Supporting Information to
"Phase transition in the economically modeled growth of a cellular nervous system"

Vincenzo Nicosia,[1,*] Petra E. Vertes,[2,*] William R. Schafer,[3] Vito Latora,[1,4] and Edward T. Bullmore[2,5,6,†]

[1] *School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, United Kingdom*
[2] *Department of Psychiatry, Behavioural and Clinical Neuroscience Institute, University of Cambridge, Cambridge CB2 0SZ, United Kingdom*
[3] *Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 0QH, United Kingdom*
[4] *Dipartimento di Fisica e Astronomia & INFN & Laboratorio sui Sistemi Complessi, Università di Catania, Via S. Sofia 61, Catania, Italy*
[5] *Cambridgeshire and Peterborough National Health Service Foundation Trust, Cambridge CB21 5EF, United Kingdom*
[6] *GlaxoSmithKline Clinical Unit Cambridge, Addenbrookes Hospital, Cambridge CB2 0QQ, United Kingdom*

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Section S1. Location of neurons born after hatching

In Fig. S-1 we show the spatial configuration of neurons before and after hatching. Notice that the majority of the neurons born before hatching are concentrated in the head and in the tail region, while most of the neurons appearing after hatching are instead placed in the body to form the ventral cord. This explains the relative higher distance from newly added neurons to existing ones observed after hatching.

Section S2. Additional one-parameter models

We present here three additional growth models which have been tested during this study, namely the Simple Spatial Growth (SSG), Spatial Growth with Elongation (SGE) and Power-law Economical Growth (PEG). We also discuss their ability to reproduce the developmental growth of the *C. elegans* neuronal network, and we will compare them with the other five models described in the main text, i.e. Barabási-Albert (BA), Binomial Accelerated Growth (BAG), Hidden-variable Accelerated Growth (HAG), Economical Spatial Growth (ESG) and Economical Spatio-Temporal Growth (ESTG). Notice that all the models considered in this study have only one free parameter. Nevertheless some of these models, and in particular the ESTG, are exceptionally accurate at reproducing the structure and development of the *C. elegans* neuronal network.

Simple Spatial Growth (SSG). This model makes the assumption that upon arrival a new node i is placed in the same position at which it appears in the adult worm. Then, node i creates an edge to each of the already existing nodes j with probability:

$$\Pi_{i\to j}^{SP} = e^{-\frac{d_{ij}^{\mathrm{ad}}}{\delta}} \tag{S-1}$$

where d_{ij}^{ad} is the distance between node i and node j in the adult worm and δ is a parameter tuning the typical edge length. Since the connection probability decreases exponentially with the distance between nodes in the adult worm, the resulting networks exhibit very few medium- and high-distance links, which are instead relatively frequent in the real *C. elegans* neuronal networks.

* These authors have equally contributed to this work
† To whom correspondence should be addressed. Email: etb23@cam.ac.uk

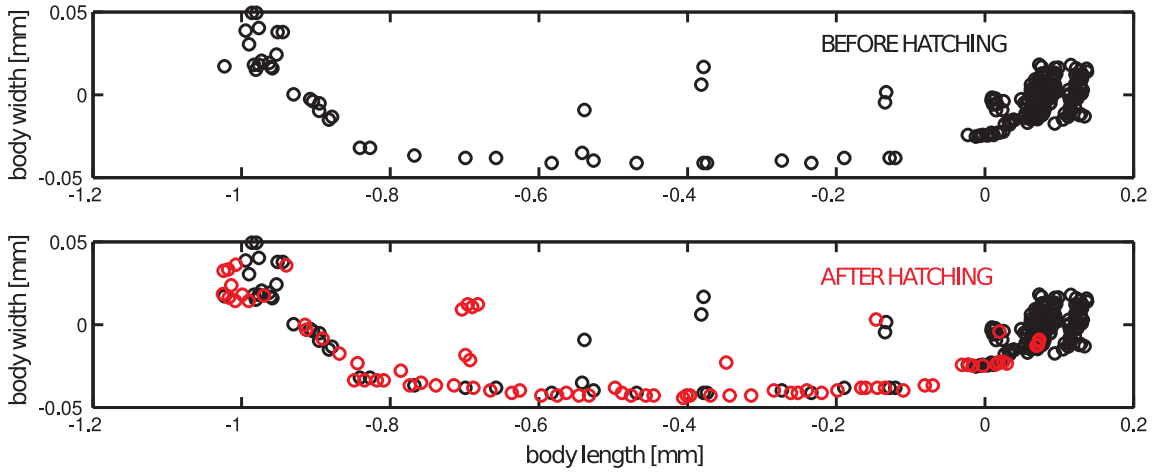

Figure S-1. **Position of neurons born after hatching.** The large majority of neurons born after hatching are located throughout the worm's body, while most of the neurons born before hatching are concentrated in the head and in the tail. The x-axis represents the distance in millimeters from the base of the head. Positive values indicate points in the worm's head, while negative values correspond to the body and the tail.

