Learning monosemantic features in multitask DNA regulatory sequence models via sparse autoencoder decomposition

Anya Korsakova[1] and David R Kelley[1,*]

[1]Calico Life Sciences LLC, South San Francisco, CA 94080, USA
[*]correspondence: drk@calicolabs.com

Abstract

Deep learning models for regulatory genomics achieve high predictive performance across diverse molecular phenotypes, yet their internal representations remain opaque. Here, we apply sparse autoencoders (SAEs) to decompose learned representations of Borzoi, a state-of-the-art CNN-transformer that predicts genome-wide transcriptional and epigenetic profiles from DNA sequence. Training TopK-SAEs on activations from Borzoi's early convolutional layers, we discover monosemantic regulatory features that correspond to transcription factor (TF) and RNA binding protein (RBP) motifs and transposable element sequences. We identify hundreds of significant position weight matrices that map SAE-discovered features to established TF binding sites through motif discovery using MEME suite against known TF databases. This work demonstrates that SAEs can systematically decompose regulatory genomics models into biologically interpretable components.

1 Introduction

Deep learning models predict genomic sequence activity with high accuracy for molecular phenotypes ranging from transcription to chromatin features [1, 2, 3, 4]. However, these models' internal representations remain largely uninterpretable. Traditional interpretability approaches—gradient-based attribution and *in silico* mutagenesis—provide local insights but scale poorly to genome-wide pattern discovery [5]. While some frameworks incorporate interpretable design features [6], they typically target specific biological processes rather than discovering emergent regulatory patterns.

Recent advances in mechanistic interpretability of natural language models using sparse autoencoders (SAEs) offer a promising solution by decomposing polysemantic neural activations into monosemantic features [7, 8, 9, 10]. Recent applications to protein language models demonstrate this potential: SAEs applied to ESM2 identified structural motifs and functional domains [11, 12, 13], while analysis of the Evo2 DNA model revealed features corresponding to tRNAs, secondary structures, and CRISPR elements [14]. In regulatory genomics, supervised training on tissue and cell-type-specific data excel over self-supervised training [2, 15, 1]; however, systematic feature discovery in these models remains unexplored.

We apply SAEs to interpret Borzoi [2], a CNN-transformer that predicts RNA transcription, splicing, and polyadenylation, as well as chromatin accessibility, histone modifications, and transcription factor binding from DNA sequence. Borzoi's hierarchical convolutional architecture and strong performance across diverse regulatory phenotypes make it an ideal interpretability target.

2 Methods

2.1 Dataset and model architecture

We extracted activations from the pretrained Borzoi model [2] using $N = 4096$ genomic sequences for SAE training. Each $524,288$bp input sequence was divided into four non-overlapping chunks of $131,072$bp segments. Every embedded position formed an example representing a sequence

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

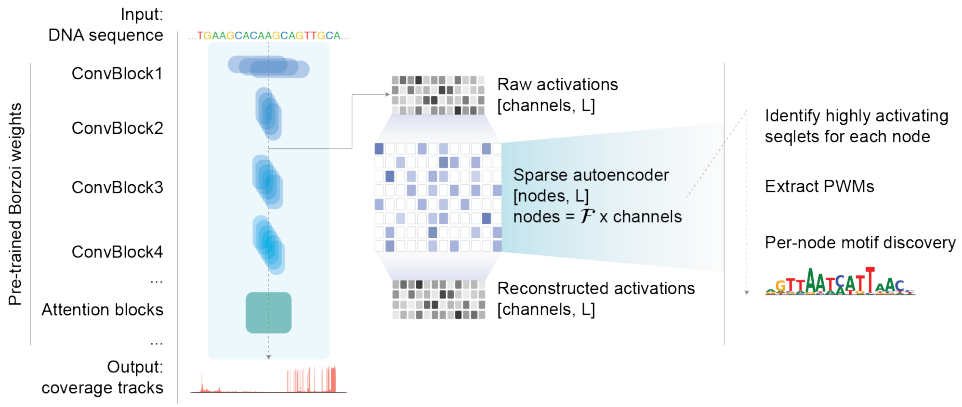

Figure 1: SAE framework for discovering monosemantic regulatory features. We extract activations from Borzoi's first four convolutional layers, train TopK-SAEs with expansion factor $\mathcal{F} = 4$ and 5% sparsity to decompose them into sparse features, then identify biological concepts by extracting top-activating sequence regions (seqlets), discovering position weight matrices (PWMs) via MEME, and matching to known TF/RBP databases and general genomic annotations.

window whose width varied by layer (16-31 bp receptive field), processed in batches. We focused on layers `conv1d_1` through `conv1d_4`, where hierarchical motif-like regulatory features emerge, instead of post-attention activations where multi-faceted functional or structural concepts arise. For evaluation, we analyzed $N = 640$ sequences from a held-out fold unseen during both Borzoi training and SAE training.

