A transformer-based language model reveals developmental constraint and network complexity during zebrafish embryogenesis

Juan F Poyatos
National Museum of Natural Sciences (MNCN-CSIC), Madrid 28006, Spain.
juanfpoyatos@csic.es

Abstract

Understanding how regulatory complexity and constraint shape organismal development remains a central challenge in biology. The developmental hourglass framework posits that mid-embryogenesis –the phylotypic stage– is a period of heightened conservation and regulatory organization. We test this using Zebraformer, a transformer-based language model trained on single-cell transcriptomic data from zebrafish embryos. Unlike models focused on prediction or classification, Zebraformer learns context-sensitive representations of genes and cells that encode temporal progression, anatomical identity, and regulatory relationships. Embeddings reflect differentiation timing and transcriptional divergence, while attention-derived gene networks reveal a transient rise in complexity during the phylotypic stage. This stage also exhibits increased perturbation sensitivity and a shift toward centralized, modular network topology. These features are supported by graph-theoretic metrics and gene ontology enrichment, offering data-driven evidence for highly structured regulation during mid-embryogenesis. Our results demonstrate that language models can extract interpretable biological structure and support longstanding developmental theory from high-dimensional data.

Significance statement

Understanding how cells coordinate to build a complex organism remains a central challenge in biology. Development is not only genetically encoded but context-dependent; shaped by dynamic interactions among genes, cells, and time. Here, we use a transformer-based language model, Zebraformer, trained on single-cell gene expression data from zebrafish embryos, to investigate how regulatory structure evolves during development. The model captures key features of organismal formation: increasing transcriptional divergence, anatomical specificity, and a transient rise in regulatory complexity and perturbation sensitivity during the conserved phylotypic stage. These findings provide data-driven support for the developmental hourglass hypothesis and demonstrate that contextual models can uncover fundamental organizational principles from biological data alone.

Introduction

Among the many processes that define living systems, organismal development is perhaps where the contextual nature of biology is most pronounced. By "context," we refer to the dynamic, situation-specific molecular and cellular states that emerge from both intrinsic genetic programs and extrinsic cues particularly during embryogenesis, when cells differentiate, migrate, and organize into tissues and organs [1].

Technological advances now allow this contextuality to be studied at scale. Single-cell RNA sequencing (RNA-seq) [2], combined with spatial and temporal resolution from high-content imaging, has opened the door to systematically mapping how gene activity varies across time, space, and cell identity [3, 4]. These tools extend the legacy of classic lineage tracing in organisms such as *Caenorhabditis elegans* [5], offering a view of development as a dynamic, high-dimensional system.

A recent example is the zebrafish (*Danio rerio*) developmental atlas by Lange et al. [6], which profiled gene expression at ten developmental timepoints (10 hours post-fertilization [hpf] to 10 days post-fertilization [dpf]) across multiple individuals. This dataset captures a rich temporal trajectory of cellular specification and morphogenesis [7], offering an opportunity to study not just cell fate trajectories, but the broader question of how regulatory structure emerges during development.

Rather than reconstruct individual lineage paths, we ask whether a contextual model trained on this data can reveal emergent principles of developmental constraint and coordination. Specifically, we test whether single-cell expression data contain latent representations that reflect how genes functionally adapt to their cellular environment, and how these representations change over time. We also explore whether these representations can uncover interactions among genes that reflect a dynamically evolving gene regulatory network (GRN) architecture [8].

We place this in the context of the developmental hourglass framework [1, 9, 10], which proposes that early and late embryogenesis are more variable across species, while a mid-embryonic "phylotypic" stage is highly conserved, marked by a shared body plan and increased developmental constraint. We assess two key predictions of this framework: (1) that the underlying regulatory architecture becomes more complex and centralized during the phylotypic stage, and (2) that this stage is more sensitive to perturbation, reflecting its functional importance.

To test these hypotheses, we introduce `Zebraformer`, a domain-specific large language model (LLM) adapted from the transformer architecture [11, 12]. LLMs, widely used in natural language processing, learn to embed inputs based on their context –an approach increasingly applied to biological systems [13–18]. Zebraformer treats each cell's transcriptome as a context and learns embeddings for genes based on their activity within that context. These embeddings encode developmental time, anatomical identity, and functional similarity. Moreover, the model's attention matrices, originally designed to track relationships among words, offer interpretable proxies for gene–gene regulatory interactions.

