Energy and Time Determine Scaling in Biological and Computer Designs

Melanie Moses,[1,2,3*] George Bezerra,[1]
Benjamin Edwards,[1] James Brown,[2,3] Stephanie Forrest[1,2,3]

[1]Department of Computer Science
University of New Mexico, Albuquerque, NM, USA.
[2]Department of Biology
The University of New Mexico, Albuquerque, NM, USA.

[3]The Santa Fe Institute, Santa Fe, NM, USA.

[*]To whom correspondence should be addressed
E-mail: melaniem@cs.unm.edu
Address: Department of Computer Science
1 University of New Mexico, Albuquerque, NM USA
Phone: 505-277-3112

April 14, 2016

Abstract

Metabolic rate in animals and power consumption in computers are analogous quantities that scale similarly with size. We analyze vascular systems of mammals and on-chip networks of microprocessors, where natural selection and human engineering respectively have produced systems that minimize both energy dissipation and delivery times. Using a simple network model that simultaneously minimizes energy and time, our analysis explains empirically observed trends in the scaling of metabolic rate in mammals and power consumption and performance in microprocessors across several orders of magnitude in size. Just as the evolutionary transitions from unicellular to multicellular animals in biology are associated with shifts in metabolic scaling, our model suggests that the scaling of power and performance will change as computer designs transition to decentralized multi-core and distributed cyber-physical systems. More generally, a single energy-time minimization principle may govern the design of many complex systems that process energy, materials, and information.

1 Introduction

Both organisms and computers have evolved from relatively simple beginnings into complex systems that vary by orders of magnitude in size and number of components. Evolution, by natural selection in organisms and by human engineering in computers, required critical features of architecture and function to be scaled up as size and complexity increased. In biology, Kleiber's law describes empirically how metabolic rate and many other traits, such as lifespan, heart rate, and number of offspring, scale with body size [29]. Similarly, computer architecture has Moore's law to describe scaling of transistor density and performance [38], Koomey's law for the energy cost per computation [32], and Rent's rule for the external communication per logic block [15].

We posit that these empirical patterns originate from a common principle: Networks that deliver resources are optimized to reduce energy dissipation and increase flow rates, expressed here as minimizing the energy-time product. That is, both living systems and computer chips are designed to maximize the rate at which resources are delivered to terminal nodes of a network and to minimize the energy dissipated as it is delivered and processed. For example, in biology the vascular network of mammals supplies oxygen and nutrients to every cell, fueling metabolism for maintenance, growth and reproduction. Since energy is a limited resource, we assume that mammals are selected to minimize the time spent and energy dissipated as oxygen is delivered through the network [55] and processed to produce ATP in the mitochondria. Similarly, computation in microprocessors relies on a network of microscopic wires that transmits bits of information between transistors on a chip. This network is designed to deliver the maximum information flow at the lowest possible energy cost.

Here, we model mammals as composed of nodes (regions of tissue) that process oxygen delivered via a hierarchical vascular network, and we model microprocessors as composed of

nodes (transistors that perform computation) that communicate bits over a network of wires. As each system scales up in size, our model identifies network designs that minimize 1) the rate at which resources are delivered by the network and processed in the nodes; and 2) the energy dissipated during these processes. Despite the obvious differences between animals and chips, we present a general model and derive energy and time scaling relations from physical principles applicable to each system. Using these relations, we express the optimal network design as a trade-off between energy cost and processing speed. This energy-time minimization model is consistent with shifts across the major evolutionary transitions, such as the transition from protists to multicellular animals and the transition from single- to multi-core computer chips. It also points to likely future trajectories of the evolution of computer architecture and to possible extensions of metabolic scaling theory to account for sociality.

Previous biological scaling models have sought either to minimize energy dissipation, e.g., [55] or to maximize resource delivery rate [3], but they have not formalized the trade-offs between these goals. By simultaneously considering energy and time minimization, our analysis helps to explain how nature and engineering are able to produce designs that approach pareto-optimality along the energy-time trade-off, a question investigated extensively in computer architecture, e.g., [26, 1]. Thus, biological evolution has produced mammals ranging in size from mice to elephants, rather than converging on a single optimal size, and computer engineers have designed processors with thousands to billions of transistors, each of which fills a specific computational niche.

In the rest of the paper, we present the unified time-energy minimization model (Section 2) and its assumptions (Section 2.1). We then use the model to derive a series of predictions about how time and energy scale with system size, first for mammals (Sections 3.1 and 3.2) and then for microprocessors (Section 3.3). We discuss new insights into previously analyzed scaling relationships in biology that we gain from the time-energy minimization framework, and

we test our scaling predictions with empirical power and performance data on computer chips. Finally, in Section 4 we discuss the implications of these results for evolutionary transitions in nature and engineering.