Spatial Growth with Elongation (SGE). This model uses information about the length of the worm at different stages. The node i arriving in the network at time t is placed in the position it occupies in the neural network at that time, and the probability for i to connect to an existing node j is defined as:

$$\Pi_{i\to j}^{SPE} = e^{-\frac{d_{ij}(t)}{\delta}} \tag{S-2}$$

where $d_{ij}(t)$ is the distance between node i and node j at time t and δ is a parameter. Notice that $d_{ij}(t)$ is a function of time, so that the probability to create an edge between a newly arrived node i and an existing node j depends on the time at which node i arrives in the network and on the relative positions of i and j at that time. This makes possible the creation of edges between nodes which are actually separated by a relatively large distance in the adult worm but have been closer in space in earlier developmental stages.

Power-law Economical Growth (PEG). This model implements a trade-off between the tendency to create edges to hubs and the relative distance of the nodes, and takes into account the elongation of the worm during development. Differently from the Economical Spatio-Temporal Growth model presented in the main text, in which the connection probability is a decreasing exponential function of distance, in this model the probability to connect to a distant node decreases as a power-law:

$$\Pi_{i\to j}^{PEG} = \frac{h_j}{h_{max}}\left[1-\left(\frac{d_{ij}(t)}{L_t}\right)^{\alpha}\right] \tag{S-3}$$

Here, h_j is the hidden degree of node j, which is set equal to the degree of node j observed in the adult worm, while h_{max} is maximum node degree in the adult neural network. As for the ESTG model, $d_{ij}(t)$ is the distance between node i and node j in the worm at time t. L_t is the total worm length at time t and α is the exponent of the power-law. Notice that the attachment probability Π^{PEG} approaches 0 when the distance $d_{ij}(t)$ is comparable with the length of the worm, while the hidden degree of the destination node plays a more important role if the two nodes are closer in space. Thanks to the preferential attachment term, based on the hidden degree of the nodes, this model tends to preserve the degree distribution of the original network.

Section S3. Parameter tuning

In this study we considered only one-parameter randomized growth models. In general, a randomized model generates an ensemble of graphs having certain characteristics. If the model has a tunable parameter, each value of the parameter generates a family of graphs sharing similar structural properties. For instance, the Binomial Accelerated Growth (BAG) model produces networks in which the number of edges grows quadratically with the number of nodes, but the expected number of edges in the final network, i.e. when the number of nodes is equal to $N = 279$, depends on the actual value of the attachment probability p.

Since a randomized one-parameter model generates a family of graphs for each value of the parameter, its ability to reproduce the structure of a given network cannot be assessed through a direct comparison of the original graph with a single realization of the model. Instead, the comparison should be performed by taking into account the expected structural properties of the ensemble of networks generated by the model, for each value of the parameter, averaging over a sufficiently large number of realizations. The first requirement of any suitable model for the *C. elegans* neural network growth is to pro-

duce networks having $N = 279$ nodes and, on average, $K = 2287$ edges. This constraint has been used to find the optimal parameter of each considered model.

We employed a two-step parameter optimization process. In the first step we used a Monte-Carlo approach to identify the interval in the parameter space for which the expected total number of edges $\tilde{K}$ of the generated networks was equal to $2287 \pm 5\%$. In this step, we considered 20 networks for each value of the parameter. In the second step we iteratively shrunk the parameter interval using the bisection method, in order to identify the value for which the difference between $\tilde{K}$ and $K = 2287$ was smaller than 1%. In this step we generated 500 networks for each value of the parameter. The optimal parameter values for each of the eight models are reported in Table S-I.