We show that this framework captures known biological structure and, critically, offers

quantitative evidence for the hourglass hypothesis from a purely data-driven, mechanistically agnostic model. In doing so, we demonstrate that LLMs can serve as hypothesis-generating tools to uncover systems-level organization in development. For background on zebrafish development, the hourglass framework, and transformer-based models in biological contexts, see Supplementary Information.

Results

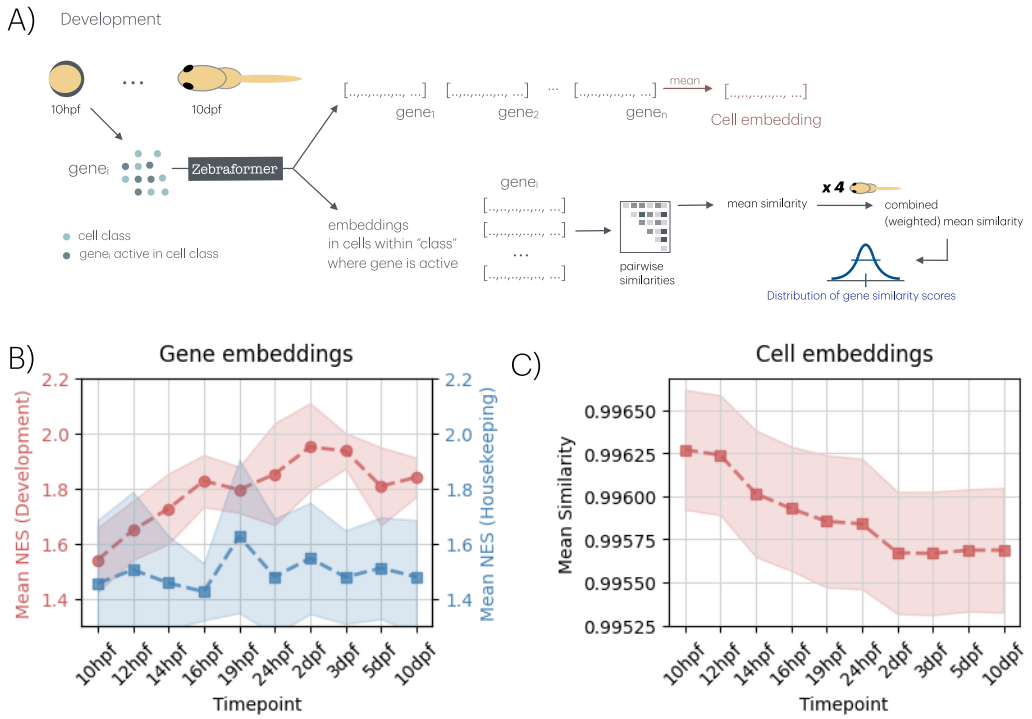

Figure 1: Gene and cell embeddings capture the zebrafish developmental axis. (A) For each developmental stage, Zebraformer generates gene embeddings for cells in a selected class (highlighted). Cell embeddings are computed as the mean of their active gene embeddings. For a gene active in multiple selected cells, pairwise embedding similarities across sibling cells are calculated, then a weighted mean is obtained using the four available sibling values. Ultimately, we obtain a distribution of these scores for all genes. (B) Gene Set Enrichment Analysis of genes ranked by embedding similarity across timepoints. Mean normalized enrichment scores (NES) are shown for the top 10 developmental (red) and housekeeping (blue) GO terms. Developmental terms show increasing enrichment, while housekeeping terms remain stable. Dots indicate mean NES; shaded areas denote standard deviation. (C) Mean embedding similarity between sibling cells, computed as the cosine similarity of cell embeddings. A decrease in similarity over time reflects transcriptional divergence and emergence of cell-type-specific programs.

Gene embeddings reflect developmental progression in zebrafish

We began by grouping cells sampled from the same developmental stage and the same sibling fish [6]. For each gene active within these groups, Zebraformer generates a contextual

embedding, a n-dimensional vector encoding the gene's behavior in its local transcriptional environment. If a gene's embeddings are highly similar across cells, this suggests consistent activity or function. To quantify this, we computed all pairwise cosine similarities for each gene's embeddings across cells, then averaged these scores within fish and across siblings (Fig. 1A, Methods).

To assess which biological processes are associated with consistent gene behavior, we ranked genes by mean similarity score and performed Gene Set Enrichment Analysis. We focused on two functional categories: developmental processes (e.g., differentiation, morphogenesis) and housekeeping functions (e.g., metabolism, transcription). For each category, we selected the top 10 enriched Gene Ontology (GO) terms and plotted normalized enrichment scores (NES) across stages (Fig. 1B, Fig. S1, Methods).

