Diversity-repeatability trade-off governs hierarchical cell fate coding in multicellular development

Authors: Chunxiuzi Liu, Yu Liu, Guoye Guan, Yu Liu, Guoye Guan

Date: 2025-12-08T00:00:00+00:00

Abstract

Multicellular development in plants and animals gives rise to a rich spectrum of cell fates performing complex physiological functions. How these cells differentiate in concert from a single-celled zygote to collectively form tissues and organs across the whole body remains a fundamental question in developmental biology. In this study, we applied the Ladderpath approach to decompose the spatially-aligned cell fate sequence of worm *Caenorhabditis elegans* embryo, which produces 671 uniquely identifiable cells with defined fates (*incl.* germline, intestine, muscle, neuron, pharynx, skin, and other). This approach automatically dissects how sub-sequence modules (Ladderons) self-repeat and mutually assemble in a hierarchical architecture, uncovering a delicate trade-off between diversity and repeatability in cell fate coding. This architecture consisting of highly diverse yet repeated Ladderons indicates that the actual cell fate sequence deviates substantially from both homogeneous and heterogeneous extremes, achieving an exceptional level of hierarchical complexity near the theoretical maximum. Genetic-algorithm-based virtual evolution further reveals such complexity as an optimization goal in development, with pseudo sequences spontaneously converging toward similar hierarchical architectures designed through moderate Ladderon numbers and lengths. Notably, the longest Ladderons highlight intercellular Notch signaling as a key mechanism for enhancing hierarchical complexity via coordinating lineal differentiation. More broadly, the repeated lineal differentiation programs serve as an essential strategy for enhancing hierarchical complexity —a pattern recurrently observed in reality. Together, this work offers new insights into design rules of cell fate coding operating from molecular to multicellular scales and provides a theoretical framework for identifying and interpreting the regulatory mechanisms beneath.

Full Text

Diversity-Repeatability Trade-Off Governs Hierarchical Cell Fate Coding in Multicellular Development

Chunxiuzi Liu^{a,b,c}, **Yu Liu**^{a,b,1}, and **Guoye Guan**^{d,e,1}

This manuscript was compiled on December 7, 2025

Multicellular development in plants and animals gives rise to a rich spectrum of cell fates performing complex physiological functions. How these cells differentiate in concert from a single-celled zygote to collectively form tissues and organs across the whole body remains a fundamental question in developmental biology. In this study, we applied the Ladderpath approach to decompose the spatially-aligned cell fate sequence of the nematode *Caenorhabditis elegans* embryo, which produces 671 uniquely identifiable cells with defined fates (including intestine, muscle, neuron, germline, pharynx, skin, and others). This approach automatically dissects how sub-sequence modules (Ladderons) self-repeat and mutually assemble in a hierarchical architecture, uncovering a delicate trade-off between diversity and repeatability in cell fate coding. This architecture consisting of highly diverse yet repeated Ladderons indicates that the actual cell fate sequence deviates substantially from both homogeneous and heterogeneous extremes, achieving an exceptional level of hierarchical complexity near the theoretical maximum. Genetic-algorithm-based virtual evolution further reveals such complexity as an optimization goal in development, with pseudo sequences spontaneously converging toward similar hierarchical architectures designed through moderate Ladderon numbers and lengths. Notably, the longest Ladderons highlight intercellular Notch signaling as a key mechanism for enhancing hierarchical complexity via coordinating lineal differentiation. More broadly, the repeated lineal differentiation programs serve as an essential strategy for enhancing hierarchical complexity—a pattern recurrently observed in reality. Together, this work offers new insights into design rules of cell fate coding operating from molecular to multicellular scales and provides a theoretical framework for identifying and interpreting the regulatory mechanisms beneath.

Significance

Multicellular life encodes a variety of cell fates—intestinal, muscular, neuronal, pharyngeal, hypodermal, etc. They all arise from a single zygote to form a physiologically functional body with multiple tissues and organs. How these complex cell fates are organized in concert is a fundamental question in developmental biology. In this study, the fate sequence of all terminal cells in the *C. elegans* embryo is decomposed, highlighting two complementary organizing modes: the mutual assembling of sub-sequence modules and their self-repeating. This diversity–repeatability trade-off yields a hierarchical architecture of these modules approaching the theoretical maximum in complexity, suggesting an outcome of evolutionary optimization. Such exceptional complexity emerges when

the organism repeats the same differentiation programs across lineages (e.g., via intercellular signaling).

Keywords: hierarchical complexity | evolutionary optimization | diversity-repeatability trade-off | cell fate | differentiation program

Life is encoded in a sequence-based manner, with complexity across multiple scales: from the associated DNA (deoxyribonucleic acid encoded with four types of nucleotide bases) [**?**], RNA (ribonucleic acid with four types of ribonucleoside bases) [**?**], and protein (polypeptide encoded with 20 types of amino acids) [**?**, **?**, **?**] to their joint molecular profiles differentiated in cells, which encode spatial sequences of cell types or cell fates (hereafter uniformly referred to as "cell fate") with specific physiological roles [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. How a pool of elements (e.g., nucleotide bases, ribonucleoside bases, amino acids, and cell fates) are organized into specific sequences within an immense design space to achieve desired functions is not only an experimental question but also a theoretical one.

The discovery of molecular-scale organization has faithfully demonstrated how biological sequences are designed to deviate from the extremes of homogeneity (repeatability) and heterogeneity (diversity) in order to achieve complexity; for example, promoter sub-sequences are repeatedly encoded in a DNA sequence for different genes (repeatability) [**?**], while other sequence domains also encode potentially repeated exons, introns, start codons, stop codons, open reading frames, other promoters, and everything of the sort (diversity) [**?**].

Author affiliations:
aDepartment of Systems Science, Faculty of Arts and Sciences, Beijing Normal University, Zhuhai 519087, China;
bInternational Academic Center of Complex Systems, Beijing Normal University, Zhuhai 519087, China;
cSchool of Systems Science, Beijing Normal University, Beijing 100875, China;
dDepartment of Systems Biology, Harvard Medical School, Boston 02115, USA;
eDepartment of Data Science, Dana-Farber Cancer Institute, Boston 02215, USA

Author contributions: Please provide details of author contributions here.

Competing interests statement: The authors declare no competing interest.

Correspondence: E-mail: yu.ernest.liu@bnu.edu.cn, guanguoye@gmail.com

After the helical structures of DNA [**?**] and RNA [**?**] were revealed in the mid-20th century, the sequencing of complete genomes—first in the worm (*Caenorhabditis elegans*, *C. elegans*) [**?**] and later in human (*Homo sapiens*, *H. sapiens*) [**?**]—was accomplished around the turn of the 21st century, further revealing a remarkable degree of conservation underlying the shared repertoire of both genes and cell fates [**?**]. Moreover, protein sequences (associated with their DNA/RNA templates' sequences) and structures have been continuously characterized [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**], revealing that proteins fold toward low

free-energy states to form canonical functional secondary and tertiary structures such as α-helices and β-sheets [?, ?]; today, their sequences, structures, and functions can be accurately predicted through molecular dynamics [?, ?, ?], physics-informed deep learning [?, ?], and more general artificial intelligence [?, ?, ?, ?], catalyzed by both experimental validation and applications [?, ?, ?, ?, ?, ?, ?, ?, ?, ?].

