Title

Pathway redistribution reveals a shared signaling backbone and context-dependent regulatory modules in RNA-binding protein 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

Understanding how regulatory architectures are reorganized across cellular contexts remains a central challenge in functional genomics. Here, we integrate co-expression–derived candidate regulatory interactions with interpretable deep learning to generate gene-level contribution scores and introduce ΔNES (normalized enrichment score difference) to quantify pathway redistribution between cellular states. Because gene expression reflects the combined effects of multiple regulatory inputs, contribution scores capture relative regulatory influence rather than transcriptional abundance itself.

Applying this framework to neural progenitor cells and K562 leukemia cells, we identify systematic redistribution of functional modules across multiple RNA-binding proteins, including PKM, HNRNPK, and NELFE. Neural System– and Immune System–associated modules are differentially positioned along the ΔNES spectrum, indicating context-dependent redistribution of regulatory influence rather than isolated pathway activation events.

At the pathway level, Signal Transduction consistently forms a shared signaling backbone across proteins and cellular contexts, while modules related to neuronal functions, immune responses, and developmental processes exhibit context-dependent redistribution. Subpathway analysis further reveals convergence on receptor-mediated signaling processes, including FGFR/RTK-, IRS-, and MAPK-related pathways.

These redistribution patterns are preserved under alternative DeepLIFT background settings despite polarity changes in contribution–expression correlations, indicating that pathway-level contrasts arise from stable rank-structure differences rather than background-dependent score artifacts.

Together, our findings demonstrate that contribution score–based pathway ranking reveals a conserved signaling backbone alongside context-dependent functional modules, providing a framework for interpreting regulatory architecture beyond expression-centric analyses.

Keywords

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

Introduction

Understanding how regulatory architectures are organized and reconfigured across cellular contexts remains a central challenge in molecular biology [1-10]. Nucleic acid-binding proteins (NABPs), including DNA-binding proteins and RNA-binding proteins (RBPs), play key roles in gene regulation; however, their functional characterization across diverse cell types remains incomplete. While experimental approaches such as ChIP-seq and eCLIP can identify binding sites, these methods are not easily scalable and often fail to capture indirect or context-dependent regulatory interactions [11], particularly for RBPs lacking well-defined binding motifs [12-14].

Computational approaches, including deep learning models, have improved the prediction of gene expression by integrating regulatory features derived from binding data. However, these methods typically rely on experimentally defined binding sites or sequence features and do not directly reveal how individual NABPs contribute to gene expression programs at the systems level. In particular, it remains unclear how regulatory

influence is organized across pathways and how it is reconfigured between cellular contexts.

Gene co-expression analysis provides a complementary perspective by capturing coordinated expression patterns across diverse biological conditions, reflecting both direct and indirect regulatory relationships [15-23]. Integrating such information into interpretable deep learning frameworks offers the potential to quantify regulatory influence at the gene level without requiring prior knowledge of binding motifs or direct interaction data.

Here, we integrate co-expression–derived regulatory interactions with interpretable deep learning to compute gene-level contribution scores and introduce ΔNES (normalized enrichment score difference) to quantify pathway redistribution between cellular contexts. Rather than focusing on individual regulatory interactions, this framework enables the analysis of regulatory architecture at the pathway level.

Applying this approach to neural progenitor cells and K562 leukemia cells, we find that regulatory influence is not organized as distinct pathway sets, but instead is structured around a shared signaling backbone, while functional modules are redistributed in a context-dependent manner. This conceptual framework provides a systems-level perspective on gene regulation, revealing how common signaling networks are differentially utilized across cellular states.

Results

Gene co-expression features enhance expression prediction accuracy

84 Replacing low-contribution DBP inputs with gene co-expression data led to a
85 substantial increase in model performance (**Fig. 1**). In human foreskin fibroblasts
86 (HFFs), substituting 302 out of 1,310 DBPs with co-expression profiles improved the
87 correlation between predicted and actual gene expression from 0.70 to 0.80 (**Fig. 2a, b**).
88 These co-expression inputs included both NABP-encoding and non-NABP mRNAs, with
89 gene symbols for DNA- and RNA-binding proteins obtained from the ENPD (Eukaryotic
90 Nucleic acid Binding Protein) database[24], and co-expression profiles sourced from
91 COXPRESdb [15]. A subsequent experiment replaced 325 low-contribution DBPs with
92 co-expression interactions involving RBPs, resulting in a further increase in prediction
93 accuracy ($r = 0.81$; **Fig. 2c**). Collectively, these findings demonstrate that integrating co-
94 expression interactions into the model enhances its ability to capture gene regulatory
95 dynamics beyond direct binding site information.

96 Contribution scores derived from co-expression–based NABP–gene interactions
97 exhibited broader distributions and higher overall magnitudes than those obtained from
98 ChIP-seq–based inputs across all four cell types examined (Wilcoxon signed-rank test, p
99 $< 10^{-2}$; **Fig. 2d** and **Supplementary Fig. S2c**; **Supplementary Table S1**). Each pair of
100 violin plots represents absolute contribution scores for co-expression–based sites (right
101 plot in each pair) and the corresponding ChIP-seq sites not included in that set (left plot).
102 **A**–**P** correspond to specific cell types and NABP classes as follows: **B**, DBP co-
103 expression–based sites in neural progenitor cells (NPCs); **A**, NPC ChIP-seq sites
104 excluding **B**. **D**, RBP co-expression–based sites in NPCs; **C**, NPC ChIP-seq sites
105 excluding **D**. **F**, DBP co-expression–based sites in HepG2 cells; **E**, HepG2 ChIP-seq sites
106 excluding **F**. **H**, RBP co-expression–based sites in K562 cells; **G**, K562 ChIP-seq sites
107 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[25]. Importantly, these trends were robust to DeepLIFT background definition: comparable differences between ChIP-seq– and co-expression–derived targets

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

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

156 background configuration enhances sensitivity to cell type–specific regulatory variation
157 for RBPs. In contrast, DNA-binding proteins (DBPs) generally showed stronger
158 correlations with gene expression under both background conditions, consistent with their
159 more direct transcriptional regulatory roles. Together, these results underscore the
160 importance of background-aware and protein class–specific interpretation of DeepLIFT
161 contribution scores.
162

163 Experimental binding data support model-inferred regulatory targets

164 To determine whether DeepLIFT-derived regulatory contribution scores reflect
165 biologically meaningful nucleic acid–binding protein (NABP) activity, we systematically
166 compared model-inferred regulatory targets with experimentally validated ChIP-seq and
167 eCLIP binding data across multiple genomic regions, threshold definitions, and cellular
168 contexts (**Fig. 3**). Experimentally validated binding sites are consistent with deep
169 learning–derived regulatory targets. Genes with ChIP-seq or eCLIP peaks for a given
170 NABP exhibited disproportionately high or low importance ranks, indicating that
171 experimentally confirmed binding sites are associated with stronger regulatory influence
172 in gene expression prediction models. This relationship supports the biological validity
173 of DeepLIFT-based contribution scores, which quantify the importance of specific
174 NABPs in driving gene expression changes.
175 To evaluate this association, we analyzed ChIP-seq data for 32 DBPs in HepG2
176 cells and eCLIP data for 29 RBPs in K562 cells—two cell lines with the most
177 comprehensive publicly available NABP binding site datasets from ENCODE and related
178 studies. A gene was considered associated with an NABP if a ChIP-seq peak was located
179 within ±5 kb of its transcription start site (TSS) or if an eCLIP peak was mapped to its

180 transcribed region. As illustrated in **Fig. 3a** and **Supplementary Fig. S1**, NABPs with
181 experimentally validated binding sites consistently showed higher absolute contribution
182 score values across their predicted target genes, reinforcing their regulatory relevance and
183 corresponding with prior interpretations of DeepLIFT scores[25,26].

184 Model-derived importance scores reliably predict DBP binding sites, as evidenced
185 by significant overlap with ChIP-seq peaks (**Fig. 3b, Supplementary Table S2** and **S3,**
186 **Supplementary Fig. S3a**). To evaluate this correspondence, we analyzed ChIP-seq data
187 for 32 DNA-binding proteins (DBPs) and quantified how frequently high-scoring gene
188 regions—defined by DeepLIFT contribution scores—coincided with experimental
189 binding evidence. Overlap ratios were calculated as the proportion of high-contribution
190 regions that intersected with ChIP-seq peaks, considering both promoter-proximal (±5
191 kb) and distal (±100 kb) regions around transcription start sites (TSSs). To account for
192 the functional heterogeneity and abundance of ChIP-seq peaks, we applied a precision-
193 at-K approach, focusing on the top-ranked or top-scoring gene regions most likely to
194 reflect regulatory relevance. Four thresholds were used to define high-contribution
195 regions: two based on contribution score rank (top/bottom 20 or 50) and two based on
196 absolute score magnitude (≥ 0.01 or ≤ -0.01; ≥ 0.001 or ≤ -0.001). A control analysis
197 using randomly selected gene regions of matched size was conducted in parallel. In all
198 tested conditions, the overlap between ChIP-seq peaks and high-contribution regions
199 exceeded that of the random control, supporting the model's ability to prioritize
200 functionally relevant regulatory sites.

201 Similarly, as illustrated in **Fig. 3c**, **Supplementary Table S4**, and **Supplementary**
202 **Fig. S3b**, the model-derived importance scores consistently prioritize biologically
203 relevant RBP–target interactions, as evidenced by the cumulative overlap between high-

204 contribution gene regions and eCLIP peaks, which was substantially greater than that
205 observed in the random control. We assessed this by analyzing eCLIP data for 29 RNA-
206 binding proteins (RBPs) in K562 cells, comparing high-contribution regions—defined
207 using the same thresholds as for DBPs—to eCLIP peaks mapped to transcribed regions
208 (from TSS to TES). Collectively, these results demonstrate that the model effectively
209 captures functionally meaningful NABP–target associations for both DNA- and RNA-
210 binding proteins.

211 Significant differences in the predicted regulatory targets of several NABPs were
212 observed between K562 and HepG2 cells based on DeepLIFT contribution scores (**Fig.
213 3d**). To assess cell type–specific regulatory activity, we compared the DeepLIFT scores
214 of genes harboring NABP binding sites, as defined by ChIP-seq or eCLIP. Because our
215 deep learning model identifies target candidates from NABP-associated co-expressed
216 gene sets, not all predicted targets necessarily reflect direct binding; rather, they represent
217 genes that participate in shared or downstream regulatory programs. RPLP0 exhibited
218 nearly identical score distributions across the two cell types, whereas ETF1, NELFE, and
219 CAMTA2 showed marked divergence. These differences indicate that the model captures
220 and adjusts NABP-dependent regulatory relationships in a cell type–specific manner.

221

222 Deep learning uncovers functionally enriched NABP target genes 223 beyond co-expression networks

224 Deep learning predictions revealed that nucleic acid–binding proteins (NABPs)
225 possess regulatory target genes with coherent and biologically specific functional
226 enrichment, extending well beyond patterns detectable from co-expression alone (**Tables
227 1, 2** and **Supplementary Note 2**). Across twelve representative NABPs[27,28], predicted

228 target sets consistently showed significant enrichment for biologically related Gene
229 Ontology (GO) terms, whereas randomly sampled gene sets of matched size displayed no
230 such enrichment (**Fig. 4**). Notably, several enriched annotations were absent from the
231 broader pools of co-expressed genes (Pearson's *z*-score > 2), indicating that the model
232 selectively extracts biologically relevant signals from large co-expression neighborhoods.
233 Functional enrichment across GO Slim, Reactome, and PANTHER Pathways further
234 confirmed this specificity (**Table 1**)[29]. A ChatGPT-based gene set interpretation method
235 independently supported these results, providing functional summaries and literature
236 consistent with GO enrichment and UniProt annotations [30] (**Table 2** and **Supplementary**
237 **Note 2**)[31].
238

239 Cell type–dependent functional specificity revealed by deep learning–
240 refined NABP targets

241 Deep learning refinement further exposed cell type–specific biological functions
242 associated with regulatory targets of several RBPs. For example, the predicted targets of
243 PKM showed distinct pathway enrichments between K562 cancer cells and neural
244 progenitor cells (NPCs). Pathways such as *Pentose phosphate pathway*, *Glycolysis*, and
245 *Hypoxia response mediated by HIF activation* were significantly enriched in K562 but
246 not in NPCs, whereas *Integrin signaling pathway*, *Glycogen metabolism*, and *Glutamate*
247 *neurotransmitter release cycle* were specifically enriched in NPCs using a background
248 defined as 0.5× the original input feature values[32-41] (**Fig. 5**, **Table 3**, **Supplementary**
249 **Table S5**). ChatGPT-based interpretation produced consistent functional assignments
250 (**Supplementary Note 2**). Notably, these cell type–dependent associations align with
251 known biology in cancer cells and NPCs.