Developmental terms showed steadily increasing enrichment, peaking around 2 dpf, when major anatomical structures are established (Supplement, Table S1). In contrast, housekeeping terms remained stable and low. These results suggest that as development proceeds, consistently expressed genes become increasingly associated with core developmental programs, supporting the idea that gene embeddings encode biologically meaningful differentiation signals.

Cell embeddings capture transcriptional divergence during development

To complement gene-level analyses, we examined cell embeddings computed as the mean of active gene embeddings per cell (Fig. 1A). We calculated pairwise similarities between cells at each developmental stage and averaged across sibling fish (Methods). We observed a clear trend: cell-cell embedding similarity decreased over developmental time (Fig. 1C). This reflects increasing transcriptional divergence as cells acquire distinct identities, reinforcing the idea that Zebraformer embeddings encode not only shared function but also progressive differentiation.

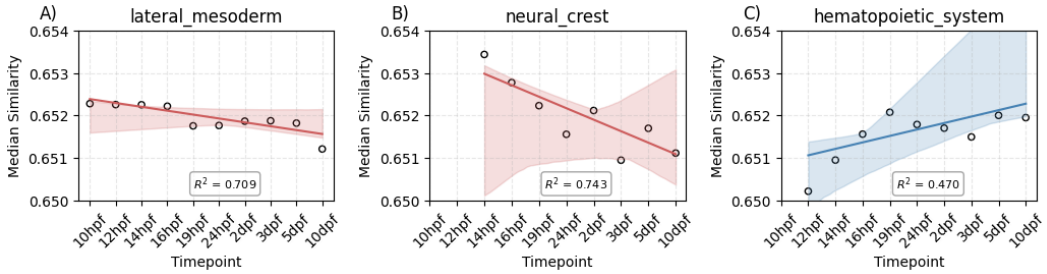

Figure 2: Gene embeddings reflect anatomical context and temporal progression. For each anatomical class, we plot regression trends of median gene embedding similarity over time. Similarity is computed per gene across cells in which it is active, then summarized at the domain level. (A) Lateral mesoderm: similarity decreases with time ($R^2 = 0.709$). (B) Neural crest: stronger decrease and higher uncertainty ($R^2 = 0.743$). (C) Hematopoietic system: increasing similarity over time ($R^2 = 0.470$). Shaded areas indicate bootstrapped uncertainty from 10,000 resamples.

Embedding similarity varies across anatomical domains.

We next asked whether embedding similarity patterns are conserved within specific anatomical regions. Using curated domain annotations (e.g., lateral mesoderm, neural crest, hematopoietic system; domains refer to *embryonic tissues* or *cell lineages* in zebrafish and vertebrates generally [19], Supplement, Table S1), we computed per-gene embedding similarities across cells within each domain, stage, and fish. Trends were analyzed across timepoints.

Figure 2 highlights three representative cases (full set in Fig.S2). Both the lateral mesoderm (LM) and neural crest (NC) showed decreasing similarity over time, while the hematopoietic system (HS) showed an increase. These patterns align with known biology. In zebrafish, the LM, including the lateral plate mesoderm, undergoes major differentiation over time, both in cell fate and spatial organization, especially from early somite stages through 1–5 dpf [20]. Both the NC and LM start as highly multipotent and give rise to diverse, unrelated tissues such as neurons, cartilage, and pigment cells for NC [21], and heart, kidney, and vessels for LM [22]. Their development follows a spatiotemporal dynamics, where both location and timing influence differentiation. This increasing contextuality likely leads to lower average embedding similarity, as gene function becomes more specific. In contrast, the HS follows a more hierarchical, tree-like path: hematopoietic stem cells gradually restrict into blood and immune lineages while maintaining multipotency over time through self-renewal [23]. These results indicate that Zebraformer embeddings capture both developmental timing and anatomical context.

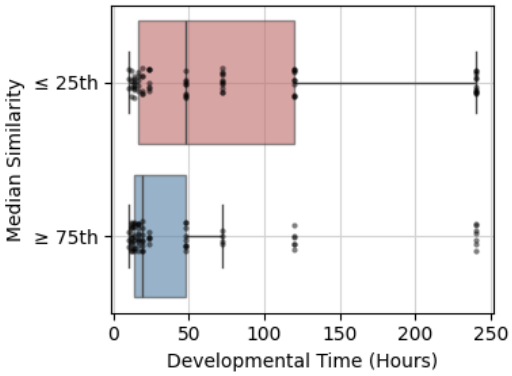

Figure 3: Cell types with lower average gene embedding similarity tend to emerge later in development. For each cell type at each timepoint, we compute the median embedding similarity of active genes across individuals. Cell types are grouped into quartiles based on similarity. Box plot shows developmental timing (hours post-fertilization) for high-similarity ($\geq$ 75th percentile) and low-similarity ($\leq$ 25th percentile) groups. Cell types with lower similarity emerge significantly later (Mann–Whitney U test, $p = 0.00676$).

