Title

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

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

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

E-mail: naokiosato11@gmail.com

*Corresponding author.

Abstract

Nucleic acid-binding proteins (NABPs) exhibit cell type–specific regulatory functions, but their target genes and biological roles remain incompletely characterized due to the limitations of current experimental approaches. Here, we present a deep learning framework that integrates gene co-expression correlations to predict NABP regulatory targets and infer their functions across diverse cellular contexts, without requiring binding site or motif information. Substituting low-informative input features with co-expression-derived interactions improved gene expression prediction accuracy. Predicted targets showed strong concordance with ChIP-seq and eCLIP binding sites, and this agreement was significantly greater than for randomly selected gene sets. Functional enrichment and ChatGPT-assisted inference revealed biologically meaningful annotations, including cell type–specific functions such as circadian regulation by AKAP8 in cancer cells and glycolytic control by PKM. Collectively, this integrative framework—combining deep learning, co-expression networks, and large language models—enables the systematic discovery of both known and previously uncharacterized NABP functions in a cell type–specific manner.

Keywords

Computational biology, Deep learning, Gene expression, Gene regulation, Protein function prediction

Introduction

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

Deep learning has substantially improved the prediction of gene expression by modeling regulatory inputs from NABPs, including transcription factors and microRNAs (miRNAs) [20]. For example, models incorporating ChIP-seq–derived NABP binding sites in promoter regions and predicted miRNA-binding sites in transcripts have achieved a high correlation (r = 0.82) between predicted and observed expression levels. In a previous study, we extended this approach by incorporating binding site data for DNA-

binding proteins within 1 Mb of transcription start sites (TSSs) in open chromatin, while excluding miRNAs and RBPs. To facilitate accurate computation of importance scores (also known as contribution scores) and to improve model interpretability, we unified the input representation by excluding a previous input layer for RNA-binding proteins and miRNAs [21]. While this modification slightly reduced predictive accuracy (r = 0.70), it enabled broader coverage of distal regulatory elements and laid the foundation for downstream functional analysis based on importance scores. These scores are later leveraged to identify regulatory targets and infer NABP functions.

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

To test the effectiveness of this approach, we substituted the binding site features of ~300 NABPs that contributed little to previous deep learning predictions with gene co-expression–based interactions. The modified model successfully predicted gene expression levels while assigning DeepLIFT-based importance scores to NABP–gene interactions. Many predicted targets overlapped with ChIP-seq or eCLIP-validated binding partners and showed higher importance scores than background interactions. Functional enrichment analyses of ten representative NABPs revealed consistent

correspondence with known annotations in UniProt [31], and several additional functions were supported by independent publications. To enhance interpretability, we also applied a ChatGPT-based method to infer gene functions and retrieve supporting literature. The results were consistent with enrichment-based findings and highlighted biologically meaningful associations such as circadian rhythm regulation (AKAP8) and glycolysis (PKM).

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

Results

Gene co-expression features enhance expression prediction accuracy

Replacing low-contribution DBP inputs with gene co-expression data led to a substantial increase in model performance (**Fig. 1**). In human foreskin fibroblasts (HFFs), substituting 302 out of 1,310 DBPs with co-expression profiles improved the correlation between predicted and actual gene expression from 0.70 to 0.80 (**Fig. 2a, b**). These co-expression inputs included both NABP-encoding and non-NABP mRNAs, with gene symbols for DNA- and RNA-binding proteins obtained from the ENPD (Eukaryotic Nucleic acid Binding Protein) database[32], and co-expression profiles sourced from COXPRESdb [22]. A subsequent experiment replaced 325 low-contribution DBPs with

108 co-expression interactions involving RBPs, resulting in a further increase in prediction
109 accuracy (r = 0.81; **Fig. 2c**). Collectively, these findings demonstrate that integrating co-
110 expression interactions into the model enhances its ability to capture gene regulatory
111 dynamics beyond direct binding site information.

112 Contribution scores calculated from co-expression–based NABP–gene interactions
113 showed increased variability and notably higher mean and median values than those of
114 the original feature set (**Fig. 2d**, **Supplementary Table S1**). Correlation analysis further
115 revealed consistent relationships between these contribution scores and $\log_2$-transformed
116 transcript expression levels across cell types (**Supplementary Fig. S2**). Example plots
117 demonstrated that the model adjusted its predictions in a cell type–specific manner, with
118 contribution scores reflecting the strength of inferred NABP regulatory influence **(Fig.**
119 **2e**).

120

121 Experimental binding data support model-inferred regulatory targets

122 Experimentally validated binding sites are consistent with deep learning–derived
123 regulatory targets. Genes with ChIP-seq or eCLIP peaks for a given NABP exhibited
124 disproportionately high or low importance ranks, indicating that experimentally
125 confirmed binding sites are associated with stronger regulatory influence in gene
126 expression prediction models. This relationship supports the biological validity of
127 DeepLIFT-based contribution scores, which quantify the importance of specific NABPs
128 in driving gene expression changes.

129 To evaluate this association, we analyzed ChIP-seq data for 32 DBPs in HepG2
130 cells and eCLIP data for 29 RBPs in K562 cells—two cell lines with the most
131 comprehensive publicly available NABP binding site datasets from ENCODE and related

132 studies. A gene was considered associated with an NABP if a ChIP-seq peak was located
133 within ±5 kb of its transcription start site (TSS) or if an eCLIP peak was mapped to its
134 transcribed region. As illustrated in **Fig. 3a** and **Supplementary Fig. S1**, NABPs with
135 experimentally validated binding sites consistently showed higher absolute contribution
136 score values across their predicted target genes, reinforcing their regulatory relevance and
137 corresponding with prior interpretations of DeepLIFT scores [20,33].

138 Model-derived importance scores reliably predict DBP binding sites, as evidenced
139 by significant overlap with ChIP-seq peaks (**Fig. 3b, Supplementary Table S2** and **S3**).
140 To evaluate this correspondence, we analyzed ChIP-seq data for 32 DNA-binding
141 proteins (DBPs) and quantified how frequently high-scoring gene regions—defined by
142 DeepLIFT contribution scores—coincided with experimental binding evidence. Overlap
143 ratios were calculated as the proportion of high-contribution regions that intersected with
144 ChIP-seq peaks, considering both promoter-proximal (±5 kb) and distal (±100 kb) regions
145 around transcription start sites (TSSs). To account for the functional heterogeneity and
146 abundance of ChIP-seq peaks, we applied a precision-at-K approach, focusing on the top-
147 ranked or top-scoring gene regions most likely to reflect regulatory relevance. Four
148 thresholds were used to define high-contribution regions: two based on contribution score
149 rank (top/bottom 20 or 50) and two based on absolute score magnitude (≥ 0.01 or $\leq -$
150 0.01; ≥ 0.001 or ≤ -0.001). A control analysis using randomly selected gene regions of
151 matched size was conducted in parallel. In all tested conditions, the overlap between
152 ChIP-seq peaks and high-contribution regions exceeded that of the random control,
153 supporting the model's ability to prioritize functionally relevant regulatory sites.

154 Similarly, as illustrated in **Fig. 3c** and **Supplementary Table S4**, the model-
155 derived importance scores consistently prioritize biologically relevant RBP–target

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

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

Model-predicted regulatory targets of nucleic acid-binding proteins (NABPs) display consistent functional enrichment patterns, highlighting the biological specificity of the deep learning approach (**Table 1, 2** and **Supplementary Note**). To evaluate the functional consistency of these predicted targets, we conducted Gene Ontology (GO) enrichment analysis on ten representative NABPs [34,35]. The selected gene sets demonstrated statistically significant enrichment for functionally related GO terms, in contrast to randomly sampled gene sets of equivalent size, which did not exhibit such enrichment (**Fig. 4**). Importantly, comparison with all co-expressed genes (Pearson's *z*-score > 2) revealed that the model-selected subsets contained enriched annotations not statistically detectable in the broader co-expression profiles. This suggests that the deep learning model filters co-expression signals to identify a biologically focused set of target genes. Functional enrichment was confirmed using the PANTHER Overrepresentation Test across multiple annotation resources, including GO Slim, Reactome, and PANTHER

180 Pathways (**Table 1**) [36]. To complement these results, we applied a ChatGPT-based gene
181 set interpretation method (**Table 2** and **Supplementary Note**) [37], which inferred
182 functions and retrieved literature consistent with the GO-based findings and UniProt
183 annotations [31]. Notably, AKAP8 was linked not only to its known role in the mitotic cell
184 cycle (UniProt) but also to circadian clock regulation, supported by literature associating
185 AKAP8 with molecular clock genes and melatonin biosynthesis (**Fig. 5** and **6**;
186 **Supplementary Table S5** and **Supplementary Note**) [38]. For this analysis, AKAP8's top
187 50 predicted targets were defined by ranking all genes according to their model-derived
188 NABP contribution scores, and selecting those for which AKAP8 received the highest
189 score in K562 cells. These results demonstrate that integrating co-expression–guided
190 deep learning with large language model–based interpretation yields robust insights into
191 both canonical and previously unrecognized biological functions of NABPs.

192 Deep learning–based refinement of NABP–target predictions enhances the
193 detection of biologically specific functional annotations. To evaluate how training
194 influences functional specificity, we compared GO term enrichment between co-
195 expressed gene sets and predicted regulatory targets for ten NABPs. Although the co-
196 expressed gene sets were larger and exhibited broader enrichment, several functionally
197 informative annotations were uniquely or more significantly enriched among the deep
198 learning–derived target genes (**Table 3**, **Supplementary Table S5** and **S6**). For instance,
199 circadian rhythm–related functions linked to AKAP8 and kinase- or glycolysis-related
200 annotations associated with PKM were observed only following model training (**Table 3**
201 and **Supplementary Table S5**) [39-42]. For PKM, we defined 332 predicted targets by
202 ranking all genes according to NABP contribution scores and selecting those for which
203 PKM ranked within the top two contributors. These results suggest that the model

204 captures biologically meaningful signals beyond simple expression correlations.
205 Importantly, both PKM and AKAP8 are RNA-binding proteins whose functional
206 predictions were derived from a deep learning model trained on gene expression data
207 from K562 cells—a human erythroleukemia cell line widely used in leukemia research.
208 This contextual consistency provides biological credibility to the inferred functions.
209 ALDOA, a glycolysis-associated partner of PKM, and AKAP8, linked to circadian gene
210 networks, have both been implicated in leukemia metabolism and clock regulation,
211 underscoring the relevance of these functions to K562 biology [38,43-46]. Furthermore, the
212 ChatGPT-based functional inference retrieved consistent biological themes—"circadian
213 clock" for AKAP8 and "glycolysis" for PKM—reinforcing the deep learning–based
214 predictions through independent, literature-derived validation (**Fig. 5** and **6**) [37].