2.2 Sparse autoencoder

We employed TopK sparse autoencoders [10] on activation vectors representing sequence patterns at each genomic position in 16-31bp windows, depending on the layer (Fig. 1). We normalized each input channel by its maximum value to balance feature importance. The SAE architecture expands the channel dimension by factor $\mathcal{F} > 1$, retains only the top k percent of autoencoder latents, and reconstructs the original activations as follows:

$$\begin{aligned} z &= \text{ReLU}(\text{TopK}(W_{enc}(x - b_{dec}) + b_{enc})) \\ \hat{x} &= W_{dec}z + b_{dec}, \end{aligned} \tag{1}$$

where z represents sparse features, $W_{enc}, b_{enc}, W_{dec}, b_{dec}$ are learned encoder and decoder weights and biases. The loss minimizes reconstruction error: $\mathcal{L} = ||x - \hat{x}||_2^2$. Top-K is applied to retain only the top k percent of autoencoder latents.

We trained SAEs for layers 1-4 (denoted L1-L4) with: expansion factor $\mathcal{F} = 4$, learning rate 1×10^{-5} (Adam optimizer), and $k = 5\%$ sparsity constraint.

2.3 Feature extraction and annotation

For each SAE node with >1,000 non-zero activations, we extracted the top 1,000 activating genomic regions (seqlets), padded to match layer-specific receptive fields (16-31 bp). We pursued two complementary annotation strategies:

Motif discovery: MEME identified one PWM per node ($p < 0.05$). TomTom compared these against J. Vierstra's non-redundant TF motif database [16] merged with the CIS-BP RNA-binding protein motif database [17], retaining matches with E-value< 0.05.

Genomic elements: BEDTools quantified overlaps between seqlets and SCREEN cCREs [18], GENCODE gene features [19], and RepeatMasker repetitive elements [20]. We assigned features hierarchically: TFs/RBPs > exons/introns > cCREs > repeats. Statistical significance for genomic overlaps was assessed using Fisher's exact test against shuffled controls (odds ratio > 2, $p < 0.05$).

3 Results

3.1 SAEs successfully decompose Borzoi's learned representations

Our hyperparameters ($\mathcal{F} = 4$, LR$= 1 \times 10^{-5}$, $k = 5\%$) balanced sensitivity for discovering transcription factors against redundancy across nodes. Higher expansion factors increased feature splitting without discovering additional unique features (Table A2). TopK-SAEs achieved strong reconstruction quality (Pearson $R = 0.84$ for L2 activations). Importantly, model outputs remained similar when using SAE-reconstructed activations ($r = 0.703$ for L2, Table A1), confirming preservation of biologically relevant information.

3.2 Discovery of diverse regulatory features

Analysis of 2,173 active SAE nodes revealed distinct regulatory categories (Fig. 2). Transposable elements covered the most nodes (1,042 nodes), followed by TF binding motifs (585 nodes), cCREs (352 nodes), and RBP motifs (42 nodes). Fundamental regulatory motifs—TATA boxes, poly-A signals, SINE/Alu elements, and GATA motifs—exhibited high mean activation values (Fig. 2), suggesting preferential signal propagation through the network. In silico shuffling mutagenesis of the top 50 activating seqlets for each node showed biological relevance; nodes discovering TFs with available ChIP tracks often had the highest L2 score for the matching TF-ChIP track. For example, perturbation of node 464 seqlets discovering NFIA motif leads to a high ChIP:NFIA L2 score, and perturbation of node 290 discovering HNF1A motif induces high ChIP:HNF1A and RNA:liver, RNA:kidney scores.

While we assigned a primary concept per node via priority ranking (Methods), nodes often recognize multiple concepts at finer motif and coarser regulatory/transposable element resolution. For instance, node 212 matched both ZNF524 ($q = 0.028$) and SINE/Alu elements (odds ratio=93.0), while node 73 recognized ASCL1 ($q = 2.0 \times 10^{-4}$) and distal enhancers (OR=3.33). This dual specificity reflects hierarchical co-occurrence patterns in the genome.