Contextual diversity increases in late-emerging cell types

To examine gene contextuality at finer resolution, we analyzed 154 annotated cell types across timepoints. For each cell type and stage, we computed median gene embedding similarity

across cells, averaged over sibling fish (Methods). This metric reflects how consistently genes behave in a given cell type at a given time.

Strikingly, cell types with lower average similarity tended to appear later in development, while those with higher similarity emerged earlier (Fig. 3, Fig. S3). Early cell types are likely defined by more generic transcriptional programs, involving reused "foundational" genes. Later-emerging cell types express more specialized or spatially restricted gene sets, resulting in more diverse embedding patterns. This finding reinforces the idea that embedding similarity serves as a proxy for transcriptional specialization.

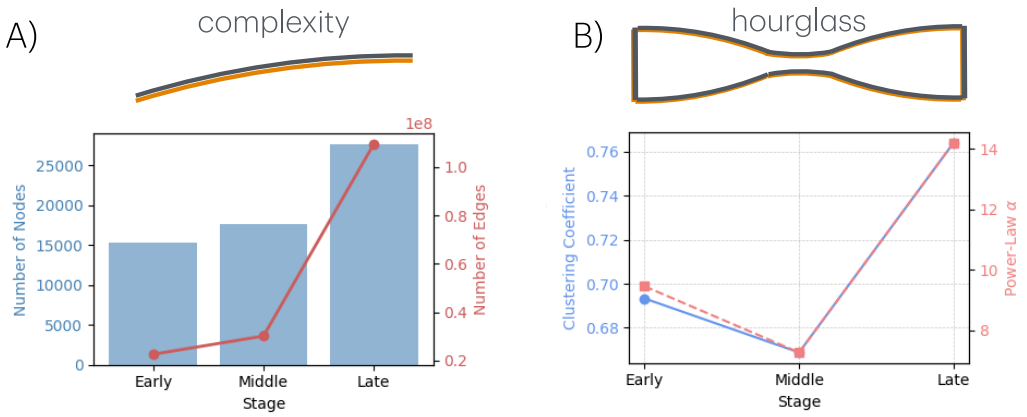

Figure 4: Attention-based gene regulatory networks evolve in a manner consistent with the hourglass framework. (A) Graph complexity across developmental time, measured by number of nodes (bars, left y-axis) and edges (line, right y-axis). (B) Network topology changes over time. Clustering coefficient (blue, left y-axis) decreases then increases; power-law exponent α (red, right y-axis) rises during late stages, indicating reduced hub dominance. These trends support a narrowing-then-expanding topology characteristic of hourglass dynamics.

Attention-based networks reveal hourglass structure of development

Can we go beyond the relational information captured by embeddings? Zebraformer's attention matrices offer another layer of information, capturing gene–gene associations learned during training [12]. These attention-derived graphs can be interpreted as data-driven GRNs, allowing us to assess regulatory architecture across developmental time.

We use this tool to assess how well Zebraformer captures the hourglass framework of development [1, 9, 10] (Supplement). We divided development into three stages: early (≤ 14 hpf), middle/phylotypic (15–19 hpf), and late (≥ 24 hpf) [24]. For each, we built attention-based GRNs and analyzed key structural features (Methods).

As expected, GRNs grew in size and connectivity across time (Fig. 4A, Fig. S4). More critically, several topological metrics revealed stage-specific patterns. The clustering coefficient increased at later stages (Fig. 4B), indicating tighter local coordination among genes. Assortativity –whether highly connected genes associate with similar nodes– became less negative over time, suggesting more modular interactions (Supplement). Modularity itself peaked during the phylotypic stage, and power-law analyses indicated that this network

contained the most pronounced hubs, consistent with a centralized regulatory architecture (Supplement, Table S2).

Together, these features reflect a regulatory "bottleneck" consistent with the hourglass framework: early and late stages exhibit flexible, distributed architectures, while the middle stage is highly organized and constrained.

Perturbation sensitivity peaks during the phylotypic stage.