Additional RBPs exhibited similar specificity. Predicted target genes of ILF2 were enriched for *G2/M DNA replication checkpoint* in K562 cells but not in NPCs[42-44] ; targets of EIF2S2 were associated with *Serine and glycine biosynthesis* in NPCs but not in K562[45,46] ; and targets of DNAJC2 were enriched for the *Ubiquitin proteasome pathway* in K562 but not in NPCs[47,48] **(Fig. 5**, **Table 3**, **Supplementary Table S5**). ChatGPT-based inference again recapitulated these cell type–dependent functional patterns (**Supplementary Note 2**), consistent with previous reports.

Together, these results demonstrate that the deep learning–based refinement of NABP target predictions enables systematic detection of cell type–dependent functions, revealing regulatory specificity not captured by co-expression alone.

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

Gene set enrichment analysis (GSEA) based on DeepLIFT-derived contribution scores revealed robust, cell type–dependent functional redistribution among NABP regulatory targets. Rather than ranking genes by expression levels, we ranked predicted targets according to their relative regulatory contributions within each cell type, enabling pathway-level interpretation based on shifts in the positional distribution of functionally related genes along contribution-ranked lists, independent of absolute score magnitude or sign (**Fig 6a**). Using this framework, PKM-associated pathways exhibited pronounced redistribution between neural progenitor cells (NPCs) and K562 leukemia cells. Multiple Reactome pathways showed large differences in normalized enrichment scores (ΔNES = NES_K562 − NES_NPC), indicating substantial divergence in inferred regulatory coupling across cellular contexts. Pathways related to synaptic organization and neuronal

system functions were enriched among positive ΔNES values (K562-biased), whereas immune-related pathways were preferentially represented among negative ΔNES values (NPC-biased). These contrasts were consistently captured across pathways ranked by ΔNES (**Fig. 6b**, **Table 4**, and **Supplementary Fig. S5**). Importantly, ΔNES reflects redistribution of contribution-ranked pathway positioning rather than direct activation or repression of lineage programs. Conceptually, ΔNES quantifies the relative positional shift of pathway-associated genes along contribution-ranked regulatory landscapes between cellular contexts.

Because DeepLIFT contribution scores depend on background definition, we evaluated alternative reference configurations (input × 0.5 and input × 2). Although background scaling altered the polarity of contribution–expression correlations and the absolute direction of individual NES values, relative ΔNES-based pathway differences between NPC and K562 were preserved across configurations (**Fig. 6**; **Supplementary Tables S7–S9**). The amplified background (input × 2) increased the number of genes with non-negligible contribution scores and generally yielded higher NES magnitudes, reflecting increased sensitivity without altering the relative redistribution patterns. These observations support the robustness of ΔNES as a comparative metric driven by stable rank-structure differences rather than background-dependent score polarity. Extending this analysis beyond PKM, similar category-level shifts were observed for additional RNA-binding proteins, including HNRNPK and NELFE (**Supplementary Tables S10–S13**). Across all three proteins, Neural System–annotated pathways were significantly overrepresented among high ΔNES terms, whereas Immune System–annotated pathways were enriched among low ΔNES terms (one-sided Fisher's exact test, $p < 0.001$; **Supplementary Table S14**). This reproducible enrichment pattern indicates systematic

300 redistribution of regulatory influence between NPC and K562 contexts that is not
301 restricted to a single protein, and remains consistent when evaluated at both pathway and
302 functional module levels.

303 Notably, Signal Transduction emerged as the dominant top-level Reactome
304 category across the ΔNES spectrum for all three RBPs (**Supplementary Table S14**).
305 Subpathway analysis revealed that this enrichment was driven primarily by broadly
306 utilized regulatory modules, including FGFR/RTK signaling, phospholipase C–mediated
307 cascades, WNT/BMP pathways, and receptor-associated second-messenger systems. At
308 the level of exact Reactome pathway identity, overlap across RBPs was limited, with only
309 a single pathway (FGFR1C ligand binding and activation) shared among PKM,
310 HNRNPK, and NELFE, and a small number of pairwise overlaps such as NF-κB
311 activation and signaling between PKM and NELFE (**Supplementary Table S16**).
312 Consistent with this observation, pairwise comparison using the Jaccard similarity index
313 revealed strong similarity between HNRNPK and NELFE ($J \approx 0.65$), whereas PKM
314 shared only limited overlap with either protein (PKM–HNRNPK: $J \approx 0.13$; PKM–
315 NELFE: $J \approx 0.15$), indicating divergence at the level of individual pathway identity.
316 However, grouping pathways into functional modules revealed a markedly different
317 structure. Despite limited exact overlap, all three RBPs converged on common receptor-
318 proximal signaling modules, particularly within the FGFR signaling axis and associated
319 phospholipase C cascades, as well as ligand–receptor and MAPK-related signaling
320 processes. These findings indicate that RBPs share a conserved signaling backbone at the
321 functional level, even when individual pathway identities differ. Furthermore, PKM
322 exhibited a broader and more diversified signaling profile, incorporating additional
323 modules related to Rho GTPase signaling, Notch signaling, and neuronal-associated

324 pathways, whereas bottom-ranked pathways were largely restricted to general signaling
325 processes such as GPCR, WNT/BMP, and ligand-binding receptor pathways, consistent
326 with leukemia/hematopoietic signaling states in K562 cells. Together, these results
327 support a hierarchical model in which a shared signaling backbone is differentially
328 extended by cell type–biased functional modules, rather than being defined by exact
329 pathway overlap.

330 To enable systematic interpretation, ΔNES-ranked pathways were grouped into
331 functional modules based on shared biological processes (see **Methods** and
332 **Supplementary Table S15**). This module-level organization revealed a clear separation
333 between pathways with positive ΔNES values (relatively enriched in K562) and those
334 with negative ΔNES values (relatively enriched in NPC) (**Fig. 6c**). Pathways with positive
335 ΔNES were predominantly associated with signaling and interaction-related modules,
336 including neuronal system–related signaling, metabolic processes, and stemness-
337 associated programs. In contrast, pathways with negative ΔNES were enriched for
338 immune system–related modules and developmental programs, such as organogenesis
339 and lineage specification. Despite this divergence, a subset of signaling pathways—
340 including receptor tyrosine kinase (RTK), MAPK, and WNT-related pathways—was
341 consistently observed across RBPs, forming a shared signaling backbone. These findings
342 indicate that RBPs do not primarily regulate distinct sets of cell-type-specific genes, but
343 instead modulate common signaling networks while redistributing functional outputs
344 across cellular contexts.

345 Consistent with established roles of PKM2 in metabolic rewiring, nuclear
346 transcriptional regulation, and proliferative state maintenance [49,50], the observed ΔNES
347 redistribution is compatible with context-dependent repositioning of metabolic and

348 neuronal regulatory modules between leukemia and neural progenitor states [36]. HNRNPK
349 and NELFE, which coordinate transcriptional and post-transcriptional regulatory
350 processes during development and malignancy, displayed parallel redistribution patterns,
351 further supporting a shared regulatory architecture modulated across cellular contexts
352 (**Supplementary Note**: Functional interpretation of ΔNES-based pathway redistribution
353 across PKM, HNRNPK, and NELFE). Together, these findings demonstrate that
354 contribution score–based GSEA captures reproducible, category-level functional
355 redistribution across cell types while revealing a conserved upstream signaling backbone
356 shared among multiple RNA-binding proteins. This framework provides an interpretable
357 and scalable strategy for dissecting context-dependent regulatory organization beyond
358 expression-centric or threshold-dependent analyses.

359 To further investigate the gene-level composition of enriched pathways, we
360 examined leading-edge genes identified by GSEA for ΔNES-ranked Reactome pathways.
361 Notably, many leading-edge genes were classified as shared functional genes rather than
362 strictly cell-type-specific genes, even within pathways associated with neuronal or
363 immune systems (**Supplementary Table S17**). These results suggest that RBPs
364 predominantly modulate broadly shared gene networks while contributing to context-
365 dependent functional differences at the pathway level.

366

367

368 ## Discussion

369 Our findings demonstrate that RNA-binding proteins (RBPs) regulate gene
370 expression programs through redistribution of regulatory influence across a shared
371 signaling backbone, rather than by controlling entirely distinct pathway sets. Across

372 multiple RBPs, including PKM, HNRNPK, and NELFE, Signal Transduction
373 consistently forms a conserved signaling backbone, while functional modules associated
374 with neuronal and immune processes are differentially positioned depending on cellular
375 context. This hierarchical organization indicates that regulatory rewiring occurs through
376 redistribution across common signaling networks.

377 Importantly, ΔNES captures this redistribution as shifts in the positional
378 distribution of pathway-associated genes along contribution-ranked lists, rather than
379 reflecting direct pathway activation or repression. Because gene expression represents the
380 combined effects of multiple regulatory inputs, contribution scores quantify relative
381 regulatory influence, enabling detection of subtle but systematic differences in regulatory
382 architecture across cell types.

383 At the pathway level, the shared signaling backbone is characterized by broadly
384 utilized receptor-mediated and growth factor–associated pathways, including RTK,
385 MAPK, and WNT-related signaling. Although individual pathway identities vary across
386 RBPs, functional module–level analysis reveals convergence on common upstream
387 signaling processes. This structural convergence distinguishes our findings from
388 conventional enrichment-based analyses, which typically emphasize pathway identity
389 rather than organization.

390 In contrast, modules associated with neuronal functions and immune responses
391 exhibit context-dependent redistribution, indicating that RBPs modulate how shared gene
392 networks are functionally deployed across cellular states. This model contrasts with
393 transcription factors, which often determine cellular identity through selective activation
394 of more restricted gene sets. Instead, RBPs appear to regulate functional allocation within
395 broadly shared signaling and metabolic networks.

The robustness of these patterns across alternative DeepLIFT background settings further supports the interpretation that pathway-level contrasts arise from stable rank-structure differences rather than background-dependent artifacts. While co-expression–based inference does not directly establish causal regulatory relationships, it captures functional coupling at the systems level, providing a scalable framework for analyzing regulatory architecture in the absence of experimentally defined binding information.

Looking forward, expanding this framework to a broader range of cell types and regulatory factors will enable systematic characterization of pathway redistribution across diverse biological contexts. Comparative analysis across transcription factors and RNA-binding proteins may further clarify their distinct and complementary roles in shaping regulatory architectures. Extending the framework to inter-individual variation and disease-associated transcriptomes, as demonstrated in related deep learning studies of gene expression regulation, may provide insight into the regulatory basis of phenotypic variation and disease susceptibility. Such analyses could facilitate the identification of early disease-associated shifts in regulatory architecture and support applications in diagnosis and preventive medicine.

Furthermore, by linking predicted regulatory targets to pathway-level redistribution patterns, this framework may contribute to the reconstruction of gene regulatory networks and enable in silico prediction of transcriptional responses to perturbations. These capabilities could support the identification of candidate therapeutic targets and guide strategies for modulating gene expression programs in a controlled manner.

419 Together, our results establish contribution score–based pathway redistribution
420 as a framework for uncovering conserved signaling architectures alongside context-
421 dependent functional modules, providing a systems-level perspective on gene regulation
422 beyond expression-centric analyses.
423
424

425 Methods

426 Predictive modeling of gene expression across diverse cell types using 427 deep learning and NABP binding information

428 To predict gene expression across multiple human cell types, we adapted a
429 previously established deep learning framework that integrates DNA-binding site data for
430 transcription factors and other regulatory proteins. In this implementation, regulatory
431 elements located both proximally and distally (up to ±1 Mb) from transcription start sites
432 (TSSs) were included as input features (**Fig. 1**)[25,51]. To minimize the effects of cellular
433 heterogeneity in bulk tissue, RNA-seq datasets were selected at the individual cell-type
434 level. Publicly available RNA-seq data were retrieved for four functionally distinct cell
435 types: human foreskin fibroblasts (HFFs), mammary epithelial cells (HMECs), neural
436 progenitor cells (NPCs), and either HepG2 hepatoma cells or K562 leukemia cells,
437 depending on the class of regulatory protein analyzed[52,53]. Gene expression values
438 (FPKM) were computed using the RSeQC toolkit based on GENCODE v19 annotations,
439 and subsequently $\log_2$-transformed after adding a small offset to avoid undefined values.
440 Following the approach described by Tasaki et al.[25,51], expression fold changes were
441 calculated relative to the median expression level across the four cell types. For each
442 model iteration, three common cell types (HFFs, HMECs, and NPCs) were used, and

443 either HepG2 (for DNA-binding protein analysis) or K562 (for RNA-binding protein
444 analysis) was included to match the availability of ChIP-seq and eCLIP validation
445 datasets. These cell types were chosen to represent diverse biological lineages and
446 regulatory contexts. For specific analyses requiring direct comparison of DeepLIFT
447 contribution scores across cell types—such as the evaluation of cell type–dependent
448 regulatory differences for the same nucleic acid-binding protein (e.g., **Fig. 3d**)—models
449 were trained using a four-cell-type configuration comprising HFFs, NPCs, HepG2, and
450 K562. This unified configuration enabled direct comparison of contribution score
451 distributions between HepG2 and K562 while maintaining a shared background of
452 common cell types.
453

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

455 To capture distal regulatory signals beyond promoter regions, we integrated DNA-
456 binding site data from putative enhancers into a deep learning framework for gene
457 expression prediction. Building upon our earlier work[25,51], this model extends the
458 architecture originally proposed by Tasaki et al. (2020)[25,51], which focused on promoter-
459 proximal regions and included RNA-binding protein and miRNA inputs. In our previous
460 study, aimed at identifying insulator-associated DNA-binding proteins, we deliberately
461 excluded RNA-binding proteins and miRNAs to focus on DNA-centered regulatory
462 elements. Instead, we combined DNA-binding sites from both promoter (-2 kb to +1 kb)
463 and enhancer-like regions (±1 Mb from the transcription start site [TSS]) into a unified
464 input matrix, enabling DeepLIFT analysis under the high-accuracy "genomics-default"
465 mode (**Supplementary Fig. S6**). 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.