Large-scale in silico experiments, particularly virtual evolution, have been instrumental in extracting key features and unveiling design rules underlying biological system optimizations, spanning molecular and multicellular scales and extending even to organismal and behavioral scales [?, ?, ?, ?, ?, ?]. For instance, at the molecular scale, amino acid sequences of polypeptides are optimized to ensure stable protein folding under a range of biophysical constraints, such as thermodynamic stability [?], structural designability [?], hydrophobic demixing [?], structural atypicality [?], kinetic accessibility [?], translational regulation [?], and so forth [?, ?]. Bridging the molecular and cellular/multicellular scales, the molecular interaction networks underlying the worm *C. elegans*' cell polarity and fruitfly *Drosophila melanogaster*' s cell fate, which are sequentially distributed along their body axis, are optimized to stabilize asymmetric interface and maximize positional information (alongside robustness) respectively [?, ?, ?, ?]. These and other cases pinpoint that the complexity of life systems originates from the dynamic interplay between functional demands and biophysical constraints, optimized over billions of years of natural selection.

However, full sequences and design rules for cell fates within multicellular organisms remain relatively under-explored, caused by the difficulty of precisely tracking individual cell fates throughout the whole body and its overwhelming complexity [?, ?, ?, ?]. Nevertheless, such multi-element, digit-like spatial sequences primarily along the anterior-posterior axis are conserved across species from higher to lower levels: (1) During somitogenesis in vertebrates represented by human, mouse, fish, and frog, binary gene expression patterns exhibit a periodic oscillation [?, ?, ?, ?]. (2) Prior to gastrulation of invertebrates represented by fruit fly and beetle, regional gene expression patterns exhibit variable region numbers, positions, and sizes, depending on the combined activations and inhibitions of dozens of genes [?, ?]. (3) Similar cell fate sequences can even realize cellular resolution and individual reproducibility in worm and ascidian –hallmarks of an invariant sequence-based lineage [?, ?] (Fig. 1). However, how sequences encompassing cell fates such as death, germline, intestine, muscle, neuron, pharynx, skin, and others are designed within a complete developing organism remains largely unclear. Unveiling its design rules will not only help us understand fundamental science in terms of biology and physics, but also provide knowledge for the fields of synthetic biology and materials science [?, ?, ?, ?].

Previous decompositions and simulations of cell fate sequences in worm and ascidian embryos revealed that self-repeating modules (corresponding to cellular differentiation programs executed upon cell division that give rise to following

cell fate modules) are favored in certain lineal parts, such as the *C. elegans* ABarapp lineage (1/64 of the complete lineage) [**?**, **?**]. To comprehensively address how cell fates are encoded to form tissues/organs differentiated from a single-celled zygote in a complete multicellular life, this study adopts *C. elegans* as a model organism, considering its well-characterized, cell-resolved differentiation programs and strong conservation with humans in genetics, genomics, and cell fates [**?**].

Decomposition of the full *C. elegans* embryonic cell fate sequence along body axes uncovers a complex hierarchical architecture consisting of diverse yet repeated sub-sequence modules, with a complexity level approaching the theoretical maximum. This exceptional hierarchical complexity is optimized quantitatively and integratively for the entire body rather than tailored to individual body parts. On one hand, in silico experiments suggest that the trade-off between diversity and repeatability among sub-sequence modules is achieved through their moderate numbers and lengths, as well as through the repeated lineal differentiation programs; on the other hand, in vivo experiments suggest that known mechanisms like intercellular Notch signaling contribute to such a trade-off and its resulting exceptional hierarchical complexity.

Together, this work presents design rules of cell fate coding in a complete multicellular life, bridging realistic mechanisms with theoretical interpretation.

Results

Full C. elegans Cell Fate Sequence Exhibits Exceptional Hierarchical Complexity with Diverse, Repeated Sub-Sequence Modules

Building on over forty years of genetic research that has fully mapped the stereotyped cell fate sequence in *C. elegans* embryos, we treat it as a natural developmental blueprint and compare it with random sequences generated artificially [**?**]. After fertilization, the *C. elegans* zygote undergoes four consecutive rounds of cell divisions, producing a lineage tree of 671 terminal cells that no longer divide during embryogenesis [**?**] (Fig. 1, Fig. S1). The cell division orientations are primarily aligned along the anterior-posterior (a-p) axis (89.44%), with fewer aligned along the left-right (l-r) axis (9.91%) and dorsal-ventral (d-v) axis (0.65%). Thus, a systematic nomenclature was devised that appends the letters “a” , “p” , “l” , “r” , “d” , or “v” to each mother cell’ s name, thereby unambiguously naming its daughter cells based on their relative initial positions. This yields a 1$\times$671 one-dimensional cell fate sequence at the end of embryogenesis, with each digit denoting death (abbr., “D” , 16.84%), germline (abbr., “G” , 0.30%), intestine (abbr., “I” , 2.98%), muscle (abbr., “M” , 12.07%), neuron (abbr., “N” , 36.07%), pharynx (abbr., “P” , 11.92%), skin (abbr., “S” , 13.26%), or other (abbr., “O” , 6.56%), ordered by well-documented lineal information (Fig. 1, Fig. S1; Table S1) and positional data [**?**, **?**, **?**].

Here, we apply the Ladderpath approach, a method specially devised to decompose a generic one-dimensional sequence into all self-repeating and mutu-

ally assembling sub-sequence modules in varying lengths (termed "Ladderons" , which together deterministically constitute a Laddergraph that intuitively visualizes their hierarchical architecture) [**?**, **?**, **?**, **?**]. The decomposed sequence is characterized by two key parameters, one previously proposed and one newly introduced (with mathematical details provided in Materials and Methods): (1) "Diversity level" n measures the number of distinct sub-sequence modules; (2) "Repeatability level" ω measures the degree to which a sequence is constituted by repeating rather than diversifying sub-sequence modules [**?**]; (3) "Hierarchical complexity" γ measures the degree to which a complex hierarchical architecture emerges from the integration of both repeating and diversifying sub-sequence modules.

Relatively, an extremely homogeneous sequence like AAAAAAAAAAAAAAAAAAAAAAAAA yields a low n, a high ω, and a low γ; an extremely heterogeneous sequence like ABCDEFGHIJKLMNOPQRSTUVWXY yields a high n, a low ω, and a low γ; most importantly, a sequence integrating both diverse and repeated sub-sequence modules yields a moderate n, a moderate ω, and a high γ, reflecting their complex hierarchical architecture.