The hourglass framework also predicts increased vulnerability to disruption during the phylotypic stage [25]. To test this, we performed *in silico* gene deletions and assessed their impact on embedding structure. Specifically, we removed 25 genes per cell across 10,000 randomly sampled cells from each stage and computed cosine distances between perturbed and wild-type embeddings (Methods).

Figure 5 shows the resulting distributions. The phylotypic stage exhibited significantly higher average embedding shifts, indicating greater perturbation sensitivity. Early and late stages were less affected. Variability in responses (distribution spread) was comparable across stages (Fig. S5), suggesting that the elevated mean is not driven by a few extreme outliers.

These results provide quantitative support for a key prediction of the hourglass hypothesis: that the middle phase of development is the most informationally dense and constrained, and thus the most sensitive to disruption.

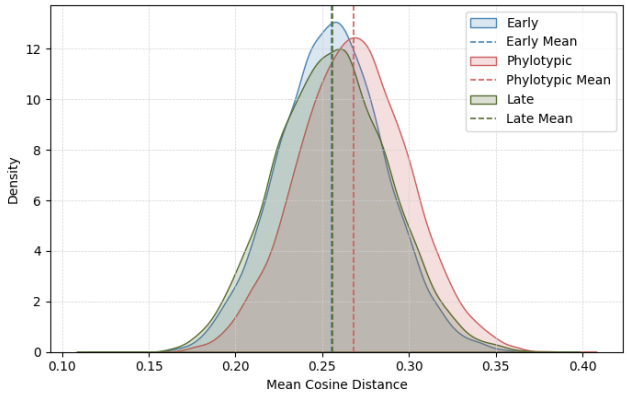

Figure 5: Sensitivity to gene perturbation peaks at the phylotypic stage. Kernel density estimates show mean cosine distances between wild-type and perturbed gene embeddings across 10,000 simulations per stage (early, middle/phylotypic, late). Vertical dashed lines indicate mean effect sizes (early: 0.256 ± 0.031; middle: 0.268 ± 0.032; late: 0.256 ± 0.033). Perturbation effects are significantly greater in the phylotypic stage, indicating increased sensitivity and constraint (KS test: early *vs.* phylotypic, p-value: 6.77e^{-116}; phylotypic *vs.* late, 1.28e^{-102}).

Discussion

Development is an inherently contextual, multi-scale process. Our goal was to ask whether a large language model –trained not to predict but to represent– could capture the regulatory

logic that underpins this process, particularly the constraints that emerge during key developmental transitions. Using Zebraformer, we show that gene embeddings learned from single-cell expression data reflect both developmental time and anatomical context. Early in development, genes exhibit high embedding similarity across cells, consistent with broad, shared programs. As development proceeds, these representations diverge, especially in later-emerging, specialized cell types. This supports the idea that Zebraformer learns not just expression levels, but contextual roles.

Beyond this, the model's attention matrices, used here as proxies for GRNs, reveal a striking pattern: network complexity, centrality, and modularity all peak during the phylotypic stage, consistent with the hourglass framework of embryogenesis [1, 9, 10]. Network topologies shift from diffuse to highly structured and back again, suggesting a transient regulatory bottleneck. Perturbation simulations further support this view: gene deletions at the phylotypic stage result in larger shifts in embedding space, suggesting greater sensitivity and constraint. Together, these results offer quantitative support for the idea that development is most fragile and most organized at its conserved midpoint.

Our attention-based network analysis also reveals functional transitions. Genes that both influence and are influenced ("common" nodes) are enriched for transcriptional and epigenetic regulation early on, but become more specialized or structural in later stages (Supplement, Table S3). This temporal reorganization of regulatory influence is consistent with developmental canalization, where early regulatory factors set the stage for downstream diversification [26, 27].

Unlike foundational models trained across many tissues and organisms (e.g., scBERT [13], Geneformer [14, 15], scGPT [16], and scFoundation [17], Supplement, Table S4), Zebraformer is tightly scoped to zebrafish embryogenesis. While this limits generalizability, it enables finer resolution and more interpretable embeddings, revealing patterns that might be diluted in broader corpora. Our aim is not universality, but biological clarity, a model that reflects the temporal and contextual logic of development.

This work also highlights the potential for LLMs as hypothesis-generating frameworks in biology. Rather than predicting annotations or clustering cells, we use the model to infer structure, both in the gene embedding space and in emergent networks. This mirrors a shift in the field, where transformer-based models are increasingly applied to uncover relational structure in biological systems [28]. Intriguingly, the attention mechanism itself may reflect a computation analogous to that of biological networks, context-aware, nonlocal, and dynamically weighted [29].

