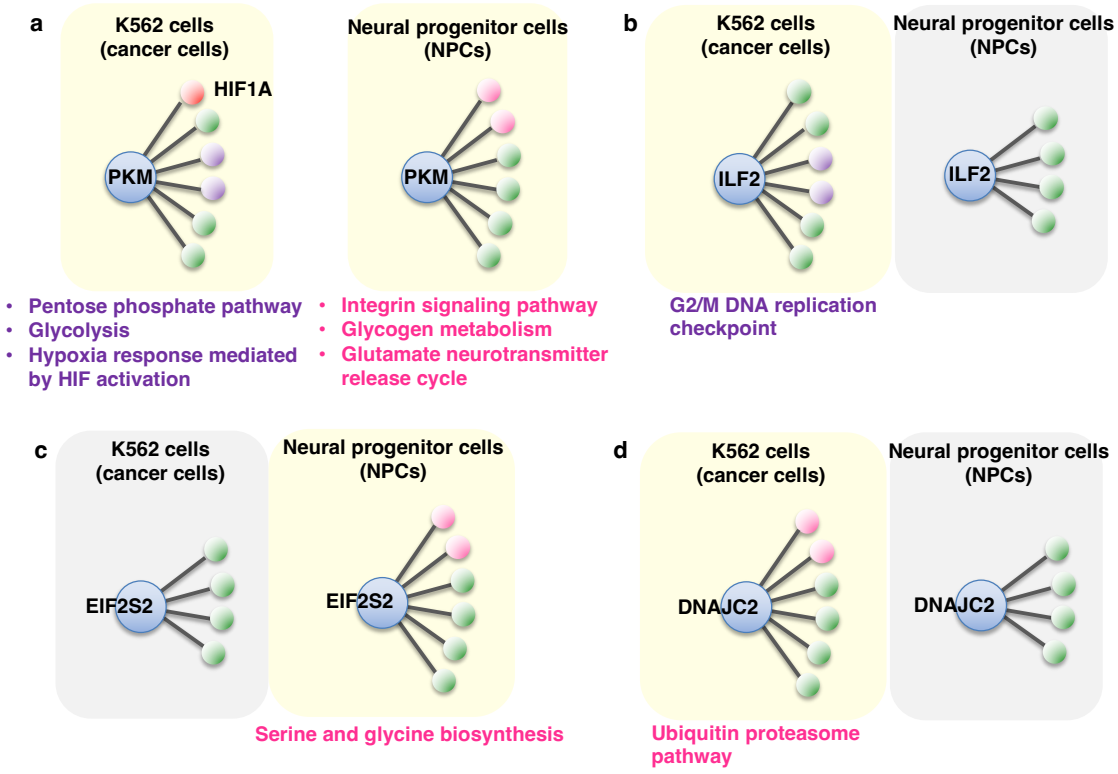
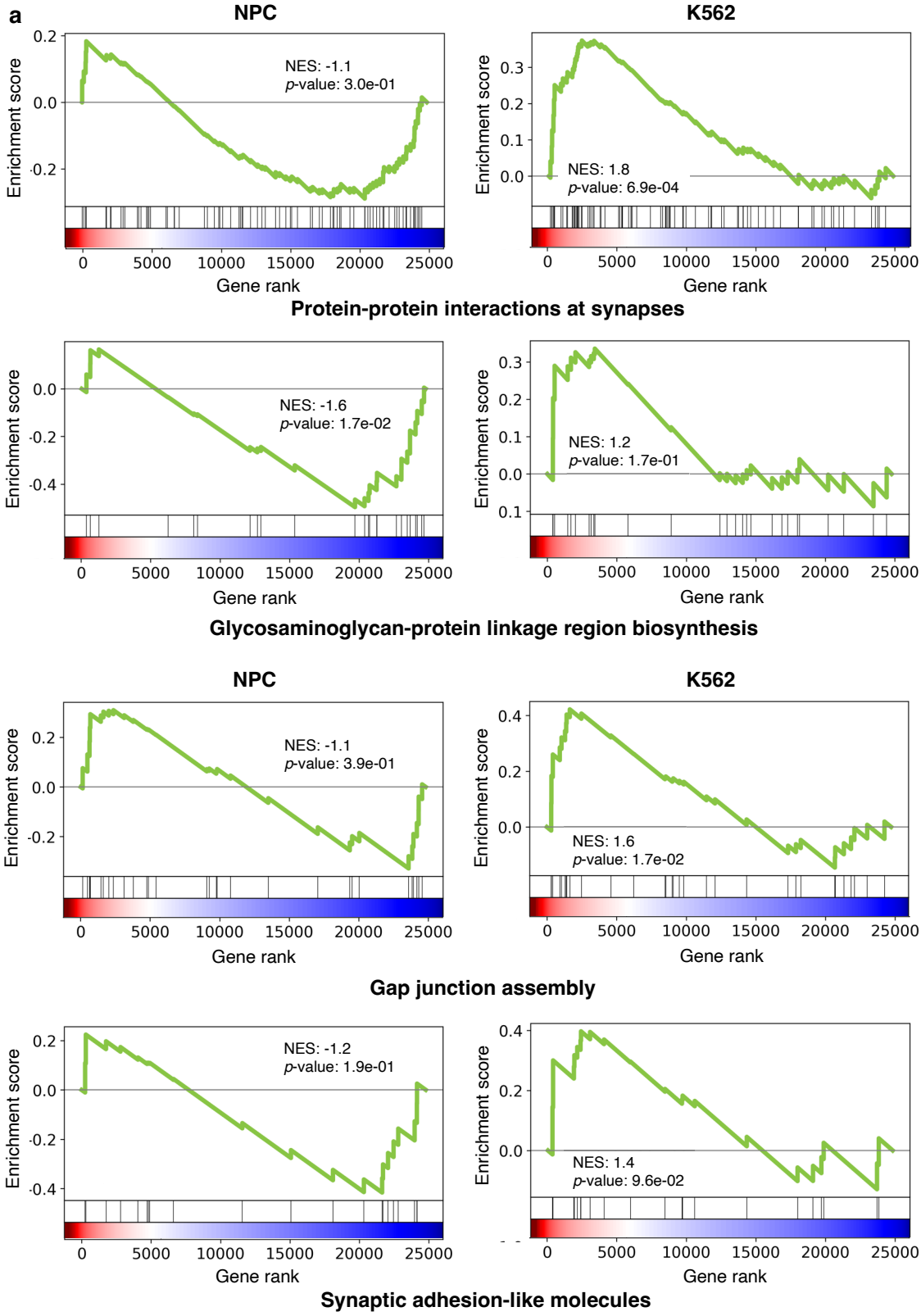
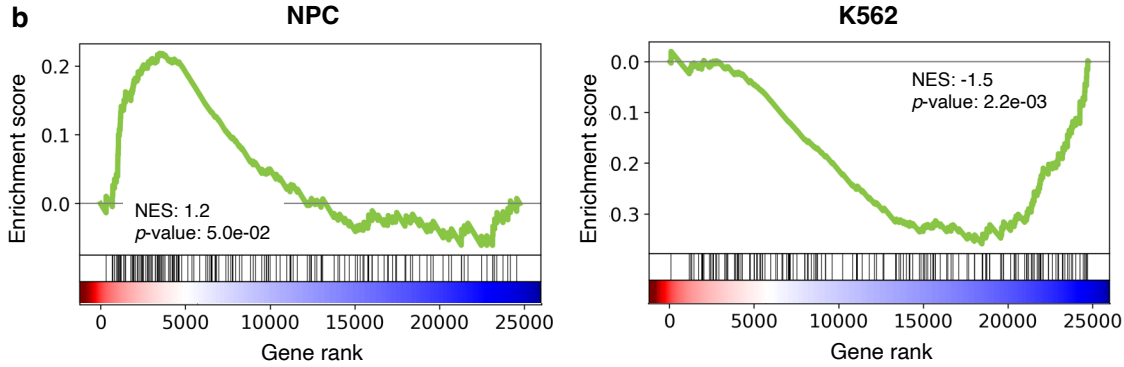