Model	**Optimal parameter**
BAG	$p = 0.0575$
HAG	$p = 0.302$
BA	$m_0 = 8, m = 8$
SSG	$\delta = 0.01365$
SGE	$\delta = 0.00235$
PEG	$\alpha = 0.0232$
ESG	$\delta = 0.0858$
ESTG	$\delta = 0.0126$

Table S-I. **Optimal model parameters.** The optimal parameter of a model guarantees the generation of networks having the same number of edges as the *C. elegans* adult neural network, with an error smaller than 1%.

Section S4. Model comparison

Since our aim was to reproduce as closely as possible the developmental growth of the *C. elegans* neuronal network, and in particular the abrupt transition in the number of edges in the graph as a function of the number of nodes, we defined a measure to quantify how closely each model matches the curve $\mathcal{K}(N)$, which indicates the number of edges in the *C. elegans* neuronal network when N nodes have been born.

We denote by $\mathcal{K}_M(N)$ the family of curves of K over N obtained using a certain model M and setting the value of the model parameter according to Table S-I. We computed, for each value of N, the expected number $\mu(\mathcal{K}_M(N))$ of edges in the network generated by model M when N nodes have been added to the graph, averaging over 500 realizations. Using this notation, $\mu(\mathcal{K}_M(100))$ is the expected number of edges in the graphs generated by model M when the first $N = 100$ nodes have been added to the graph.

In Fig. S-2 we report the average curve $\mu(\mathcal{K}_M(N))$ for each of the eight considered models, together with the original curve $\mathcal{K}(N)$ corresponding to the growth of the *C. elegans* neural network. By visual inspection, we conclude that the model which best fits the developmental growth of the original network and the phase transition at hatching is the Economical Spatio-Temporal Growth.

In order to quantify the discrepancy between $\mathcal{K}(N)$ and $\mathcal{K}_M(N)$ we computed, for each model and for each value of N, the difference $\xi(N)$:

$$\xi(N) = |\mathcal{K}(N) - \mu(\mathcal{K}_M(N))| \tag{S-4}$$

and we considered the expected value $\mu[\xi(N)]$ and the standard deviation $\sigma[\xi(N)]$ of $\xi(N)$. In Table S-II we report the values of $\mu[\xi(N)]$ and $\sigma[\xi(N)]$ for the eight models considered. In general, smaller values of $\mu[\xi(N)]$ and $\sigma[\xi(N)]$ indicate a closer match of the original growth curve. In agreement with the conclusions drawn after visual inspection of Fig S-2, which suggested that ESTG was the model which most closely reproduced the growth curve, the smallest values of $\mu[\xi(N)]$ and $\sigma[\xi(N)]$ are indeed obtained by the Economical Spatio-Temporal Growth model. The networks generated by all the other models fail to follow the original growth curve by a large extent, and they consequently exhibit larger values of $\mu[\xi(N)]$ and $\sigma[\xi(N)]$.

Model	$\mu[\xi(N)]$	$\sigma[\xi(N)]$
BAG	154.2	123.7
HAG	154.2	123.7
BA	216.7	150.7
SSG	205.2	167.1
SGE	89.5	73.7
PEG	209.4	168.4
ESG	215.6	172.9
ESTG	**37.3**	**31.6**

Table S-II. **Quality of growth fit.**Average and standard deviation of the point-to-point difference between the observed growth curve $\mathcal{K}(N)$ and the average curve corresponding to each of the eight considered models. The model parameters are set according to Table S-I. The smaller the value of $\mu[\xi(N)]$, the more closely a model can reproduce the growth of the *C. elegans* neural network. The Barabasi-Albert model (BA) exhibits the highest average point-to-point distance, while the Economical Spatio-Temporal Growth model (ESTG) largely outperforms all the other models.

Section S5. Node degree, edge length and node efficiency

Here we compare the structure of the networks produced by each of the eight models described in this study with that observed in the adult *C. elegans* neural network, by using three classical network metrics. The first metric is the degree distribution. Given an undirected graph $G(V, E)$ associated with the symmetric adjacency matrix $A = \{a_{ij}\}$, the degree of a node i is defined as the number of edges incident on i, and is denoted by $k_i = \sum_j a_{ij}$. The degree distribution $P(k)$ of the graph indicates, for each value of k, the probability of finding a node whose degree is equal to k. The second metric is the distribution of connection distances. Given two directly connected nodes i and j of a spatially-embedded network, we define the distance of the edge (i, j) as the Euclidean distance d_{ij} separating node i and node j. The