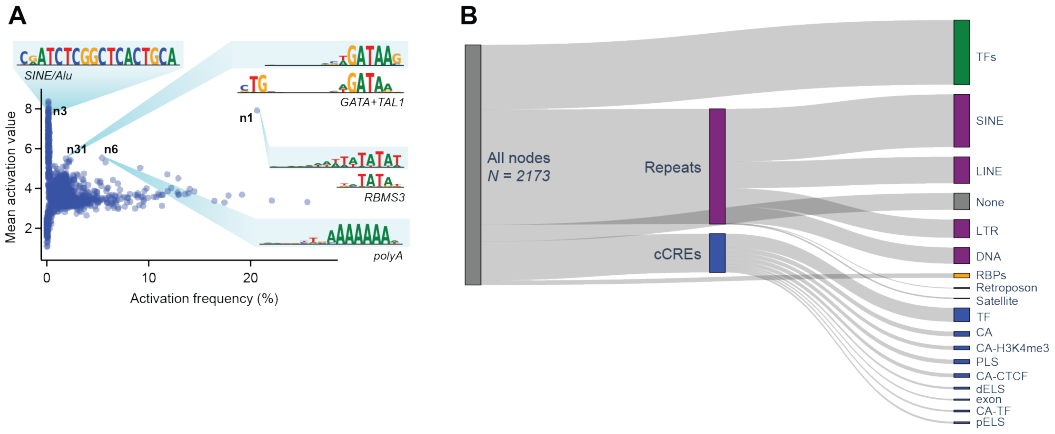

Figure 2: Landscape of SAE-discovered regulatory features. (A) Feature importance plot showing activation frequency versus mean activation strength for L2 nodes. High-activation nodes capture fundamental regulatory elements: node 1 (TATA/RBMS3, $E = 0.0065$), node 3 (SINE/Alu), node 6 (poly-A), and node 31 (GATA1+TAL1 composite element). (B) Sankey diagram showing distribution of 2,173 discovered features across biological categories, revealing comprehensive regulatory encoding from single motifs to repetitive elements. "None" category indicates nodes that were not mapped to any concept. cCRE legend: PLS – promoter-like signature; dELS – distal enhancer-like signature; pELS – proximal enhancer-like signature; CA – chromatin accessibility; CA-H3K4me3 – chromatin accessibility + H3K4me3; TF – transcription factor; CA-TF – chromatin accessibility + transcription factor.

3.3 Feature redundancy reflects biological specificity

Multiple nodes often recognize similar patterns, a phenomenon known as feature splitting, where increasing SAE width leads to replacing a general concept with more specialized concepts [8]. We frequently discovered similar TF/RBP motifs from distinct nodes. The most frequently discovered motif in L2 was ZNF384 (51 nodes), whose PWM resembles a poly-A homopolymer. Examining redundancy revealed biological specificity. Trivially, separate nodes discovered forward and reverse strand motif versions, demonstrating strand equivariance. For same strand motif nodes, we observed minimal seqlet overlap between nodes (mean Jaccard index across stranded motif-node pairs $=$ 0.005 with $p = 0.05$ significance cutoff), indicating that underlying sequences activating the nodes are largely non-overlapping. For example, despite 49 L2 nodes recognizing GATA1+TAL1, pairwise seqlet Jaccard indices were low (mean: 0.0012; max: 0.1828, Fig. 3A).