Binding site coordinates were obtained from the GTRD database (v20.06)[54], and regulatory associations were inferred using cis-eQTLs from GTEx v8. A ChIP-seq peak that overlapped a 100-bp window centered on an expression-associated SNP was considered linked to the corresponding transcript. Trans-eQTLs were excluded, and coordinates were converted to the hg19 assembly using UCSC LiftOver[55]. To limit feature dimensionality while maximizing genomic coverage, we computed peak lengths for each DNA-binding protein (DBP) in 100-bp bins across ±1 Mb around the TSS, then retained the maximum peak value within each 10-kb window, yielding up to 200 enhancer bins. These were concatenated with 30 promoter-region bins to form a 230-bin unified input layer. This single-input-layer structure was necessary because DeepLIFT's genomics-default configuration ("RevealCancel rule on dense layers and Rescale rule on conv layers") does not support models with multiple input layers. Concatenating all DBP-related features into one matrix thus maintained compatibility while enhancing interpretability.

Model training followed the protocol described by Tasaki et al. and applied in our prior work[25,51]. Specifically, 2,705 genes from chromosome 1 were used for testing, while 22,251 genes from other chromosomes were randomly assigned to the training set and 2,472 to the validation set. Models were implemented in TensorFlow 2.6.0 and trained on an NVIDIA RTX 3080 GPU, typically completing in 2 to 4 hours. Feature importance was assessed using DeepLIFT, which assigned contribution scores to input features indicating their relative influence on gene expression prediction.

490 **Predicting NABP regulatory targets using gene co-expression–informed**
491 **substitutions**

492 To uncover regulatory targets of NABPs without relying on direct binding site
493 assays, we implemented a co-expression–driven modification of the deep learning input.
494 Specifically, DNA-binding proteins with low contribution scores in the original model
495 were replaced with NABPs whose regulatory interactions were inferred from gene co-
496 expression data. Of the 1,310 DBPs initially included, 302 were substituted with other
497 DBPs, and in a separate RNA-binding protein analysis, 325 were replaced with RNA-
498 binding proteins (RBPs) identified from the ENPD database[24]. Putative NABP–target
499 relationships were established using co-expression profiles provided by COXPRESdb[15],
500 where a Pearson's *z*-score > 2 between a NABP-encoding gene and a non-NABP gene
501 indicated a strong co-expression link. These co-expression pairs served as proxies for
502 regulatory interactions and were used to generate new input features for the model. To
503 integrate these proxies into the deep learning framework, hypothetical NABP binding
504 sites were assigned at the transcription start site (TSS) and 2 kb upstream of each co-
505 expressed gene. The model was then retrained using this modified input, and DeepLIFT
506 was applied to calculate feature-level importance scores (hereafter referred to as
507 contribution scores), which quantify the influence of each NABP–gene interaction on the
508 model's predicted expression output. These contribution scores represent feature-wise
509 attributions derived from the trained deep learning model and do not correspond directly
510 to measured gene expression levels, which reflect the integrated effects of multiple
511 regulatory inputs. This approach enables the functional characterization of NABPs with
512 unknown or uncharacterized binding motifs by leveraging gene co-expression as a proxy
513 for regulatory targeting.

514

515 **Validation of predicted NABP targets using experimentally determined**
516 **binding**

517 To assess the biological relevance of the predicted regulatory targets, we
518 compared deep learning–inferred NABP–gene associations with experimentally
519 validated binding data. Specifically, we evaluated the spatial overlap between predicted
520 target genes and ChIP-seq peaks for 32 DNA-binding proteins (within ±5 kb of
521 transcription start sites [TSSs]) as well as eCLIP peaks for 29 RNA-binding proteins
522 (within the transcribed region from TSS to transcription end site [TES]). Gene structure
523 annotations were derived from the GENCODE v19 release[56], and binding site data were
524 obtained from the ENCODE portal and prior publications (**Supplementary Table S6**)[57].
525 This overlap analysis provides orthogonal validation for the predicted targets and
526 supports the utility of gene co-expression–based modeling for inferring NABP–target
527 relationships.

528

529 **Prioritization of NABP regulatory targets based on importance score**
530 **rankings**

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

538 of high-confidence predictions. While a contribution score rank cutoff of five was
539 sufficient in most cases, the threshold was relaxed in specific cases—for example, up to
540 rank ten for proteins, which initially yielded a smaller number of predicted targets.

541 Functional enrichment and AI-based interpretation confirm the biological
542 relevance of NABP regulatory targets. To evaluate the biological coherence of predicted
543 NABP regulatory targets, we performed functional enrichment analysis using the
544 Statistical Overrepresentation Test from the PANTHER database alongside custom in-
545 house scripts[29]. Gene Ontology (GO) terms from biological process, molecular function,
546 and cellular component categories were analyzed for enrichment across three gene sets:
547 (1) primary regulatory target genes identified by the model, (2) all co-expressed genes
548 (Pearson's z-score > 2), and (3) randomly selected genes of matched size. Enrichment
549 was assessed using Fisher's exact test with Benjamini–Hochberg false discovery rate
550 (FDR) correction, and FDR < 0.05 was considered statistically significant. For this
551 analysis, human gene symbols were mapped to Ensembl Gene IDs using the HGNC
552 database (https://www.genenames.org/), and GO annotations were obtained from the
553 Gene Ontology resource[27,28]. Target gene IDs served as the test set, while the full
554 expression dataset used in the deep learning model formed the reference background.

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

562 normalized contribution values specific to the DBP or RBP of interest. The ranked lists
563 were used as input for GSEA with the Reactome (v2025.1) and Gene Ontology
564 (Molecular Function, Biological Process, and Cellular Component, 2021 releases) gene
565 set collections. Analyses were performed with 10,000 permutations, weighted enrichment
566 statistics (weight = 1), and a minimum gene-set size of 1 to retain all functional categories.
567 Enrichment significance was assessed using normalized enrichment scores (NES),
568 nominal *p*-values, and false-discovery rates (FDR). This procedure enabled the
569 identification of pathways enriched among genes with high positive or negative
570 DeepLIFT contributions, thereby revealing regulatory roles and cell type–specific
571 pathways associated with each nucleic acid–binding protein. Pairwise similarity between
572 Signal Transduction subpathway sets was quantified using the Jaccard similarity index,
573 defined as the size of the intersection divided by the size of the union of the compared
574 sets. Unlike conventional pathway activity methods such as GSVA [58] or AUCell [59], which
575 rank genes based on expression levels, our framework ranks genes using deep learning–
576 derived contribution scores, enabling pathway analysis in terms of inferred regulatory
577 influence.

578 To complement the statistical analyses, we applied a ChatGPT-based functional
579 inference method[31]. This approach used curated prompts—detailed in *Extended Data*
580 *Tables 1 and 5*—to predict functional roles based on gene sets, and retrieved relevant
581 literature using a modified version of the "4.Reference search and validation.ipynb"
582 script[31]. This script is publicly available on Code Ocean and GitHub
583 (https://doi.org/10.24433/CO.7045777.v1 ;
584 https://github.com/idekerlab/llm_evaluation_for_gene_set_interpretation). The

585 ChatGPT-based results provided consistent and interpretable biological themes aligned
586 with the enrichment findings.
587

588 ## Definition of functional modules for Reactome pathways

589 To facilitate interpretation of pathway-level results, individual Reactome
590 pathways were grouped into higher-order functional modules based on shared biological
591 processes. Functional modules were manually curated by integrating pathway annotations,
592 pathway names, and known biological relationships among pathways. Pathways
593 belonging to the same top-level Reactome category were first examined and further
594 subdivided into functionally coherent modules (e.g., synaptic signaling, receptor tyrosine
595 kinase signaling, immune receptor signaling, and metabolic sub-pathways). This
596 grouping was performed to reduce redundancy among closely related pathways and to
597 enable more interpretable visualization of pathway enrichment patterns. Each pathway
598 was assigned to a single functional module, and a corresponding abbreviated module
599 name ("module short") was defined for compact representation in figures. The complete
600 mapping between Reactome pathways and functional modules is provided in
601 **Supplementary Table S15**. Although this module definition involves manual curation,
602 the grouping was guided by established biological knowledge and Reactome pathway
603 structure, ensuring consistency across categories and RBPs.
604

605 ## Statistical analysis

606 Functional enrichment was assessed using Fisher's exact test, with multiple testing
607 correction applied via the Benjamini–Hochberg false discovery rate (FDR) procedure.
608 Statistical significance was defined as FDR < 0.05. Distributions of DeepLIFT

609 contribution scores for nucleic acid-binding sites derived from gene co-expression data
610 and ChIP-seq data were compared using the Mann–Whitney U test (also known as the
611 Wilcoxon rank-sum test). A p-value < 0.05 was considered statistically significant. GSEA
612 (gene set enrichment analysis) was performed using DeepLIFT contribution scores of
613 DBPs and RBPs. The normalized enrichment score (NES) was used to quantify whether
614 regulatory target genes associated with specific biological functions were preferentially
615 located at the high or low end of the ranked gene list. Statistical significance was assessed
616 using permutation-based p-values, as implemented in the GSEA framework.
617

618 **Data availability**

619 RNA-seq data for human foreskin fibroblast (HFF) cells were obtained from the Gene
620 Expression Omnibus (GEO) under accession number GSM941744. RNA-seq data for
621 human mammary epithelial cells (HMEC) were retrieved from ENCODE via the UCSC
622 Genome Browser (http://genome.ucsc.edu/encode/) under the file name
623 *wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam*. RNA-seq data for human neural
624 progenitor cells (NPC) were obtained from GEO under accession number GSM915326.
625 RNA-seq data for HepG2 and K562 cells were obtained from the ENCODE portal under
626 accession numbers ENCFF670LIE and ENCFF584WML, respectively. mRNA co-
627 expression data were downloaded from COXPRESdb v7
628 (https://zenodo.org/record/6861444). The list of nucleic acid-binding proteins was
629 obtained from the Eukaryotic Nucleic acid Binding Protein (ENPD) database
630 (http://qinlab.sls.cuhk.edu.hk/ENPD/). ChIP-seq data for transcription factors were
631 downloaded from the GTRD database (v20.06 release; https://gtrd.biouml.org). eQTL
632 data (GTEx_Analysis_v8_eQTL.tar) were obtained from the GTEx portal (v8 release;

633 https://gtexportal.org). eCLIP data for RNA-binding proteins were retrieved from the
634 ENCODE portal (accession ENCSR456FVU) and from Supplementary Data 2 of Van
635 Nostrand et al., *Nature* (2020), https://doi.org/10.1038/s41586-020-2077-3.
636