Model	$D_{KL}\left(P(k), P_M(k)\right)$	$D_{KL}\left(P(d), P_M(d)\right)$	$D_{KL}\left(P(E_i), P_M(E_i)\right)$
BAG	0.875	0.346	0.966
HAG	0.301	0.290	0.545
BA	0.309	0.176	0.226
SSG	0.710	0.884	0.611
SGE	0.428	0.269	1.447
PEG	0.149	0.322	0.214
ESG	0.708	0.685	0.361
ESTG	0.143	0.099	0.223

Table S-III. **Kullback-Leibler divergence.** The symmetrized Kullback-Leibler divergence between the degree, edge length and node efficiency distributions of the adult *C. elegans* neural network and the corresponding average distributions of the networks generated through each of the eight models. Smaller values of symmetrized divergence indicate higher similarity between the two distributions. The best and second-best values are highlighted in green and yellow, respectively, while the worst and second-worst are marked in red and orange, respectively. BAG and SSG exhibit the worst values of divergence. Interestingly, besides being the best model at fitting the developmental growth of the *C. elegans* neural network (as shown in Fig. S-2 and in Table S-II) ESTG performs more consistently than any of the other models in reproducing the structural properties of the adult worm's nervous system.

associated distance distribution $P(d)$ is the probability of finding an edge whose distance is exactly equal to d. The third metric is node efficiency. Given an undirected and unweighted graph G, the efficiency of a node is defined as:

$$E_i = \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{1}{\lambda_{ij}} \tag{S-5}$$

where λ_{ij} is the distance between node i and node j, measured as the number of edges in the shortest path connecting i to j. The node efficiency of i measures how easy it is to reach any other node in the graph by starting from i and traveling across shortest paths. In general, the smaller the distance between i and j, the higher the contribution of j to the efficiency of node i. If the graph is not connected and node i and j belong to two different connected components then there exists no path connecting them. In this case, the distance λ_{ij} is conventionally set to ∞, and the contribution of node j to the efficiency of i is equal to $1/\infty \equiv 0$.

In Fig. S-3, S-4 and S-5 we show, respectively, the average degree distribution, length distribution and node efficiency distribution of the networks generated by each of the eight models, together with those observed in the adult *C. elegans* neural network (reported in each panel in shaded grey). By visual inspection, we notice that ESTG seems to be the model which most closely reproduces all these distributions.

In order to quantify the difference between the distributions of degree, length and node efficiency of synthetic graphs with those of the *C. elegans* neural network we used the Kullback-Leibler divergence. Given two probability distributions $P = \{p_i\}$ and $Q = \{q_i\}$, the Kullback-Leibler divergence of Q from P is defined as:

$$D_{KL}(P||Q) = \sum_i p_i \log \frac{p_i}{q_i} \tag{S-6}$$

The Kullback-Leibler divergence measures the information lost when Q is used as an approximation of P, and is non-symmetric, i.e. $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. Since we are interested in measuring the similarity between two distributions, and not the relative information lost when using one of them as a predictor of the other, we opted for the symmetrized Kullback-Leibler divergence, which is defined as follows:

$$D_{KL}(P,Q) = \frac{D_{KL}(P||Q) + D_{KL}(Q||P)}{2} \tag{S-7}$$

In general, the smaller the value of $D_{KL}(P,Q)$, the more similar the two distributions P and Q. If we denote by $P(c)$ the distribution of the generic quantity c in the *C. elegans* neural network and by $P_M(c)$ the distribution of the same quantity c in networks generated through model M, the symmetrized Kullback-Leibler divergence between $P(c)$ and $P_M(c)$ is denoted as $D_{KL}\left(P(c), P_M(c)\right)$. In Table S-III we report, for each model, the values of the symmetrized Kullback-Leibler divergence between the degree, edge length and node efficiency distributions of the adult *C. elegans* neural network and the networks generated by each of the eight models, which are respectively denoted by $D_{KL}\left(P(k), P_M(k)\right)$, $D_{KL}\left(P(d), P_M(d)\right)$ and $D_{KL}\left(P(E_i), P_M(E_i)\right)$. The best and the second-best value of $D_{KL}(P,Q)$ for each metric are highlighted in green and yellow, respectively. Notice that the smallest values of the symmetrized Kullback-Leibler divergence are consistently obtained by the ESTG model, with the only exception being node efficiency for which PEG outperforms ESTG by a small amount.

