Capturing Context in Organismal Development with a Large Language Model

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

Abstract

Development, the process by which cells differentiate into specific fates to build the adult organism, is arguably the most context-dependent phenomenon in biology, shaped by constant interactions among neighboring cells and signals. Here, we investigate how a large language model (LLM) can leverage such contextuality to capture the dynamics of cellular differentiation. Using single-cell RNA activity data from developing zebrafish embryos, we train a domain-specific LLM, Zebraformer. We demonstrate that the gene and cell embeddings, numerical representations encoding biological meaning, produced by Zebraformer reflect the progression of zebrafish organismal formation, including the emergence of major anatomical domains and cell types. Furthermore, by using the model's attention matrices, which capture how strongly genes influence each other, we observe patterns consistent with the developmental hourglass framework. This framework proposes that the middle phase of embryogenesis is more conserved across species than the early or late phases. At this phylotypic stage, the attention matrices reveal a rise in gene regulatory network complexity, reflected in increased node and edge counts and a narrowing-then-broadening network topology. This stage also shows greater sensitivity to genetic perturbations, as indicated by shifts in gene embeddings, pointing to heightened constraint consistent with the hourglass assumption. Our findings show that language models, when properly adapted, can offer powerful new ways to represent and understand the intricate choreography that enables cells to coordinate and build a living organism.

Introduction

Among the various processes that define living systems, organismal development is perhaps where the contextual nature of biology is most evident. By "contextual nature," we refer to the dynamic, situation-specific molecular and cellular events that arise in response to intrinsic genetic programs, environmental cues, and the changing developmental stages of the organism [1].

To study these context-dependent effects, we now benefit from technological advances that allow increasingly precise observation and analysis. Most notably, the emergence of single-cell technologies combined with high-resolution imaging methods offers an unprecedented view into how cellular lineages organize into functional structures –tissues, organs, and, ultimately, entire organisms– under varying conditions [2, 3]. These innovations mark a turning point, enabling us to systematically dissect the influence of context on development in ways that were previously only possible in landmark efforts, such as the cell-by-cell lineage tracing in the nematode *Caenorhabditis elegans*, which laid the groundwork for modern developmental biology [4].

A case study illustrating these new approaches is the recent zebrafish (*Danio rerio*) multimodal developmental atlas by Lange *et al.* [5]. In this work, four fish were sampled at each of ten developmental stages, and their gene expression was profiled using RNA sequencing (RNA-seq) [6]. These stages range from the end of gastrulation (10 hours post-fertilization, hpf, when the embryo transitions from a simple ball of cells to a layered structure that will give rise to all organs and tissues) to 10 days post-fertilization (dpf), covering the emergence of organ systems [7]. The transcriptomic data were complemented by cellular-resolution imaging, together providing insight into cell differentiation trajectories.

Our focus here will not primarily be on such developmental trajectories, but rather on exploring a few specific questions related to the contextual aspects of cellular differentiation. In this light, it is natural to ask to what extent a gene's "action" is shaped by the identity of the cell in which it is expressed and how this influence changes over the course of development. We can also ask whether, by integrating these gene actions, we can distinguish individual cells and whether such cellular individuality becomes more pronounced as the organism matures, as one might expect. Finally, we hope to gain insight into how these gene actions interact with one another, constituting what is commonly referred to as a gene regulatory network (GRN) [8]. We examine GRNs in reference to the *hourglass* framework of organismal formation. The framework suggests that this process tend to diverge between species – belonging to the same *phylum*, or major group sharing a common body plan– at the early and late stages, but converge during a highly conserved middle phase known as the phylotypic stage [1, 9, 10]. We evaluate two key aspects of this framework: first, the complexity of the corresponding GRN and second, the sensitivity to genetic perturbations.