215 In neuronal cells, however, AKAP8 exhibited minimal predicted regulatory activity,
216 in contrast to its circadian-linked role in cancer. The model identified very few AKAP8
217 targets in the neuronal context, which limited downstream functional annotation by both
218 PANTHER and ChatGPT-based interpretation. This was based on a contribution score–
219 based ranking across all genes, where only four genes had AKAP8 ranked as the top
220 contributor, and two genes had AKAP8 ranked among the six lowest-scoring NABPs.
221 These findings suggest that AKAP8 may have limited transcriptional influence in neurons,
222 reinforcing its potential role as a context-dependent regulator.

223 PKM exhibits distinct regulatory profiles in cancer and neuronal cells, reflecting
224 divergent metabolic adaptations. In K562 cancer cells, PKM is associated with a
225 Warburg-like metabolic reprogramming, as indicated by the significant enrichment of
226 aldolase activity among its predicted targets, notably including ALDOC and TALDO1
227 (**Supplementary Table S5**) [47,48]. TALDO1, a key enzyme in the non-oxidative branch of

228 the pentose phosphate pathway, contributes to nucleotide biosynthesis and redox balance,
229 supporting the proliferative and survival needs of cancer cells. This metabolic shift is
230 correlates with the activation of the hypoxia-inducible factor (HIF) pathway, involving
231 HIF1A, VEGFA, BNIP3, and DDIT4, which collectively enhance glycolysis,
232 angiogenesis, and adaptation to hypoxic conditions (**Supplementary Note**) [49,50]. In
233 contrast, in neuronal cells, ALDOC—but not TALDO1 or ALDOA—was among the
234 predicted PKM-regulated genes, and aldolase activity was not significantly enriched
235 (**Supplementary Table S5**) [51]. For this analysis, 326 predicted targets were selected
236 based on NABP contribution scores, where PKM ranked among the top two contributors
237 per gene. This suggests that neuronal PKM (predominantly PKM1) maintains a stable
238 glycolytic flux to support basal ATP production, rather than engaging in metabolic
239 reprogramming [52,53]. Furthermore, although several HIF-responsive genes (e.g., EGLN1,
240 BNIP3, MIR210HG) were among the predicted targets, HIF1A itself was not, indicating
241 a context-dependent modulation of the HIF axis in neurons. Collectively, these findings
242 suggest that PKM is repurposed across cell types, integrating core metabolic programs
243 with distinct transcriptional partners tailored to tissue-specific functional demands.
244 Finally, the integration of deep learning-based co-expression modeling with AI-
245 assisted literature mining enabled functional predictions even for less thoroughly
246 characterized regulators such as AKAP8. This framework supports both hypothesis
247 generation and validation across cell types, and highlights the utility of scalable,
248 interpretable tools for deciphering context-specific gene regulatory logic.
249
250
251

Discussion

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

Gene expression correlations alone can serve as a powerful basis for uncovering the regulatory roles of nucleic acid-binding proteins (NABPs), even in the absence of experimental binding data. Leveraging transcriptomic profiles from diverse public RNA-seq datasets, our deep learning framework inferred regulatory targets of NABPs based solely on co-expression patterns. Unlike binding site assays such as ChIP-seq or eCLIP—which are technically demanding and limited by the availability of high-quality antibodies—this computational approach is scalable and applicable across NABPs with poorly characterized binding motifs. Functional enrichment analysis of the predicted target genes revealed biologically meaningful roles, many of which were in agreement with existing annotations in public databases. Additionally, several novel functional associations—not found in UniProt or GO but supported by published literature—were uncovered. These findings were further corroborated using a ChatGPT-assisted literature mining method, which retrieved consistent biological themes and references. Together, these results underscore the biological relevance of NABP–target predictions based on gene expression correlations and demonstrate the value of integrating deep learning with

275 AI-assisted interpretation. This strategy offers a scalable path toward elucidating NABP
276 functions and discovering new regulatory mechanisms across diverse cellular contexts.
277 In cancer cells, AKAP8 appears to serve as a molecular integrator of transcriptional
278 regulation under environmental stress. Our analyses, combined with recent reports,
279 suggest that AKAP8 may support hypoxic adaptation through its interaction with DDX5,
280 a multifunctional RNA helicase known to promote HIF-1 transcriptional activity (**Fig. 6**)
281 [54]. DDX5 enhances the recruitment of HIF-1α/β heterodimers to hypoxia-responsive
282 elements (HREs), thus facilitating the expression of genes involved in glycolysis,
283 angiogenesis, and cell survival under low oxygen conditions [55]. Notably, AKAP8 has
284 been shown to physically interact with DDX5 in lung carcinoma cells, forming a
285 chromatin-associated complex that modulates R-loop stability and transcriptional
286 dynamics [56]. Given that DDX5 levels decrease under hypoxia, yet their re-expression
287 restores aberrant R-loop accumulation, AKAP8 may function as a stabilizing co-factor
288 that compensates for transcriptional stress in hypoxic tumor microenvironments. In
289 addition to hypoxia-associated functions, AKAP8 is also implicated in circadian rhythm
290 regulation. In our study, AKAP8 was identified as a potential regulator of circadian genes
291 in K562 cancer cells, and ChatGPT-based inference retrieved functional annotations such
292 as "circadian clock" and "rhythm." Prior studies have linked the disruption of circadian
293 timing systems to tumor progression [57], suggesting that AKAP8 may contribute to
294 temporal dysregulation in cancer. Thus, AKAP8 may play dual roles by coupling
295 chromatin-associated RNA processing to both metabolic and temporal regulatory circuits
296 in malignancies. Taken together, these findings position AKAP8 as a context-dependent
297 regulatory hub that coordinates hypoxia response and circadian control through
298 interaction with DDX5 and possibly other nuclear partners. Targeting such

299 multifunctional scaffolds may offer new strategies for modulating tumor cell plasticity
300 and transcriptional resilience in the face of environmental stress.

301 A key feature distinguishing cancer cells from non-proliferative cells such as
302 neurons is their ability to reprogram core metabolic pathways to support growth under
303 hypoxic and nutrient-limited conditions. Our analysis supports the presence of a PKM2–
304 HIF-1−ALDOC/TALDO1 regulatory axis in cancer cells (**Fig. 6**) [58-61], which coordinates
305 glycolytic and pentose phosphate pathway gene expression to enable metabolic flexibility.
306 Notably, although ALDOA is a well-established glycolytic enzyme and a known PKM
307 interactor in human and mouse (STRING database), it was not among the PKM co-
308 expressed genes in our K562 transcriptomic data (Pearson's z-score < 2), whereas
309 ALDOC—typically regarded as a brain-specific isoform—emerged as a strongly co-
310 expressed and PKM-regulated gene. This suggests that ALDOC may functionally
311 substitute for or complement ALDOA in hematopoietic contexts. In agreement with this,
312 transcriptomic profiles from the Human Protein Atlas indicate that both ALDOA and
313 ALDOC are expressed in K562 cells, raising the possibility that PKM2 may preferentially
314 regulate ALDOC under hypoxia, while ALDOA may operate independently or through
315 non-PKM-mediated control. Functionally, ALDOC contributes to anaerobic glycolysis
316 and has been implicated in chemoresistance and tumor cell survival, while TALDO1, a
317 key enzyme in the non-oxidative branch of the pentose phosphate pathway, facilitates
318 carbon flux recycling and supports NADPH and nucleotide biosynthesis—functions
319 essential for redox balance and proliferation under stress. These roles are consistent with
320 their upregulation in various cancer models under hypoxic conditions. In contrast,
321 neuronal cells predominantly express the PKM1 isoform, which sustains a high-
322 efficiency glycolytic flux coupled to mitochondrial oxidative phosphorylation. In our

323 neuronal cell model, ALDOC—but not ALDOA or TALDO1—was predicted as a PKM-
324 regulated gene, and no significant enrichment for aldolase activity was detected. This
325 underscores the metabolic priorities of neurons, which emphasize consistent ATP supply
326 over biosynthetic plasticity, and exhibit minimal reliance on the HIF-1 pathway. The
327 absence of HIF1A among PKM targets in neurons further supports the notion of a
328 transcriptionally uncoupled hypoxic response in these cells. Together, these findings
329 reveal how PKM isoform usage and gene network context shape distinct metabolic wiring
330 across cell types. In cancer, PKM2 engages in metabolic reprogramming through the
331 selective regulation of genes like ALDOC and TALDO1 to support hypoxia-adaptive and
332 proliferative states. In neurons, PKM1 maintains homeostatic energy production under
333 physiological conditions. The selective engagement of ALDOC in both contexts—but
334 through divergent regulatory logic—highlights a potential metabolic node of interest,
335 while the lack of ALDOA co-expression with PKM suggests a context-dependent
336 separation of biochemical and transcriptional control. These insights may inform future
337 strategies to exploit tumor-selective metabolic dependencies.

338 Contribution scores derived from gene expression correlations capture both direct
339 and indirect regulatory interactions, offering insights beyond what ChIP-seq alone can
340 provide. While ChIP-seq detects thousands of DNA-binding events per protein, many of
341 these sites do not contribute directly to gene expression regulation and may reflect the
342 multifunctionality of transcription factors [62,63]. Consequently, ChIP-seq–derived binding
343 maps and DeepLIFT-based predictions show only partial overlap, underscoring both
344 methodological differences and the inherent complexity of regulatory networks. In this
345 study, gene co-expression analysis revealed fewer target genes than ChIP-seq but
346 identified subsets that are likely to be more functionally relevant. Our previous work on

347 insulator-associated factors linked such proteins to enhancer–promoter insulation and
348 chromatin organization, including phase separation and loop formation [21]. That study
349 incorporated eQTL data to anchor binding sites to expression outcomes. Here, by contrast,
350 we focused on NABPs with unknown motifs and excluded motif-based or chromatin
351 accessibility data such as DNase-seq or ATAC-seq, to assess the capacity of gene
352 expression correlations to reveal regulatory roles. Notably, the distribution of
353 contribution scores for co-expression–based inputs included both high and low values,
354 implying that gene expression patterns reflect not only direct binding but also indirect
355 regulation mediated through broader network effects. This highlights the value of
356 contribution scores as indicators of both proximal and distal regulatory influences, even
357 when binding motifs are unknown or experimental binding data are unavailable.