637 **Code availability**

638 The deep learning model implemented in this study is available at
639 https://github.com/reposit2/coexgene.
640

Figures

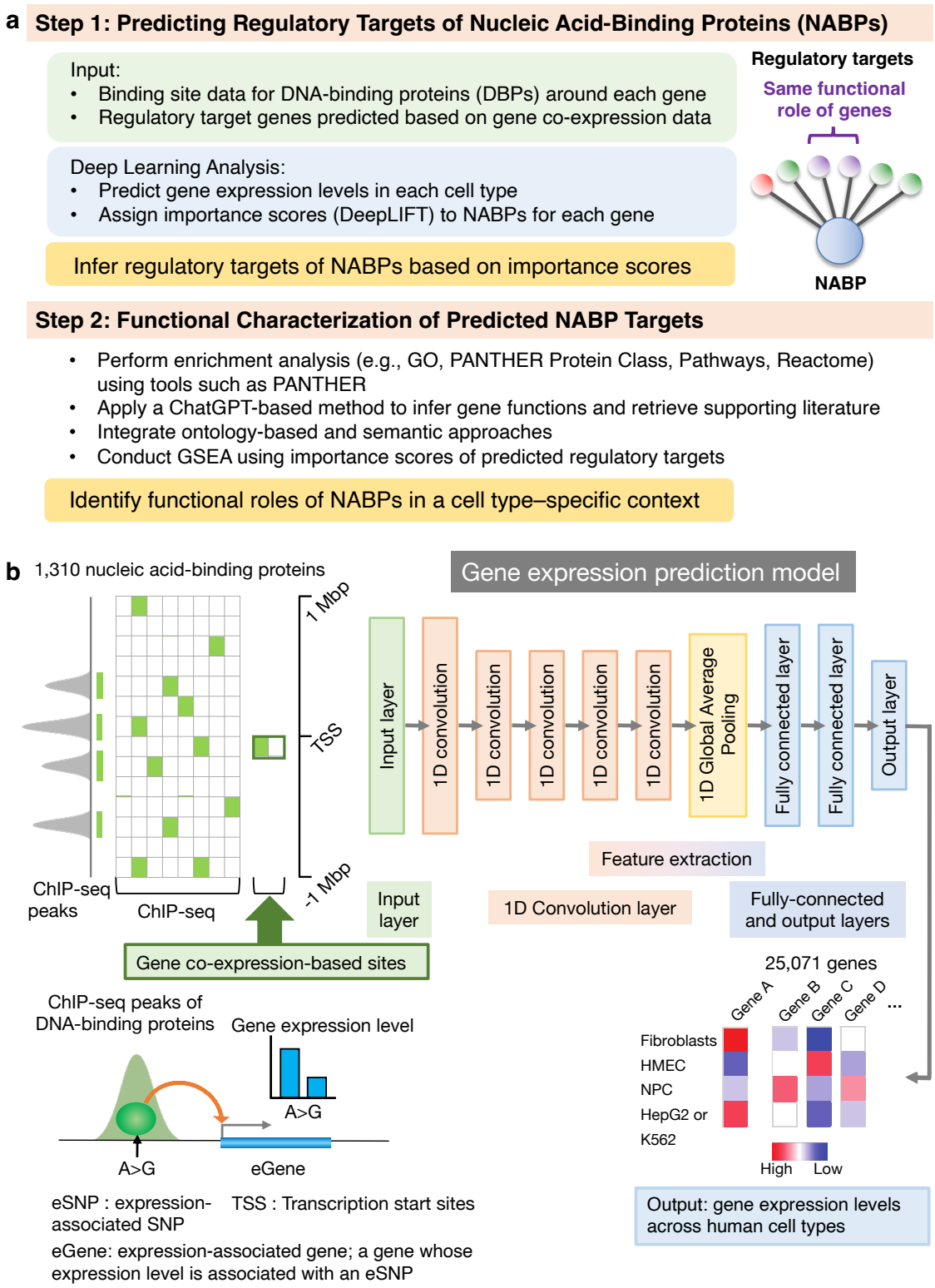

a
Step 1: Predicting Regulatory Targets of Nucleic Acid-Binding Proteins (NABPs)
Input:
• Binding site data for DNA-binding proteins (DBPs) around each gene
• Regulatory target genes predicted based on gene co-expression data
Deep Learning Analysis:
• Predict gene expression levels in each cell type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

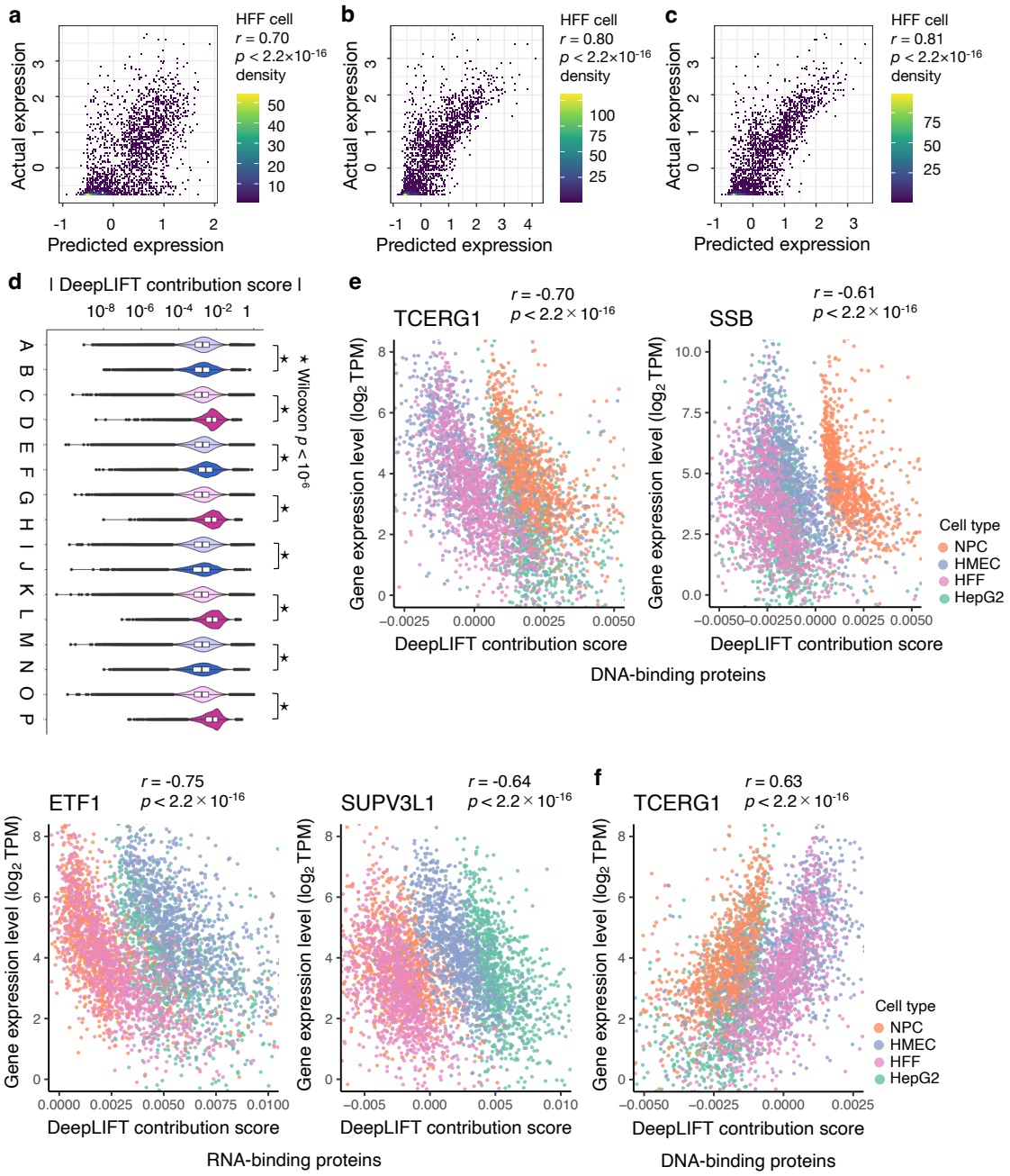
Figure 1. Overview of the deep learning framework for predicting and functionally characterizing NABP regulatory targets.

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

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

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

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


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

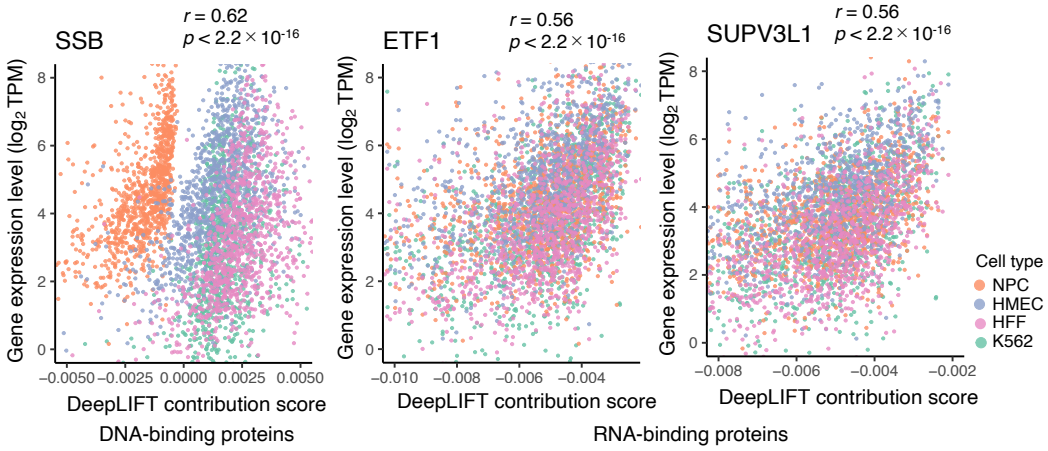


Figure 2. Relationship between gene expression levels and contribution scores.

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.

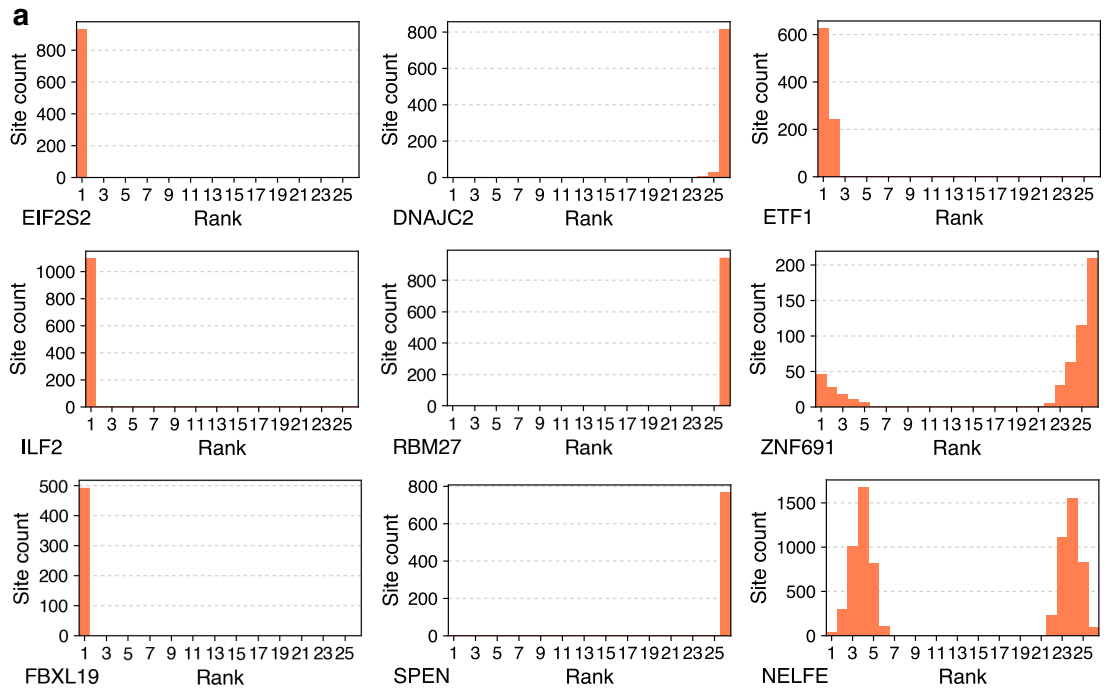
a. DNA-binding sites of 1,310 proteins derived from ChIP-seq data.

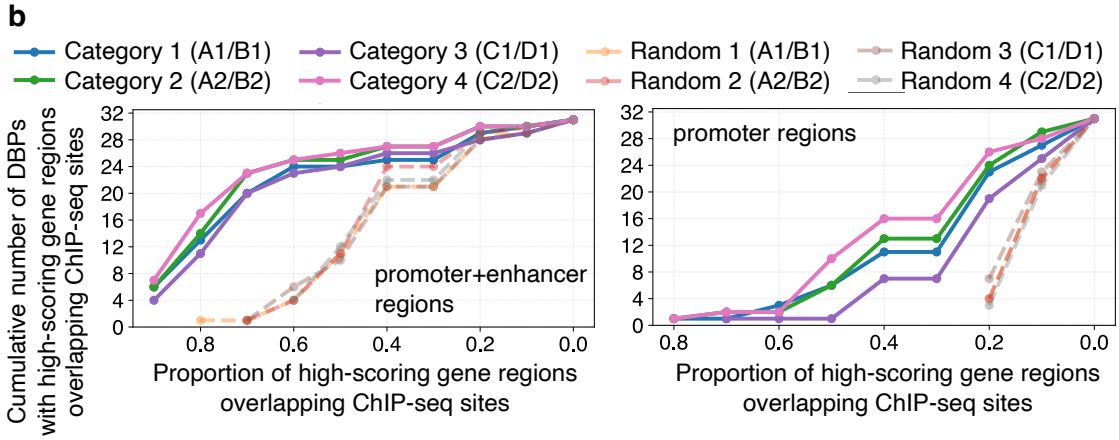
b. Replacement of 302 low-contribution DNA-binding proteins with co-expression–derived DNA-binding proteins.