While our work builds on long-standing questions in developmental biology, its novelty lies in how we redefine the concept of gene "action." Instead of viewing gene action as a somehow static measure –such as a gene's expression level– we propose a dynamic interpretation: gene action is represented as a position in a high-dimensional space, derived from a domain-

specific large language model (LLM) we call `Zebraformer` (Figure 1A). LLMs, originally developed for natural language processing, use transformer architectures to learn word meaning from contextual relationships within sentences [11]. These models now represent a catalytic advance in artificial intelligence: not only do they challenge our understanding of how learning systems work [12], but they are also driving innovation across many scientific disciplines, e.g., [13].

In our case, Zebraformer applies this idea to genes, modeling their behavior based on their context within developmental processes. By leveraging the transformer architecture's exceptional sensitivity to relational patterns [14–19], we generate rich, dynamic representations of gene activity that capture how genes respond to and vary with the developmental state of the cell. This approach is designed to address the questions posed above (for more background, see the Supplement for details on zebrafish development, the hourglass framework, transformers, and LLMs in single-cell data).

Results

Zebrafish developmental progression is captured by gene embeddings

We first grouped cells that were collected at the same developmental timepoint (e.g., 10 hpf, 12 hpf, etc.) and derived from the same sibling fish [5], forming distinct classes. For each gene that is active in a cell within these classes, Zebraformer generates an embedding –a vector that represents the gene's *contextual* activity in that specific cellular environment. If these embeddings are similar across cells, it suggests that the gene shows consistent behavior, either in its expression pattern or functional role. To quantify this, we compute the average of all pairwise similarity scores between the gene's embeddings across the group. We then calculate a weighted average of the scores obtained from each sibling fish, resulting in a single combined mean similarity score *per* gene (Fig. 1A, Methods).

Which genes exhibit more consistent behavior across cells at a specific timepoint, and which biological processes are they enriched for? We ranked genes by their combined mean similarity scores, where a higher score indicates more consistent behavior across the cell class. Using this ranked list, we applied gene set enrichment analysis to determine whether genes involved in particular biological processes –as defined by Gene Ontology (GO) terms– are disproportionately represented among the most consistent genes. We focused on two complementary categories of biological function: developmental processes (e.g., differentiation, morphogenesis) and housekeeping functions (e.g., transcription, metabolism). For each category, we selected the top 10 most enriched GO terms and calculated the average and standard deviation of their normalized enrichment scores (NES, Methods, Fig. S1).