Ladderpath decomposition of the full *C. elegans* cell fate sequence outputs a complex, hierarchical Laddergraph (Fig. 2(a)) consisting of 68 sub-sequence modules that both self-repeat and mutually assemble frequently (Table S2). Strikingly, its hierarchical complexity ($\gamma \approx 0.936$) approaches the theoretical maximum ($\gamma = 1$). By comparison, such complexity is gone ($\gamma = 0$) if the sub-sequence modules only self-repeat (extremely homogeneous sequence) or only mutually assemble (extremely heterogeneous sequence) while preserving the element types, numbers, and proportions of *C. elegans* (Table 1). This strongly supports that the *C. elegans* cell fate sequence is designed to achieve an exceptional hierarchical complexity. Note that by definition, each sub-sequence module appears at least twice within the terminal cell fate sequence. Therefore, on one side, self-repeating symbolizes the same differentiation program appearing across close lineage positions—for example, the progenies of the E lineage end as only intestine and the progenies of the D lineage end as only muscle; on the other side, mutually assembling symbolizes different differentiation programs appearing across distinct lineage positions—for example, the progenies of the MS lineage end as death, muscle, pharynx, and other fates (Fig. 1, Fig. S1; Table S1) [**?**, **?**, **?**]. Apparently, the integration of these two modes enables a concurrent expansion of both cell number and cell fate across the entire body of an embryo.

Hierarchical Complexity of C. elegans Cell Fate Sequence is Optimized at the Whole-Organism Scale, Not for Individual Body Parts

Whereas *C. elegans* exhibits exceptional hierarchical complexity (γ) near the theoretical maximum, far from that of either an extremely homogeneous sequence or an extremely heterogeneous sequence, we are curious whether this optimization participates not only in the entire body but also in individual

body parts derived from the somatic founder cell lineages. Hence, we examine the hierarchical complexity of partial cell fate sequences derived from all the somatic founder cell lineages, each originating from the first to fourth rounds of consecutive asymmetric cell divisions (driven by molecular interaction networks capable of cell polarization): AB (with P0 as mother and P1 as sister), EMS (with P1 as mother and P2 as sister), C (with P2 as mother and P3 as sister), and D (with P3 as mother and P4 as sister) (Fig. 1, Fig. S1; Table S1) [**?**, **?**, **?**]. Their cell fate sequences exhibit hierarchical complexity ($\gamma = 0$ to $\gamma \sim 0.942$) spanning nearly the range from the theoretical minimum ($\gamma = 0$) to the theoretical maximum ($\gamma = 1$) (Table 1). Comparable to the full lineage, the AB lineage (deriving cells destined for death, muscle, pharynx, and skin) and C lineage (deriving cells destined for death, muscle, neuron, and skin) also exhibit exceptional hierarchical complexity $\gamma = 0.918$ and 0.942 respectively, as both accommodate at least four distinct cell fates. However, the D lineage (deriving cells destined for muscle only) exhibits zero hierarchical complexity as expected, attributed to its repeatability but no diversity. Last but not least, despite the EMS lineage accommodating at least five cell fates as well, it exhibits only moderate hierarchical complexity $\gamma = 0.586$, suggesting that the way cell fates are encoded —beyond their mere element numbers, types, and proportions—is influential to their collective hierarchical architecture; although both the EMS and C lineages accommodate multiple cell fates, the EMS lineage with over twice the element number still exhibits a lower diversity level n and a higher repeatability level ω, consistent with previous observations that certain *C. elegans* lineal parts (e.g., the ABarapp lineage) tend toward the repeatability mode to avoid redundancy in differentiation programs [**?**, **?**]. To sum up, while the hierarchical complexity of the full cell fate sequence represents an evolutionary optimization goal for a complete multicellular life as a whole, the organism' s individual body parts can adjust their developmental manners to fulfill specific physiological functions required in reality.

Cell Fate Sequence Evolves Exceptional Hierarchical Complexity through an Optimal Trade-Off between Sub-Sequence Module Diversity and Repeatability

Inspired by the *C. elegans* cell fate sequence that contains an abundance of self-repeating and mutually assembling sub-sequence modules and near-maximum hierarchical complexity, we hypothesize that hierarchical complexity serves as a quantitative metric or developmental constraint orchestrating these two modes in concert. Hereafter, we refer to them as the repeatability and diversity modes. We employ a genetic algorithm to evolve a one-dimensional sequence while preserving the element types, numbers, and proportions of *C. elegans* (see Materials and Methods). A population of 4,000 in silico cell fate sequences is initialized from each of two extreme cases with zero hierarchical complexity: an extremely homogeneous sequence with element types arranged sequentially and an extremely heterogeneous sequence with element types mixed randomly (Fig. 3(a)).

For each iterative round, 1,200 (30%) sequences with the highest γ values are retained; the remaining 2,800 (70%) are first formed into 1,400 pairs of parents subjected to crossover, with a crossover probability of 0.8 to exchange a randomly-picked digit between them, and then subjected to a mutation probability of 0.03 to reset a randomly-picked digit into one of the other fates, accompanied by another randomly picked digit proceeding with the opposite resetting. Updating the pool of 4,000 samples through 3,000 iterative rounds, each case continuously evolves toward a hierarchical complexity of one as set (Fig. 3(b)); the terminal Laddergraphs consistently exhibit complex hierarchical architectures in which sub-sequence modules self-repeat and mutually assemble frequently, mirroring that of *C. elegans* in nature (Fig. S2). Amazingly, such continuous evolution of hierarchical complexity γ is equivalent to the product of two terms positively linear with diversity level n and repeatability level ω respectively, indicating that the optimization goal in natural evolution is the strong bonding of these two modes in trade-off, rather than a bias toward either of them alone (Fig. 3(c), Fig. S5). More specifically regarding sub-sequence module (Ladderon) organization, it is noteworthy that both evolutionary trajectories converge on a moderate Ladderon number, meaning that a moderate level of diversity mode is required (Fig. 3(d)); coherently, the maximum Ladderon length is also moderate, meaning that a moderate level of repeatability mode is required (Fig. 3(e)). What is more, the dynamics of the above parameters along evolutionary trajectories initiated from the two extreme cases demonstrate a clear trade-off between the diversity and repeatability modes, where an increase in one is inherently coupled with a decrease in the other, and vice versa. To our surprise, the optimal Ladderon number obtained from virtual evolution is 78.5 ± 4.5, very close to the *C. elegans* value of 76 in the real world, providing strong evidence that hierarchical complexity governs the design of *C. elegans* cell fate sequence by enforcing a delicate trade-off between diversity and repeatability.