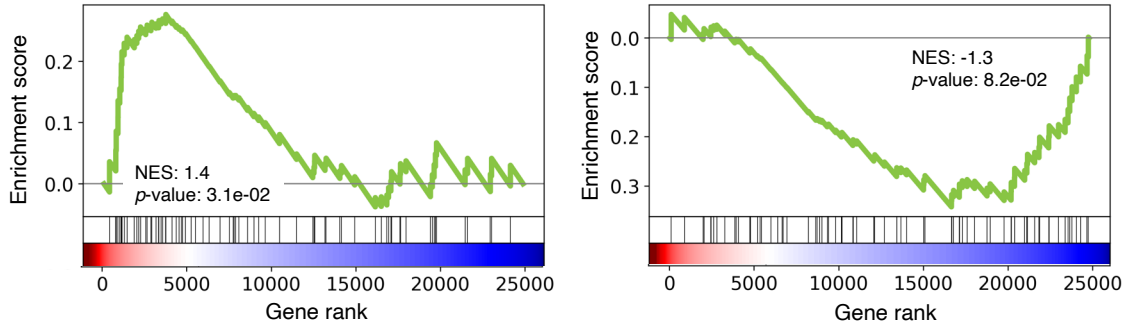
Figure 5. Functional mapping of nucleic acid-binding proteins reveals context-dependent regulatory roles in cancer and neural progenitor cells.