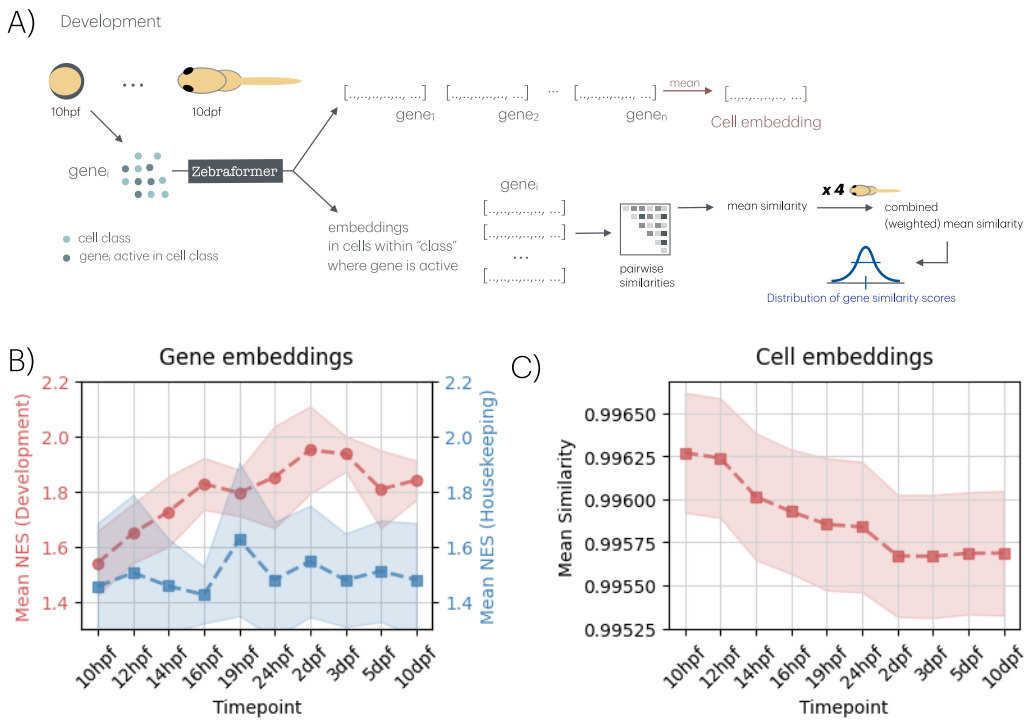

Figure 1: Gene and cell embeddings capture the zebrafish developmental axis. A) Top. For each developmental stage, Zebraformer generates embeddings for genes active in the selected class of cells (dark cells). Cell embeddings are computed as the mean of the embeddings of active genes within each cell. Bottom. For a given gene active in multiple selected cells, we compute pairwise similarities between its embeddings across cells, then calculate a weighted mean using the four available sibling values. Ultimately, we obtain a distribution of these scores for all genes. B) Mean normalized enrichment scores (NES) for the top 10 developmental (red) and housekeeping (blue) GO terms across timepoints, based on gene set enrichment analysis of genes ranked by embedding similarity. Dots show the mean NES; shaded areas indicate standard deviation (sd). Developmental terms show increasing enrichment, while housekeeping terms remain relatively stable. C) Mean (and sd) similarity of cell embeddings computed by averaging the embeddings of active genes per cell across sibling cells at each timepoint. Cell–cell similarity decreases with developmental time, reflecting progressive transcriptional divergence and the emergence of specialized cell identities. See main text and Methods for details.

As shown in Fig. 1B, developmental GO terms (red) exhibit progressively higher enrichment scores, peaking around 2 dpf when the body plan is established (Supplement, Table S1), suggesting that genes with consistent behavior are increasingly associated with core developmental programs. In contrast, housekeeping GO terms (blue) remain relatively stable and lower in enrichment throughout, indicating less association between cell-to-cell consistency and general maintenance functions. This trend supports the idea that increasing similarity among developmentally important genes reflects coordinated functional programs during embryogenesis.

Zebrafish development is also reflected in cell embeddings

We can instead ask about the overall transcriptional state of individual cells that we can score by means of the *cell* embedding [15]. These embeddings are computed by averaging those embeddings of all genes active in each cell, thus capturing a summary of the cell's functional identity (Fig. 1A). As with gene embeddings, we calculated pairwise similarities between cell embeddings for each fish at a given developmental timepoint, followed by a combined similarity measure across siblings (Methods). Notably, these cell–cell similarities decrease over developmental time (Fig. 1C). This trend suggests that as development proceeds, cells increasingly diverge in their transcriptional programs, reflecting the emergence of distinct cell types and lineages. This supports the idea that embedding-based similarity not only captures shared function among genes, but also highlights developmental divergence at the cellular level.

Gene embeddings encode features associated with major anatomical domains.

Beyond the previous "global" perspective, we now examine the extent to which gene embeddings encode biological information on major anatomical domains. Domains refer to *embryonic tissues* or *cell lineages* in zebrafish (and vertebrates generally [20]) and explicitly include in the considered dataset examples like the the endoderm (one of the three germ layers along with mesoderm and ectoderm) or the neural crest (a multipotent, migratory cell population arising at the border of neural and non-neural ectoderm, Supplement, Table S1).