In sum, our findings demonstrate that a transformer model trained purely on expression context can recapitulate key features of developmental constraint, supporting long-standing theoretical models from a new, data-driven angle. As models like Zebraformer become more integrated with experimental approaches, we anticipate that contextual language models will play a growing role in developmental and systems biology not as black-box predictors, but as interpretable lenses onto the logic of living systems.

Methods

Large language model. We apply masked language modeling (MLM) following the Geneformer approach [14], see also [13]. To this aim, we use the BertForMaskedLM model from Hugging Face's Transformers library [30], a widely adopted toolkit that provides access to pretrained transformer models. BertForMaskedLM is a variant of BERT [31] fine-tuned for MLM, i.e., predicting missing tokens in input text (Supplement, Figs. S6-S8).

Developmental dataset. We downloaded the single cell RNA-seq data of the zebrafish developmental trajectory from Zebrahub (Supplement). We retain cells with at least 100 expressed genes and genes detected in at least 3 cells, excluding those with extensive mitochondrial or noncoding expression. This results in a final dataset of 111,249 cells (acting as "sentences") and 28,677 genes (that will be part of the "tokens", the small units that serve as the input for the LLM).

Rank value encoding and tokenization. Each gene's activity, as captured by the single-cell transcriptome, is represented by a rank value. Genes are ranked by their expression level in each cell normalized by the non-zero median expression of each gene across the full developmental dataset. The protocol implies that specific genes, e.g., transcription factors, and housekeeping genes receive higher or lower prioritization in the rank-based encoding [14]. This ordering is analogous to the word order in a sentence and represents the input provided to the LLM after tokenization. Tokenization is done in terms of a dictionary of length 28,679 (28,677 genes and two special tokens: $\langle$pad$\rangle$ as 0 and $\langle$mask$\rangle$ as 1). The number of genes for each cell is limited to what is called the sentence/sequence length, e.g., sequence length is 2,048 in Geneformer [14] and 512 in standard BERT [31]; we took 512 to limit computational overloading, see below).

Zebraformer training. We used Optuna, an open-source Python library for hyperparameter optimization, especially designed for machine learning and deep learning models [32]. The selected model has an embedding dimension of 384, 2 transformer layers, 3 attention heads and a maximum input size of 512 tokens. For training batch size and learning rate are set with memory constraints in mind. We used AdamW optimizer with weight decay.

Gene and cell embeddings. We select the penultimate hidden layer of the model to extract *gene* embeddings. While the final layer is typically more closely tied to the model's training objective, the penultimate layer tends to be more representation-focused, capturing contextual information in a way that is more transferable across tasks [33]. To compute a representative embedding for each *cell*, we aggregated gene-level embeddings across all genes active in the particular cell, normalized by the number of contributing genes. This yielded a mean embedding vector for each cell that reflects the average of its active gene embeddings (Fig. 1A).

Weighted gene similarities. For each gene i, we computed a similarity score based on the cellular contexts in which it is active and the broad class of cells considered. Specifically, we defined three broad classes based on the following groupings: (i) by timepoint alone, (ii) by anatomical domain within a specific timepoint, or (iii) by cell type within a specific timepoint. Within each class, we further partitioned cells by sibling fish identity. For each sibling group k, we computed the mean pairwise *cosine* similarity μ_i across all pairs of gene embeddings $\mathbf{e}_a$ and $\mathbf{e}_b$, for all cell pairs a and b in the group, where similarity is defined as $s^i_{ab} = (\mathbf{e}^i_a \cdot \mathbf{e}^i_b)/(\|\mathbf{e}^i_a\| \, \|\mathbf{e}^i_b\|)$. We then calculated the weighted average similarity across all sibling groups as $\bar{\mu}^i = \sum_k n_k \mu_k / \sum_k n_k$, where n_k is the number of contributing cells in each group, and $\sum_k n_k$ gives the total number of contributing cells (Fig. 1A). To ensure robust estimates, we included only genes that were active in more than four cells (on average) in at least three different fish, and had a pooled standard deviation $s_{\text{pooled}} < 0.05$, computed by combining within-fish variance and between-fish deviation from the overall mean. Moreover, to quantify the average of all gene similarity scores, we alternatively use the median as a descriptor of the distribution when they are not normal (Figs. 1–2).

Weighted cell similarities. To quantify transcriptional consistency within individual fish over developmental time, we computed pairwise cosine similarity scores between cell embeddings for each fish and timepoint. For each class, we recorded both the number of cells and the mean pairwise similarity. As in the gene-level analysis, we aggregated *per*-fish statistics by computing a weighted average similarity. These aggregated values enable longitudinal comparisons across developmental stages.