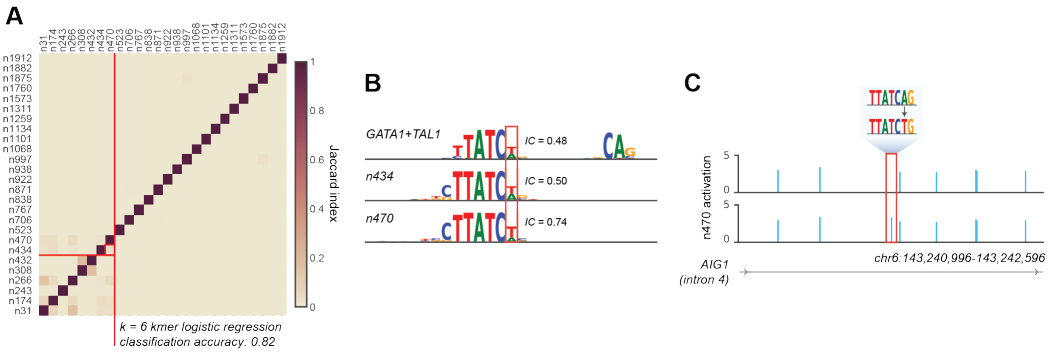

Figure 3: Biological basis of apparent GATA1+TAL1 redundancy. (A) Low pairwise Jaccard indices among 49 nodes recognizing reverse-strand GATA1+TAL1 motif, filtered for high information content PWMs ($IC_{any} > 1$). (B) Distinct flanking sequence preferences between nodes 434 and 470, distinguishable by 6-mer composition via logistic regression (82% accuracy). Red box highlights flanking thymines with varying information content that affects activation recognition by the node. (C) Single-nucleotide mutation (A $\rightarrow$ T) in *AIG1* intron selectively activates node 470, demonstrating context-specific recognition via motif-flanking sequence. Red box highlights introduction of the node 470 activation peak upon mutation.

To investigate this specificity, we analyzed two similar reverse-strand GATA1+TAL1 nodes (434, 470). A logistic regression classifier distinguished their activating sequences with 82% accuracy using 6-mer composition revealing consistent differences in flanking nucleotide preferences (Fig. 3B). Mutating a single nucleotide (A $\rightarrow$ T) flanking a GATA motif in the *AIG1* gene's fourth intron specifically activated node 470 without affecting node 434 (Fig. 3C), confirming context-dependent recognition beyond core motifs.

3.4 Progressive refinement across network depth

Regulatory concepts propagate through network layers with progressive refinement: 25.5% of TF motifs appear in L1, increasing to 67.5% in L3 and 75.9% in L4. 66.05% of motifs discovered in L2 moved to L3. Tracking the kidney-specific transcription factor HNF1A revealed this refinement process. L2 contained three HNF1A nodes with varying orientations and information content (Fig. A1). By L4, only two reverse-oriented nodes remained; loss of the forward-strand version illustrates the challenge of maintaining complete motif recall in deeper layers with this framework. Near the kidney-expressed *UMOD* gene, L4 preserved L2's strongest activation peaks while suppressing spurious signals. We visualized sliding activations for HNF1A nodes in L2 and L4, alongside in silico mutagenesis nucleotide attributions for the two top activation peaks, revealing portions of the HNF1A motif. The noise reduction likely results from L4's expanded 31bp receptive field enabling better context integration to prioritize biologically relevant motifs.

3.5 Interactive visualization platform

We developed an interactive Streamlit application (`motifscout.com`) enabling real-time exploration of SAE features across layers L1-L4. The platform provides node-level views with PWMs, TomTom matches, and similarity heatmaps, plus motif-level analysis with strand-specific node matches and seqlet overlap quantification (Fig. A2).

4 Conclusion

This work demonstrates that sparse autoencoders successfully decompose deep learning representations in regulatory genomics models into interpretable biological features. Our analysis of Borzoi revealed over 2,000 monosemantic features corresponding to TF/RBP motifs, transposable elements, and other coarser regulatory elements. Apparent redundancy often encoded biologically meaningful specificity for motif-flanking sequences and context-dependent recognition. Progressive refinement across network depth suggests hierarchical learning of increasingly specific and relevant regulatory features.

Although we focused our analysis on TF/RBP motifs, the observation that transposable elements are comprehensively recognized by Borzoi warrants further investigation. Most sequencing experiments struggle to align reads and determine true activity signal on these repetitive regions. Borzoi may be learning these elements in order to predict low coverage caused by the loss of ambiguous lost alignments. Alternatively, TEs have extensively rewired regulatory networks and may influence nearby transcription and chromatin [21].

Future work could extend this approach to deeper layers and attention mechanisms, but increased feature complexity (e.g. motifs combinations) will challenge interpretability. Discovered features may enable targeted model perturbation to analyze motif influence on predictions. Improved methods to enhance sensitivity while distinguishing technical from biological redundancy would strengthen this framework. Our interactive visualization platform and open-source code provide foundations for mechanistic understanding of how deep learning models encode regulatory biology (`github.com/calico/sae-borzoi` and `motifscout.com`).