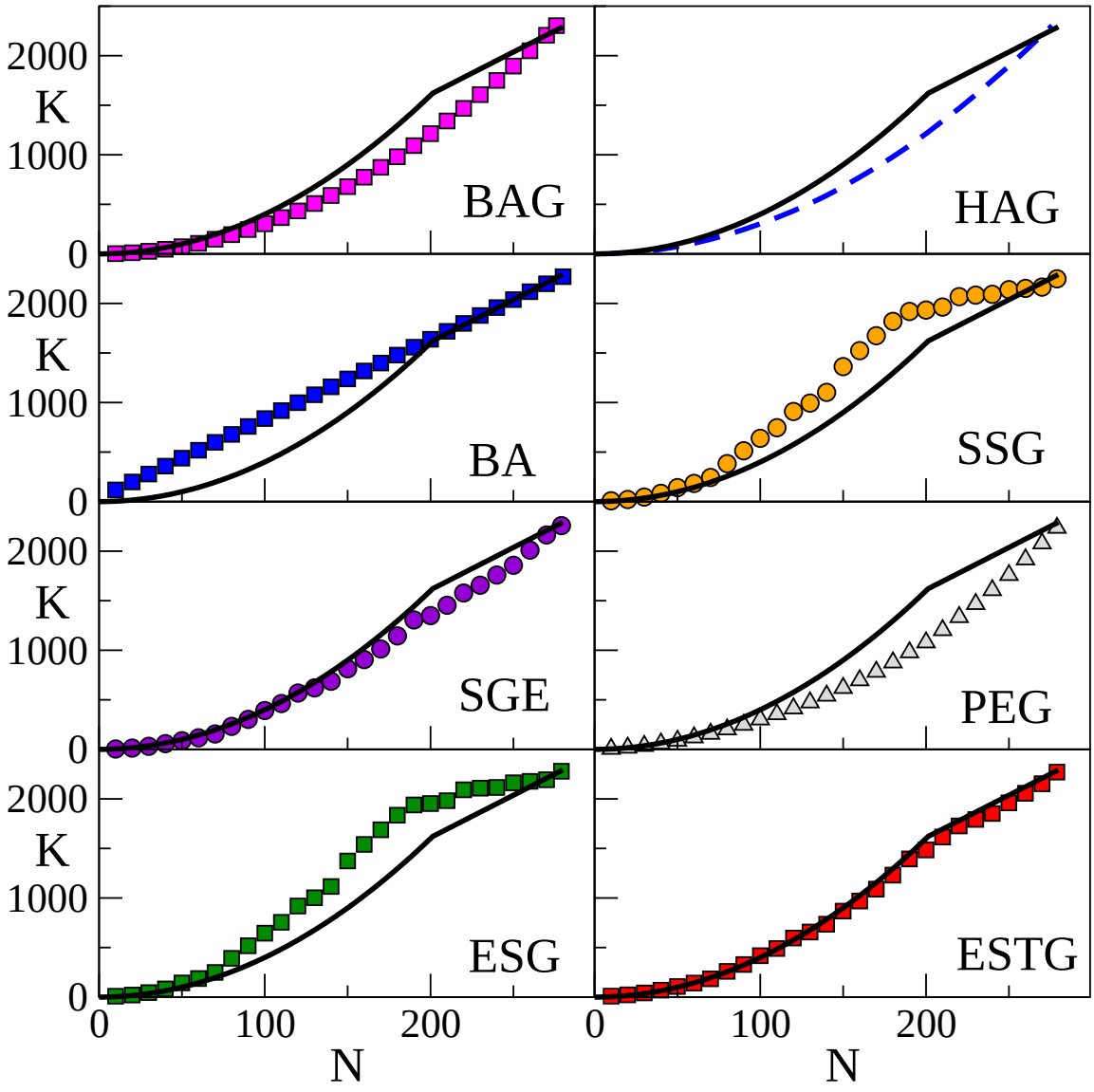

Figure S-2. **Growth curves.** The average total number of edges $\mathcal{K}_M(N)$ as a function of N for each of the eight models. The original growth curve of the *C. elegans* neural network is reported for reference in each panel, as a solid black line. The SSG, SGE, ESG and ESTG models exhibit a transition from a quadratic to a linear increasing regime, but only ETSG is able to closely match the growth curve observed in the original graph.