Normalized enrichment score. We perform gene set enrichment analysis to assess whether gene ontology (GO) terms related to development and housekeeping functions are associated with stronger gene embedding similarity as time progresses. To this end, we rank the gene similarity scores for each timepoint and test whether a given biological process term (GO Biological Process 2021 database) is concentrated near the top (strong similarity) or bottom (weak similarity) of the list using the official Python 3 interface to the g:Profiler toolkit [34]. This is quantified by the normalized enrichment score (NES), with positive values indicating enrichment. We select specific developmental or housekeeping terms (Fig. S1) and compute their mean NES and standard deviation of the top 10 instances. These values are shown in Fig. 1B.

Similarities and anatomical domais. Each gene can be active in a number of cells associated with an anatomical term (e.g., the Jun dimerization protein 2b, jdp2b, is active in the central nervous system related cells at timepoint 10hpf). We take the mean and sd of the gene embeddings similarities on such set and compute again a combined (weighted) mean and pooled sd as before (considering the number of cells on each fish; and the filtering for robust estimates discussed above). For each anatomical domain, we calculate a robust trend analysis using Theil–Sen regression by considering timepoints (categorical timepoints encoded as integers, e.g., "24hpf" $\rightarrow$ 0, "48hpf" $\rightarrow$ 1, etc.,) and the associated median

similarity values for those timepoints.

Attention matrices. We computed attention scores for each of the three developmental stages described by the hourglass framework. The early stage includes data from cells collected at 10hpf to 14hpf; the phylotypic (middle) stage includes 16hpf and 19hpf; and the late stage includes 24hpf to 10dpf. For each stage, we computed how much attention each gene gives to other genes. This attention is aggregated across all corresponding sampled cells, attention heads in the model, and the top 5 most-attended target genes per source gene. The result is a combined, simplified gene–gene attention map that highlights genes which consistently appear among the top attention targets. While this provides a useful approximation, more detailed analyses could take into account attention patterns specific to individual heads, use averaging rather than summing, or apply different weights to different attention heads.

In silico perturbation. To quantify the impact of individual gene perturbations, we applied the following *in silico* perturbation protocol: For each cell, we began with its sequence of ranked genes and randomly "delete" one by removing it from the ranked list. Using this modified sequence, we computed updated gene embeddings using the pretrained model. For each gene in the perturbed sequence, we calculated the cosine distance (1-cosine similarity) between its original (*wild-type*) and perturbed embeddings. These distances were then aggregated across all genes in the sequence to generate a perturbation score for that specific mutation. To obtain a robust estimate of perturbation sensitivity for each cell, we repeated this process 25 times, each time deleting a different gene, and averaged the resulting aggregate scores. This final value represents the mean perturbation impact for that cell. This procedure was applied independently to 10,000 cells, sampled across hourglass developmental timepoints and cell types.

Acknowledgements

This work was supported by grant PID2023-151289NB-I00 funded by MICIU/AEI/10.13039/501100011033 and by "ERDF/EU". We thank Joaquin Torres for discussions on LLMs, Merlin Lange and Yang-Joon Kim for assistance with the Zebrahub dataset, and the Computational Systems Biology Group at the Spanish National Biotechnology Center for providing access to their computational resources. This work was done with a single GPU unit.

Code availability

Code available at github.com/juanfpoyatos/Zebraformer. The author used OpenAI's ChatGPT to assist with text improvement and code debugging. The author reviewed and edited all content and is fully responsible for the final manuscript.

References

1. Rudolf A. Raff. *The Shape of Life: Genes, Development, and the Evolution of Animal Form*. University of Chicago Press, Chicago, IL, 2000.

2. Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

3. Linlin Tian, Fei Chen, and Evan Z. Macosko. The expanding vistas of spatial transcriptomics. *Nature Biotechnology*, 41:773–782, 2023.

4. Jan Huisken, Jim Swoger, Filippo Del Bene, Joachim Wittbrodt, and Ernst H. K. Stelzer. Optical sectioning deep inside live embryos by selective plane illumination microscopy. *Science*, 305(5686):1007–1009, 2004.

5. J.E. Sulston, E. Schierenberg, J.G. White, and J.N. Thomson. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental Biology*, 100(1):64–119, 1983.