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.

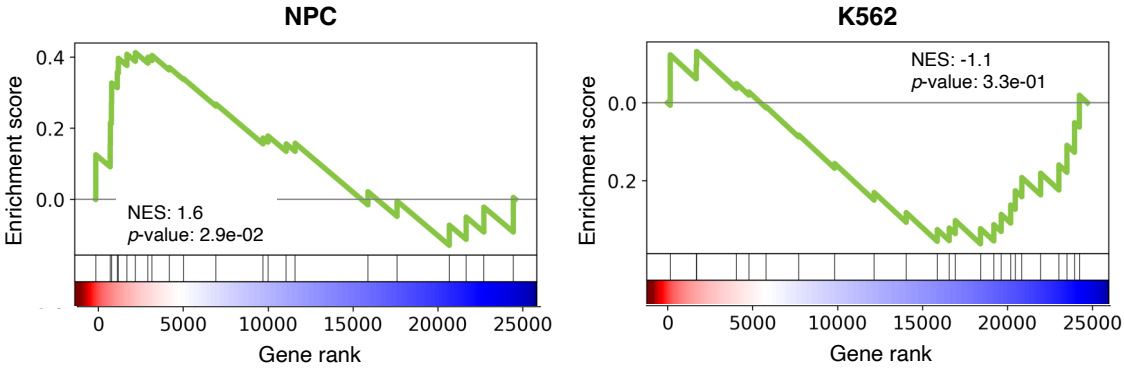

a
NPC
K562
Enrichment score
Gene rank
NES: -1.1
p-value: 3.0e-01
NES: 1.8
p-value: 6.9e-04
Protein-protein interactions at synapses
NES: -1.6
p-value: 1.7e-02
NES: 1.2
p-value: 1.7e-01
Glycosaminoglycan-protein linkage region biosynthesis
NES: -1.1
p-value: 3.9e-01
NES: 1.6
p-value: 1.7e-02
Gap junction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



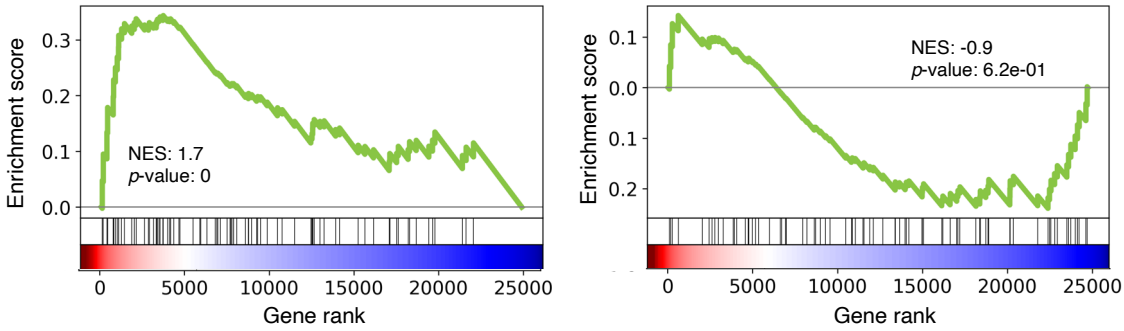
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell



Role of phospholipids in phagocytosis



WNT ligand biogenesis and trafficking



FCERI mediated MAPK activation

832 **Figure 6. Cell type–specific pathway associations inferred from contribution score–**
833 **based GSEA.**
834 GSEA (gene set enrichment analysis) was performed using DeepLIFT contribution scores
835 of nucleic acid-binding proteins (NABPs). This analysis revealed that genes with high or
836 low contribution scores for a given NABP are preferentially associated with biological
837 pathways relevant to each cell type.
838 The x-axis represents genes ranked by their DeepLIFT contribution scores for each NABP,
839 with higher-scoring genes placed toward the left. PKM is shown here as a representative
840 RNA-binding protein to illustrate contribution score–based pathway analysis. The y-axis
841 indicates the running enrichment score, which increases when pathway-associated genes
842 are concentrated among genes with higher contribution scores and decreases when they
843 are enriched among genes with lower contribution scores. The normalized enrichment
844 score (NES) summarizes the strength and direction of this enrichment across pathways.
845 For example, genes annotated to "Protein–protein interactions at synapses" were
846 preferentially enriched among genes with lower PKM contribution scores in NPCs but
847 among genes with higher contribution scores in K562 cells, indicating a marked cell type–
848 specific difference in the inferred association between PKM and this pathway.
849 Conversely, genes associated with "Immunoregulatory interactions between a Lymphoid
850 and a non-Lymphoid cell" showed contrasting enrichment patterns between K562 and
851 NPCs, highlighting differential pathway associations inferred from PKM contribution
852 scores across the two cell types.
853
854
855 **Tables**