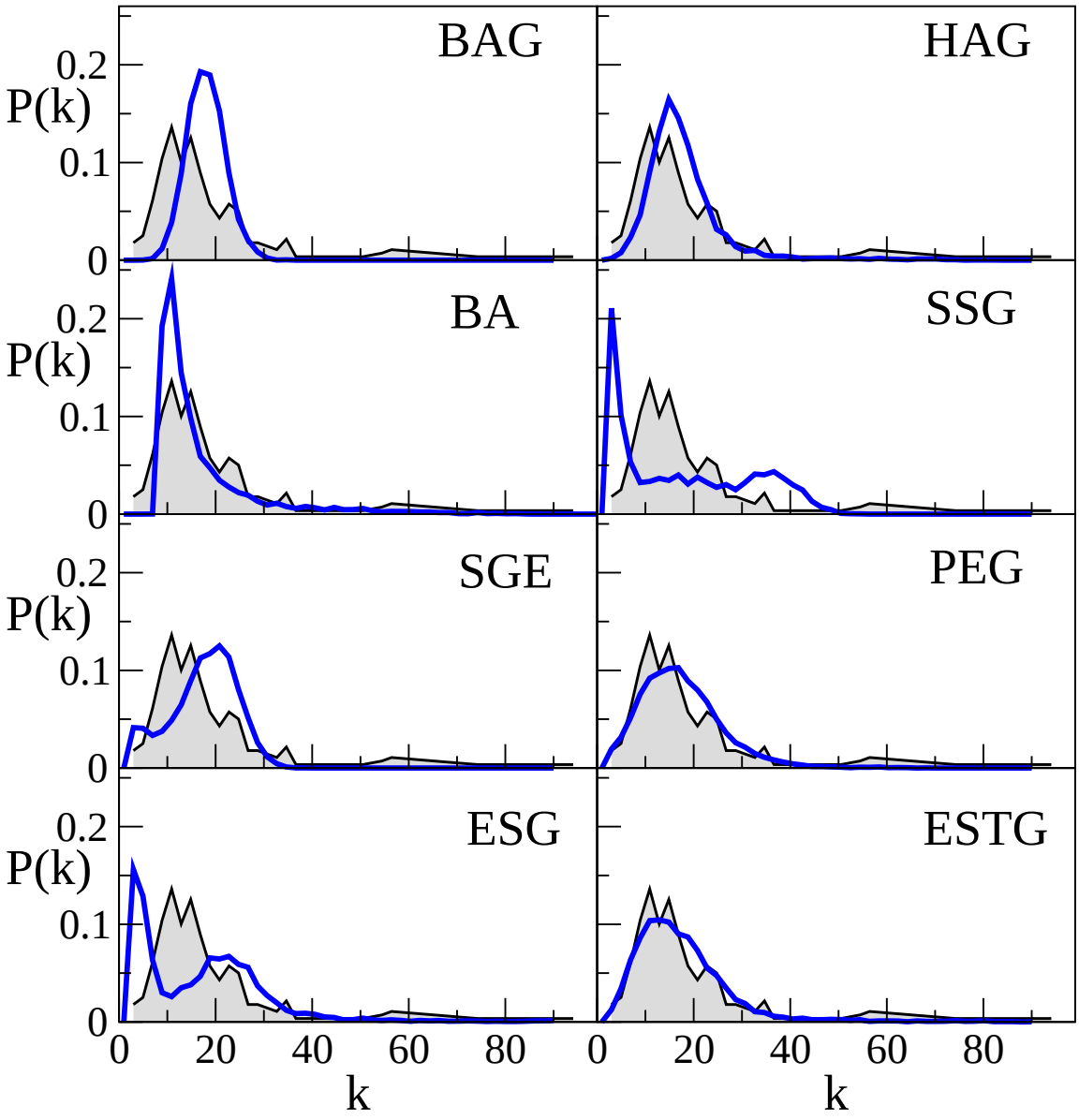

Figure S-3. **Degree distributions.** The average degree distribution of the networks generated by each of the eight models. The degree distribution of the adult *C. elegans* neural networks is reported in each panel in shaded gray, for comparison. Only the models based on hidden-variables, i.e. HAG, PEG and ETSG, are able to reproduce the degree distribution of the worm more closely. In the BA, SSG and ESG models low-degree nodes are over-represented, while in the BAG and SGE models low-degree nodes are substantially under-represented.