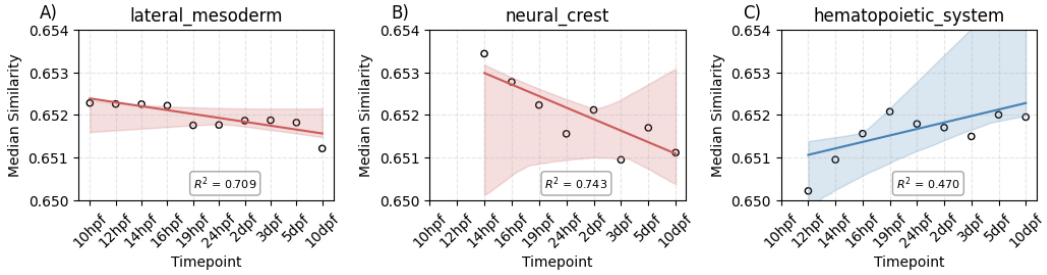

Figure 2: Gene embeddings and context across anatomical domains. Regression trends of median gene embedding similarity per anatomy class (some of them emerge later than others, Suppplement), with shaded regions indicating bootstrapped uncertainty (10,000 samples). For each gene, similarity scores are computed across cells where the gene is active; the median similarity across all such genes in a given anatomical domain and timepoint reflects the overall contextual consistency of gene representations. The lateral mesoderm (A) shows slight decrease in similarity over time, with low uncertainty and a strong fit ($R^2 = 0.709$). The neural crest (B) also decreases, a bit more steeply but showing greater uncertainty with still a strong fit ($R^2 = 0.743$). The hematopoietic system (C) increases similarity over time with moderate uncertainty and a weaker fit ($R^2 = 0.470$).

We thus computed the embedding similarity of genes active in cells associated with specific anatomical domains, which now serve as the new cell classes. As before, we pooled the data for each fish and timepoint to obtain a distribution of gene similarities. We then asked whether the central value of this distribution changes across developmental stages (Methods). There are indeed some cases where this is so. Figure 2 shows three exemplary

cases (full set in Fig. S2). Lateral mesoderm (LM), neural crest (NC), and the hematopoietic system (HS) show robust tendencies to decrease or increase (hematopoietic), respectively the consistency of gene embedding representations.

This appears to reflect the underlying 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 [21]. 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 [22], and heart, kidney, and vessels for LM [23]. Their development follows a spatiotemporal model, 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 [24].

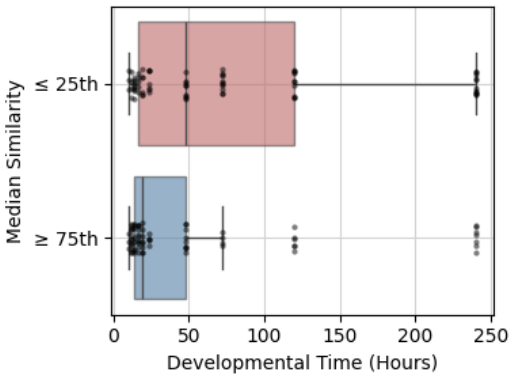

Figure 3: Cell types with lower average gene embedding similarity tend to emerge later in development. For each gene active in cells of a given type and timepoint, we compute its embedding similarity across fish, then take the median across genes to obtain an average gene embedding similarity *per* cell. Cell types in the upper ($\geq$ 75th percentile) and lower ($\leq$ 25th percentile) quartiles of this distribution are compared. Box plot shows developmental time (hours post-fertilization) for these two groups; cell types with higher similarity appear earlier, while those with lower similarity emerge later (Mann–Whitney U, $p = 0.00676$).

Zebraformer records a gradient of contextual diversity

Finally, we refined our analysis by focusing on a more specific definition of cell class: individual cell types at each developmental timepoint (there are 154 types in the dataset). For each gene active in these fine-grained classes, we again computed the combined gene embedding similarity –averaged across sibling fish as before– and then took the average (median) similarity across genes within each cell type at a given timepoint (Methods). This score reflects how contextually consistent genes are within a given cell type: in other words, how similarly genes behave across cells of the same type and stage. Analyzing the distribution of these average similarity scores across all cell types, we observe a manifest

pattern: cell types with lower similarity scores (below the first quartile) tend to emerge later in development, whereas those with higher scores (above the third quartile) are associated with earlier developmental stages (Fig. 3, Fig. S3).