358 Functional roles of NABPs can be inferred even from small gene sets, provided
359 that the target genes exhibit shared or functionally related patterns. In this study, ten
360 NABPs showed functional enrichments consistent with known annotations, supporting
361 the validity of the deep learning–based predictions. These enrichments were identified
362 using Fisher's exact test with false discovery rate (FDR) correction via the PANTHER
363 database, which typically requires at least 50 genes for robust statistical significance. For
364 proteins with fewer than 50 predicted targets, enrichment may be harder to detect
365 depending on the functional composition of the gene set. Relaxing the ranking thresholds
366 increased the number of target genes and enhanced statistical power. Alternatively, more
367 stringent thresholds could be retained if the contribution scores were sufficiently strong,
368 offering flexibility depending on analysis goals. Additional features—such as gene
369 expression levels or co-expression correlations—could further refine enrichment analyses,
370 particularly for borderline cases. Notably, enrichment analysis of both high- and low-

371 ranking targets revealed candidate novel functions for specific NABPs, suggesting that
372 genes with large-magnitude contribution scores are biologically informative. Moreover,
373 a complementary ChatGPT-based approach demonstrated utility in interpreting small
374 gene sets. Unlike standard enrichment tools, this method can infer biological roles even
375 when gene counts fall below conventional significance thresholds, provided that genes
376 with related functions are present. This capacity helps overcome statistical limitations and
377 broadens the interpretability of regulatory predictions in a cell type–specific context.

378 Strategic use of co-expression data enhances functional inference, while
379 indiscriminate replacement can diminish biological relevance. The selective integration
380 of gene co-expression data improved target prediction and the functional interpretability
381 of the model. In contrast, replacing all input features with co-expression-derived
382 interactions led to a decrease in enriched annotations, and known functional roles were
383 no longer recovered. Excessive replacement of input features with gene co-expression
384 data further diminished the biological relevance of model predictions. When all NABP–
385 target inputs were substituted with co-expression-based interactions, the number of
386 enriched functional annotations declined, and previously identified functions from the
387 UniProt database were no longer recovered. This suggests that indiscriminate
388 replacement may cause the model to capture non-informative or biologically spurious
389 associations. A similar trend was observed in a prior control analysis, where substituting
390 true regulatory targets with randomly selected genes led to a gradual loss of enrichment
391 as randomness increased. Based on these observations, the normalized count of enriched
392 functional terms—adjusted by the number of predicted targets—may serve as a useful
393 proxy for evaluating the accuracy and biological validity of NABP–target predictions [64].
394 These findings also raise intriguing questions about how the deep learning model

395 prioritizes informative input features during training. Future work exploring the model's
396 internal representation dynamics—such as how it distinguishes signal from noise when
397 presented with diverse regulatory interactions—may provide valuable insights into the
398 mechanisms of transcriptional regulation and model interpretability. Investigating how
399 the model converges on biologically meaningful NABP–target relationships during
400 training—possibly through attention mechanisms or layer-wise relevance propagation—
401 could deepen our understanding of both gene regulation and model interpretability.

402 Cell type specificity and cooperative binding dynamics may underlie the absence
403 of detectable regulatory signals for certain NABPs. While some NABPs showed no
404 high- or low-ranking target genes in HepG2 or K562 cells, such absence does not rule out
405 regulatory relevance. These proteins may exhibit cell type–specific regulatory activity
406 detectable only in other cellular contexts, where functional enrichment analysis could
407 reveal distinct biological roles. Moreover, many transcription factors and DNA-binding
408 proteins act in concert, forming multi-protein complexes that bind cooperatively to shared
409 promoters [65-67]. In such cases, individual components may not always yield strong
410 contribution scores due to their distributed regulatory influence across the complex. This
411 highlights the potential of co-binding site analysis as a complementary strategy to uncover
412 cooperative regulatory mechanisms and to disentangle the functional interplay of NABPs
413 within transcriptional networks.

414 This study establishes a scalable computational framework for uncovering the
415 functional roles of nucleic acid-binding proteins (NABPs) using gene expression
416 correlations and deep learning. By leveraging co-expression–based interactions and
417 contribution scoring, the model successfully identified both canonical and novel
418 regulatory associations across cell types—even in the absence of experimentally defined

binding motifs. The results highlight how expression-driven inference, complemented by AI-assisted functional interpretation, can advance our understanding of NABP biology beyond conventional binding site–centered approaches. Although the present analysis focused on a representative subset of proteins and cell types, the framework is readily adaptable to broader transcriptomic datasets and larger NABP cohorts. As such, it offers a path forward for systematic annotation of NABPs, particularly those with poorly characterized regulatory roles. Together, these findings provide a foundation for future studies to decode complex gene regulatory architectures and integrate transcriptomic and AI-driven predictions into functional genomics research at scale.

Methods

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

To predict gene expression across multiple human cell types, we adapted a previously established deep learning framework that integrates DNA-binding site data for transcription factors and other regulatory proteins. In this implementation, regulatory elements located both proximally and distally (up to ±1 Mb) from transcription start sites (TSSs) were included as input features (**Fig. 1**) [20,21]. To minimize the effects of cellular heterogeneity in bulk tissue, RNA-seq datasets were selected at the individual cell-type level. Publicly available RNA-seq data were retrieved for four functionally distinct cell types: human foreskin fibroblasts (HFFs), mammary epithelial cells (HMECs), neural progenitor cells (NPCs), and either HepG2 hepatoma cells or K562 leukemia cells, depending on the class of regulatory protein analyzed [68,69]. Gene expression values

443 (FPKM) were computed using the RSeQC toolkit based on GENCODE v19 annotations,
444 and subsequently $\log_2$-transformed after adding a small offset to avoid undefined values.
445 Following the approach described by Tasaki et al. [20,21], expression fold changes were
446 calculated relative to the median expression level across the four cell types. For each
447 model iteration, three common cell types (HFFs, HMECs, and NPCs) were used, and
448 either HepG2 (for DNA-binding protein analysis) or K562 (for RNA-binding protein
449 analysis) was included to match the availability of ChIP-seq and eCLIP validation
450 datasets. These cell types were chosen to represent diverse biological lineages and
451 regulatory contexts.
452
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 [20,21], this model extends the
458 architecture originally proposed by Tasaki et al. (2020) [20,21], 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.

466 Binding site coordinates were obtained from the GTRD database (v20.06) [70], and
467 regulatory associations were inferred using cis-eQTLs from GTEx v8. A ChIP-seq peak
468 that overlapped a 100-bp window centered on an expression-associated SNP was
469 considered linked to the corresponding transcript. Trans-eQTLs were excluded, and
470 coordinates were converted to the hg19 assembly using UCSC LiftOver [71]. To limit
471 feature dimensionality while maximizing genomic coverage, we computed peak lengths
472 for each DNA-binding protein (DBP) in 100-bp bins across ±1 Mb around the TSS, then
473 retained the maximum peak value within each 10-kb window, yielding up to 200 enhancer
474 bins. These were concatenated with 30 promoter-region bins to form a 230-bin unified
475 input layer. This single-input-layer structure was necessary because DeepLIFT's
476 genomics-default configuration ("RevealCancel rule on dense layers and Rescale rule on
477 conv layers") does not support models with multiple input layers. Concatenating all DBP-
478 related features into one matrix thus maintained compatibility while enhancing
479 interpretability.

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

487

488

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

491 To uncover regulatory targets of NABPs without relying on direct binding site
492 assays, we implemented a co-expression–driven modification of the deep learning input.
493 Specifically, DNA-binding proteins with low contribution scores in the original model
494 were replaced with NABPs whose regulatory interactions were inferred from gene co-
495 expression data. Of the 1,310 DBPs initially included, 302 were substituted with other
496 DBPs, and in a separate RNA-binding protein analysis, 325 were replaced with RNA-
497 binding proteins (RBPs) identified from the ENPD database [32]. Putative NABP–target
498 relationships were established using co-expression profiles provided by COXPRESdb [22],
499 where a Pearson's *z*-score > 2 between a NABP-encoding gene and a non-NABP gene
500 indicated a strong co-expression link. These co-expression pairs served as proxies for
501 regulatory interactions and were used to generate new input features for the model. To
502 integrate these proxies into the deep learning framework, hypothetical NABP binding
503 sites were assigned at the transcription start site (TSS) and 2 kb upstream of each co-
504 expressed gene. The model was then retrained using this modified input, and DeepLIFT
505 was applied to calculate importance scores, evaluating the influence of each NABP–gene
506 interaction on predicted expression levels. This approach enables the functional
507 characterization of NABPs with unknown or uncharacterized binding motifs by
508 leveraging gene co-expression as a proxy for regulatory targeting.
509
510
511

512 **Prioritization of NABP regulatory targets based on importance score**
513 **rankings**

514 To identify the primary regulatory targets of each NABP, genes were ranked
515 based on the importance scores assigned to NABP binding sites by the deep learning
516 model. Genes with binding sites consistently ranked in the top five or bottom five across
517 all NABPs were selected as putative targets, reflecting high positive or negative
518 regulatory influence. These targets represent genes most strongly associated with a given
519 NABP's regulatory activity. To ensure sufficient statistical power for downstream
520 enrichment analysis, rank thresholds were adjusted for each NABP based on the number
521 of high-confidence predictions. While a contribution score rank cutoff of five was
522 sufficient in most cases, the threshold was relaxed in specific cases—for example, up to
523 rank eight for proteins such as RPLP0, which initially yielded a smaller number of
524 predicted targets.