5 Acknowledgments

This work was funded by Calico Life Sciences LLC. The funder had no role in study design, data collection or analysis. Publication of the manuscript was approved after an internal scientific review process. We thank Benjamin Auerbach, Johannes Linder, Divyanshi Srivastava, Fanny Huang, Han Yuan, and David Wang for helpful discussions and valuable feedback.

References

[1] Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. *bioRxiv*, page 2025.06.25.661532, July 2025.

[2] Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R. Kelley. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nat. Genet.*, pages 1–13, January 2025.

[3] Geoff Fudenberg, David R. Kelley, and Katherine S. Pollard. Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods*, 17:1111–1117, November 2020.

[4] Jian Zhou. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat. Genet.*, 54:725–734, May 2022.

[5] Jacob Schreiber, Surag Nair, Akshay Balsubramani, and Anshul Kundaje. Accelerating in silico saturation mutagenesis using compressed sensing. *Bioinformatics*, 38(14):3557–3564, 06 2022.

[6] Kseniia Dudnyk, Donghong Cai, Chenlai Shi, Jian Xu, and Jian Zhou. Sequence basis of transcription initiation in the human genome. *Science*, 384(6694), April 2024.

[7] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models. *arXiv*, September 2023.

[8] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning, June 2024. [Online; accessed 9. Jun. 2025].

[9] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet, March 2024. [Online; accessed 21. Jul. 2025].

[10] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv*, June 2024.

[11] Elana Simon and James Zou. InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders. *bioRxiv*, page 2024.11.14.623630, November 2024.

[12] Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From Mechanistic Interpretability to Mechanistic Biology: Training, Evaluating, and Interpreting Sparse Autoencoders on Protein Language Models. *bioRxiv*, page 2025.02.06.636901, February 2025.

[13] Nithin Parsan, David J. Yang, and John J. Yang. Towards Interpretable Protein Structure Prediction with Sparse Autoencoders. *arXiv*, March 2025.

[14] Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakshi, Anton

Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Etowah Adams, Stephen A. Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R. K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, Jonathan C. Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modeling and design across all domains of life with Evo 2. *bioRxiv*, page 2025.02.18.638918, February 2025.

[15] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[16] Jeff Vierstra, John Lazar, Richard Sandstrom, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, Douglas Dunn, Fidencio Neri, Eric Haugen, Eric Rynes, Alex Reynolds, Jemma Nelson, Audra Johnson, Mark Frerker, Michael Buckley, Rajinder Kaul, Wouter Meuleman, and John A. Stamatoyannopoulos. Global reference mapping of human transcription factor footprints. *Nature*, 583:729–736, July 2020.

[17] Debashish Ray, Hilal Kazan, Kate B. Cook, Matthew T. Weirauch, Hamed S. Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, Hong Na, Manuel Irimia, Leah H. Matzat, Ryan K. Dale, Sarah A. Smith, Christopher A. Yarosh, Seth M. Kelly, Behnam Nabet, Desirea Mecenas, Weimin Li, Rakesh S. Laishram, Mei Qiao, Howard D. Lipshitz, Fabio Piano, Anita H. Corbett, Russ P. Carstens, Brendan J. Frey, Richard A. Anderson, Kristen W. Lynch, Luiz O. F. Penalva, Elissa P. Lei, Andrew G. Fraser, Benjamin J. Blencowe, Quaid D. Morris, and Timothy R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499:172–177, July 2013.

[18] Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A. Davis, Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L. Van Nostrand, Peter Freese, David U. Gorkin, Yin Shen, Yupeng He, Mark Mackiewicz, Florencia Pauli-Behn, Brian A. Williams, Ali Mortazavi, Cheryl A. Keller, Xiao-Ou Zhang, Shaimae I. Elhajjajy, Jack Huey, Diane E. Dickel, Valentina Snetkova, Xintao Wei, Xiaofeng Wang, Juan Carlos Rivera-Mulia, Joel Rozowsky, Jing Zhang, Surya B. Chhetri, Jialing Zhang, Alec Victorsen, Kevin P. White, Axel Visel, Gene W. Yeo, Christopher B. Burge, Eric Lécuyer, David M. Gilbert, Job Dekker, John Rinn, Eric M. Mendenhall, Joseph R. Ecker, Manolis Kellis, Robert J. Klein, William S. Noble, Anshul Kundaje, Roderic Guigó, Peggy J. Farnham, J. Michael Cherry, Richard M. Myers, Bing Ren, Brenton R. Graveley, Mark B. Gerstein, Len A. Pennacchio, Michael P. Snyder, Bradley E. Bernstein, Barbara Wold, Ross C. Hardison, Thomas R. Gingeras, John A. Stamatoyannopoulos, and Zhiping Weng. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583:699–710, July 2020.