Intercellular Notch Signaling and Downstream Differentiation Programs Contribute to Hierarchical Complexity of C. elegans

Since hierarchical complexity arises from a delicate trade-off between repeating and diversifying cell fate sub-sequence modules (Ladderons), we wonder how this is realized across a cleaving lineage (Fig. 1; Fig. S1; Table S1). The inspiring, representative natural developmental blueprint in *C. elegans* is known to have roughly half of its cell fate coding realized by intercellular signalings [**?**, **?**, **?**]. Among them, Notch signaling is highly conserved across species of varying complexity (e.g., worm [**?**, **?**], fly [**?**], frog [**?**], zebrafish [**?**], mouse [**?**], human [**?**]). It comprises a set of cell-membrane-bound molecules (i.e., ligands and receptors) that transmit signals from a signaling cell to a responding cell via physical contact, after which downstream differentiation programs induce the formation of various tissues/organs (e.g., the head and kidney in *C. elegans*). Within the longest Ladderons of *C. elegans* (Fig. 2; Fig. S2; Table S2), we identified two overlapping with downstream differentiation programs of the 3rd and 4th Notch signaling events (corresponding to the fifth- and first-longest

Ladderons respectively); namely, the ABplaaa lineage differentiates from the ABpraaa lineage after receiving the 3rd Notch signaling for head development and the ABplpapp lineage differentiates from the ABprpapp lineage after receiving the 4th Notch signaling for kidney development [**?**] (Fig. 4(a)). To study the role of these two signaling-modulated Ladderons, we implement three types of cell fate sub-sequence module disruptions: 1. the left lineage is replaced by the right lineage; 2. the right lineage is replaced by the left lineage; 3. the left and right lineages are swapped (Fig. 4(a)). Note that prior in vivo experiments have already verified that blocking the 3rd Notch signaling event results in a Type 1 disruption and blocking the 4th Notch signaling event results in a Type 2 disruption [**?**, **?**, **?**]. Intriguingly, while none of the Type 1-3 disruptions elevates the hierarchical complexity γ, those matching experimentally verified signaling-blocking outcomes cause the most pronounced reduction (Fig. 4(b)).

Further, we ask whether the full cell fate sequence—that is known to be shaped by numerous differentiation programs including not only intercellular Notch signaling, but also intercellular Wnt signaling [**?**, **?**], cell polarization [**?**, **?**], cell size segregation [**?**, **?**], among others—is globally organized under the principle of hierarchical complexity. To answer this question, we implement a more general sequence disruption via stochastic digit mutations. When the digit at each position is assigned an equal probability $p = 0.5$ of being switched to one of the other cell fates, the hierarchical complexity γ declines to zero through 1,000 iterative rounds (Fig. 4(c)). This strongly suggests that the exceptional hierarchical complexity of *C. elegans* relies not only on the intercellular Notch signaling just probed, but also on the coordinated action of differentiation programs operating throughout embryogenesis. As with Ladderons 1 and 5 linked to Notch signalings, other Ladderons may be linked to additional intercellular signalings or differentiation programs, warranting further investigation and serving as a valuable resource for identifying underexplored differentiation programs or regulatory mechanisms (Table 6).

In Real Developmental Context, Cellular Differentiation Programs May or May Not Be Repeated Across Lineages, Roughly in a Half-Half Ratio

Provided that downstream differentiation programs modulated by intercellular Notch signalings give rise to cell fate sub-sequence modules (Ladderons) repeated across distinct lineage positions (Table S2), we next inspect whether other differentiation programs are repeated as well (Table S3). To this end, we define the stemness of any cell in the complete lineage by all its progenies' fates (represented as a $1\$\times 8 vector in which 0 and 1 denote the absence and presence of each unique fate), then its cellular differentiation pr$ matrix in which the 1st row corresponds to the mother, while the 2nd and 3rd rows correspond to the anterior/left/dorsal and posterior/right/ventral daughters respectively) (Fig. 5(a)). Remarkably, we identify a diverse pool of 118 non-identical cellular differentiation programs across 670 non-identical

dividing cells (Fig. 1; Fig. S1; Table S1), of which 58.47% are repeated in at least two cells, proposing that the delicate diversity-repeatability trade-off operates not only in the sub-sequence modules of the full cell fate sequence but also in the cellular differentiation programs of the complete lineage that derives them (Table S3).

Moreover, we introduce two parameters to assess how cellular differentiation programs are repeated from one cell to another across distinct lineage positions: 1. "Lineage Distance" *LD* assesses the degree of separation between two cells with respect to the complete lineage, based on the generation of their most recent common ancestor (e.g., the lineage distance between Da and Dp is as low as 1, whereas that between ABalapappaa and ABplapapppp is as high as 9). The observed lineage distances span the theoretically allowed range, indicating that a cellular differentiation program can be repeated immediately in two daughter cells following cell division or spontaneously in two distant cells that have presumably undergone substantial differentiation before somehow converging on the same fate (Fig. 5(b)). 2. "Lineage Coherence" *LC* assesses the degree of symmetry between two distant cells with respect to the complete lineage, based on the difference of their lineage positions (e.g., the lineage coherence between ABalapaaa and ABprpppppp is low, whereas that between ABalapaaa and ABplapaaa is high). The observed lineage coherence spans the theoretically allowed range, indicating that cellular differentiation programs can either preserve or break symmetric lineage positions (just like those obeying or disobeying anterior-posterior, left-right, and dorsal-ventral lineage symmetry) [**?**, **?**]. The ones with high lineage coherence appear to involve single fates repeated (e.g., N $\rightarrow$ N & N, S $\rightarrow$ S & S) or diversified (e.g., N/D $\rightarrow$ N & D), while the ones with low lineage coherence appear to be modulated by mechanisms such as Notch signalings and are hardly repeated (e.g., N/D/P/S $\rightarrow$ D/P/S & N/D/P/S in the ABalpa and ABara that receive Notch signaling to lose their originally high lineage coherence) [**?**, **?**] (Fig. 5(c)). Overall, both lineage distance and coherence collectively suggest that while a part of cellular differentiation programs take place only once or a few times, the other part is widespread across the complete lineage.

Lineal Differentiation Program Repeatability Significantly Enhances Hierarchical Complexity in a Cleaving Lineage from Zygotic Stemness to Differentiated Fates

Given that our top-down analysis of *C. elegans* data reveals how differentiation programs are orchestrated (Figure 5 [Figure 5: see original paper]), can a bottom-up analysis of a general cleaving lineage better elucidate how such programs might emerge from scratch? To test this idea, we numerically sample cell lineage cleavages along with random cellular differentiation programs at large scale, so as to construct a comprehensive lineage-to-sequence design space (see Materials and Methods).