856 **Table 1. Functional roles inferred for nucleic acid-binding proteins: integration of**
857 **UniProt annotations, ChatGPT-based predictions, and confidence scores.**
858 Functional annotations obtained from UniProt entries were systematically compared with
859 significantly enriched biological processes among the model-predicted regulatory targets,
860 as identified by the PANTHER Overrepresentation Test (see **Supplementary Table S5**).
861 Shared or overlapping functions between each nucleic acid-binding protein (NABP) and
862 its predicted targets are indicated. Functional roles inferred using a ChatGPT-based
863 interpretation method are highlighted in yellow, with confidence scores—defined as the
864 proportion of annotated genes within the gene set—provided in parentheses. Confidence
865 levels are categorized as high (0.87–1.00), medium (0.82–0.86), low (0.01–0.81), or none
866 (0). For methodological reference, see Hu M. et al., *Nat. Methods*, 2025
867 (https://doi.org/10.1038/s41592-024-02525-x) [37].
868

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

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

877 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

878

879

880

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

882 **acid-binding proteins.**

883 Functional enrichment analysis was performed using the PANTHER database. This table
884 provides selected examples of annotations that were statistically enriched among
885 regulatory target genes predicted by the deep learning model trained on gene co-
886 expression data. '**Reference**' indicates the total number of human genes; '**Analyzed**'
887 denotes the number of regulatory target genes examined; '**Expected**' represents the
888 number of genes expected to have the annotation by chance; '**Fold Enrichment**' reflects
889 the observed-to-expected ratio; '**Raw *p*-value**' reports the result of Fisher's exact test; and
890 '**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

891

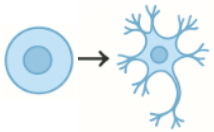
892

893

894 **Table 4. Cell type−dependent enrichment of functional pathways identified by**
895 **GSEA**

896 NES denotes the normalized enrichment score obtained from gene set enrichment
897 analysis (GSEA). ΔNES represents the difference in NES between K562 and NPC
898 ($\Delta NES = NES_{K562} - NES_{NPC}$). This table summarizes pathway enrichment results for
899 PKM, shown as a representative RNA-binding protein consistent with **Fig. 6**. The
900 twenty Reactome pathways with the largest positive and negative ΔNES values are
901 listed in the upper and lower sections, respectively. Pathways with large positive ΔNES

902 values were predominantly annotated to high-level Reactome categories related to
903 neural development and differentiation, whereas pathways with large negative ΔNES
904 values were enriched for categories associated with immune regulation, cell cycle
905 control, and genome maintenance, reflecting contrasting functional associations
906 between NPCs and K562 cells. Corresponding analyses performed using an amplified
907 DeepLIFT background (input × 2) yielded qualitatively consistent pathway-level
908 contrasts, with generally higher enrichment scores due to an increased number of
909 predicted PKM-associated target genes (**Supplementary Table S8**).
910
911