525 Functional enrichment and AI-based interpretation confirm the biological
526 relevance of NABP regulatory targets. To evaluate the biological coherence of predicted
527 NABP regulatory targets, we performed functional enrichment analysis using the
528 Statistical Overrepresentation Test from the PANTHER database alongside custom in-
529 house scripts [36]. Gene Ontology (GO) terms from biological process, molecular function,
530 and cellular component categories were analyzed for enrichment across three gene sets:
531 (1) primary regulatory target genes identified by the model, (2) all co-expressed genes
532 (Pearson's z-score > 2), and (3) randomly selected genes of matched size. Enrichment
533 was assessed using Fisher's exact test with Benjamini–Hochberg false discovery rate
534 (FDR) correction, and FDR < 0.05 was considered statistically significant. For this
535 analysis, human gene symbols were mapped to Ensembl Gene IDs using the HGNC

536 database (https://www.genenames.org/), and GO annotations were obtained from the
537 Gene Ontology resource [34,35]. Target gene IDs served as the test set, while the full
538 expression dataset used in the deep learning model formed the reference background. To
539 complement the statistical analysis, we applied a ChatGPT-based functional inference
540 method [37]. This approach used curated prompts—detailed in *Extended Data Tables 1 and*
541 *5*—to predict functional roles based on gene sets, and retrieved relevant literature using
542 a modified version of the "4.Reference search and validation.ipynb" script [37]. This script
543 is publicly available on Code Ocean and GitHub
544 (https://doi.org/10.24433/CO.7045777.v1 ;
545 https://github.com/idekerlab/llm_evaluation_for_gene_set_interpretation). The
546 ChatGPT-based results provided consistent and interpretable biological themes aligned
547 with the enrichment findings.
548

549 **Validation of predicted NABP targets using experimentally determined**
550 **binding**

551 To assess the biological relevance of the predicted regulatory targets, we
552 compared deep learning–inferred NABP–gene associations with experimentally
553 validated binding data. Specifically, we evaluated the spatial overlap between predicted
554 target genes and ChIP-seq peaks for 32 DNA-binding proteins (within ±5 kb of
555 transcription start sites [TSSs]) as well as eCLIP peaks for 29 RNA-binding proteins
556 (within the transcribed region from TSS to transcription end site [TES]). Gene structure
557 annotations were derived from the GENCODE v19 release [72], and binding site data were
558 obtained from the ENCODE portal and prior publications (**Supplementary Table S7**) [73].
559 This overlap analysis provides orthogonal validation for the predicted targets and

560 supports the utility of gene co-expression–based modeling for inferring NABP–target
561 relationships.
562

563 **Statistical analysis**

564 Functional enrichment was assessed using Fisher's exact test, with multiple testing
565 correction applied via the Benjamini–Hochberg false discovery rate (FDR) procedure.
566 Statistical significance was defined as FDR < 0.05. Distributions of DeepLIFT
567 contribution scores for nucleic acid-binding sites derived from gene co-expression data
568 and ChIP-seq data were compared using the Mann–Whitney U test (also known as the
569 Wilcoxon rank-sum test). A p-value < 0.05 was considered statistically significant.
570

571 **Data availability**

572 RNA-seq data for human foreskin fibroblast (HFF) cells were obtained from the Gene
573 Expression Omnibus (GEO) under accession number GSM941744. RNA-seq data for
574 human mammary epithelial cells (HMEC) were retrieved from ENCODE via the UCSC
575 Genome Browser (http://genome.ucsc.edu/encode/) under the file name
576 *wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam*. RNA-seq data for human neural
577 progenitor cells (NPC) were obtained from GEO under accession number GSM915326.
578 RNA-seq data for HepG2 and K562 cells were obtained from the ENCODE portal under
579 accession numbers ENCFF670LIE and ENCFF584WML, respectively. mRNA co-
580 expression data were downloaded from COXPRESdb v7
581 (https://zenodo.org/record/6861444). The list of nucleic acid-binding proteins was
582 obtained from the Eukaryotic Nucleic acid Binding Protein (ENPD) database
583 (http://qinlab.sls.cuhk.edu.hk/ENPD/). ChIP-seq data for transcription factors were

584 downloaded from the GTRD database (v20.06 release; https://gtrd.biouml.org). eQTL
585 data (GTEx_Analysis_v8_eQTL.tar) were obtained from the GTEx portal (v8 release;
586 https://gtexportal.org). eCLIP data for RNA-binding proteins were retrieved from the
587 ENCODE portal (accession ENCSR456FVU) and from Supplementary Data 2 of Van
588 Nostrand et al., *Nature* (2020), https://doi.org/10.1038/s41586-020-2077-3.
589

590 **Code availability**

591 The deep learning model implemented in this study is available at
592 https://github.com/reposit2/coexgene.
593
594

Figures

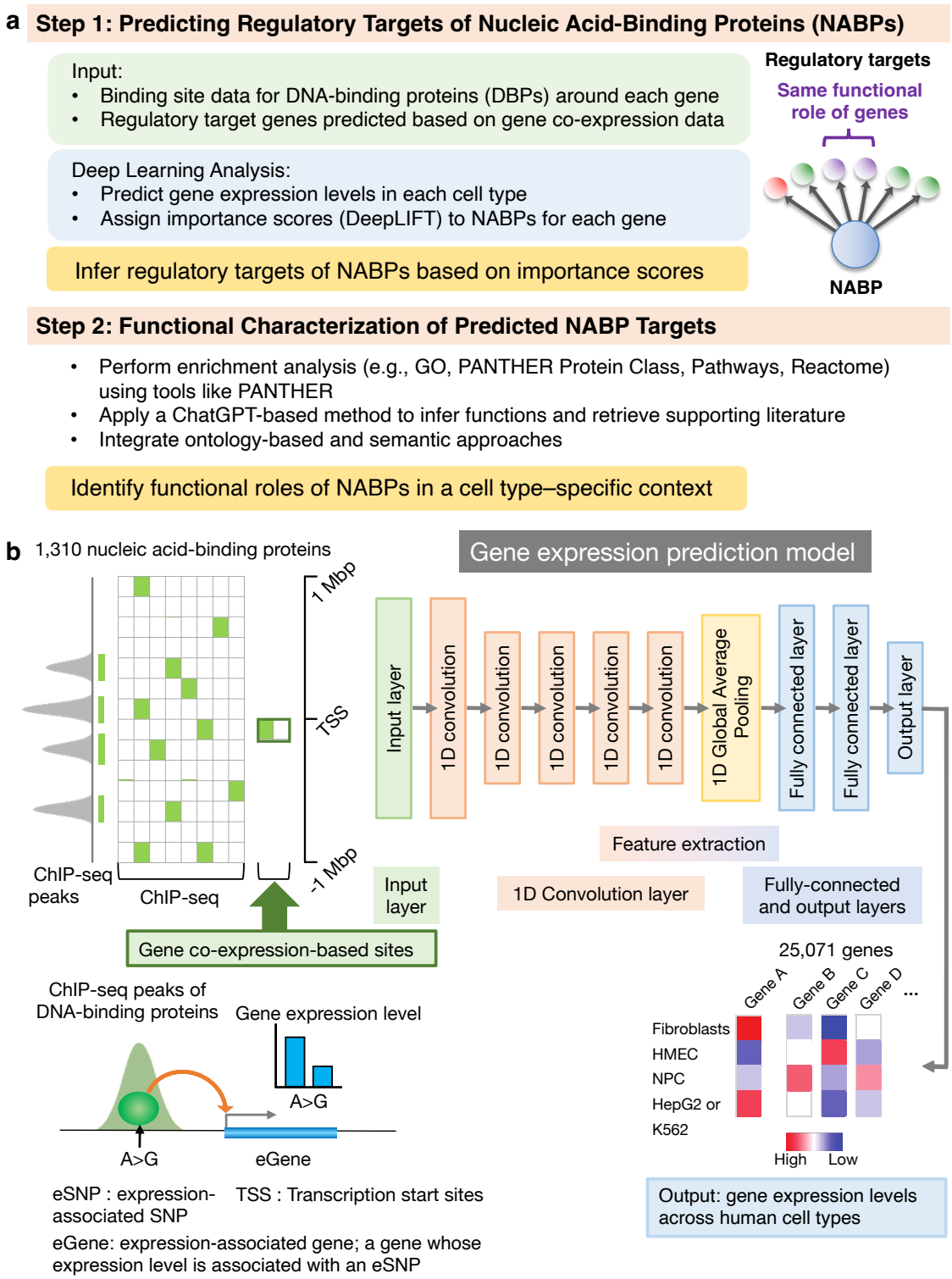


Figure 1. Overview of the deep learning framework for predicting and functionally characterizing NABP regulatory targets.

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 both conventional enrichment analysis and 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.

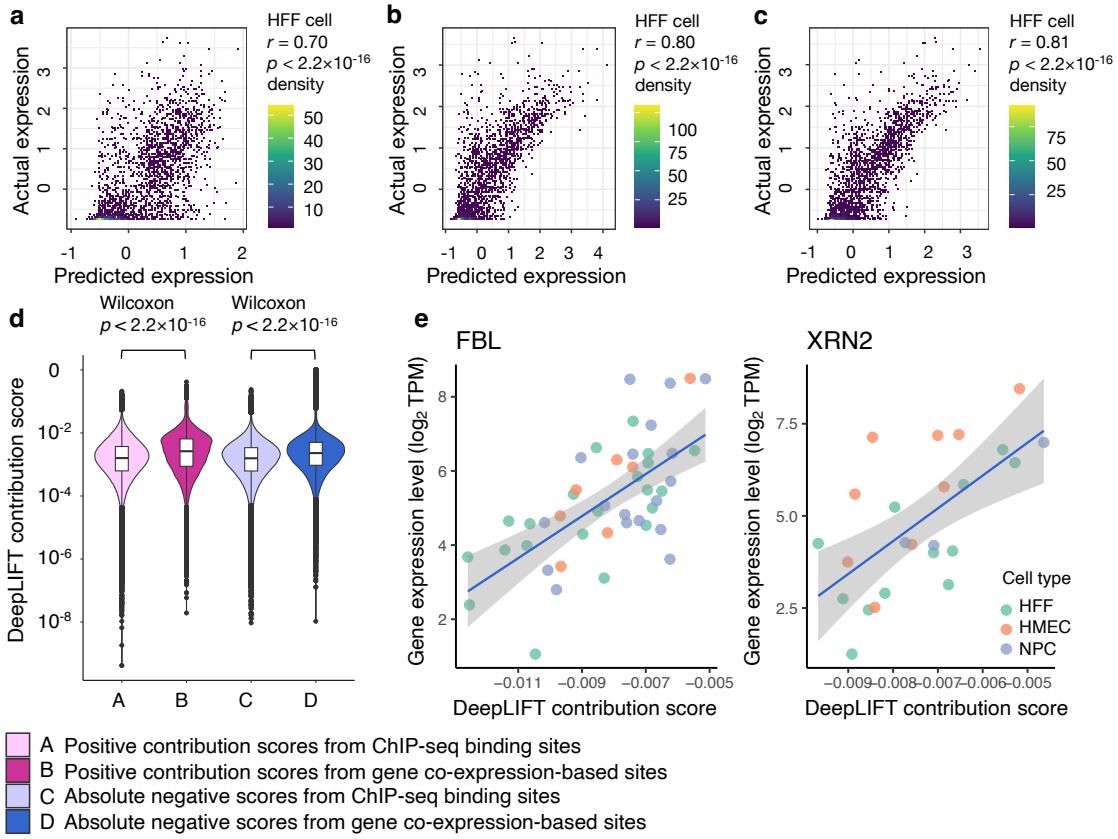


Figure 2. Relationship between gene expression levels and contribution scores. Gene expression levels ($\log_2$ TPM) predicted by a deep learning model were compared with actual expression levels in HFF cells using Spearman's rank correlation.

a. Input features included DNA-binding sites of 1,310 proteins obtained from ChIP-seq data.