[19] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, 22(9):1760–1774, September 2012.

[20] AFA Smit, R Hubley, and P. Green. Repeatmasker open-4.0, 2013-2015.

[21] Arsala Ali, Kyudong Han, and Ping Liang. Role of transposable elements in gene regulation in the human genome. *Life (Basel)*, 11(2):118, February 2021.

A Appendix

A.1 Supplementary tables

Table A1: SAE reconstruction quality across Borzoi layers.

Layer	Pearson r (activations)	Pearson r (output tracks)
L1	0.851	0.592
L2	0.840	0.703
L3	0.838	0.722
L4	0.871	0.704

Table A2: L2-SAE motif discovery metrics under different hyperparameters.

Learning rate	Top K %	$\mathcal{F}$ (exp. factor)	% motifs discovered	% non-redundant motifs
1e-05	0.1	8	90.56	10.82
1e-05	0.05	8	82.87	15.47
1e-05	0.1	4	82.52	16.01
1e-05	0.05	4	75.17	17.80
1e-05	0.1	2	67.13	24.39
1e-05	0.05	2	61.89	27.23
0.0001	0.1	8	91.96	8.07
0.0001	0.05	8	90.21	7.88
0.0001	0.1	4	80.77	13.60
0.0001	0.05	4	80.77	13.01
0.0001	0.1	2	67.48	21.01
0.0001	0.05	2	70.28	20.61

A.2 Supplementary figures

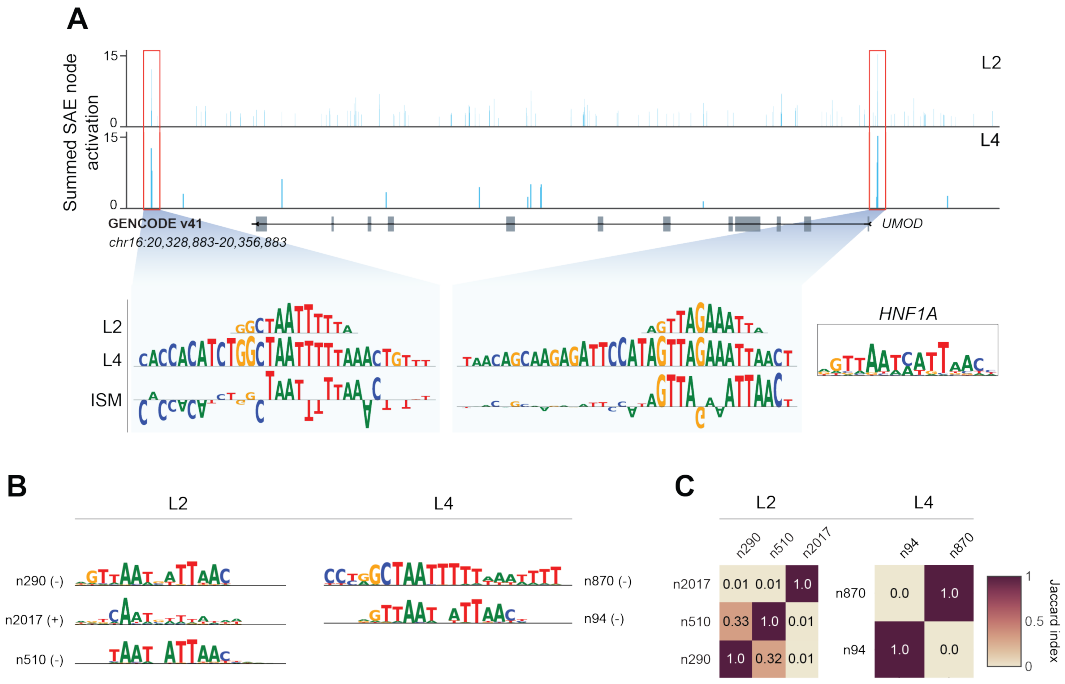

Figure A1: Progressive refinement of HNF1A motif recognition across layers. (A) Summed activations of HNF1A-discovering nodes near the kidney-specific *UMOD* gene. L4 preserves L2's strongest signals while reducing noise. Sliding window activations and in silico mutagenesis (ISM) attributions shown below. (B) Layer-specific HNF1A PWMs showing information content evolution. (C) Low Jaccard indices confirm distinct sequence contexts for each HNF1A node.