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	Reactome top pathway description	Reactome top pathway ID
Protein-protein interactions at synapses	R-HSA-6794362	1.79	-1.11	2.89	Neuronal System	R-HSA-112316
Glycosaminoglycan-protein linkage region biosynthesis	R-HSA-1971475	1.23	-1.56	2.79	Metabolism	R-HSA-1430728
Golgi-to-ER retrograde transport	R-HSA-8856688	1.62	-1.15	2.77	Vesicle-mediated transport	R-HSA-5653656
Attachment and Entry	R-HSA-9694614	1.39	-1.37	2.76	Disease	R-HSA-1643685
Defective EXT2 causes exostoses 2	R-HSA-3656237	1.39	-1.36	2.75	Disease	R-HSA-1643685
Defective B4GALT7 causes EDS, progeroid type	R-HSA-3560783	1.13	-1.60	2.73	Disease	R-HSA-1643685
Phase 0 - rapid depolarisation	R-HSA-5576892	1.45	-1.24	2.69	Muscle contraction	R-HSA-397014
Gap junction assembly	R-HSA-190861	1.62	-1.06	2.68	Vesicle-mediated transport	R-HSA-5653656
Diseases associated with glycosaminoglycan metabolism	R-HSA-3560782	1.07	-1.61	2.68	Disease	R-HSA-1643685
Extracellular matrix organization	R-HSA-1474244	1.46	-1.21	2.67	Extracellular matrix organization	R-HSA-1474244
Synaptic adhesion-like molecules	R-HSA-8849932	1.37	-1.24	2.61	Neuronal System	R-HSA-112316
Neuronal System	R-HSA-112316	1.63	-0.96	2.59	Neuronal System	R-HSA-112316
Degradation of the extracellular matrix	R-HSA-1474228	1.50	-1.08	2.59	Extracellular matrix organization	R-HSA-1474244
IRE1alpha activates chaperones	R-HSA-381070	1.25	-1.31	2.56	Cellular responses to stimuli	R-HSA-8953897
Smooth Muscle Contraction	R-HSA-445355	1.51	-1.04	2.54	Muscle contraction	R-HSA-397014
FGFR2c ligand binding and activation	R-HSA-190375	1.29	-1.23	2.52	Signal Transduction	R-HSA-162582
Receptor-type tyrosine-protein phosphatases	R-HSA-388844	1.62	-0.88	2.50	Neuronal System	R-HSA-112316
Muscle contraction	R-HSA-397014	1.57	-0.92	2.49	Muscle contraction	R-HSA-397014
LGI-ADAM interactions	R-HSA-5682910	1.53	-0.97	2.49	Developmental Biology	R-HSA-1266738
Respiratory syncytial virus (RSV) attachment and entry	R-HSA-9820960	1.39	-1.1	2.49	Disease	R-HSA-1643685
Activation of anterior HOX genes in hindbrain development during early embryogenesis	R-HSA-5617472	-1.11	1.50	-2.62	Developmental Biology	R-HSA-1266738
Role of LAT2/NTAL/LAB on calcium mobilization	R-HSA-2730905	-1.12	1.50	-2.62	Immune System	R-HSA-168256
FCERI mediated MAPK activatio	R-HSA-2871796	-0.92	1.72	-2.64	Immune System	R-HSA-168256
WNT ligand biogenesis and trafficking	R-HSA-3238698	-1.10	1.56	-2.66	Signal Transduction	R-HSA-162582
Role of phospholipids in phagocytosis	R-HSA-2029485	-1.31	1.38	-2.69	Immune System	R-HSA-168256
Nucleotide Excision Repair	R-HSA-5696398	-1.30	1.41	-2.71	DNA Repair	R-HSA-73894
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	R-HSA-198933	-1.52	1.24	-2.75	Immune System	R-HSA-168256
Transcription-Coupled Nucleotide Excision Repair (TC-NER)	R-HSA-6781827	-1.47	1.33	-2.79	DNA Repair	R-HSA-73894
Creation of C4 and C2 activators	R-HSA-166786	-1.13	1.70	-2.83	Immune System	R-HSA-168256
CD22 mediated BCR regulation	R-HSA-5690714	-1.11	1.72	-2.83	Immune System	R-HSA-168256
Initial triggering of complement	R-HSA-166663	-1.23	1.64	-2.87	Immune System	R-HSA-168256
FCGR activation	R-HSA-2029481	-1.17	1.79	-2.96	Immune System	R-HSA-168256
Scavenging of heme from plasma	R-HSA-2168880	-1.26	1.74	-3.00	Vesicle-mediated transport	R-HSA-5653656
Mitotic G1 phase and G1/S transition	R-HSA-453279	-1.27	1.73	-3.01	Cell Cycle	R-HSA-1640170
Response to metal ions	R-HSA-5660526	-1.29	1.77	-3.06	Cellular responses to stimuli	R-HSA-8953897
Mitotic Metaphase and Anaphase	R-HSA-2555396	-1.27	1.79	-3.06	Cell Cycle	R-HSA-1640170
M Phase	R-HSA-68886	-1.27	1.84	-3.11	Cell Cycle	R-HSA-1640170
DNA Replication	R-HSA-69306	-1.30	1.94	-3.23	DNA Replication	R-HSA-69306
Synthesis of DNA	R-HSA-69239	-1.29	1.98	-3.28	DNA Replication	R-HSA-69306
Metallothioneins bind metals	R-HSA-5661231	-1.04	2.61	-3.65	Cellular responses to stimuli	R-HSA-8953897

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.

938 # Supplementary Note and Tables

939

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

942 **Contribution score ranking and binning**

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

948

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

950 High- and low-contribution regions were defined using both rank-based and score-
951 based thresholds to assess the robustness of enrichment results. Rank-based thresholds
952 included upper and lower rank cutoffs (e.g., top or bottom 20 and 50 ranks), while score-
953 based thresholds were defined using absolute contribution score cutoffs (e.g., ≥ 0.01 or ≥
954 0.001). All threshold combinations used in the analyses are summarized in **Tables S2** and
955 **S3.**