6. Merlin Lange, Alejandro Granados, Shruthi VijayKumar, Jordão Bragantini, Sarah Ancheta, Yang-Joon Kim, Sreejith Santhosh, Michael Borja, Hirofumi Kobayashi, Erin McGeever, Ahmet Can Solak, Bin Yang, Xiang Zhao, Yang Liu, Angela M. Detweiler, Sheryl Paul, Ilan Theodoro, Honey Mekonen, Chris Charlton, Tiger Lao, Rachel Banks, Sheng Xiao, Adrian Jacobo, Keir Balla, Kyle Awayan, Samuel D'Souza, Robert Haase, Alexandre Dizeux, Olivier Pourquie, Rafael Gómez-Sjöberg, Greg Huber, Mattia Serra, Norma Neff, Angela Oliveira Pisco, and Loïc A. Royer. A multimodal zebrafish developmental atlas reveals the state-transition dynamics of late-vertebrate pluripotent axial progenitors. *Cell*, 187(23):6742–6759.e17, 2024.

7. Charles B. Kimmel, William W. Ballard, Sarah R. Kimmel, Bonnie Ullmann, and Thomas F. Schilling. Stages of embryonic development of the zebrafish. *Developmental Dynamics*, 203(3):253–310, 1995.

8. Eric H. Davidson. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*. Academic Press, Burlington, MA, 2nd edition, 2006.

9. Jonathan MW Slack, Peter WH Holland, and Charles F Graham. The zootype and the phylotypic stage. *Nature*, 361:490–492, 1993.

10. Denis Duboule. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Development: Supplement*, pages 135–142, 1994.

11. Melanie Mitchell and David C. Krakauer. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

12. Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition, 2025.

13. Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

14. Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

15. Han Chen, Madhavan S. Venkatesh, Javier Gómez Ortega, Siddharth V. Mahesh, Tarak N. Nandi, Ravi K. Madduri, Karin Pelka, and Christina V. Theodoris. Quantized multi-task learning for context-specific representations of gene network dynamics. *bioRxiv*, 2024.

16. Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024.

17. Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21:1481–1491, 2024.

18. Artur Szałata, Karin Hrovatin, Sören Becker, Alejandro Tejada-Lapuerta, Haotian Cui, Bo Wang, and Fabian J. Theis. Transformers in single-cell omics: a review and new perspectives. *Nature Methods*, 21(8):1430–1443, 2024.

19. Lewis Wolpert, Cheryll Tickle, Alfonso Martinez Arias, Peter Lawrence, and James Locke. *Principles of Development*. Oxford University Press, Oxford, UK, 6th edition, 2023.

20. Lindsey M. F. Mao, Erin A. T. Boyle Anderson, and Robert K. Ho. Anterior lateral plate mesoderm gives rise to multiple tissues and requires tbx5a function in left-right asymmetry, migration dynamics, and cell specification of late-addition cardiac cells. *Developmental Biology*, 472:52–66, 2021.

21. Marianne E. Bronner and Nicole M. LeDouarin. Development and evolution of the neural crest: an overview. *Developmental Biology*, 366(1):2–9, 2012.

22. Karin D. Prummel, Susan Nieuwenhuize, and Christian Mosimann. The lateral plate mesoderm. *Development*, 147(12):dev175059, 2020.

23. Eirini Laurenti and Berthold Göttgens. From haematopoietic stem cells to complex differentiation landscapes. *Nature*, 553(7689):418–426, 2018.

24. Kai Schmidt and J. Matthias Starck. Developmental variability during early embryonic development of zebra fish, *Danio rerio*. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 302(5):446–457, 2004.

25. Frietson Galis and Johan A. J. Metz. Testing the vulnerability of the phylotypic stage: on modularity and evolutionary conservation. *Proceedings: Biological Sciences*, 268(1477):181–188, 2001.

26. Günter P. Wagner. The developmental genetics of homology. *Nature Reviews Genetics*, 8(6):473–479, 2007.

27. James DiFrisco, Alan C. Love, and Günter P. Wagner. Character identity mechanisms: A conceptual model for comparative–mechanistic biology. *Biology and Philosophy*, 35(4):1–32, 2020.

28. Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723), 2024.

29. Leo Kozachkov, Ksenia V Kastanenka, and Dmitry Krotov. Building transformers from neurons and astrocytes. *Proceedings of the National Academy of Sciences*, 120(19):e2219150120, 2023.

30. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, 2020.

31. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1*, pages 4171–4186. Association for Computational Linguistics, 2019.

32. Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

33. Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 55–65. Association for Computational Linguistics, 2019.

34. Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, 51(W1):W207–W212, 2023.