Here, we initiate round-by-round bifurcation starting from one cell and expand-

ing to 512 cells through 9 rounds of divisions, during which each cell division may undergo a random differentiation (Fig. 6(a)). While the initial zygote possesses all stemness (represented as a 1$\times$8 vector with all ones denoting all stemness to give rise to fate A, B, C, D, E, F, G, and H, where the element number and type are adjustable in principle), every fate is independently deleted with a differentiation probability q in a daughter cell after division. By setting q from 10^{-7} to 0.9 through both magnitude and equidistance changes, we obtain a total of 16,000 sampled lineages, each with a clear record of all cell fate codes across 1023 cells and cellular differentiation programs across 511 dividing cells. To more quantitatively assess how "cellular differentiation programs" are orchestrated across distinct lineage positions, likewise, we extend this concept from a focal cell and its two daughters to all its progenies by appending rows of binary cell fate codes for all its anterior/left/dorsal and posterior/right/ventral progenies, generation by generation—called "lineal differentiation program". Meanwhile, the terminal cell fate sequence can be derived by non-identical cell fate codes, from which the hierarchical complexity γ is subsequently calculated.

As the differentiation probability for a single cell fate (q) varies from low to high, hierarchical complexity γ exhibits a pronounced pulse-like pattern (Figure 6 Figure 6: see original paper). This pulse matches the results mentioned above: a low q value produces lineages with an extremely homogeneous cell fate sequence (repeatability) and a high q value produces lineages with an extremely heterogeneous cell fate sequence (diversity). Unexpectedly, a similar pulse also exists for the cell proportion with repeated lineal differentiation program R (Figure 6(c)). Their joint relationship directly shows the cell proportion with repeated lineal differentiation program R is positively correlated to the hierarchical complexity γ and monotonically raises the lower bound of hierarchical complexity γ, with values already touching the theoretical maximum when $S = 0.2$ (Fig. 6(d); Fig. S6). Interestingly, *C. elegans* makes full use of this strategy—by operating S as high as 0.694, effectively pushing its hierarchical complexity to $\gamma = 0.972$.

Discussion

How biological sequences are designed to fulfill specific functions across scales has remained a central question for dozens of years—from DNA sequence at the molecular scale (which are transcribed into RNA and then translated into functional protein) to cell fate sequence at the multicellular scale (that forms functional tissue, organ, and body). In this study, we aim to address a fundamental question in developmental biology: how are cell fates encoded in development to form a complete multicellular life? We took advantage of the Ladderpath approach to systematically decompose the full cell fate sequence in the worm *C. elegans*, revealing a complex, hierarchical architecture consisting of diverse yet repeated sub-sequence modules (Fig. 1, Fig. 2). These modes of diversity and repeatability symbolize two compensatory ways in which the full cell fate sequence is constituted: mutual assembling between modules and self-repeating of modules. Further virtual evolution based on genetic algorithm indicated

that the near-maximum hierarchical complexity is achieved by a delicate trade-off between diversity and repeatability (with moderate Ladderon numbers and lengths) that maximizes the product of their levels in a simple mathematical relationship (Fig. 3).

In silico disruption of realistic *C. elegans* lineage demonstrated that the previously reported Notch signalings transduced between neighboring cells, together with their downstream differentiation programs, substantially contribute to the hierarchical complexity observed in vivo (Fig. 4). In addition, cellular differentiation programs were broadly distributed across the complete lineage, with roughly half being repeated in other cells and the remaining half being isolated (exemplified by cases modulated by Notch signaling) (Fig. 5). Motivated by the statistics of cellular differentiation programs above, further large-scale lineage sampling that models these programs suggested that repeating lineal differentiation programs (including all cellular differentiation programs within the lineage derived from an ancestor cell) across multiple cells is essential to achieve an exceptional hierarchical complexity—a strategy naturally exploited by *C. elegans* (Fig. 6).

The above-mentioned findings highlight promising future research directions worth investigation. First, although the *C. elegans* cell fate sequence is proven to be near-optimal in theory, many alternative sequence designs can achieve comparable fitness. Perhaps the current Ladderpath approach decomposes only the cell fate sequence itself, while neglecting other important information in space (e.g., how differentiated cells are positioned in space to form functional tissues, organs, and body) or resource (e.g., how limited polarity and signaling molecules are positioned in space to trigger cellular differentiation programs). Apart from this, more finely specified cell fates, like those grouped under "Other" and sub-fates within one major group conceal deeper hierarchical architectures and hide additional design rules.

In the future, Ladderon deployment and lineage bifurcations could be mapped onto the three-dimensional geometry of the embryo to link cell position cues with cell fate modules and gain insight into how spatial patterning is coordinated, potentially alongside optimization goals proposed before in this scheme (e.g., cell volume segregation ratio and cell migration). Second, because the 3rd and 4th Notch signaling events are found to be related to the longest Ladderons, it is worth investigating whether other Ladderons are related to other regulatory mechanisms or differentiation programs, including but not limited to those relevant to Wnt signaling and cell polarization (Table S2). This might provide useful clues for identifying their theoretical roles, especially when analyzed jointly with cell-lineage–resolved gene-expression profiles from thousands of publicly available embryo samples [**?**, **?**, **?**, **?**]. Third, supported by the exceptional hierarchical complexity in *C. elegans* as well as the contribution of Notch signaling, whether the cell lineage and fate sequence is also designed for complexity and with a delicate trade-off between diversity and repeatability, is conserved among organisms and serves as a general principle, is worth studying.

Public datasets with cell fate documentation in the lineage tree (e.g., other nematodes and ascidians) or in spatial distribution (e.g., spatial transcriptomics in vivo [**?**, **?**] or in vitro [**?**, **?**]) may also be amenable to exploration.

The combinatorial bottom-up simulation (virtual evolution and random sampling) and top-down statistics (*C. elegans* cell fate and lineage) used in this study suggest biological sequence analyses need not be limited to the typical molecular scale; instead, they allow exploration into vast design spaces beyond those accessible in nature on computer. These methods may also be applicable to other sequence studies, such as DNA, RNA, and protein ones, which likewise involve self-repeating and mutually assembling sub-sequence modules. Besides, whether the optimal diversity-repeatability trade-off principle is also at work in other, similar systems remains a fascinating question. For example, in cooperative systems involving teamwork among agents (e.g., academic labs or research institutions), it may likewise be optimal to distribute members across a moderate range of research directions and collaborations (diversity) with certain resources in each (repeatability), rather than concentrating everyone on a single project or dispersing them into unrelated pursuits. This principle may also operate in broader social and cultural systems, which could be explored through a generalized agent-based model (instead of an intact sequence), where agents differ in identity yet cooperate to achieve both collective and individual functions.

Materials and Methods

Ladderpath Approach for Sequence Decomposition

Among various sequence analytical approaches [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**], the Ladderpath approach provides a transparent, white-box methodology that explicitly decomposes a sequence into self-repeating and mutually assembling components of varying lengths, thereby quantifying their repeatability and diversity in a principled manner [**?**, **?**]. This decomposition intuitively displays how these components collectively constitute a full sequence with an underlying hierarchical structure. Owing to its ability to handle sequences of arbitrary length and arbitrary numbers of element types, the Ladderpath approach has been successfully applied in design rule discovery and engineering guidance across a wide range of scenarios, including amino acid sequences in protein synthesis and parameter sequences in artificial neural networks [**?**, **?**, **?**].