This suggests that early cell types express more "foundational" genes, genes that are reused across many early-stage contexts, leading to more consistent embedding patterns and thus higher similarity. In contrast, later-emerging cell types are more specialized, expressing genes in more unique or variable combinations, including tissue-specific or temporally restricted genes. This increased diversity in gene context results in lower similarity scores, as the same gene may behave differently across later cell types. This confirms the pattern observed before with cell embeddings.

Attention-based GRNs reproduce the hourglass developmental framework

Can we go beyond the relational information captured by embeddings? A key advantage of using LLMs is their ability to identify gene-gene relationships through what are called *attention matrices*, a mechanism originally designed to capture associations between words in a sentence [11]. In our context, these matrices reveal how strongly the model associates different genes with one another. We use this tool to assess how well Zebraformer captures the hourglass framework of development [1, 9, 10] (Supplement). Thus, we defined three developmental periods: (i) early embryogenesis (up to 14 hpf in zebrafish), (ii) middle embryogenesis (the phylotypic stage, approximately 15–19 hpf [25]), and (iii) late embryogenesis (from 24 hpf to 10 dpf). We first evaluate a key aspect of the framework: the complexity of the GRNs. To assess changes in regulatory complexity across stages, we compute an attention-based GRN for each stage and analyze its structural properties (Methods).

Across developmental time, GRNs grow substantially in both size and connectivity as one might expect given the increasing complexity of the organism (Fig. 4A, Fig. S4). Beyond this basic expansion, several more nuanced network properties reveal informative patterns (Table S2). We first quantify the likelihood that groups of genes interact closely (e.g., as part of the same biological pathway or developmental module), finding that the so-called clustering coefficient increases during the late developmental stages (Fig. 4B). This suggests tighter, more localized gene coordination. We can also evaluate whether genes with many connections tend to link to similarly connected genes (positive "assortativity") or to less-connected ones (negative). Assortativity becomes progressively less negative what also confirms the hourglass framework (Supplement).

Moreover, modularity, a measure of how well the network separates into distinct communities, reaches its highest value during the phylotypic stage, consistent with the idea of tightly coordinated, functional regulatory programs during this conserved developmental window. Analysis of power-law behavior further shows that the phylotypic stage network contains more prominent regulatory hubs, reflecting a centralized and possibly more constrained regulatory regime (Fig. 4B, Supplement). Altogether, these observations align with the developmental hourglass notion: early and late stages exhibit more flexible and dispersed regulatory architectures, while the middle (phylotypic) stage is marked by high structural organization and control, a phase likely essential for robust and conserved developmental

outcomes.