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

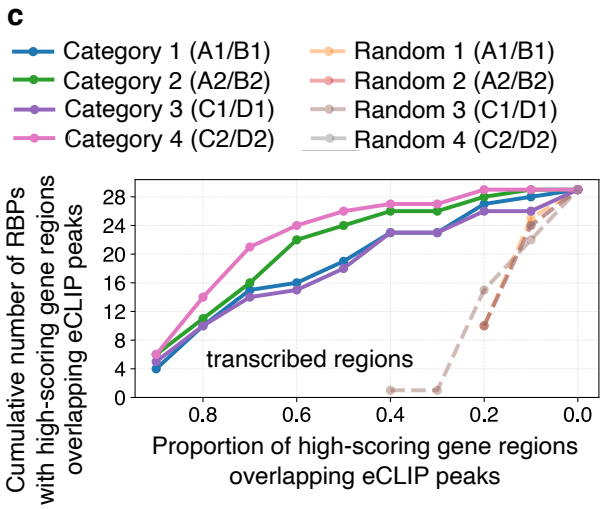
d. Comparison of absolute DeepLIFT contribution scores for predicted regulatory sites derived from gene co-expression versus ChIP-seq data, using a background defined as 0.5× the original input feature values. Each pair of violin plots represents co-expression–based sites (right) and corresponding ChIP-seq 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

701


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

702

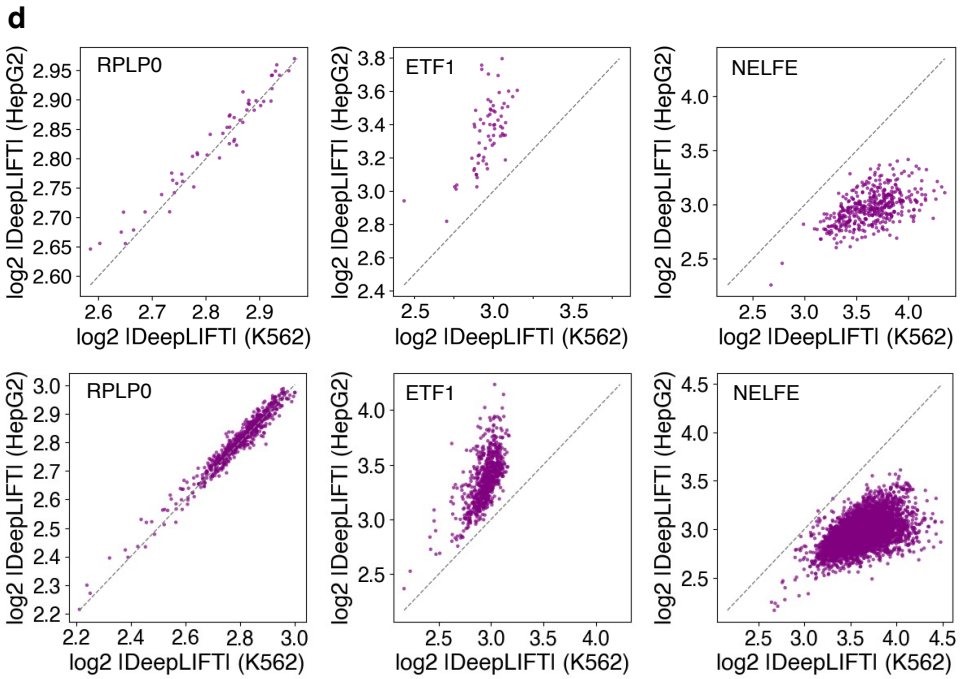


Figure 3. Comparison of predicted regulatory target genes with experimental ChIP-seq and eCLIP binding data. This figure illustrates the correspondence between gene-level regulatory contribution scores inferred by deep learning and experimentally identified nucleic acid-binding protein (NABP) binding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

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

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

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

735

736

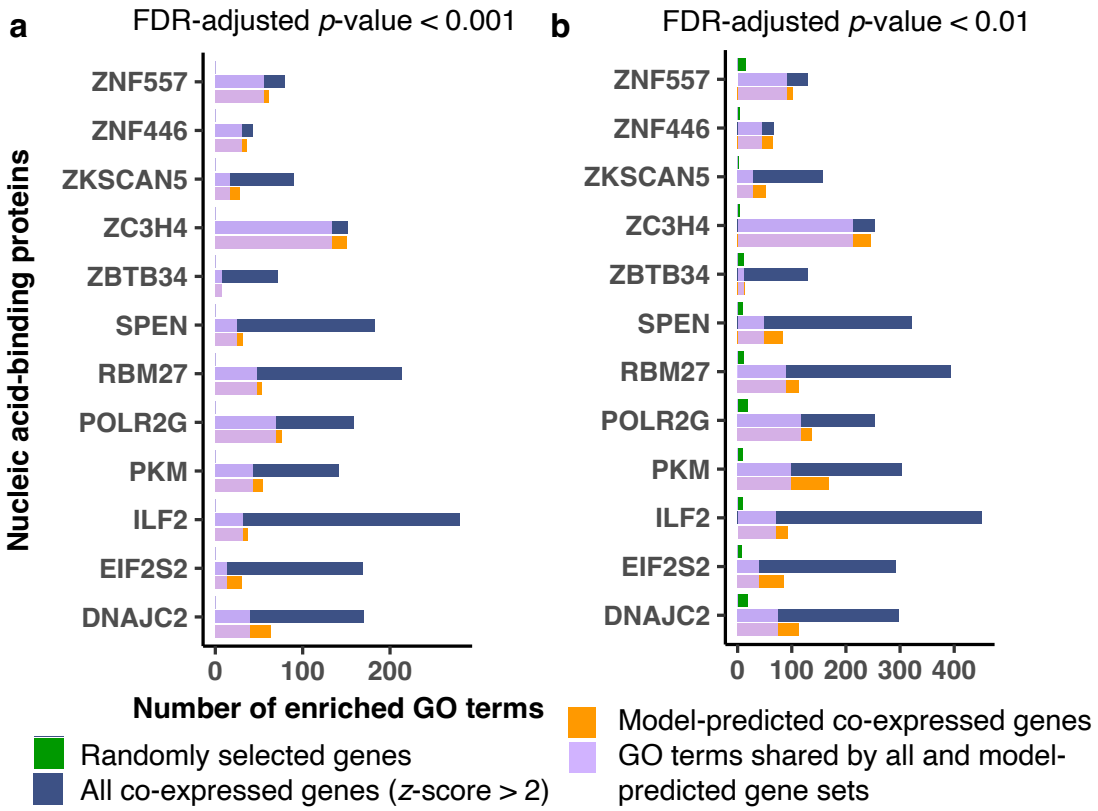


Figure 4. Overrepresented functional annotation terms among target genes of nucleic acid-binding proteins.

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).

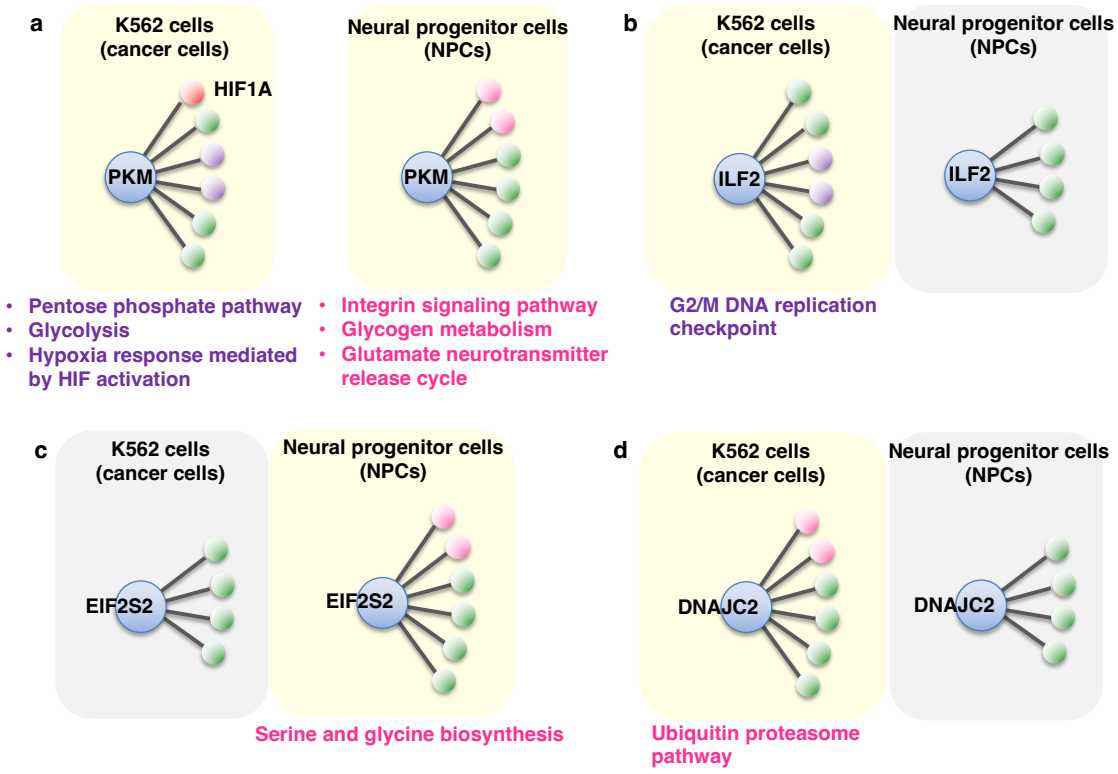
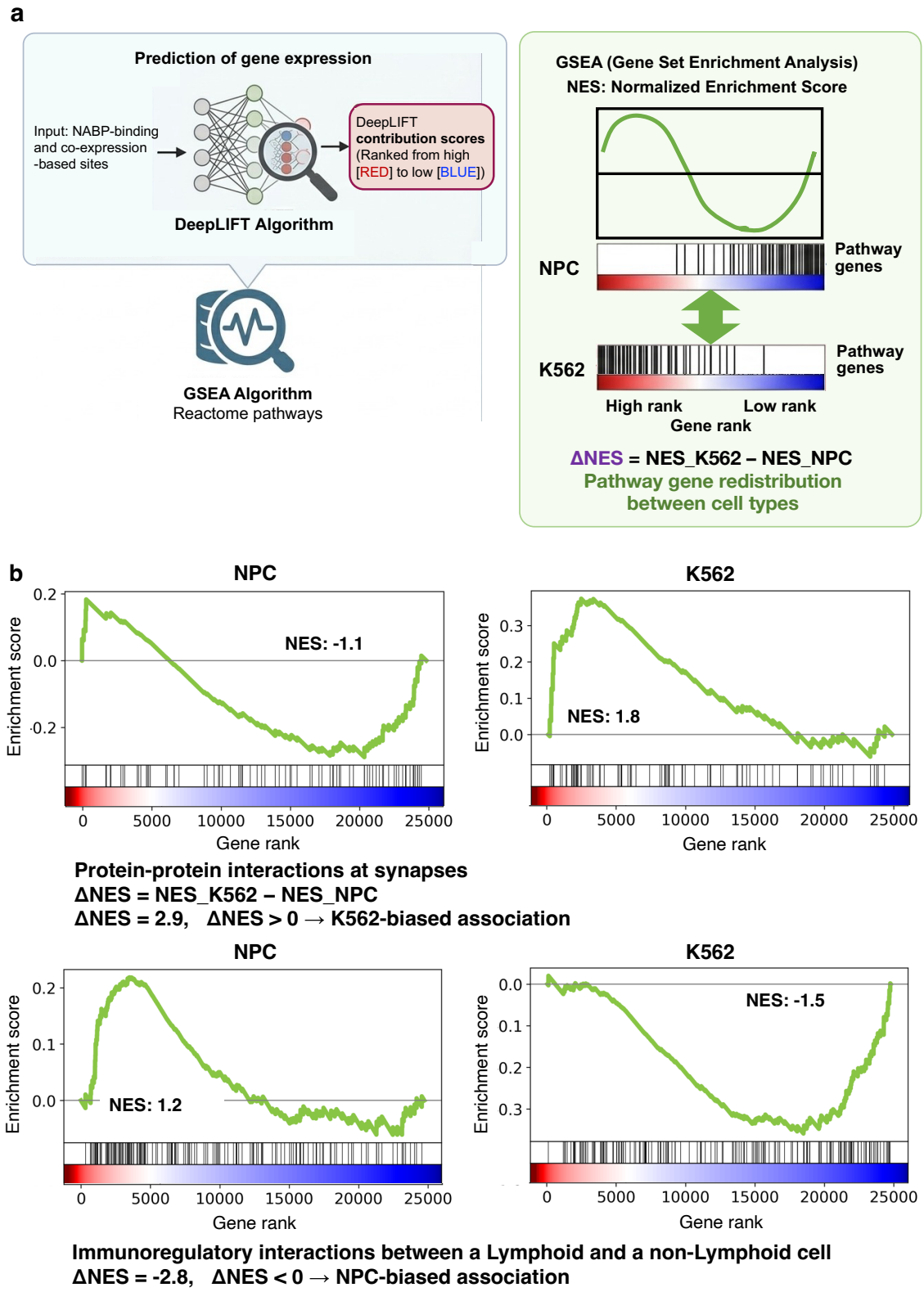


Figure 5. Functional mapping of nucleic acid-binding proteins reveals context-dependent regulatory roles in cancer and neural progenitor cells.