956

957 **Genomic regions used for overlap analyses**

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

962 spanning from the TSS to the transcription end site (TES), including both exonic and
963 intronic intervals.
964
965 **Random controls**
966 Random control regions were generated by sampling genomic intervals matched
967 for size and genomic context, excluding annotated binding sites, to estimate the expected
968 overlap by chance. All enrichment and overlap analyses were repeated using these
969 controls for comparison.
970
971 **Background configuration for DeepLIFT analyses**
972 Unless otherwise noted, DeepLIFT contribution scores were computed using a
973 background defined as 0.5× the original input feature values. Analyses using an
974 alternative background configuration (input × 2) yielded qualitatively similar results and
975 are presented in **Supplementary Figures S2** and **S3.**
976
977 **Cell type configurations used for model training and comparative analyses**
978 To accommodate differences in experimental data availability and analytical
979 objectives, multiple cell type configurations were used across analyses. For the primary
980 analyses of DNA-binding proteins (DBPs), models were trained using expression data
981 from HFFs, HMECs, NPCs, and HepG2, enabling validation against ChIP-seq datasets
982 available for HepG2 cells. For RNA-binding protein (RBP) analyses, models were trained
983 using HFFs, HMECs, NPCs, and K562 to facilitate validation using eCLIP datasets
984 available for K562 cells. For analyses requiring direct comparison of regulatory
985 contribution patterns between HepG2 and K562 for the same nucleic acid-binding

986 proteins (e.g., **Fig. 3d**), a unified four–cell-type configuration comprising HFFs, NPCs,
987 HepG2, and K562 was employed. This configuration enabled direct comparison of
988 DeepLIFT contribution score distributions across these two cell types while maintaining
989 a shared transcriptomic background. In all cases, cell type selection was driven by
990 analytical requirements and the availability of independent experimental validation data,
991 and no validation datasets were used as direct inputs during model training.
992

993 Gene co-expression–based predictions align with ChIP-seq–defined 994 DNA-binding targets

995 High-ranking predicted binding regions frequently coincide with ChIP-seq peaks,
996 particularly in promoter–enhancer intervals. Deep learning–derived DNA-binding site
997 predictions based on gene co-expression data show substantial concordance with ChIP-
998 seq results, especially when broader genomic regions encompassing promoters and
999 putative enhancers are considered. When overlap was assessed within ±100 kb of
1000 transcription start sites (TSSs)—a window supported by prior enhancer–gene mapping
1001 studies [19,67]—the mean and median overlap ratios reached 0.70 and 0.77, respectively
1002 (Ratio A2 / B2; **Supplementary Tables S2 and S3**). In contrast, overlap was lower for
1003 promoter regions alone (±5 kb), with corresponding mean and median values of 0.35 and
1004 0.34. Several DNA-binding proteins, including AKAP8, POGK, SPEN, THAP7, ZNF142,
1005 and ZNF691, showed notably high promoter-region overlap ratios (>0.5), suggesting
1006 potential direct regulatory activity. Nevertheless, not all ChIP-seq peaks necessarily
1007 reflect functionally relevant transcriptional binding events. For RNA-binding proteins,
1008 overlap was examined within the transcribed regions of RNA molecules, and the mean
1009 and median overlap ratios were 0.70 and 0.76, respectively (Ratio A2 / B2;

1010 **Supplementary Table S4**). These findings indicate that model-inferred high-scoring
1011 sites do not universally coincide with experimentally determined binding regions,
1012 underscoring the complexity and context dependence of regulatory interactions.
1013
1014

1015 Supplementary Table S1. Co-expression–based selection enhances the 1016 contribution scores of NABP binding sites.

1017 Binding sites identified through gene co-expression analysis exhibited higher
1018 absolute DeepLIFT contribution scores than those selected solely from ChIP-seq data
1019 without co-expression information. This trend was consistently observed for DNA-
1020 binding proteins and was significantly stronger for RNA-binding proteins across the
1021 analyzed cell types. For each group, the table presents the mean and median values of
1022 both positive and negative contribution scores. These results indicate that gene co-
1023 expression–based selection preferentially highlights functionally relevant regulatory
1024 interactions. The upper and lower sections of the table correspond to analyses performed
1025 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

1029 **Supplementary Table S2. Enrichment of high-contribution promoter**
1030 **regions with experimentally validated DBP binding sites.** (See
1031 Supplementary Excel file)

1032 This table summarizes the proportion of promoter regions (±5 kb from transcription
1033 start sites [TSSs]) that were predicted to be functionally important by DeepLIFT
1034 contribution scores and validated by ChIP-seq data for the corresponding DNA-binding
1035 proteins (DBPs). Overlap ratios were calculated across four thresholds—two based on
1036 rank and two based on absolute score—to assess the robustness of the predictions.