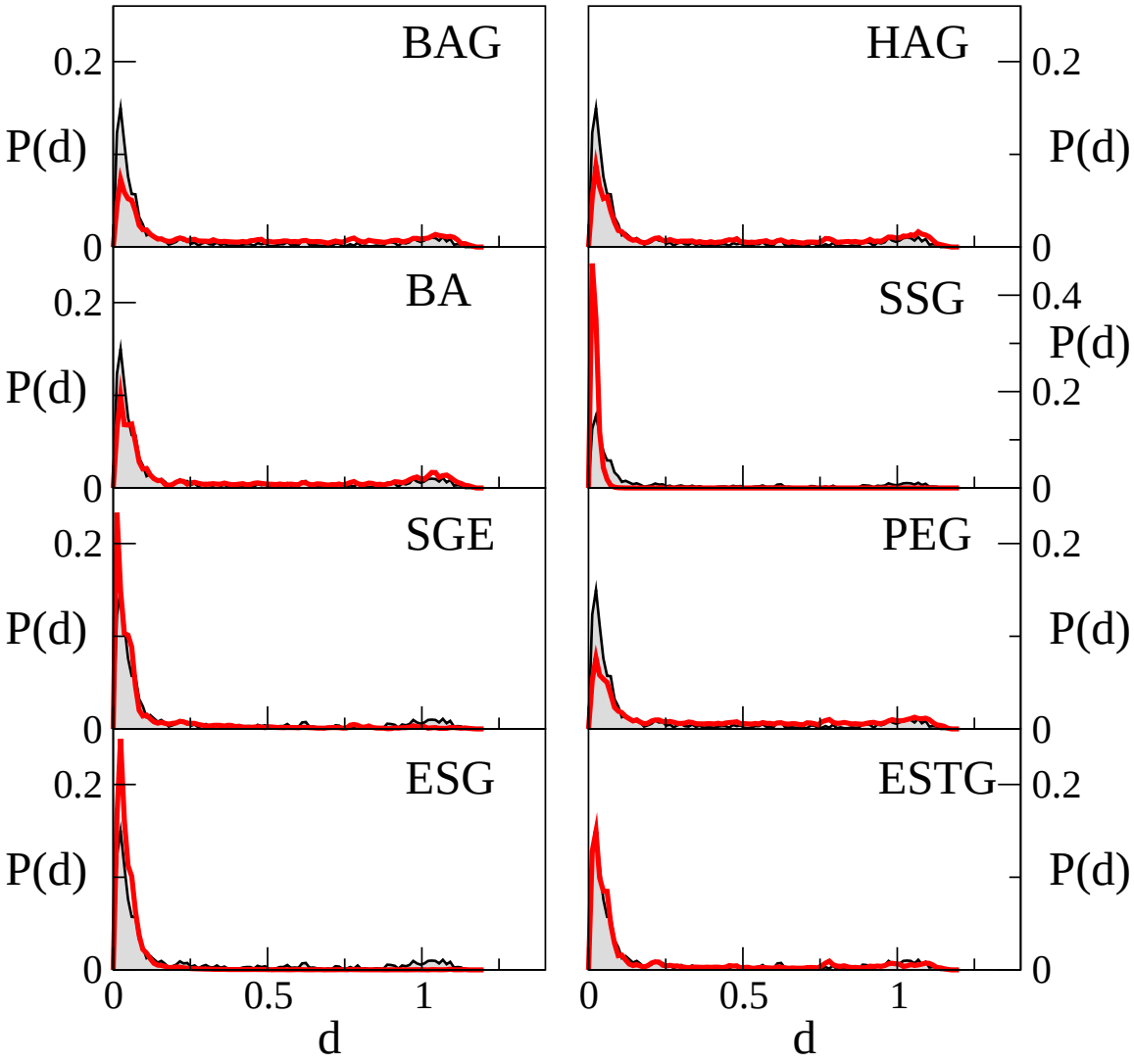

Figure S-4. **Edge length distributions.** The average distribution of edge length in the networks generated by each of the eight models, compared with the distribution of edge length observed in the adult *C. elegans* network (reported in shaded gray). BAG, HAG, BA and PEG produce networks with substantially longer links, while SSG, SGE and ESG exhibit a substantially larger percentage of short links (notice the different scale of the y-axis in the SSG panel). The only model which closely matches the distribution of edge length is ESTG.