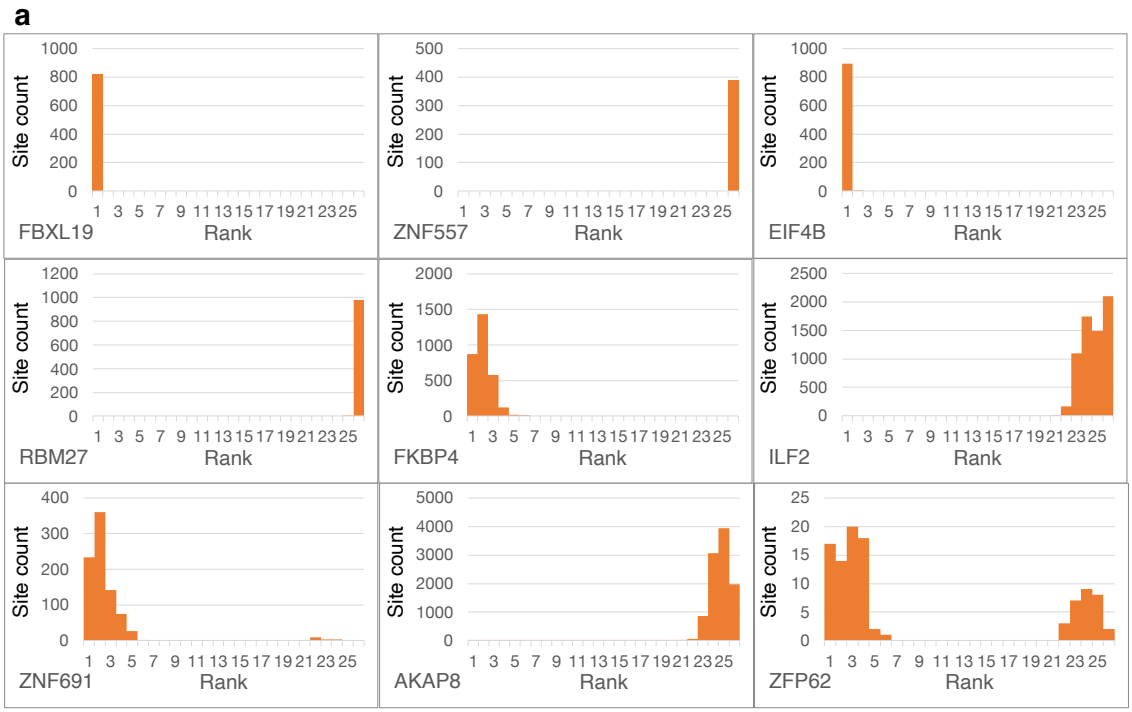
b. Among these, 302 proteins with low contribution scores were replaced by an equal number of DNA-binding proteins identified based on gene co-expression.

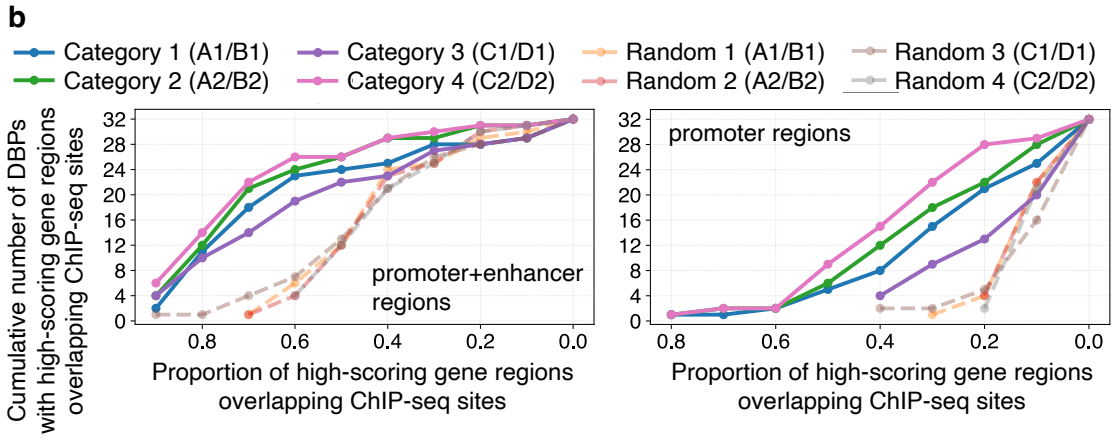
c. Similarly, 325 DNA-binding proteins were replaced with RNA-binding proteins identified from co-expression relationships.

d. DeepLIFT contribution scores were significantly higher for binding sites derived from gene co-expression data compared to those from ChIP-seq data (Wilcoxon signed-rank

test, $p < 2.2 \times 10^{-16}$). Panels a and b show distributions of positive contribution scores, while C and D show absolute values of negative contribution scores.

e. Representative examples illustrating the correlation between a regulator's DeepLIFT contribution score and the expression levels of its predicted target genes. Linear regression lines with 95% confidence intervals are shown.





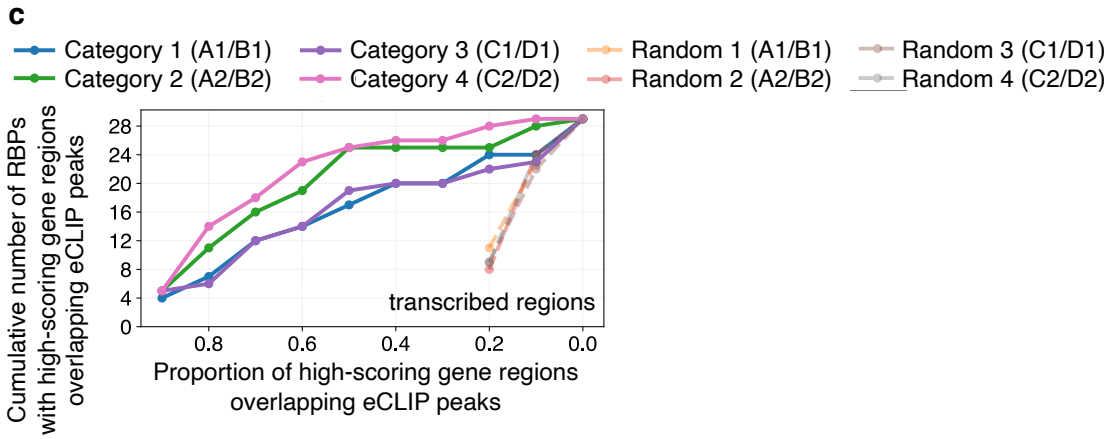


Figure 3. Comparison of predicted regulatory target genes with experimental ChIP-seq and eCLIP binding data. This figure illustrates how predicted gene-level regulatory contributions derived from deep learning correspond to experimentally identified nucleic acid-binding protein targets across genomic regions.

a. Distribution of contribution scores across ChIP-seq– and eCLIP–validated binding sites. The x-axis shows the rank of contribution scores assigned to nucleic acid-binding sites used for gene expression prediction. Scores were ranked from 1 (highest positive contribution, left) to 1,310 (lowest negative contribution, right) and binned into 50-rank intervals. Rank 1 corresponds to ranks 1–50, rank 2 to ranks 51–100, and rank 26 to ranks 1,251–1,310. Nucleic acid-binding sites were categorized as located within promoters or transcribed regions of genes experimentally bound by the nucleic acid-binding protein shown in each panel. These sites exhibited higher or lower contribution scores compared to those across all ChIP-seq and eCLIP binding sites. The y-axis indicates the number of sites in each 50-rank bin. Positive contribution scores indicate gene activation, whereas negative scores indicate gene repression.

b. Proportion of high-contribution gene regions overlapping ChIP-seq binding sites in promoter + enhancer and promoter-only regions, relative to all high-contribution

662 **regions.** Labels indicating "promoter + enhancer regions" and "promoter-only regions"
663 are shown above the corresponding panels. The proportion represents the number of high-
664 contribution regions within promoter + enhancer intervals (±100 kb of the transcription
665 start site [TSS]; left panel) and promoter-only intervals (±5 kb of the TSS; right panel)
666 that overlap with ChIP-seq binding sites, divided by the total number of high-contribution
667 regions in each category. To define high- and low-contribution regions, multiple
668 thresholds were applied to the 1,310 DNA-binding proteins analyzed: **Rank-based**
669 **thresholds** — (A1 and B1): rank ≥ 20 or ≤ 1,290; (A2 and B2): rank ≥ 50 or ≤ 1,260.
670 **Score-based thresholds** — (C1 and D1): contribution score ≥ 0.01 or ≤ -0.01; (C2 and
671 D2): contribution score ≥ 0.001 or ≤ -0.001 (see **Supplementary Note and Tables S2**
672 **and S3**).
673 **c. Proportion of high-contribution gene regions overlapping eCLIP peaks in**
674 **transcribed regions, relative to all high-contribution regions.** The cumulative number
675 of RNA-binding proteins (RBPs) is shown as a function of the proportion of high-
676 contribution gene regions overlap with experimentally determined eCLIP peaks within
677 gene transcribed regions (from transcription start site [TSS] to transcription end site
678 [TES]). High- and low-contribution regions were defined using the same thresholds as in
679 panel (b) applied to the 29 RBPs analyzed in K562 cells. Random controls represent the
680 overlap expected by chance. "Transcribed regions" refer to the full gene body, including
681 both exonic and intronic regions. See Supplementary Note for additional details.
682
683