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
Prediction of gene expression
Input: NABP-binding and co-expression -based sites
DeepLIFT contribution scores (Ranked from high [RED] to low [BLUE])
DeepLIFT Algorithm
GSEA Algorithm
Reactome pathways
GSEA (Gene Set Enrichment Analysis)
NES: Normalized Enrichment Score
NPC
Pathway genes
K562
Pathway genes
High rank
Low rank
Gene rank
ΔNES = NES_K562 – NES_NPC
Pathway gene redistribution between cell types
b
NPC
K562
Enrichment score
NES: -1.1
NES: 1.8
Gene rank
Protein-protein interactions at synapses
ΔNES = NES_K562 – NES_NPC
ΔNES = 2.9, ΔNES > 0 → K562-biased association
NES: 1.2
NES: -1.5
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
ΔNES = -2.8, ΔNES < 0 → NPC-biased association

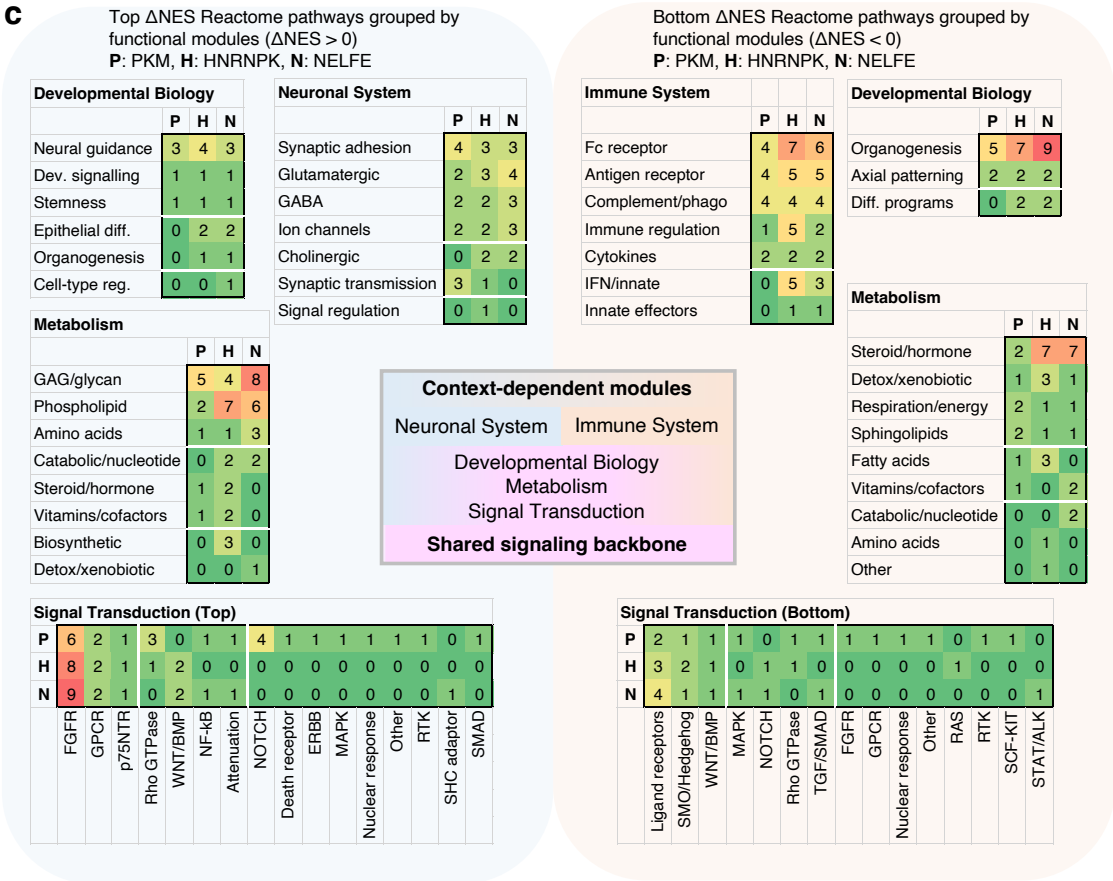


Figure 6. Pathway redistribution reveals a shared signaling backbone and context-dependent functional modules across RNA-binding proteins.

a. Conceptual framework of ΔNES-based pathway analysis. Gene-level contribution scores were computed using DeepLIFT and used to rank genes for each RNA-binding protein (RBP) and cellular context. Gene set enrichment analysis (GSEA) was applied to Reactome pathways, and pathway redistribution between cell types was quantified as ΔNES (NES_K562 − NES_NPC), reflecting shifts in the positional distribution of pathway-associated genes along contribution-ranked lists.

b. Representative GSEA enrichment plots illustrating pathway redistribution. Positive ΔNES values indicate relative enrichment in K562 cells, whereas negative ΔNES values indicate relative enrichment in neural progenitor cells (NPCs). These opposing

775 patterns demonstrate redistribution of pathway-associated genes across contribution-
776 ranked gene lists rather than direct pathway activation or repression. Other examples are
777 shown in **Supplementary Fig. S5**.

778 **c. Functional module–level organization of ΔNES-ranked pathways across RBPs**
779 **(PKM, HNRNPK, and NELFE).** A conserved signaling backbone is observed across all
780 RBPs, alongside context-dependent redistribution of functional modules. Top (ΔNES >
781 0) and bottom (ΔNES < 0) pathways were grouped into functional modules based on
782 shared biological processes. Numbers in the heatmap indicate the number of Reactome
783 pathways assigned to each functional module. Color scale indicates pathway counts or
784 enrichment distribution. Across all RBPs, broadly utilized signaling pathways—
785 including receptor tyrosine kinase (RTK), MAPK, and WNT-related signaling—form a
786 conserved signaling backbone shared across cellular contexts. In contrast, modules
787 associated with neuronal functions (e.g., synaptic signaling) and immune processes are
788 differentially enriched, indicating context-dependent functional redistribution. These
789 results support a hierarchical model in which RBPs regulate common signaling networks
790 while modulating cell type–specific functional outputs through redistribution of pathway-
791 associated genes.

792

793

794 ## Tables

795 **Table 1. Functional roles inferred for nucleic acid-binding proteins: integration of**
796 **UniProt annotations, ChatGPT-based predictions, and confidence scores.**

797 Functional annotations obtained from UniProt entries were systematically compared with
798 significantly enriched biological processes among the model-predicted regulatory targets,

799 as identified by the PANTHER Overrepresentation Test (see **Supplementary Table S5**).
800 Shared or overlapping functions between each nucleic acid-binding protein (NABP) and
801 its predicted targets are indicated. Functional roles inferred using a ChatGPT-based
802 interpretation method are highlighted in yellow, with confidence scores—defined as the
803 proportion of annotated genes within the gene set—provided in parentheses. Confidence
804 levels are categorized as high (0.87–1.00), medium (0.82–0.86), low (0.01–0.81), or none
805 (0). For methodological reference, see Hu M. et al., *Nat. Methods*, 2025
806 (https://doi.org/10.1038/s41592-024-02525-x)[31].

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

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

815 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

816

817

818

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

820 **acid-binding proteins.**

821 Functional enrichment analysis was performed using the PANTHER database. This table
822 provides selected examples of annotations that were statistically enriched among
823 regulatory target genes predicted by the deep learning model trained on gene co-
824 expression data. '**Reference**' indicates the total number of human genes; '**Analyzed**'
825 denotes the number of regulatory target genes examined; '**Expected**' represents the
826 number of genes expected to have the annotation by chance; '**Fold Enrichment**' reflects
827 the observed-to-expected ratio; '**Raw *p*-value**' reports the result of Fisher's exact test; and
828 '**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

829

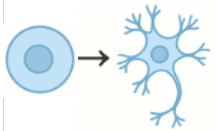
830

831

832 **Table 4. Cell type−dependent enrichment of functional pathways identified by**
833 **GSEA**

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

840 values were predominantly annotated to high-level Reactome categories related to
841 neural development and differentiation, whereas pathways with large negative ΔNES
842 values were enriched for categories associated with immune regulation, cell cycle
843 control, and genome maintenance, reflecting contrasting functional associations
844 between NPCs and K562 cells. Corresponding analyses performed using an amplified
845 DeepLIFT background (input × 2) yielded qualitatively consistent pathway-level
846 contrasts, with generally higher enrichment scores due to an increased number of
847 predicted PKM-associated target genes (**Supplementary Table S8** and **S9**).
848
849



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	1.71	-1.16	2.87	Neuronal System	R-HSA-112316
Attachment and Entry	R-HSA-9694614	1.40	-1.36	2.76	Disease	R-HSA-1643685
Golgi-to-ER retrograde transport	R-HSA-8856688	1.58	-1.17	2.75	Vesicle-mediated transport	R-HSA-5653656
Glycosaminoglycan-protein linkage region biosynthesis	R-HSA-1971475	1.17	-1.58	2.75	Metabolism	R-HSA-1430728
Defective EXT2 causes exostoses 2	R-HSA-3656237	1.39	-1.35	2.74	Disease	R-HSA-1643685
Defective B4GALT7 causes EDS, progeroid type	R-HSA-3560783	1.12	-1.60	2.72	Disease	R-HSA-1643685
Phase 0 - rapid depolarisation	R-HSA-5576892	1.46	-1.24	2.70	Muscle contraction	R-HSA-397014
Asparagine N-linked glycosylation	R-HSA-446203	1.55	-1.13	2.68	Metabolism of proteins	R-HSA-392499
HS-GAG degradation	R-HSA-2024096	1.17	-1.50	2.68	Metabolism	R-HSA-1430728
Diseases associated with glycosaminoglycan metabolism	R-HSA-3560782	1.06	-1.60	2.66	Disease	R-HSA-1643685
Regulation of CDH1 posttranslational processing and trafficking to plasma membrane	R-HSA-9768727	1.34	-1.31	2.65	Cell-Cell communication	R-HSA-1500931
LGI-ADAM interactions	R-HSA-5682910	1.66	-0.94	2.60	Developmental Biology	R-HSA-1266738
Extracellular matrix organization	R-HSA-1474244	1.37	-1.23	2.60	Extracellular matrix organization	R-HSA-1474244
Synaptic adhesion-like molecules	R-HSA-8849932	1.23	-1.36	2.60	Neuronal System	R-HSA-112316
Neuronal System	R-HSA-112316	1.57	-1.00	2.56	Neuronal System	R-HSA-112316
Signaling by Rho GTPases, Miro GTPases and RHOBTB3	R-HSA-9716542	1.48	-1.06	2.54	Signal Transduction	R-HSA-162582
Smooth Muscle Contraction	R-HSA-445355	1.45	-1.08	2.54	Muscle contraction	R-HSA-397014
Ion channel transport	R-HSA-983712	1.33	-1.15	2.48	Transport of small molecules	R-HSA-382551
Receptor-type tyrosine-protein phosphatases	R-HSA-388844	1.60	-0.88	2.48	Neuronal System	R-HSA-112316
Respiratory syncytial virus (RSV) attachment and entry	R-HSA-9820960	1.38	-1.09	2.48	Disease	R-HSA-1643685
Binding and Uptake of Ligands by Scavenger Receptors	R-HSA-2173782	-1.18	1.46	-2.64	Vesicle-mediated transport	R-HSA-5653656
WNT ligand biogenesis and trafficking	R-HSA-3238698	-1.10	1.56	-2.66	Signal Transduction	R-HSA-162582
Complement cascade	R-HSA-166658	-1.09	1.56	-2.66	Immune System	R-HSA-168256
Role of phospholipids in phagocytosis	R-HSA-2029485	-1.32	1.34	-2.67	Immune System	R-HSA-168256
Nucleotide Excision Repair	R-HSA-5696398	-1.28	1.39	-2.67	DNA Repair	R-HSA-73894
Translation	R-HSA-72766	-1.14	1.60	-2.73	Metabolism of proteins	R-HSA-392499
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	R-HSA-198933	-1.52	1.23	-2.75	Immune System	R-HSA-168256
Transcription-Coupled Nucleotide Excision Repair (TC-NER)	R-HSA-6781827	-1.44	1.32	-2.76	DNA Repair	R-HSA-73894
Creation of C4 and C2 activators	R-HSA-166786	-1.13	1.68	-2.81	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	1.61	-2.84	Immune System	R-HSA-168256
Aerobic respiration and respiratory electron transport	R-HSA-1428517	-1.23	1.67	-2.90	Metabolism	R-HSA-1430728
Response to metal ions	R-HSA-5660526	-1.23	1.69	-2.92	Cellular responses to stimuli	R-HSA-8953897
FCGR activation	R-HSA-2029481	-1.18	1.78	-2.95	Immune System	R-HSA-168256
Mitotic G1 phase and G1/S transition	R-HSA-453279	-1.24	1.72	-2.96	Cell Cycle	R-HSA-1640170
M Phase	R-HSA-68886	-1.32	1.73	-3.05	Cell Cycle	R-HSA-1640170
Mitotic Metaphase and Anaphase	R-HSA-2555396	-1.27	1.81	-3.08	Cell Cycle	R-HSA-1640170
DNA Replication	R-HSA-69306	-1.27	1.82	-3.09	DNA Replication	R-HSA-69306
Scavenging of heme from plasma	R-HSA-2168880	-1.44	1.67	-3.11	Vesicle-mediated transport	R-HSA-5653656
Metallothioneins bind metals	R-HSA-5661231	-1.05	2.61	-3.66	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.