1037 High- and low-contribution regions were classified using the following thresholds
1038 for the 1,310 NABPs analyzed:

1039 **Rank-based thresholds**:

1040 (A1 and B1) proteins ranked in the top or bottom 20 (i.e., rank ≤ 20 or ≥ 1,290)

1041 (A2 and B2) proteins ranked in the top or bottom 50 (i.e., rank ≤ 50 or ≥ 1,260)

1042 **Score-based thresholds**:

1043 (C1 and D1) contribution score ≥ 0.01 or ≤ –0.01

1044 (C2 and D2) contribution score ≥ 0.001 or ≤ –0.001

1045

1046

1047 **Supplementary Table S3. Enrichment of high-contribution promoter**
1048 **and enhancer regions with ChIP-seq–validated DBP binding sites.**

1049 This table summarizes the proportion of predicted high-contribution regions—
1050 defined within ±100 kb of transcription start sites (TSSs), encompassing both promoter
1051 and distal enhancer regions—that overlap with ChIP-seq binding sites for the
1052 corresponding DNA-binding proteins (DBPs). Strong concordance between model-

1053 derived predictions and ChIP-seq peaks across multiple thresholds highlights the
1054 reliability of the contribution scoring method for identifying functionally relevant
1055 regulatory sites.

1056 High- and low-contribution regions were classified using the following thresholds
1057 for the 1,310 NABPs analyzed:

1058 **Rank-based thresholds**:

1059 (A1 and B1) proteins ranked in the top or bottom 20 (i.e., rank ≤ 20 or ≥ 1,290)

1060 (A2 and B2) proteins ranked in the top or bottom 50 (i.e., rank ≤ 50 or ≥ 1,260)

1061 **Score-based thresholds**:

1062 (C1 and D1) contribution score ≥ 0.01 or ≤ –0.01

1063 (C2 and D2) contribution score ≥ 0.001 or ≤ –0.001

1064

1065

1066 Supplementary Table S4. Enrichment of high-contribution transcribed 1067 regions with eCLIP–validated RBP binding sites.

1068 This table summarizes the proportion of predicted high-contribution regions—
1069 defined within the transcribed regions of genes (from transcription start site [TSS] to
1070 transcription end site [TES])—that overlap with eCLIP peaks for the corresponding
1071 RNA-binding proteins (RBPs). Strong concordance between model-based predictions
1072 and eCLIP peaks across multiple thresholds underscores the reliability of the contribution
1073 scoring method for identifying functionally relevant regulatory sites.

1074 High- and low-contribution regions were classified using the following
1075 thresholds applied to the 1,310 NABPs analyzed:

1076 **Rank-based thresholds**:

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

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

Score-based thresholds:

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

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

References

1 Belz, G. T. & Nutt, S. L. Transcriptional programming of the dendritic cell network. *Nature reviews. Immunology* **12**, 101-113 (2012). https://doi.org/10.1038/nri3149

2 Miller, J. C. *et al.* Deciphering the transcriptional network of the dendritic cell lineage. *Nature immunology* **13**, 888-899 (2012). https://doi.org/10.1038/ni.2370

3 Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744-752 (2010). https://doi.org/10.1016/j.cell.2010.01.044

4 Gertz, J. *et al.* Distinct properties of cell-type-specific and shared transcription factor binding sites. *Molecular cell* **52**, 25-36 (2013). https://doi.org/10.1016/j.molcel.2013.08.037

5 Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665 (2018). https://doi.org/10.1016/j.cell.2018.01.029

6 Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics* **10**, 252-263 (2009). https://doi.org/10.1038/nrg2538

7 Pierson, E. *et al.* Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLoS Comput Biol* **11**, e1004220 (2015). https://doi.org/10.1371/journal.pcbi.1004220

8 Sonawane, A. R. *et al.* Understanding Tissue-Specific Gene Regulation. *Cell Rep* **21**, 1077-1088 (2017). https://doi.org/10.1016/j.celrep.2017.10.001

9 Saha, A. *et al.* Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome research* **27**, 1843-1858 (2017). https://doi.org/10.1101/gr.216721.116

10 Boyle, A. P. *et al.* Comparative analysis of regulatory information and circuits across

distant species. *Nature* **512**, 453-456 (2014). https://doi.org/10.1038/nature13668

11 Ray, D. *et al.* RNA-binding proteins that lack canonical RNA-binding domains are rarely sequence-specific. *Sci Rep* **13**, 5238 (2023). https://doi.org/10.1038/s41598-023-32245-9

12 Zeke, A. *et al.* Deep structural insights into RNA-binding disordered protein regions. *Wiley Interdiscip Rev RNA* **13**, e1714 (2022). https://doi.org/10.1002/wrna.1714

13 Zaharias, S. *et al.* Intrinsically disordered electronegative clusters improve stability and binding specificity of RNA-binding proteins. *J Biol Chem* **297**, 100945 (2021). https://doi.org/10.1016/j.jbc.2021.100945

14 Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J. & Mann, R. S. Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annu Rev Cell Dev Biol* **35**, 357-379 (2019). https://doi.org/10.1146/annurev-cellbio-100617-062719

15 Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**, 1274-1286 (2012). https://doi.org/10.1016/j.cell.2012.04.040