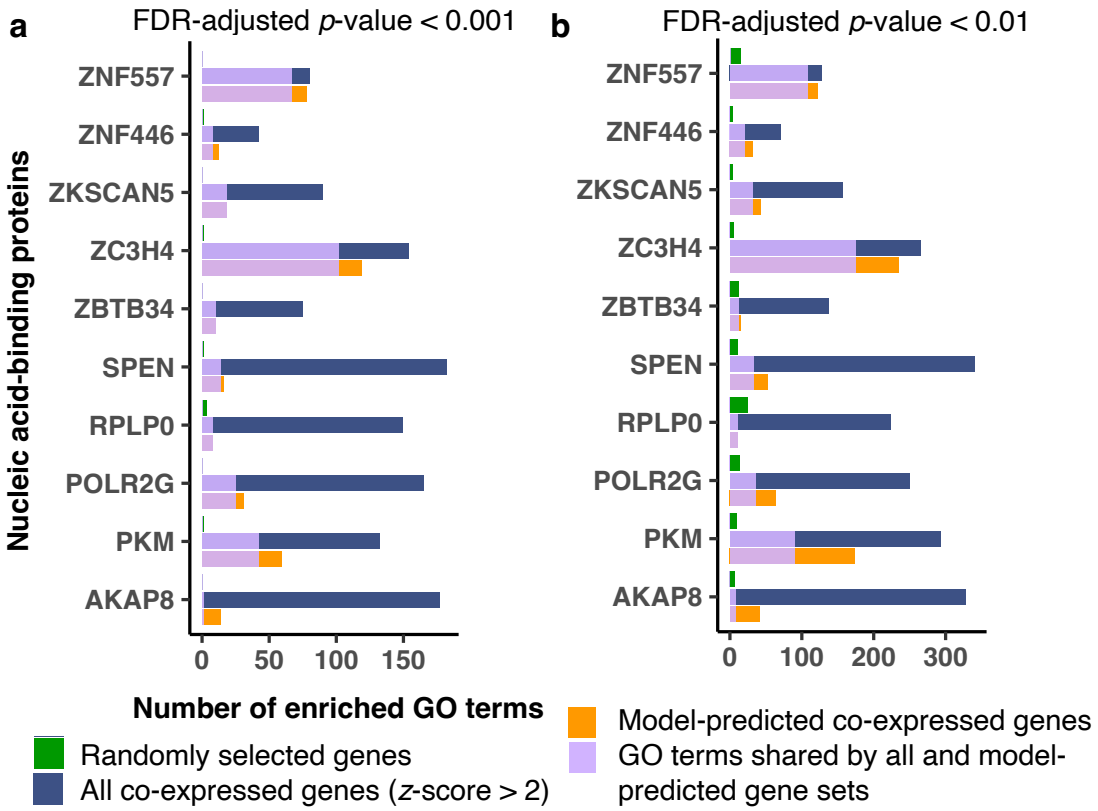


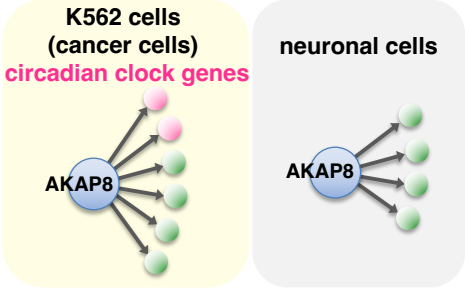
Figure 4. Overrepresented functional annotation terms among target genes of nucleic acid-binding proteins.

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).

AKAP8 (K562 cells)

1. Circadian Control of Metabolic Regulation

This system includes multiple components of the circadian clock (*PER2, PER3, CRY2, NR1D1*), which regulate daily oscillations of gene expression in metabolic tissues. These transcriptional regulators control the rhythmic expression of key metabolic enzymes and transporters, thereby orchestrating the timing of energy storage, mobilization, and hormonal signaling across the light-dark cycle. Their co-presence with nutrient transporters and signaling proteins suggests functional integration in time-dependent metabolic homeostasis.



PKM (K562 cells)

1. Hypoxia Sensing and Adaptive Response

This system prominently features components regulated by or regulating **hypoxia-inducible factors (HIFs).** *HIF1A*, *EGLN1*, *EGLN3*, *MIR210HG*, *VEGFA*, *BNIP3*, and *DDIT4* are core mediators of the transcriptional and metabolic response to low oxygen conditions. *EGLNs* hydroxylate HIF1A under normoxia, while *HIF1A* activates genes that enhance glycolytic flux, angiogenesis (*VEGFA*), and survival during hypoxic stress (*BNIP3*, *MIR210HG*). *DDIT4* connects hypoxia to mTOR inhibition, reducing energy expenditure. These factors establish a functional axis adapting metabolism and signaling to oxygen availability.

2. Metabolic Reprogramming Toward Glycolysis

Glycolytic genes including *PFKP*, *PFKL*, *ENO1*, *ENO2*, *ENO3*, *ALDOC*, *LDHA*, and *HK2* are key effectors of the hypoxia-induced metabolic shift. HIF1A upregulates these enzymes, favoring glycolysis and lactate production over mitochondrial respiration. This shift is supported by regulatory components like *PDK1*, *PDK3*, and *SLC16A3* (lactate transporter), which further suppress pyruvate entry into mitochondria and facilitate acid-base balance under anaerobic metabolism.

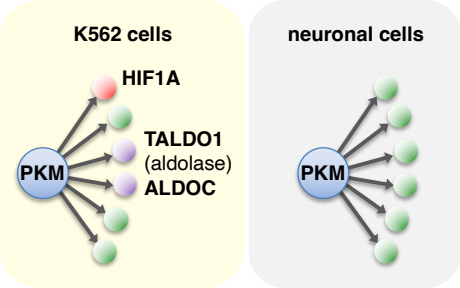


Figure 5. Functional mapping of nucleic acid-binding proteins reveals context-dependent regulatory roles in cancer and neuronal cells.

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. The resulting schematic highlights the distinct functional roles of AKAP8 and PKM in K562 leukemia cells—circadian regulation and metabolic reprogramming under hypoxia, respectively—in contrast to their more limited or divergent activities in neuronal cells.

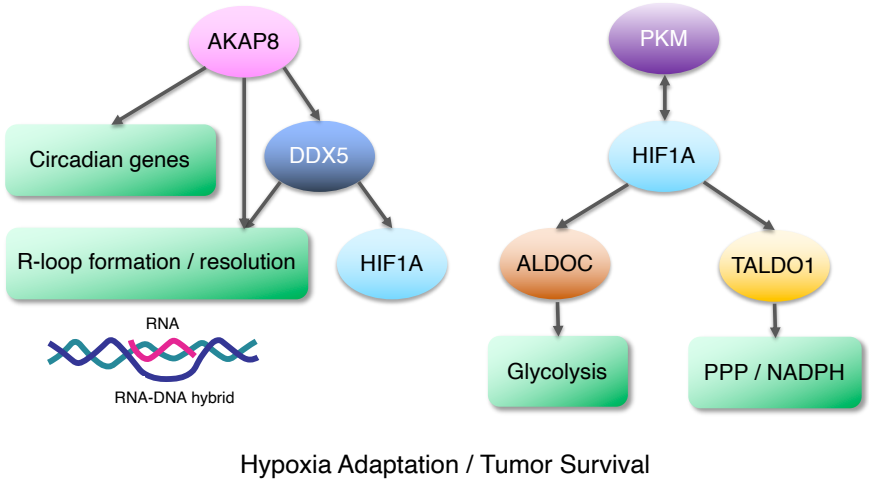


Figure 6. Integrated hypoxia-responsive regulatory network orchestrated by AKAP8 and PKM in cancer cells. This model illustrates how two nucleic acid-binding proteins (NABPs)—AKAP8 and PKM—regulate distinct yet interconnected transcriptional programs to support tumor adaptation under hypoxic stress. AKAP8 acts through DDX5 to influence R-loop resolution [74], circadian gene expression, and the stabilization of HIF1A transcriptional activity. In parallel, PKM modulates HIF1A-driven metabolic reprogramming, promoting glycolysis via ALDOC and redox homeostasis via TALDO1. The predicted regulatory links were inferred from contribution scores calculated by a deep learning framework and are supported by published functional studies in cancer models. PPP: Pentose Phosphate Pathway; NADPH: Nicotinamide Adenine Dinucleotide Phosphate (reduced form).

726 # Tables

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

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

739

AKAP8
negative regulation of RNA biosynthetic process
Circadian clock system
Cry-Per complex
Circadian-regulated energy metabolism and insulin signaling (0.74)

PKM
kinase
glycolytic process
ATP metabolic process
Hypoxia-adaptive metabolic remodeling and cell migration (0.79)

POLR2G
RNA polymerase II transcription regulatory region sequence-specific DNA binding
general transcription factor
translational protein
Ribosome biogenesis and RNA maturation at the interface of translation and nuclear export (0.91)

RPLP0
large ribosomal subunit
ribonucleoprotein complex
cytosolic large ribosomal subunit
Mitochondrial-linked ribosome function and RNA-guided surveillance in translational homeostasis (0.84)

SPEN
mRNA binding
chromatin binding
transcription coregulator activity
Chromatin-associated transcriptional regulation and nuclear RNA export integration (0.83)

ZBTB34
C2H2 zinc finger transcription factor
DNA-binding transcription activator activity, RNA polymerase II-specific
RNA polymerase II transcription regulatory region sequence-specific DNA binding
Transcriptional regulation by C2H2 zinc finger proteins and signaling modulation (0.72)

ZC3H4
chromatin/chromatin-binding, or -regulatory protein
transcription regulator activity
metal ion binding
RNA-binding protein–mediated transcriptional integration and pre-mRNA processing (0.89)

ZKSCAN5
DNA-binding transcription factor activity
C2H2 zinc finger transcription factor
RNA polymerase II cis-regulatory region sequence-specific DNA binding
metal ion binding
RNA processing, ribonucleoprotein assembly, and transcription-coupled quality control (0.83)

ZNF446
C2H2 zinc finger transcription factor
extracellular space
Transcription-coupled RNA processing and mitochondrial RNA metabolism (0.81)

ZNF557
DNA-binding transcription activator activity, RNA polymerase II-specific
RNA polymerase II transcription regulatory region sequence-specific DNA binding
ion binding, metal ion binding
Transcriptional regulation and RNA metabolic coordination via zinc finger proteins and RNA-associated complexes (0.88)

Table 2. Functional characterization of predicted regulatory targets of nucleic acid-binding proteins using a ChatGPT-based inference framework.