To illustrate how the Ladderpath approach decomposes a sequence, let' s use an arbitrary 25-bit sequence as an example: ABACDA CDA B DB ABACDA D DB CDAB. First, the single-bit components (i.e., the individual element types) constitute the most foundational layer, formalized as {A, B, C, D}. Next, all multi-bit components that repeat at least twice are regarded as "Ladderons" and placed on upper hierarchical layers according to the following rules: if a single-bit component self-repeats or multiple single-bit components assemble mutually, the second layer is constituted, formalized after the first layer as: {A,

B, C, D // CDA, DB} (here, CDA self-repeats four times and DB self-repeats twice in the full sequence); the remaining Ladderons—those mutually assembled from Ladderons in lower layers—are then placed onto successively higher layers, so the third layer is constituted, formalized after the second layer as: {A, B, C, D // CDA, DB // ABACDA, CDAB} (here, ABACDA self-repeats twice and CDAB self-repeats twice), and so on and so forth. To represent the hierarchical structure, each Ladderon in the formula is annotated with the cumulative number of times (>1) it self-repeats or mutually assembles with another Ladderon to form a higher-level Ladderon or the full sequence, which is equal to the number of arrows emanating from that Ladderon in the Laddergraph (a graph that intuitively visualizes how all Ladderons self-repeat and mutually assemble to constitute the full sequence); to avoid redundancy and because every multi-bit Ladderon is already repeated at least twice by definition, we omit annotating a repeat number of 2. Eventually, the formula becomes {A(3), B(3), C, D (3) // CDA, DB // ABACDA, CDAB} (Fig. 7). While the sum of the annotated numbers (3+3+1+3+1+1+1+1=14) represents how many operations (including self-repeating or mutually assembling) are required to constitute the full sequence, subtracting it from the sequence length ($\omega = 25 - 14 = 11$) represents the extent to which the sequence is constituted through repeating rather than diversifying its components.

The ω value reflects how the hierarchical structure of sequence components gains complexity as it deviates from these two extreme cases: in the case of an absolutely homogeneous sequence exemplified with $S = 25$ (i.e., AAAAAAAAAAAAAAAAAAAAAAAAA), its Ladderpath representation A(4) // AAA // AAAAAA // AAAAAAAAAAAA yields a large $\omega = 25$, indicating intensive self-repeating mode (repeatability) (Fig. 7(b)); in the case of an absolutely heterogeneous sequence exemplified with $S = 25$ (i.e., ABCDEFGHIJKLMNOPQRSTUVWXY), its Ladderpath representation A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y yields a large $\omega = 0$, indicating intensive mutual-assembling mode (diversity) (Fig. 7(c)). Normalizing ω to the 0-1 range allows us to define a new metric, also ranging from 0 to 1, that evaluates the degree to which a sequence integrates the two modes cooperatively to gain hierarchical structural complexity:

$$\gamma = 1 - \left| 2\frac{\omega(x) - \omega_0(S)}{\omega_{\max}(S) - \omega_0(S)} - 1 \right|$$

where $\omega_0(S)$ and $\omega_{\max}(S)$ denote, respectively, the average and maximum ω values obtained from numerous random sequences of length S sharing the same digit pool.

Assessing the Lineage Coherence between Two Cells with the Same Cellular Differentiation Program

The central idea is to assess how closely the lineage positions of two cells or their names align after excluding their shared prefix, since a lineage-based nomenclature was implemented. To quantify their difference of lineage positions, we define a difference vector $\mathbf{d}$, which encodes mismatches between corresponding characters of the two strings (s_1 and s_2) of length L in a binary form:

$$d_i = \begin{cases} 1 & \text{if } s_1(L-i) \neq s_2(L-i) \\ 0 & \text{otherwise} \end{cases}$$

where i denotes the $i+1$-th character of the string counted from the left; $s(k)$ denotes the k-th character of the string counted from the right.

The lineage coherence between two cells s_1 and s_2 is then computed as:

$$\text{LineageCoherence}(s_1, s_2) = \frac{\sum_{i=0}^{L-2} d_i \cdot 2^i}{\sum_{i=0}^{L-2} 2^i}$$

The denominator in the formula scales the value between 0 and 1, and the numerator measures the cumulative impact of character mismatches as a binary-weighted sum of differences. If two cells share an ancestor many generations above but locate in similar or symmetric lineal positions (i.e., ABalaaa and ABplaaa) and share the same differentiation program, they are evaluated with high lineage coherence. For example, the lineage symmetry between 'ABalaaaa' and 'ABplaaaa' is $\frac{2^1}{2^6-2} = 1$, representing maximum symmetry. Conversely, if two cells are in deviated lineal positions (i.e., ABalaaa and ABprppp) and share the same differentiation program, they are evaluated with low lineage coherence. For example, 'ABprpppp' is very different from 'ABalaaaa' and the lineage symmetry is $\frac{2^5+2^4+2^3+2^2+2^1+2^0}{2^6\cdot(2^6-2)} = 0.0615$.

Data, Materials, and Software Availability

The codes generated and analyzed during the current study are available on GitHub. All other study data are included in the article and/or supporting information.

Acknowledgments

We thank Prof. Ke Zhang and Prof. He Liu for their fruitful discussion on this project. We also thank all members in ecsLab. The large-scale computation was partly conducted using approximately 200 CPU cores in the Interdisciplinary Intelligent Supercomputing Center of Beijing Normal University at Zhuhai. This work was supported by the National Natural Science Foundation

of China (Grant No. 12205012 to Y.L.) and Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515012923 to Y.L.).

References

1. J Watson, F Crick, Molecular structure of nucleic acid. *Nature* 171, 737-738 (1953).

2. MW Nirenberg, JH Matthaei, The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci.* 47, 1588-1602 (1961).

3. A Ambrogelly, S Palioura, D Söll, Natural expansion of the genetic code. *Nat. Chem. Biol.* 3, 29-35 (2007).

4. C Nüsslein-Volhard, E Wieschaus, Mutations affecting segment number and polarity in *Drosophila*. *Nature* 287, 795-801 (1980).

5. CC Fowlkes, et al., A quantitative spatiotemporal atlas of gene expression in the *Drosophila* blastoderm. *Cell* 133, 364-374 (2008).

6. JO Dubuis, G Tkačik, EF Wieschaus, T Gregor, W Bialek, Positional information, in bits. *Proc. Natl. Acad. Sci.* 110, 16301-16308 (2013).