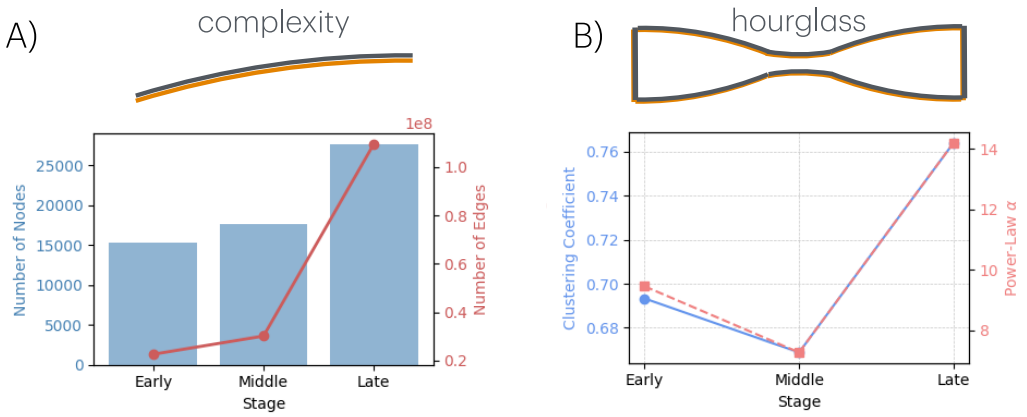

Figure 4: Attention-based gene regulatory networks and their evolution across developmental stages of the hourglass framework. A) Complexity: The number of nodes (bars, left y-axis) and edges (line, right y-axis) increase from early to late stages, indicating rising network complexity over time. B) Hourglass structure: Clustering coefficient (blue line, left y-axis) initially declines and then increases, while the power-law exponent α (red line, right y-axis; α denotes the shape of the degree distribution with lower values indicating a heavier tail, i.e., more hubs) sharply rises at the late stage, reflecting a narrowing and subsequent expansion in network connectivity consistent with an hourglass framework.

Sensitivity to perturbations also exhibit key features of the hourglass architecture.

A second important feature of the hourglass architecture is how differently each developmental stage responds to disruptions, with the middle (phylotypic) one being the most sensitive [26]. To explore this, we use Zebraformer to simulate genetic perturbations using the embedding signals. Specifically, we delete a ranked gene in a given cell and then run the *altered* cell through the model to generate new gene embeddings. We then measure the average difference between these embeddings and those from the unaltered ("wild-type") cell, providing a quantitative view of the impact of the perturbation [15].

We analyzed 10,000 randomly selected cells from each hourglass stage and simulated 25 *individual* gene deletions *per* cell. To quantify the impact of each perturbation on the embedding space, we computed the cosine distance between the original and perturbed gene embeddings –where a larger distance indicates a greater effect (Methods). Figure 5 displays the Kernel Density Estimate (KDE) of the distribution of mean cosine distances resulting from the *in silico* perturbations, comparing the three developmental stages of the hourglass framework.

The phylotypic stage has a higher average distance, suggesting that perturbations in this stage have stronger effects on the model's embeddings. The early and late stages show slightly lower sensitivity (lower mean distances), consistent with the hourglass hypothesis that the middle stage is the most constrained or information-dense (distribution spread is comparable, indicating similar variability in perturbation effects across stages, Fig. S5). This results supports a core prediction of the hourglass model of development: that the phylotypic

(middle) stage is the most constrained.

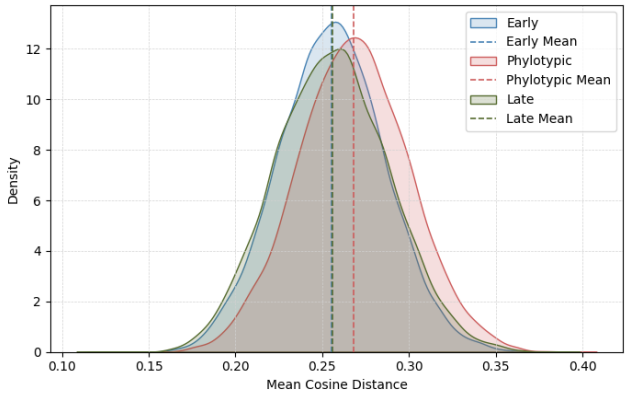

Figure 5: Sensitivity to gene perturbations across the hourglass developmental stages quantified by embedding changes. Kernel density estimates of $10,000$ mean cosine distances between *wild-type* and *perturbed* gene embeddings are shown for early (blue), middle/phylotypic (red), and late (green) stages (see main text for details). Dashed vertical lines indicate the mean distance for each stage (early: 0.256 ± 0.031, middle: 0.268 ± 0.032, late: 0.256 ± 0.033). The phylotypic stage shows a higher mean perturbation effect, reflecting increased sensitivity and constraint consistent with the hourglass model of development. Early *vs.* Phylotypic: Kolmogorov–Smirnov (KS) p-value: 6.7741×10^{-116}; Phylotypic *vs* Late: KS p-value: 1.2817×10^{-102}.