743 Representative biological functions were inferred for each NABP by applying a large
744 language model–based gene set interpretation method. Functional themes are
745 summarized in bold for each NABP, while numbered subcategories denote distinct
746 functional roles associated with individual gene subsets.
747

AKAP8 (K562 cells)
Circadian-regulated energy metabolism and insulin signaling (0.74)
1. Circadian Control of Metabolic Regulation
2. Insulin Signaling and Glucose Transport
3. Fatty Acid and Lipid Homeostasis
4. Metabolite Transport and Mitochondrial Function
5. Signal Integration and Vesicle Trafficking
6. Developmental and Structural Modulators

PKM (K562 cells)
Hypoxia-adaptive metabolic remodeling and cell migration (0.79)
1. Hypoxia Sensing and Adaptive Response
2. Metabolic Reprogramming Toward Glycolysis
3. Extracellular Matrix Remodeling and Migration
4. Vesicular and Membrane Trafficking
5. Mitochondrial Function and ROS Homeostasis
6. Transcriptional and Epigenetic Regulation

PKM (neuronal cells)
Hypoxia-driven metabolic and structural reprogramming (0.87)
1.Hypoxia Sensing and Transcriptional Response
2.Glycolytic Shift and Mitochondrial Modulation
3.Cytoskeletal Remodeling and Motility Enhancement
4.Endocytosis, Vesicle Transport, and Membrane Dynamics
5.RNA Processing and Translation Control
6.Ubiquitination and Proteostasis
7.Stress Signaling and Nuclear Regulation

POLR2G (K562 cells)
Ribosome biogenesis and RNA maturation at the interface of translation and nuclear export (0.91)
1. Cytoplasmic and Nuclear Ribosomal Component Assembly
2. Ribosomal RNA Processing and Modification
3. Translational Initiation and Ribosome Recycling
4. Proteasome-Linked Ribosome Quality Control
5. Nuclear Export, mRNP Surveillance, and RNA Chaperoning

RPLP0 (K562 cells)
Mitochondrial-linked ribosome function and RNA-guided surveillance in translational homeostasis (0.84)
1. Ribosomal Structural Components and Translational Core
2. RNA Modification and Surveillance by Small Non-Coding RNAs
3. RNA Surveillance and Nuclear Export Quality Control
4. Mitochondrial and Redox-Linked Translational Interface
5. Membrane Trafficking and Endoplasmic Reticulum Integration
6. Zinc Finger Regulators and DNA Repair Coupling

SPEN (K562 cells)
Chromatin-associated transcriptional regulation and nuclear RNA export integration (0.83)
1. Transcriptional Regulation and Chromatin Remodeling
2. RNA Binding, Processing, and Export
3. Protein Quality Control and Ubiquitin-Mediated Regulation
4. Signal-Responsive Transcription Integration
5. RNA Surveillance and Stability
748 6. Nuclear Structural and Trafficking Components

ZBTB34 (HepG2 cells)
Transcriptional regulation by C2H2 zinc finger proteins and signaling modulation (0.72)
1. Transcriptional Regulation by Zinc Finger Proteins
2. Signaling and Transcriptional Crosstalk
3. Modulators of Chromatin and RNA Stability
4. Functional Integration and Cell Fate Control

ZC3H4 (HepG2 cells)
RNA-binding protein–mediated transcriptional integration and pre-mRNA processing (0.89)
1. RNA-Binding Protein Network and Pre-mRNA Processing
2. Transcription-Coupled Regulatory Scaffold
3. Spliceosome Assembly and Dynamic Coordination
4. mRNA Export and Quality Control
5. Epigenetic and Histone Regulatory Interface
6. RNA Modification and Stability Regulation
7. Nucleolar and Ribosome Biogenesis Linkage

ZKSCAN5 (HepG2 cells)
RNA processing, ribonucleoprotein assembly, and transcription-coupled quality control (0.83)
1. RNA Processing and mRNA Surveillance
2. Transcriptional Regulation and Chromatin Remodeling
3. Ribosome Biogenesis and Mitochondrial Translation
4. Protein Quality Control and Ubiquitin-Dependent Regulation
5. Nucleocytoplasmic Transport and Organelle Coordination
6. ZNF and ZBTB Family-Mediated Regulatory Modules

ZNF446 (HepG2 cells)
Transcription-coupled RNA processing and mitochondrial RNA metabolism (0.81)
1. Transcriptional Regulation and Initiation
2. Splicing and RNA Processing
3. Mitochondrial RNA Metabolism and Translation
4. RNA Transport and Nuclear Export
5. Ubiquitin-Proteasome and RNA-Protein Complex Quality Control
6. Chromatin-Linked Transcription Modulators

ZNF557 (HepG2 cells)
Transcriptional regulation and RNA metabolic coordination via zinc finger proteins and RNA-associated complexes (0.88)
1. Transcriptional Regulation by Zinc Finger Proteins
2. General Transcription Machinery and RNA Polymerase Assembly
3. RNA Splicing, Processing, and Surveillance
4. rRNA Processing and Ribosome Biogenesis
5. Epigenetic and Chromatin-Level Regulation
6. Coupling with RNA Modification and Translation Control

749

750

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

752 **acid-binding proteins.**

753 Functional enrichment analysis was conducted using the PANTHER database. This table

754 lists annotations that were statistically enriched in regulatory target genes predicted by a

755 deep learning model trained on gene co-expression data. These annotations were not

756 significantly overrepresented in the full set of co-expressed genes (Pearson's *z*-score > 2)

757 that included the nucleic acid-binding protein itself. '**Reference**' indicates the total

758 number of human genes; '**Analyzed**' denotes the number of regulatory target genes

759 examined; '**Expected**' represents the number of genes expected to have the annotation by

760 chance; '**Over/Under**' specifies whether the annotation is overrepresented or

761 underrepresented; '**Fold Enrichment**' reflects the observed-to-expected ratio; '**Raw P-**

762 **value**' reports the result of Fisher's exact test; and '**FDR**' refers to the Benjamini–

763 Hochberg false discovery rate.

AKAP8							
PANTHER Pathways							
PANTHER Pathways	Reference (17130)	Analyzed (49)	Expected	Over/ Under	**Fold Enrichment**	raw P-value	FDR
Circadian clock system (P00015)	10	3	0.03	+	> 100	2.60E-06	4.16E-04
PANTHER GO-Slim Biological Process							
PANTHER GO-Slim Biological Process	Reference (17130)	Analyzed (49)	Expected	Over/ Under	**Fold Enrichment**	raw P-value	FDR
circadian regulation of gene expression (GO:0032922)	16	3	0.05	+	65.55	1.20E-05	1.27E-02
response to light stimulus (GO:0009416)	21	3	0.06	+	49.94	2.82E-05	1.99E-02
circadian rhythm (GO:0007623)	21	3	0.06	+	49.94	2.82E-05	1.49E-02
rhythmic process (GO:0048511)	24	3	0.07	+	43.7	4.27E-05	1.81E-02
GO cellular component complete							
GO cellular component complete	Reference (17130)	Analyzed (49)	Expected	Over/ Under	**Fold Enrichment**	raw P-value	FDR
Cry-Per complex (GO:1990512)	3	2	0.01	+	> 100	2.40E-05	4.77E-02
PKM							
PANTHER Protein Class							
PANTHER Protein Class	Reference (17130)	Analyzed (326)	Expected	Over/ Under	**Fold Enrichment**	raw P-value	FDR
aldolase (PC00044)	5	2	0.1	+	21.02	3.48E-03	5.68E-02
kinase (PC00137)	134	9	2.55	+	3.53	1.07E-03	2.62E-02
PANTHER GO-Slim Molecular Function							
PANTHER GO-Slim Molecular Function	Reference (17130)	Analyzed (326)	Expected	Over/ Under	**Fold Enrichment**	raw P-value	FDR
enzyme inhibitor activity (GO:0004857)	107	8	2.04	+	3.93	1.01E-03	4.68E-02
kinase activity (GO:0016301)	562	23	10.7	+	2.15	7.08E-04	3.57E-02
PANTHER GO-Slim Biological Process							
PANTHER GO-Slim Biological Process	Reference (17130)	Analyzed (326)	Expected	Over/ Under	**Fold Enrichment**	raw P-value	FDR
endoplasmic reticulum tubular network organization (GO:0071786)	12	3	0.23	+	13.14	1.32E-03	3.68E-02
regulation of cellular component biogenesis (GO:0044087)	139	10	2.65	+	3.78	3.31E-04	1.05E-02

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.

794 **Supplementary Note and Tables**

795

796 **Gene co-expression–based predictions align with ChIP-seq–defined**
797 **DNA-binding targets**

798 High-ranking predicted binding regions frequently coincide with ChIP-seq peaks,
799 particularly in promoter–enhancer intervals. Deep learning–derived DNA-binding site
800 predictions based on gene co-expression data show substantial concordance with ChIP-
801 seq results, particularly when broader genomic regions encompassing promoters and
802 putative enhancers are considered. When overlap was assessed within ±100 kb of
803 transcription start sites (TSSs)—a window supported by prior enhancer–gene mapping
804 studies [19,75]—the mean and median overlap ratios reached 0.69 and 0.76, respectively
805 (Ratio A2 / B2; **Supplementary Tables S2 and S3**). In contrast, overlap was lower for
806 promoter regions alone (±5 kb), with corresponding mean and median values of 0.33 and
807 0.34. Some DNA-binding proteins, including FBXL19, POGK, SPEN, THAP7, ZNF142,
808 and ZNF800, showed notably high promoter-region overlap ratios (>0.5), indicating
809 potential direct regulatory activity. Nevertheless, not all ChIP-seq peaks necessarily
810 reflect functionally relevant transcriptional binding events. Consequently, model-inferred
811 high-scoring sites do not universally coincide with experimentally determined binding
812 regions, underscoring the complexity and context dependence of regulatory interactions.