7. JA Briggs, et al., The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* 360, eaar5780 (2018).

8. DE Wagner, et al., Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* 360, 981-987 (2018).

9. B Pijuan-Sala, et al., A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 566, 490-495 (2019).

10. N Karaiskos, et al., The *Drosophila* embryo at single-cell transcriptome resolution. *Science* 358, 194-199 (2017).

11. X Liao, et al., Repetitive DNA sequence detection and its role in the human genome. *Commun. Biol.* 6, 954 (2023).

12. R Hunt-Newbury, et al., High-throughput in vivo analysis of gene expression in *Caenorhabditis elegans*. *PLoS Biol.* 5, e237 (2007).

13. JI Murray, et al., Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. *Nat. Methods* 5, 703-709 (2008).

14. JS Packer, et al., A lineage-resolved molecular atlas of *C. elegans*

embryogenesis at single-cell resolution. *Science* 365, eaax1971 (2019).

15. S Arnott, et al., X-ray diffraction studies of double helical ribonucleic acid. *Nature* 211, 227–232 (1966).

16. C elegans Sequencing Consortium, *Genome sequence of the nematode* C. elegans*: a platform for investigating biology.* Science* 282, 2012–2018 (1998).

17. ES Lander, et al., Initial sequencing and analysis of the human genome. *Nature* 409, 860–921 (2001).

18. EA Baker, A Woollard, How weird is the worm? Evolution of the developmental gene toolkit in *Caenorhabditis elegans. J. Dev. Biol.* 7, 19 (2019).

19. Y Xu, et al., Structure of the protein phosphatase 2A holoenzyme. *Cell* 127, 1239–1251 (2006).

20. Z Wang, et al., Molecular and structural basis of the dual regulation of the polycystin-2 ion channel by small-molecule ligands. *Proc. Natl. Acad. Sci.* 121, e2316230121 (2024).

21. Y Zhao, et al., Cryo-EM structures of apo and antagonist-bound human Cav3.1. *Nature* 576, 492–497 (2019).

22. X Yao, X Fan, N Yan, Cryo-EM analysis of a membrane protein embedded in the liposome. *Proc. Natl. Acad. Sci.* 117, 18497–18503 (2020).

23. T Wu, X Yang, X Jin, N Yan, Z Li, Critical role of extracellular loops in differential modulations of TTX-sensitive and TTX-resistant Nav channels. *Proc. Natl. Acad. Sci.* 122, e2510355122 (2025).

24. TN Sasaki, H Cetin, M Sasai, A coarse-grained Langevin molecular dynamics approach to de novo protein structure prediction. *Biochem. Biophys. Res. Commun.* 369, 500–506 (2008).

25. T Wang, et al., CryoSeek II: Cryo-EM analysis of glycofibrils from freshwater reveals well-structured glycans coating linear tetrapeptide repeats. *Proc. Natl. Acad. Sci.* 122, e2423943122 (2025).

26. T Tlusty, A Libchaber, Life sets off a cascade of machines. *Proc. Natl. Acad. Sci.* 122, e2418000122 (2025).

27. J Jumper, et al., Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).

28. J Wu, B Liu, Y Cui, Are sutural structures in biology the optimal topological design? *Compos. Struct.* 354, 118825 (2025).

29. JC Lee, et al., Instructional materials that control cellular activity through synthetic Notch receptors. *Biomaterials* 297, 122099 (2023).

30. T Hayes, et al., Simulating 500 million years of evolution with a language model. *Science* 387, 850–858 (2025).

31. N Chou, et al., Impact-absorbing helmet design inspired by walnut texture reaction-diffusion mechanisms. *Acta Biomater.* 196, 244–256 (2025).

32. T Yamada, et al., Synthetic organizer cells guide development via spatial and biochemical instructions. *Cell* 188, 778–795 (2025).

33. MW Nirenberg, JH Matthaei, The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci.* 47, 1588–1602 (1961).

34. A Ambrogelly, S Palioura, D Söll, Natural expansion of the genetic code. *Nat. Chem. Biol.* 3, 29–35 (2007).

35. JA Briggs, et al., The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* 360, eaar5780 (2018).

36. DE Wagner, et al., Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* 360, 981–987 (2018).

37. B Pijuan-Sala, et al., A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 566, 490–495 (2019).

38. N Karaiskos, et al., The *Drosophila* embryo at single-cell transcriptome resolution. *Science* 358, 194–199 (2017).

39. X Liao, et al., Repetitive DNA sequence detection and its role in the human genome. *Commun. Biol.* 6, 954 (2023).

40. R Hunt-Newbury, et al., High-throughput in vivo analysis of gene expression in *Caenorhabditis elegans*. *PLoS Biol.* 5, e237 (2007).

41. JI Murray, et al., Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. *Nat. Methods* 5, 703–709 (2008).

42. JS Packer, et al., A lineage-resolved molecular atlas of *C. elegans*

embryogenesis at single-cell resolution. *Science* 365, eaax1971 (2019).

43. L Li, et al., Spatiotemporal single-cell architecture of gene expression in the *Caenorhabditis elegans* germ cells. *Cell Discov.* 11, 26 (2025).

44. P Xie, et al., Digital reconstruction of full embryos during early mouse organogenesis. *Cell* (2025).

45. YB Tzur, et al., Spatiotemporal gene expression analysis of the *Caenorhabditis elegans* germline uncovers a syncytial expression switch. *Genetics* 210, 587–605 (2018).

46. AK Corsi, B Wightman, M Chalfie, A transparent window into biology: a primer on *Caenorhabditis elegans*. *Genetics* 200, 387–407 (2015).

47. X Ma, et al., A 4D single-cell protein atlas of transcription factors delineates spatiotemporal patterning during embryogenesis. *Nat. Methods* 18, 893–902 (2021).

48. G Guan, et al., Cell lineage-resolved embryonic morphological map reveals signaling associated with cell fate and size asymmetry. *Nat. Commun.* 16, 3700 (2025).

49. Y Liu, Z Di, P Gerlee, Ladderpath approach: how tinkering and reuse increase complexity and information. *Entropy* 24, 1082 (2022).

50. Z Zhang, et al., Evolutionary tinkering enriches the hierarchical and nested structures in amino acid sequences. *Phys. Rev. Res.* 6, 023215 (2024).

51. S Li, et al., Discovery of highly bioactive peptides through hierarchical structural information and molecular dynamics simulations. *J. Chem. Inf. Model.* 64, 8164–8175 (2024).

52. Z Xu, et al., Correlating measures of hierarchical structures in artificial neural networks with their performance. *npj Complex.* 1, 15 (2024).

53. L Rose, P Gönczy, Polarity establishment, asymmetric division and segregation of fate determinants in early *C. elegans* embryos. *WormBook* (2014).

54. L Hubatsch, et al., A cell-size threshold limits cell polarity and asymmetric division potential. *Nat. Phys.* 15, 1078–1085 (2019).