878 # Supplementary Note and Tables

879

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

882 **Contribution score ranking and binning**

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

888

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

890 High- and low-contribution regions were defined using both rank-based and score-
891 based thresholds to assess the robustness of enrichment results. Rank-based thresholds
892 included upper and lower rank cutoffs (e.g., top or bottom 20 and 50 ranks), while score-
893 based thresholds were defined using absolute contribution score cutoffs (e.g., ≥ 0.01 or ≥
894 0.001). All threshold combinations used in the analyses are summarized in **Tables S2** and
895 **S3.**

896

897 **Genomic regions used for overlap analyses**

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

spanning from the TSS to the transcription end site (TES), including both exonic and intronic intervals.

Random controls

Random control regions were generated by sampling genomic intervals matched for size and genomic context, excluding annotated binding sites, to estimate the expected overlap by chance. All enrichment and overlap analyses were repeated using these controls for comparison.

Background configuration for DeepLIFT analyses

Unless otherwise noted, DeepLIFT contribution scores were computed using a background defined as 0.5× the original input feature values. Analyses using an alternative background configuration (input × 2) yielded qualitatively similar results and are presented in **Supplementary Figures S2** and **S3.**

Cell type configurations used for model training and comparative analyses

To accommodate differences in experimental data availability and analytical objectives, multiple cell type configurations were used across analyses. For the primary analyses of DNA-binding proteins (DBPs), models were trained using expression data from HFFs, HMECs, NPCs, and HepG2, enabling validation against ChIP-seq datasets available for HepG2 cells. For RNA-binding protein (RBP) analyses, models were trained using HFFs, HMECs, NPCs, and K562 to facilitate validation using eCLIP datasets available for K562 cells. For analyses requiring direct comparison of regulatory contribution patterns between HepG2 and K562 for the same nucleic acid-binding

926 proteins (e.g., **Fig. 3d**), a unified four–cell-type configuration comprising HFFs, NPCs,
927 HepG2, and K562 was employed. This configuration enabled direct comparison of
928 DeepLIFT contribution score distributions across these two cell types while maintaining
929 a shared transcriptomic background. In all cases, cell type selection was driven by
930 analytical requirements and the availability of independent experimental validation data,
931 and no validation datasets were used as direct inputs during model training.
932

933 Gene co-expression–based predictions align with ChIP-seq–defined 934 DNA-binding targets

935 High-ranking predicted binding regions frequently coincide with ChIP-seq peaks,
936 particularly in promoter–enhancer intervals. Deep learning–derived DNA-binding site
937 predictions based on gene co-expression data show substantial concordance with ChIP-
938 seq results, especially when broader genomic regions encompassing promoters and
939 putative enhancers are considered. When overlap was assessed within ±100 kb of
940 transcription start sites (TSSs)—a window supported by prior enhancer–gene mapping
941 studies[60,61]—the mean and median overlap ratios reached 0.70 and 0.77, respectively
942 (Ratio A2 / B2; **Supplementary Tables S2 and S3**). In contrast, overlap was lower for
943 promoter regions alone (±5 kb), with corresponding mean and median values of 0.35 and
944 0.34. Several DNA-binding proteins, including AKAP8, POGK, SPEN, THAP7, ZNF142,
945 and ZNF691, showed notably high promoter-region overlap ratios (>0.5), suggesting
946 potential direct regulatory activity. Nevertheless, not all ChIP-seq peaks necessarily
947 reflect functionally relevant transcriptional binding events. For RNA-binding proteins,
948 overlap was examined within the transcribed regions of RNA molecules, and the mean
949 and median overlap ratios were 0.70 and 0.76, respectively (Ratio A2 / B2;

950 **Supplementary Table S4**). These findings indicate that model-inferred high-scoring
951 sites do not universally coincide with experimentally determined binding regions,
952 underscoring the complexity and context dependence of regulatory interactions.
953
954
955 **Functional interpretation of ΔNES-based pathway redistribution across**
956 **PKM, HNRNPK, and NELFE**
957 To contextualize the ΔNES-based redistribution patterns described in the main
958 text, we examined established biological roles of PKM, HNRNPK, and NELFE in
959 relation to the pathway categories identified by contribution score–based GSEA.
960
961 **PKM**
962 PKM2, the predominant isoform in proliferative and stem-like states, is widely
963 recognized as a central mediator of oncogenic metabolic rewiring and has been shown to
964 exert non-canonical nuclear functions that coordinate transcriptional programs during
965 tumorigenesis [49,50]. Beyond its canonical glycolytic role, PKM2 regulates gene expression
966 through protein–protein interactions with transcriptional regulators and chromatin-
967 associated factors. Isoform switching between PKM2 and PKM1 accompanies metabolic
968 transitions during differentiation, including neuronal maturation, where PKM1-enriched
969 states are associated with oxidative metabolism and terminal differentiation [36]. PKM has
970 also been identified as a ribosome-associated factor capable of modulating mRNA
971 translation [62], and accumulating evidence suggests that metabolic enzymes, including
972 pyruvate kinase, can function as non-canonical RNA-binding proteins [63].

973 In addition to its roles in metabolism and transcriptional regulation, PKM2 has
974 emerged as a key regulator of immunometabolic processes. PKM2-dependent metabolic
975 reprogramming has been shown to modulate cytokine production, including both pro-
976 inflammatory and anti-inflammatory responses, and to promote immune cell proliferation
977 and activation [64-66]. These findings indicate that PKM2 links metabolic state to immune
978 signaling outputs, consistent with the enrichment of immune-associated pathways
979 observed in ΔNES-based analyses.

980 PKM2 is also tightly integrated with major signaling pathways. For example,
981 receptor tyrosine kinase signaling, including fibroblast growth factor receptor 1 (FGFR1),
982 directly phosphorylates PKM2, inducing a shift from an active tetrameric form to a less
983 active dimeric state and thereby promoting the Warburg effect [67]. In addition, PKM2 has
984 been functionally connected to NOTCH signaling, where coordinated regulation of
985 PKM2 and NOTCH1 influences tumor growth and metastasis [68], as well as to GPCR-
986 associated pathways, such as the GPR3–β-arrestin2–PKM2 axis, which regulates
987 glycolysis and liver pathophysiology [69]. PKM2 further contributes to TGF-β signaling by
988 stabilizing TGF-β receptor complexes, thereby promoting fibrotic responses [70].

989 Beyond these pathways, PKM2 participates in diverse context-dependent
990 processes, including lipid metabolism, angiogenesis, and metastatic progression. PKM2
991 has been implicated as a regulator of lipid metabolic programs, particularly in cardiac
992 tissues [71,72], and promotes tumor angiogenesis by modulating endothelial cell junction
993 dynamics and ATP production [73,74]. Oxidative modification of PKM2 has also been linked
994 to enhanced metastatic potential in cancer cells [75].

995 Together, these multifunctional properties position PKM2 at the interface of
996 metabolism, signaling, and gene regulation. This integrated regulatory capacity is

997 consistent with the ΔNES-inferred redistribution of neuronal, immune, metabolic, and
998 signaling-associated pathways across neural progenitor and leukemia contexts,
999 supporting a model in which PKM2 contributes to context-dependent repositioning of
1000 shared functional networks rather than driving strictly lineage-specific pathway activation
1001 [76-78].
1002
1003 **HNRNPK**
1004 HNRNPK is a multifunctional RNA-binding protein that integrates
1005 transcriptional and post-transcriptional regulatory processes. Haploinsufficiency of
1006 HNRNPK predisposes to hematologic malignancies and disrupts proliferation and
1007 differentiation programs in myeloid lineages [79]. In stem and progenitor contexts,
1008 HNRNPK coordinates transcriptional activation of proliferation-associated genes,
1009 including MYC and growth factor signaling components, while simultaneously
1010 regulating RNA stability and processing [80]. These dual transcriptional and post-
1011 transcriptional roles support a model in which HNRNPK contributes to lineage
1012 specification and tumor-associated transcriptional remodeling. Accordingly, the
1013 redistribution of immune- and neural-associated pathway annotations across cellular
1014 contexts observed here is compatible with a broader function of HNRNPK in modulating
1015 differentiation–proliferation balance rather than directing a single lineage-specific
1016 program.
1017
1018 **NELFE**
1019 NELFE is a subunit of the negative elongation factor (NELF) complex that
1020 regulates promoter-proximal RNA polymerase II pausing, a metazoan-specific

1021 mechanism critical for rapid transcriptional responsiveness and developmental gene
1022 regulation [81,82]. Oncogenic activation of NELFE has been shown to enhance MYC-
1023 associated transcriptional programs and promote tumor progression [83]. In hematopoietic
1024 differentiation, dynamic modulation of NELF protein abundance alters genome-wide
1025 transcriptional pausing and controls granulocytic lineage commitment [84]. Furthermore,
1026 NELF regulates progenitor expansion and stem cell pool maintenance in regenerative
1027 contexts through modulation of p53-dependent transcriptional programs [85]. Collectively,
1028 these findings indicate that NELFE functions as a context-dependent regulator of
1029 transcriptional architecture, influencing lineage transitions through control of elongation
1030 dynamics. The parallel ΔNES redistribution patterns observed for NELFE and HNRNPK
1031 are therefore consistent with convergence of distinct RNA-binding proteins on shared
1032 regulatory infrastructures governing differentiation and proliferative states.
1033
1034 **Shared signaling architecture**
1035 Across all three proteins, Signal Transduction emerged as the dominant top-level
1036 Reactome category. Subpathway analysis revealed enrichment of broadly utilized
1037 regulatory modules, including FGFR/RTK signaling, IRS-mediated cascades,
1038 WNT/BMP pathways, and receptor-associated second-messenger systems. Although
1039 neuronal receptor–related pathways were present, they constituted a minority of enriched
1040 subpathways (**Fig. 6c**). This distribution argues against a simple reflection of neuron-
1041 specific transcriptional signatures and instead supports the presence of a conserved
1042 upstream signaling backbone that is differentially interpreted across cellular contexts.
1043 Importantly, the ΔNES-based framework does not imply direct activation or
1044 repression of lineage programs by individual proteins. Rather, it captures hierarchical

1045 shifts in the positioning of pathway-associated genes along contribution-ranked lists,
1046 reflecting redistribution of inferred regulatory influence across cellular states. The
1047 convergence of redistribution patterns across metabolically (PKM), transcriptionally
1048 integrative (HNRNPK), and elongation-regulatory (NELFE) proteins supports the
1049 broader interpretation that contribution-based pathway ranking can reveal conserved
1050 regulatory scaffolds underlying context-dependent functional organization.
1051
1052

1053 **Supplementary Table S1. Co-expression–based selection enhances the**
1054 **contribution scores of NABP binding sites.**

1055 Binding sites identified through gene co-expression analysis exhibited higher
1056 absolute DeepLIFT contribution scores than those selected solely from ChIP-seq data
1057 without co-expression information. This trend was consistently observed for DNA-
1058 binding proteins and was significantly stronger for RNA-binding proteins across the
1059 analyzed cell types. For each group, the table presents the mean and median values of
1060 both positive and negative contribution scores. These results indicate that gene co-
1061 expression–based selection preferentially highlights functionally relevant regulatory
1062 interactions. The upper and lower sections of the table correspond to analyses performed
1063 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

1067 **Supplementary Table S2. Enrichment of high-contribution promoter**
1068 **regions with experimentally validated DBP binding sites.** (See
1069 Supplementary Excel file)

1070 This table summarizes the proportion of promoter regions (±5 kb from transcription
1071 start sites [TSSs]) that were predicted to be functionally important by DeepLIFT
1072 contribution scores and validated by ChIP-seq data for the corresponding DNA-binding
1073 proteins (DBPs). Overlap ratios were calculated across four thresholds—two based on
1074 rank and two based on absolute score—to assess the robustness of the predictions.