813

814

815

816

Context-specific predictions reveal known and emerging functional roles of NABPs

The deep learning–based predictions successfully recapitulate both established and potentially novel biological functions of NABPs. For example, AKAP8—previously annotated in UniProt with canonical cell cycle functions (e.g., "GO:0000278 mitotic cell cycle" and "GO:0044839 cell cycle G2/M phase transition")—was also predicted to be associated with circadian rhythm regulation in this study. Supporting literature retrieved from PubMed indicates a potential role for AKAP8 in circadian rhythm control: one study identified AKAP8 as strongly co-expressed with core clock genes such as *CRY1*, while another reported its involvement in melatonin biosynthesis, a pathway that is often disrupted in individuals with autism spectrum disorders [76,77].

PKM (pyruvate kinase M) was functionally linked to glycolytic regulation, consistent with literature-based protein–protein interaction data, including its experimentally validated interaction with ALDOA in both human and mouse systems [39,40]. However, co-expression analysis in our dataset revealed a notable context-specific pattern: ALDOA and ALDOB were not significantly co-expressed with PKM (Pearson's z-score < 2), whereas ALDOC was identified as a prominently co-expressed gene prior to deep learning model training. This result suggests that ALDOC may serve as a context-dependent effector of PKM-regulated glycolysis, particularly in hematopoietic settings such as K562 cells. Indeed, ALDOC was not only among the predicted regulatory targets of PKM, but also appeared in gene sets annotated with hypoxia-responsive functions involving HIF1A. Moreover, transcriptomic data from the Human Protein Atlas confirmed that both ALDOA and ALDOC are expressed in K562 cells, further supporting the possibility that PKM may selectively regulate ALDOC in specific cellular contexts,

841 despite the known biochemical role of ALDOA in glycolysis. These findings exemplify
842 how deep learning–guided predictions can refine our understanding of tissue-specific
843 metabolic wiring and highlight functional divergence within enzyme families.
844
845
846 **Supplementary Table S1. Co-expression–based selection enhances the**
847 **contribution scores of NABP binding sites.**
848 Binding sites identified through gene co-expression exhibited higher absolute
849 DeepLIFT contribution scores compared to those selected solely from ChIP-seq data
850 without co-expression information. This trend was observed for both DNA-binding
851 proteins in HepG2 cells and RNA-binding proteins in K562 cells. For each group, the
852 table reports the mean and median values of positive and negative contribution scores.
853 These findings suggest that gene co-expression–based selection preferentially identifies
854 functionally impactful regulatory interactions.

DNA-binding proteins (HepG2 cell)			
Co-expression–based		**Non-co-expression–based**	
	Plus score		Plus score
Mean	0.005673	Mean	0.003391
Median	0.003829	Median	0.001893
	Minus score		Minus score
Mean	-0.017000	Mean	-0.002502
Median	-0.002286	Median	-0.001583

RNA-binding proteins (K562 cell)			
Co-expression–based		**Non-co-expression–based**	
	Plus score		Plus score
Mean	0.004916	Mean	0.002962
Median	0.002422	Median	0.001689
	Minus score		Minus score
Mean	-0.011987	Mean	-0.002485
Median	-0.001521	Median	-0.001414

855
856
857
858
859

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

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

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

Rank-based thresholds:

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

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

Score-based thresholds:

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

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

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

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

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

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

889 **Rank-based thresholds**:

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

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

892 **Score-based thresholds**:

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

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

895

896

897 Supplementary Table S4. Enrichment of high-contribution transcribed 898 regions with eCLIP–validated RBP binding sites.

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

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

907 **Rank-based thresholds**:

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

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

Score-based thresholds:

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

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

References

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

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

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

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

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

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

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

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

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

10 Boyle, A. P. *et al.* Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**, 453-456, doi:10.1038/nature13668 (2014).

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

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

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

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

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

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

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

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

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

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

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

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

23 Yang, S. *et al.* COEXPEDIA: exploring biomedical hypotheses via co-expressions associated with medical subject headings (MeSH). *Nucleic acids research* **45**, D389-D396, doi:10.1093/nar/gkw868 (2017).

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

25 Miller, H. E. & Bishop, A. J. R. Correlation AnalyzeR: functional predictions from gene

co-expression correlations. *BMC Bioinformatics* **22**, 206, doi:10.1186/s12859-021-04130-7 (2021).

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

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

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

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

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

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

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

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

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

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

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

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

38 Farshadi, E., van der Horst, G. T. J. & Chaves, I. Molecular Links between the Circadian Clock and the Cell Cycle. *J Mol Biol* **432**, 3515-3524, doi:10.1016/j.jmb.2020.04.003

(2020).

39 Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research* **28**, 3442-3444, doi:10.1093/nar/28.18.3442 (2000).

40 Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research* **51**, D638-D646, doi:10.1093/nar/gkac1000 (2023).

41 Tian, W. *et al.* Bioinformatics analysis of the role of aldolase A in tumor prognosis and immunity. *Sci Rep* **12**, 11632, doi:10.1038/s41598-022-15866-4 (2022).

42 Zhang, F. *et al.* Elevated transcriptional levels of aldolase A (ALDOA) associates with cell cycle-related genes in patients with NSCLC and several solid tumors. *BioData Min* **10**, 6, doi:10.1186/s13040-016-0122-4 (2017).

43 Zlacka, J. & Zeman, M. Glycolysis under Circadian Control. *Int J Mol Sci* **22**, doi:10.3390/ijms222413666 (2021).

44 Yang, Y., Pu, J. & Yang, Y. Glycolysis and chemoresistance in acute myeloid leukemia. *Heliyon* **10**, e35721, doi:10.1016/j.heliyon.2024.e35721 (2024).

45 Puram, R. V. *et al.* Core Circadian Clock Genes Regulate Leukemia Stem Cells in AML. *Cell* **165**, 303-316, doi:10.1016/j.cell.2016.03.015 (2016).

46 Lee, Y. Roles of circadian clocks in cancer pathogenesis and treatment. *Exp Mol Med* **53**, 1529-1538, doi:10.1038/s12276-021-00681-0 (2021).

47 Liberti, M. V. & Locasale, J. W. The Warburg Effect: How Does it Benefit Cancer Cells? *Trends Biochem Sci* **41**, 211-218, doi:10.1016/j.tibs.2015.12.001 (2016).

48 Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science (New York, N.Y.)* **324**, 1029-1033, doi:10.1126/science.1160809 (2009).

49 Semenza, G. L. Hypoxia-inducible factors: mediators of cancer progression and targets for cancer therapy. *Trends Pharmacol Sci* **33**, 207-214, doi:10.1016/j.tips.2012.01.005 (2012).

50 Keith, B., Johnson, R. S. & Simon, M. C. HIF1alpha and HIF2alpha: sibling rivalry in hypoxic tumour growth and progression. *Nat Rev Cancer* **12**, 9-22, doi:10.1038/nrc3183 (2011).

51 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science (New York, N.Y.)* **347**, 1260419, doi:10.1126/science.1260419 (2015).

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

53 Bluemlein, K. *et al.* No evidence for a shift in pyruvate kinase PKM1 to PKM2 expression

during tumorigenesis. *Oncotarget* **2**, 393-400, doi:10.18632/oncotarget.278 (2011).

54 Wang, X. & Liu, L. AKAP8 interacting with DDX5 to regulate R-loop balance involves in lung carcinoma cell growth. *bioRxiv*, doi:10.1101/2023.11.28.568985 (2024).

55 Shirai, Y., Suwa, T., Kobayashi, M., Koyasu, S. & Harada, H. DDX5 enhances HIF - 1 activity by promoting the interaction of HIF - 1α with HIF - 1β and recruiting the resulting heterodimer to its target gene loci. *Biology of the Cell* **116**, doi:10.1111/boc.202300077 (2023).

56 Leszczynska, K. B. *et al.* Hypoxia - mediated regulation of DDX5 through decreased chromatin accessibility and post - translational targeting restricts R - loop accumulation. *Molecular Oncology* **17**, 1173-1191, doi:10.1002/1878-0261.13431 (2023).

57 Farshadi, E., van der Horst, G. T. J. & Chaves, I. Molecular Links between the Circadian Clock and the Cell Cycle. *Journal of Molecular Biology* **432**, 3515-3524, doi:10.1016/j.jmb.2020.04.003 (2020).

58 Ran, J. *et al.* Hypoxia regulates glycolysis through the HIF-1α/BMAL1/ALDOC axis to reduce oxaliplatin sensitivity in colorectal cancer. *Journal of Cancer* **16**, 2503-2515, doi:10.7150/jca.108582 (2025).

59 De Vitis, C. *et al.* ALDOC- and ENO2- driven glucose metabolism sustains 3D tumor spheroids growth regardless of nutrient environmental conditions: a multi-omics analysis. *Journal of Experimental & Clinical Cancer Research* **42**, doi:10.1186/s13046-023-02641-0 (2023).

60 Minchenko, O. H., Garmash, I. A. & Minchenko, D. O. Inhibition of IRE1 modifies hypoxic regulation of G6PD, GPI, TKT, TALDO1, PGLS and RPIA genes expression in U87 glioma cells. *The Ukrainian Biochemical Journal* **89**, 38-49, doi:10.15407/ubj89.01.038 (2017).

61 Obeidat, N. *et al.* Effects of cyclic acute and chronic hypoxia on the expression levels of metabolism related genes in a pancreatic cancer cell line. *Biomedical Reports* **17**, doi:10.3892/br.2022.1564 (2022).

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

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

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

65 Burley, S. K. & Kamada, K. Transcription factor complexes. *Curr Opin Struct Biol* **12**, 225-

230, doi:10.1016/s0959-440x(02)00314-7 (2002).

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

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

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

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

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

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

72 Mudge, J. M. *et al.* GENCODE 2025: reference gene annotation for human and mouse. *Nucleic acids research* **53**, D966-D975, doi:10.1093/nar/gkae1078 (2025).

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

74 Young Cheon, N., Kim, S. & Yil Lee, J. in *Gene Expression* Ch. Chapter 3, (IntechOpen, 2022).

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

76 Tan, X. *et al.* Novel Insights into the Circadian Rhythms Based on Long Noncoding and Circular RNA Profiling. *Int J Mol Sci* **25**, doi:10.3390/ijms25021161 (2024).

77 Poelmans, G., Franke, B., Pauls, D. L., Glennon, J. C. & Buitelaar, J. K. AKAPs integrate genetic findings for autism spectrum disorders. *Transl Psychiatry* **3**, e270, doi:10.1038/tp.2013.48 (2013).