16 Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature methods* **13**, 366-370 (2016). https://doi.org/10.1038/nmeth.3799

17 Fagny, M. *et al.* Exploring regulation in tissues with eQTL networks. *Proc Natl Acad Sci U S A* **114**, E7841-E7850 (2017). https://doi.org/10.1073/pnas.1707375114

18 Agarwal, V. *et al.* Massively parallel characterization of transcriptional regulatory elements. *Nature* (2025). https://doi.org/10.1038/s41586-024-08430-9

19 Avsec, Z. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods* **18**, 1196-1203 (2021). https://doi.org/10.1038/s41592-021-01252-x

20 Tasaki, S., Gaiteri, C., Mostafavi, S. & Wang, Y. Deep learning decodes the principles of differential gene expression. *Nat Mach Intell* **2**, 376-386 (2020). https://doi.org/10.1038/s42256-020-0201-6

21 Osato, N. & Hamada, M. Systematic discovery of directional regulatory motifs associated with human insulator sites. *bioRxiv* (2025). https://doi.org/https://doi.org/10.1101/2024.01.20.573595

22 Obayashi, T., Kodate, S., Hibara, H., Kagaya, Y. & Kinoshita, K. COXPRESdb v8: an animal gene coexpression database navigating from a global view to detailed investigations. *Nucleic acids research* **51**, D80-D87 (2023). https://doi.org/10.1093/nar/gkac983

23 Yang, S. *et al.* COEXPEDIA: exploring biomedical hypotheses via co-expressions associated with medical subject headings (MeSH). *Nucleic acids research* **45**, D389-D396 (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

1217 51 Huang, X. *et al.* D-Serine regulates proliferation and neuronal differentiation of neural
1218 stem cells from postnatal mouse forebrain. *CNS Neurosci Ther* **18**, 4-13 (2012).
1219 https://doi.org/10.1111/j.1755-5949.2011.00276.x
1220 52 El-Hattab, A. W. Serine biosynthesis and transport defects. *Mol Genet Metab* **118**, 153-
1221 159 (2016). https://doi.org/10.1016/j.ymgme.2016.04.010
1222 53 Kaida, A. & Iwakuma, T. Regulation of p53 and Cancer Signaling by Heat Shock Protein
1223 40/J-Domain Protein Family Members. *Int J Mol Sci* **22** (2021).
1224 https://doi.org/10.3390/ijms222413527
1225 54 Sterrenberg, J. N., Blatch, G. L. & Edkins, A. L. Human DNAJ in cancer and stem cells.
1226 *Cancer Lett* **312**, 129-142 (2011). https://doi.org/10.1016/j.canlet.2011.08.019
1227 55 Ladomery, M. Multifunctional proteins suggest connections between transcriptional and
1228 post-transcriptional processes. *Bioessays* **19**, 903-909 (1997).
1229 https://doi.org/10.1002/bies.950191010
1230 56 Kim, S., Yu, N. K. & Kaang, B. K. CTCF as a multifunctional protein in genome regulation
1231 and gene expression. *Exp Mol Med* **47**, e166 (2015).
1232 https://doi.org/10.1038/emm.2015.33
1233 57 Osato, N. Characteristics of functional enrichment and gene expression level of human
1234 putative transcriptional target genes. *BMC Genomics* **19**, 957 (2018).
1235 https://doi.org/10.1186/s12864-017-4339-5
1236 58 Burley, S. K. & Kamada, K. Transcription factor complexes. *Curr Opin Struct Biol* **12**, 225-
1237 230 (2002). https://doi.org/10.1016/s0959-440x(02)00314-7
1238 59 Goos, H. *et al.* Human transcription factor protein interaction networks. *Nat Commun* **13**,
1239 766 (2022). https://doi.org/10.1038/s41467-022-28341-5
1240 60 Berenson, A. *et al.* Paired yeast one-hybrid assays to detect DNA-binding cooperativity
1241 and antagonism across transcription factors. *Nat Commun* **14**, 6570 (2023).
1242 https://doi.org/10.1038/s41467-023-42445-6
1243 61 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic*
1244 *acids research* **41**, D991-995 (2013). https://doi.org/10.1093/nar/gks1193
1245 62 Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene
1246 expression and hybridization array data repository. *Nucleic acids research* **30**, 207-210
1247 (2002). https://doi.org/10.1093/nar/30.1.207
1248 63 Kolmykov, S. *et al.* GTRD: an integrated view of transcription regulation. *Nucleic acids*
1249 *research* **49**, D104-D111 (2021). https://doi.org/10.1093/nar/gkaa1057
1250 64 Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic acids*
1251 *research* **34**, D590-598 (2006). https://doi.org/10.1093/nar/gkj144
1252 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