55. G Guan, MK Wong, Z Zhao, LH Tang, C Tang, Volume segregation programming in a nematode' s early embryogenesis. *Phys. Rev. E* 104,

054409 (2021).

56. JE Sulston, E Schierenberg, JG White, JN Thomson, The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* 100, 64–119 (1983).

57. Z Du, et al., The regulatory landscape of lineage differentiation in a metazoan embryo. *Dev. Cell* 34, 592–607 (2015).

58. R Hunt-Newbury, et al., High-throughput in vivo analysis of gene expression in *Caenorhabditis elegans*. *PLoS Biol.* 5, e237 (2007).

59. JI Murray, et al., Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. *Nat. Methods* 5, 703–709 (2008).

60. JS Packer, et al., A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* 365, eaax1971 (2019).

61. L Li, et al., Spatiotemporal single-cell architecture of gene expression in the *Caenorhabditis elegans* germ cells. *Cell Discov.* 11, 26 (2025).

62. P Xie, et al., Digital reconstruction of full embryos during early mouse organogenesis. *Cell* (2025).

63. YB Tzur, et al., Spatiotemporal gene expression analysis of the *Caenorhabditis elegans* germline uncovers a syncytial expression switch. *Genetics* 210, 587–605 (2018).

64. A Camproux, P Tuffery, J Chevrolat, J Boisvieux, S Hazout, Hidden Markov model approach for identifying the modular framework of the protein backbone. *Protein Eng.* 12, 1063–1073 (1999).

65. TL Bailey, N Williams, C Misleh, WW Li, MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, W369–W373 (2006).

66. RD Finn, J Clements, SR Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37 (2011).

67. QY Tang, K Kaneko, Long-range correlation in protein dynamics: confirmation by structural data and normal mode analysis. *PLoS Comput. Biol.* 16, e1007670 (2020).

68. G Rix, et al., Continuous evolution of user-defined genes at 1 million times the genomic mutation rate. *Science* 386, eadm9073 (2024).

69. J Wu, B Liu, Y Cui, Are sutural structures in biology the optimal topological design? *Compos. Struct.* 354, 118825 (2025).

70. B Seelig, IA Chen, Intellectual frameworks to understand complex biochemical systems at the origin of life. *Nat. Chem.* 17, 11–19 (2025).

71. C Xia, Z Zhang, X Guan, Q Tang, Protein structural bioinformatics empowered by statistical physics and artificial intelligence. *Synth. Biol. J.* 6, 547 (2025).

72. Y Liu, et al., Exploring and mapping chemical space with molecular assembly trees. *Sci. Adv.* 7, eabj2465 (2021).

73. KM Mickey, CC Mello, MK Montgomery, A Fire, JR Priess, An inductive interaction in 4-cell stage *C. elegans* embryos involves apx-1 expression in the signaling cell. *Development* 122, 1791–1798 (1996).

74. CC Mello, BW Draper, JR Priess, The maternal genes apx-1 and glp-1 and establishment of dorsal-ventral polarity in the early *C. elegans* embryo. *Cell* 77, 95–106 (1994).

75. G Weinmaster, VJ Roberts, G Lemke, A homolog of *Drosophila* notch expressed during mammalian development. *Development* 113, 199–205 (1991).

76. C Coffman, W Harris, C Kintner, Xotch, the Xenopus homolog of *Drosophila* notch. *Science* 249, 1438–1441 (1990).

77. A Raya, et al., Activation of notch signaling pathway precedes heart regeneration in zebrafish. *Proc. Natl. Acad. Sci.* 100, 11889–11895 (2003).

78. J Luis de la Pompa, et al., Conservation of the notch signaling pathway in mammalian neurogenesis. *Development* 124, 1139–1148 (1997).

79. AL Penton, LD Leonard, NB Spinner, Notch signaling in human development and disease in *Seminars in Cell and Developmental Biology*. (Elsevier), Vol. 23, pp. 450–457 (2012).

80. IP Moskowitz, JH Rothman, lin-12 and glp-1 are required zygotically for early embryonic cellular interactions and are regulated by maternal glp-1 signaling in *Caenorhabditis elegans*. *Development* 122, 4105–4117 (1996).

81. EJ Lambie, J Kimble, Two homologous regulatory genes, lin-12 and glp-1, have overlapping functions. *Development* 112, 231–240 (1991).

82. AL Zacharias, T Walton, E Preston, JI Murray, Quantitative differences in nuclear β-catenin and TCF pattern embryonic cells in *C. elegans*. *PLoS Genet.* 11, e1005585 (2015).

83. LR Girard, et al., WormBook: the online review of *Caenorhabditis elegans* biology. *Nucleic Acids Res.* 35, D472–D475 (2007).

84. YW Lim, FL Wen, P Shankar, T Shibata, F Motegi, A balance between antagonizing PAR proteins specifies the pattern of asymmetric and symmetric divisions in *C. elegans* embryogenesis. *Cell Reports* 36 (2021).

85. G Guan, MK Wong, Z Zhao, LH Tang, C Tang, Volume segregation programming in a nematode' s early embryogenesis. *Phys. Rev. E* 104, 054409 (2021).

86. JE Sulston, E Schierenberg, JG White, JN Thomson, The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* 100, 64–119 (1983).

87. Z Du, et al., The regulatory landscape of lineage differentiation in a metazoan embryo. *Dev. Cell* 34, 592–607 (2015).

88. R Hunt-Newbury, et al., High-throughput in vivo analysis of gene expression in *Caenorhabditis elegans*. *PLoS Biol.* 5, e237 (2007).

89. JI Murray, et al., Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. *Nat. Methods* 5, 703–709 (2008).

90. JS Packer, et al., A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* 365, eaax1971 (2019).

91. L Li, et al., Spatiotemporal single-cell architecture of gene expression in the *Caenorhabditis elegans* germ cells. *Cell Discov.* 11, 26 (2025).

92. P Xie, et al., Digital reconstruction of full embryos during early mouse organogenesis. *Cell* (2025).

93. YB Tzur, et al., Spatiotemporal gene expression analysis of the *Caenorhabditis elegans* germline uncovers a syncytial expression switch. *Genetics* 210, 587–605 (2018).

94. A Camproux, P Tuffery, J Chevrolat, J Boisvieux, S Hazout, Hidden Markov model approach for identifying the modular framework of the protein backbone. *Protein Eng.* 12, 1063–1073 (1999).

95. TL Bailey, N Williams, C Misleh, WW Li, MEME: discovering and

analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, W369–W373 (2006).

96. RD Finn, J Clements, SR Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37 (2011).

97. QY Tang, K Kaneko, Long-range correlation in protein dynamics: confirmation by structural data and normal mode analysis. *PLoS Comput. Biol.* 16, e1007670 (2020).

Note: Figure translations are in progress. See original paper for figures.