1075 High- and low-contribution regions were classified using the following thresholds
1076 for the 1,310 NABPs analyzed:

1077 **Rank-based thresholds**:

1078 (A1 and B1) proteins ranked in the top or bottom 20 (i.e., rank ≤ 20 or ≥ 1,290)

1079 (A2 and B2) proteins ranked in the top or bottom 50 (i.e., rank ≤ 50 or ≥ 1,260)

1080 **Score-based thresholds**:

1081 (C1 and D1) contribution score ≥ 0.01 or ≤ –0.01

1082 (C2 and D2) contribution score ≥ 0.001 or ≤ –0.001

1083

1084

1085 **Supplementary Table S3. Enrichment of high-contribution promoter**
1086 **and enhancer regions with ChIP-seq–validated DBP binding sites.**

1087 This table summarizes the proportion of predicted high-contribution regions—
1088 defined within ±100 kb of transcription start sites (TSSs), encompassing both promoter
1089 and distal enhancer regions—that overlap with ChIP-seq binding sites for the
1090 corresponding DNA-binding proteins (DBPs). Strong concordance between model-

1091 derived predictions and ChIP-seq peaks across multiple thresholds highlights the
1092 reliability of the contribution scoring method for identifying functionally relevant
1093 regulatory sites.

1094 High- and low-contribution regions were classified using the following thresholds
1095 for the 1,310 NABPs analyzed:

1096 **Rank-based thresholds**:

1097 (A1 and B1) proteins ranked in the top or bottom 20 (i.e., rank ≤ 20 or ≥ 1,290)

1098 (A2 and B2) proteins ranked in the top or bottom 50 (i.e., rank ≤ 50 or ≥ 1,260)

1099 **Score-based thresholds**:

1100 (C1 and D1) contribution score ≥ 0.01 or ≤ –0.01

1101 (C2 and D2) contribution score ≥ 0.001 or ≤ –0.001

1102

1103

1104 Supplementary Table S4. Enrichment of high-contribution transcribed 1105 regions with eCLIP–validated RBP binding sites.

1106 This table summarizes the proportion of predicted high-contribution regions—
1107 defined within the transcribed regions of genes (from transcription start site [TSS] to
1108 transcription end site [TES])—that overlap with eCLIP peaks for the corresponding
1109 RNA-binding proteins (RBPs). Strong concordance between model-based predictions
1110 and eCLIP peaks across multiple thresholds underscores the reliability of the contribution
1111 scoring method for identifying functionally relevant regulatory sites.

1112 High- and low-contribution regions were classified using the following
1113 thresholds applied to the 1,310 NABPs analyzed:

1114 **Rank-based thresholds**:

1115 (A1 and B1) proteins ranked in the top or bottom 20 (i.e., rank ≤ 20 or ≥ 1,290)

1116 (A2 and B2) proteins ranked in the top or bottom 50 (i.e., rank ≤ 50 or ≥ 1,260)

1117 **Score-based thresholds**:

1118 (C1 and D1) contribution score ≥ 0.01 or ≤ –0.01

1119 (C2 and D2) contribution score ≥ 0.001 or ≤ –0.001

1120

1121

1122 **References**

1123 1 Belz, G. T. & Nutt, S. L. Transcriptional programming of the dendritic cell network.
1124 *Nature reviews. Immunology* **12**, 101–113 (2012). https://doi.org/10.1038/nri3149
1125 2 Miller, J. C. *et al.* Deciphering the transcriptional network of the dendritic cell lineage.
1126 *Nature immunology* **13**, 888–899 (2012). https://doi.org/10.1038/ni.2370
1127 3 Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man.
1128 *Cell* **140**, 744–752 (2010). https://doi.org/10.1016/j.cell.2010.01.044
1129 4 Gertz, J. *et al.* Distinct properties of cell-type-specific and shared transcription factor
1130 binding sites. *Molecular cell* **52**, 25–36 (2013).
1131 https://doi.org/10.1016/j.molcel.2013.08.037
1132 5 Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
1133 https://doi.org/10.1016/j.cell.2018.01.029
1134 6 Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of
1135 human transcription factors: function, expression and evolution. *Nature reviews. Genetics*
1136 **10**, 252–263 (2009). https://doi.org/10.1038/nrg2538
1137 7 Pierson, E. *et al.* Sharing and Specificity of Co-expression Networks across 35 Human
1138 Tissues. *PLoS Comput Biol* **11**, e1004220 (2015).
1139 https://doi.org/10.1371/journal.pcbi.1004220
1140 8 Sonawane, A. R. *et al.* Understanding Tissue-Specific Gene Regulation. *Cell Rep* **21**,
1141 1077–1088 (2017). https://doi.org/10.1016/j.celrep.2017.10.001
1142 9 Saha, A. *et al.* Co-expression networks reveal the tissue-specific regulation of transcription
1143 and splicing. *Genome research* **27**, 1843–1858 (2017).
1144 https://doi.org/10.1101/gr.216721.116
1145 10 Boyle, A. P. *et al.* Comparative analysis of regulatory information and circuits across
1146 distant species. *Nature* **512**, 453–456 (2014). https://doi.org/10.1038/nature13668

11 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

12 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

13 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

14 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

15 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

16 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

17 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

18 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

19 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

20 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

21 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

22 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

23 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

24 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

25 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

26 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).

27 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

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

29 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

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

31 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

32 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

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

34 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

35 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

36 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

37 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

38 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

39 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

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

41 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

42 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

43 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

44 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

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

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

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

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

49 Yang, W. *et al.* ERK1/2-dependent phosphorylation and nuclear translocation of PKM2 promotes the Warburg effect. *Nat Cell Biol* **14**, 1295–1304 (2012). https://doi.org/10.1038/ncb2629

50 Israelsen, W. J. & Vander Heiden, M. G. Pyruvate kinase: Function, regulation and role in cancer. *Semin Cell Dev Biol* **43**, 43–51 (2015). https://doi.org/10.1016/j.semcdb.2015.08.004

51 Osato, N. & Hamada, M. Systematic discovery of directional regulatory motifs associated with human insulator sites. *bioRxiv* (2025).

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

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

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

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

56 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

57 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

58 Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013). https://doi.org/10.1186/1471-2105-14-7

59 Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nature methods* **14**, 1083–1086 (2017). https://doi.org/10.1038/nmeth.4463

60 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

61 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

62 Kejiou, N. S. *et al.* Pyruvate Kinase M (PKM) binds ribosomes in a poly-ADP ribosylation dependent manner to induce translational stalling. *Nucleic acids research* **51**, 6461–6478 (2023). https://doi.org/10.1093/nar/gkad440

63 Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012). https://doi.org/10.1016/j.cell.2012.04.031

64 Toller-Kawahisa, J. E. *et al.* Metabolic reprogramming of macrophages by PKM2 promotes IL-10 production via adenosine. *Cell Rep* **44**, 115172 (2025). https://doi.org/10.1016/j.celrep.2024.115172

65 Yang, P. *et al.* Pyruvate kinase M2 accelerates pro-inflammatory cytokine secretion and

cell proliferation induced by lipopolysaccharide in colorectal cancer. *Cell Signal* **27**, 1525–1532 (2015). https://doi.org/10.1016/j.cellsig.2015.02.032

66 Liu, Z., Le, Y., Chen, H., Zhu, J. & Lu, D. Role of PKM2-Mediated Immunometabolic Reprogramming on Development of Cytokine Storm. *Front Immunol* **12**, 748573 (2021). https://doi.org/10.3389/fimmu.2021.748573

67 Kachel, P. *et al.* Phosphorylation of pyruvate kinase M2 and lactate dehydrogenase A by fibroblast growth factor receptor 1 in benign and malignant thyroid tissue. *BMC Cancer* **15**, 140 (2015). https://doi.org/10.1186/s12885-015-1135-y

68 Wang, J. *et al.* Down-regulation of NOTCH1 and PKM2 can inhibit the growth and metastasis of colorectal cancer cells. *Am J Transl Res* **14**, 5455–5465 (2022).

69 Dong, T. *et al.* Activation of GPR3-beta-arrestin2-PKM2 pathway in Kupffer cells stimulates glycolysis and inhibits obesity and liver pathogenesis. *Nat Commun* **15**, 807 (2024). https://doi.org/10.1038/s41467-024-45167-5

70 Gao, J., Zhao, Y., Li, T., Gan, X. & Yu, H. The Role of PKM2 in the Regulation of Mitochondrial Function: Focus on Mitochondrial Metabolism, Oxidative Stress, Dynamic, and Apoptosis. PKM2 in Mitochondrial Function. *Oxid Med Cell Longev* **2022**, 7702681 (2022). https://doi.org/10.1155/2022/7702681

71 Liu, Z., Le, Y. & Lu, D. PKM2: A novel helmsman of lipid metabolism. *Cell Signal* **134**, 111967 (2025). https://doi.org/10.1016/j.cellsig.2025.111967

72 Lee, K. C. Y. *et al.* PKM2 is a key regulator of cardiac lipid metabolism in mice. *Mitochondrion* **85**, 102070 (2025). https://doi.org/10.1016/j.mito.2025.102070

73 Li, L., Zhang, Y., Qiao, J., Yang, J. J. & Liu, Z. R. Pyruvate kinase M2 in blood circulation facilitates tumor growth by promoting angiogenesis. *J Biol Chem* **289**, 25812–25821 (2014). https://doi.org/10.1074/jbc.M114.576934

74 Gomez-Escudero, J. *et al.* PKM2 regulates endothelial cell junction dynamics and angiogenesis via ATP production. *Sci Rep* **9**, 15022 (2019). https://doi.org/10.1038/s41598-019-50866-x

75 He, D. *et al.* Methionine oxidation activates pyruvate kinase M2 to promote pancreatic cancer metastasis. *Molecular cell* **82**, 3045–3060 e3011 (2022). https://doi.org/10.1016/j.molcel.2022.06.005

76 Zhang, X., Lei, Y., Zhou, H., Liu, H. & Xu, P. The Role of PKM2 in Multiple Signaling Pathways Related to Neurological Diseases. *Mol Neurobiol* **61**, 5002–5026 (2024). https://doi.org/10.1007/s12035-023-03901-y

77 Venkatesh, H. S. *et al.* Electrical and synaptic integration of glioma into neural circuits. *Nature* **573**, 539–545 (2019). https://doi.org/10.1038/s41586-019-1563-y

78 Yu, L., Chen, X., Sun, X., Wang, L. & Chen, S. The Glycolytic Switch in Tumors: How

Many Players Are Involved? *J Cancer* **8**, 3430–3440 (2017). https://doi.org/10.7150/jca.21125

79 Gallardo, M. *et al.* hnRNP K Is a Haploinsufficient Tumor Suppressor that Regulates Proliferation and Differentiation Programs in Hematologic Malignancies. *Cancer Cell* **28**, 486–499 (2015). https://doi.org/10.1016/j.ccell.2015.09.001

80 Li, J. *et al.* HNRNPK maintains epidermal progenitor function through transcription of proliferation genes and degrading differentiation promoting mRNAs. *Nat Commun* **10**, 4198 (2019). https://doi.org/10.1038/s41467-019-12238-x

81 Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature reviews. Genetics* **13**, 720–731 (2012). https://doi.org/10.1038/nrg3293

82 Abuhashem, A., Garg, V. & Hadjantonakis, A. K. RNA polymerase II pausing in development: orchestrating transcription. *Open Biol* **12**, 210220 (2022). https://doi.org/10.1098/rsob.210220

83 Dang, H. *et al.* Oncogenic Activation of the RNA Binding Protein NELFE and MYC Signaling in Hepatocellular Carcinoma. *Cancer Cell* **32**, 101–114 e108 (2017). https://doi.org/10.1016/j.ccell.2017.06.002

84 Liu, X. *et al.* Dynamic Change of Transcription Pausing through Modulating NELF Protein Stability Regulates Granulocytic Differentiation. *Blood Adv* **1**, 1358–1367 (2017). https://doi.org/10.1182/bloodadvances.2017008383

85 Robinson, D. C. L. *et al.* Negative elongation factor regulates muscle progenitor expansion for efficient myofiber repair and stem cell pool repopulation. *Dev Cell* **56**, 1014–1029 e1017 (2021). https://doi.org/10.1016/j.devcel.2021.02.025