Discussion

We focused on development as a dynamic, relational process well suited to LLMs, which learn meaning from interactions between elements rather than isolated inputs. Our model, Zebraformer, processes gene activity within individual cells to uncover how it relates to the surrounding cellular environment. Trained on single-cell RNA-seq data from zebrafish embryogenesis [5], Zebraformer produces gene representations and uses attention mechanisms to highlight relationships we interpret as GRNs [8].

Our analysis identifies two main aspects of this learning. First, it captures differentiation along the developmental axis, showing increasing specificity in gene representations across time and tissue. Using embedding similarity across developmental stages, anatomical domains, and cell types, the model consistently reconstructs aspects of organismal development from transcriptomic patterns, Figs. 1, 2, and 3. Second, we tested whether the model captures known features of gene-gene interaction by focusing on the hourglass framework of development. In particular, we asked whether Zebraformer could identify the unique regulatory properties of the *phylotypic* stage, which is thought to represent a highly conserved and constrained period in vertebrate development [1, 9, 10].

The learned network features support this idea: the middle stage is shaped by a core set of pleiotropic regulatory genes conceptually linked to the idea of character identity networks [27, 28]. These genes are highly connected, tightly constrained, and central to body

plan specification, Fig. 4 (see below). This is further supported by perturbation tests across developmental stages, showing that disruptions during the phylotypic stage are more likely to be harmful, Fig. 5, as shown in previous studies [26].

Complementing this, we analyzed attention scores across the three hourglass stages. Genes were categorized by whether they predominantly influence others, are influenced, or do both (common). Enrichment analysis of these groups showed that early and middle-stage "common" genes are enriched for epigenetic and transcriptional regulation. The "influencing" genes shift from broad to specialized regulators, while "influenced" genes evolve from general transcriptional targets to structural effectors. The middle stage stands out for its cohesive regulatory signature, again aligning with the hourglass hypothesis, where development is most constrained and centrally orchestrated (Supplement, Table S3).

While our model is limited in data dimensionality and scope, it demonstrates a proof of concept: LLMs can extract meaningful developmental structure from raw single-cell data. At the same time, our focused approach contrasts with the foundation models recently developed for more generalized single-cell tasks. Models such as scBERT [14], Geneformer [15, 16], scGPT [17], and scFoundation [18] are trained on tens of millions of human single-cell profiles across diverse tissues and conditions. These models aim for broad applicability across annotation, trajectory inference, perturbation prediction, and more (Supplement, Table S4). In contrast, Zebraformer demonstrates how a specialized model, trained specifically on developmental data, can reveal fine-grained temporal and regulatory patterns that might be diluted in large, pan-tissue corpora, an approach inspired in part by earlier work in natural language studies, e.g., [29].

Overall, this work reflects a broader trend emerging across biology, where transformer-based models are increasingly used to uncover patterns that conventional analyses often miss (interestingly, biological networks themselves could be implementing transformer-like computations, e.g., [30]). We believe developmental biology is particularly well-suited to benefit from these approaches, precisely because of its deep reliance on context, timing, and hierarchical regulation. As such, we expect that many others will follow this path contributing to a new synthesis of computational and developmental science.

Methods

Large language model. We apply masked language modeling (MLM) following the Geneformer approach [15], see also [14]. To this aim, we use the BertForMaskedLM model from Hugging Face's Transformers library [31], a widely adopted toolkit that provides access to pretrained transformer models. BertForMaskedLM is a variant of BERT [32] 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 [15]. 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 [15] and 512 in standard BERT [32]; 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 [33]. 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 [34]. 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 [35]. 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. Linlin Tian, Fei Chen, and Evan Z. Macosko. The expanding vistas of spatial transcriptomics. *Nature Biotechnology*, 41:773–782, 2023.

3. 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.

4. 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.

5. 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.

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

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. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

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

21. 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.

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

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

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

25. 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.

26. 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.

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

28. 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.

29. Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.