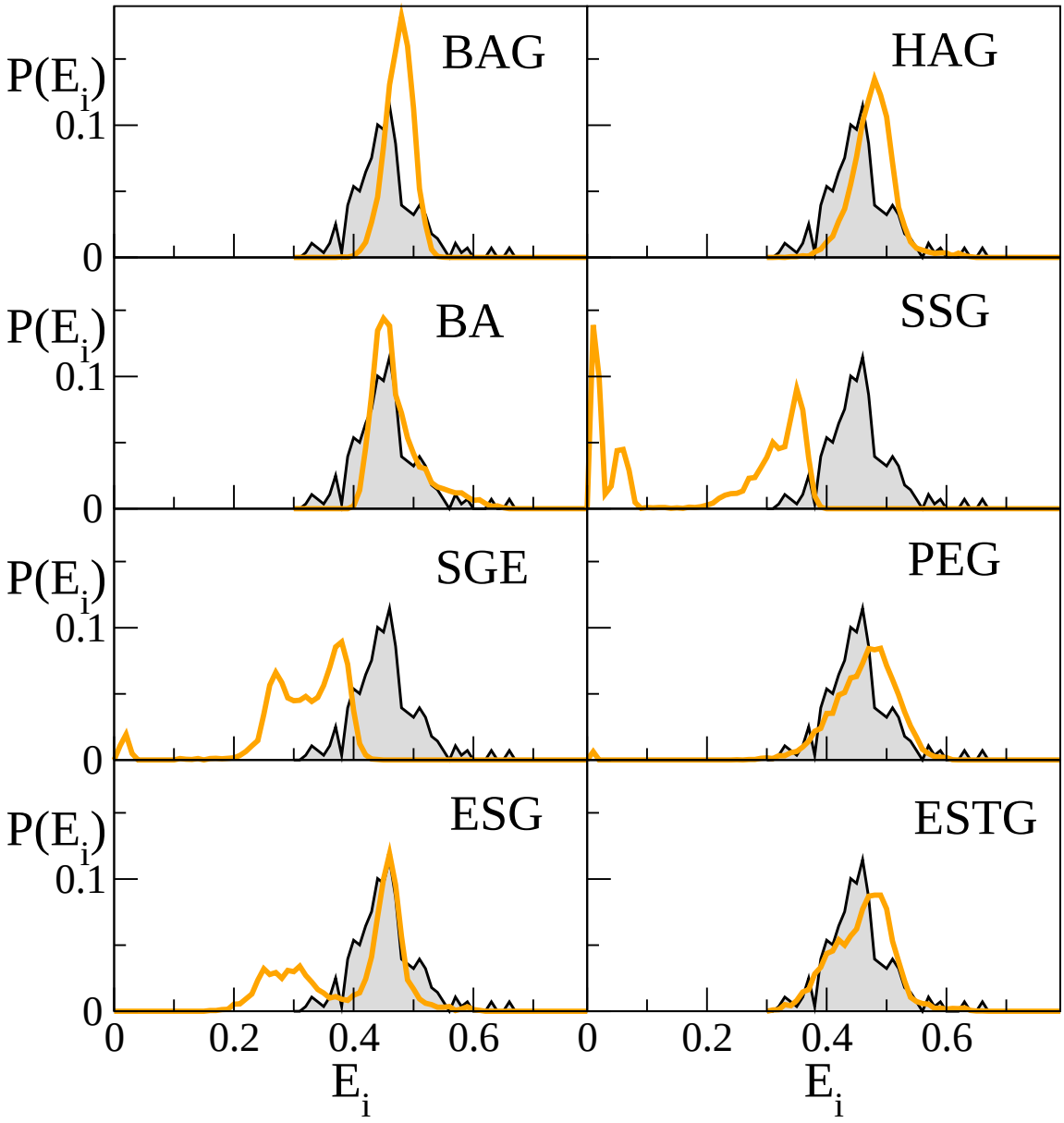

Figure S-5. **Node efficiency distribution.** The distribution of node efficiency of networks generated with each of the eight models, compared with that observed in the adult *C. elegans* (shaded gray). Both BAG and HAG produce binomial distributions of edge efficiency; for SSG, SGE and ESG models the distribution of efficiency is skewed towards smaller values while BA is able to capture the peak around $E_i = 0.47$. PEG and ESTG reproduce the original distribution in a more balanced way, even if nodes with efficiency around 0.5 are substantially over-represented while the peak around 0.47 is missing.