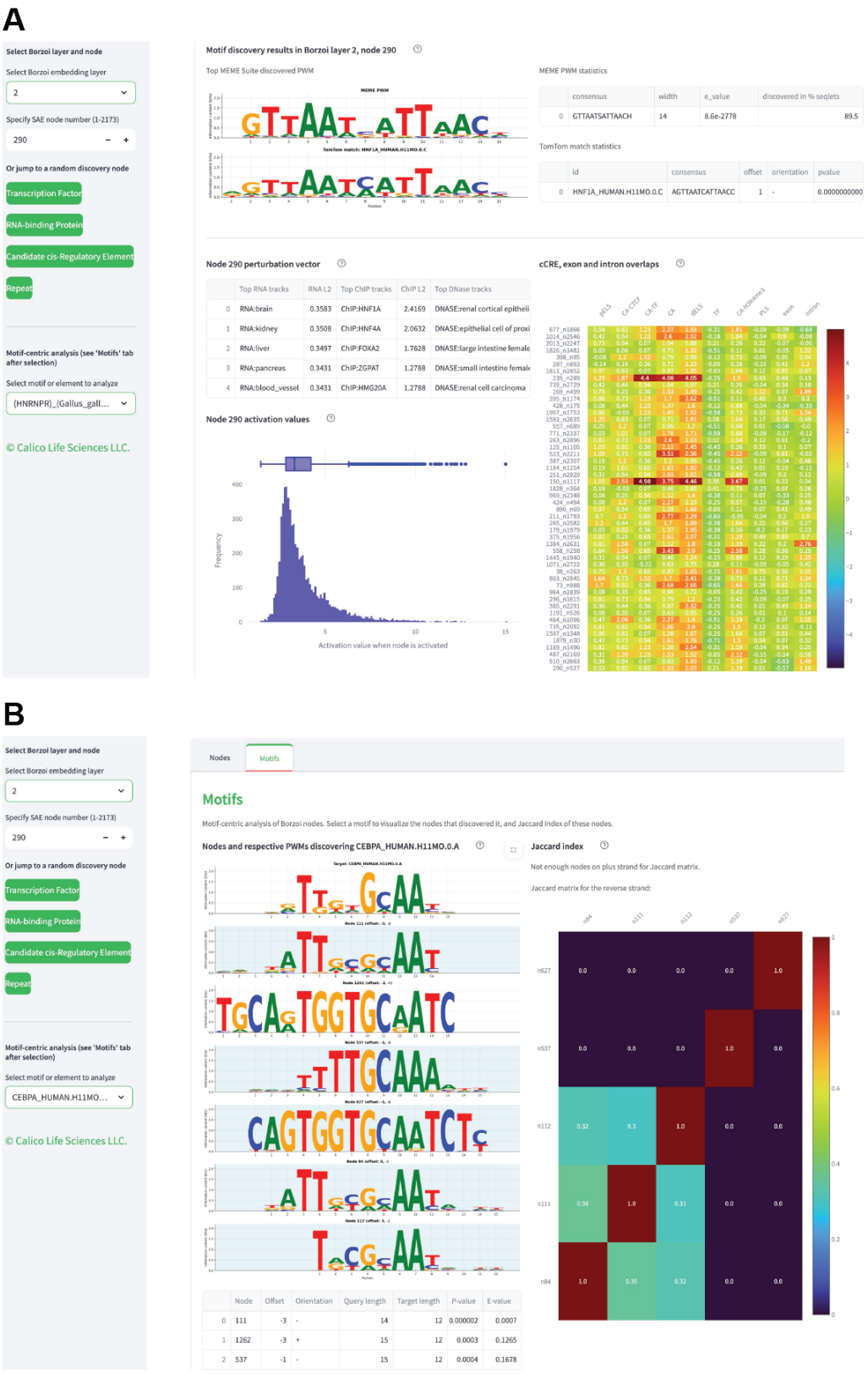

Figure A2: Motif Scout visualization server layout. A) Node view, where user can select Borzoi layer and visualize a particular node number with respective PWM, TomTom match, perturbation vector, cCRE and RMSK overlaps, where available. B) Motif view allows to compare node PWMs of a given motif if it was discovered by multiple nodes.