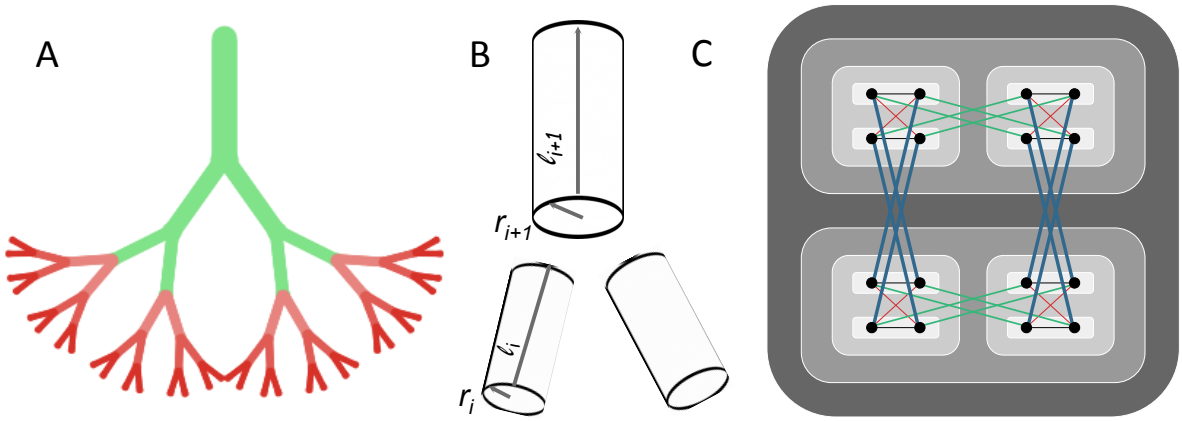

Figure 1: Idealized branching models in biology (Panel A) and computers (Panel C). Panel A shows a cardiovascular tree with branching factor $\lambda = 2$, $H = 5$ hierarchical branchings, and $N = 32$ terminal branches at level 0 that represent capillaries. Panel B shows the radius and length of successive branches: D_r defines the relative radius and D_l defines the relative length of pipe or wire between successive hierarchical levels (i and $i + 1$) in both biology (Panel A) and computers (Panel C). Panel C shows the semi-hierarchical branching of logic wires on a computer chip. Each module within a hierarchical level is shaded the same color. The purple, red, green and blue wires cross 0, 1, 2 and 3 modules respectively. The wire lengths and widths increase as they cross more levels according to D_l and D_r. D_w defines the number of wires, determined by the ratio of internal (intra-module) communication per node to external (inter-module) communication per node. Here $D_w = 2$ so that a node is connected to all nodes within a module (in this case only 1) by a purple wire, 1/2 of the nodes in the next hierarchical level by red wires, 1/4 of the nodes in the next level by green wires, and 1/8 of the nodes in the next level by blue wires.

2 Unified Model of Network Scaling

Vascular systems are hierarchical branching networks where blood vessels (pipes) become thicker and longer through the hierarchy from the capillaries to the aorta. Similarly, micropro-

cessor chips are organized hierarchically into a nested structure of modules and submodules, where wires become longer and thicker as the hierarchical level of a module increases (Figure 1). These wires are organized into metal layers, where short, thin wires are routed on the lowest layers, and long, thick wires are placed on the top layers. We model the scaling of length (l) and thickness (r) of both pipes and wires as:

$$l_i = l_0 \lambda^{\frac{i}{D_l}} \tag{1}$$

and

$$r_i = r_0 \lambda^{\frac{i}{D_r}}, \tag{2}$$

where i is the hierarchical level of a branch or module, λ is the branching factor, and D_l and D_r are the length and thickness dimensions. This model resembles the hierarchical pipe model of vascular systems proposed in [55], where $\lambda^{\frac{1}{D_r}}$ and $\lambda^{\frac{1}{D_l}}$ correspond to β and γ respectively in [55] (note that in [55], the aorta or top of the network is labelled as level 0, while here the smallest branches, the capillaries, are labelled as level 0).

In vascular networks, r represents the radius of cylindrical pipes, and in computer interconnect, r represents the width of wires with aspect ratio 1. D_r describes the relative radius of pipes between successive hierarchical levels. The smallest edges occur at $i = 0$, and have constant radius, r_0, but length, l_0, that scales with system size [3].

The length parameter D_l is determined by the spatial dimension occupied by the nodes of the network [35]. For chips, $D_l = 2$, since transistors are placed on a single two-dimensional layer [20]. For three-dimensional organisms, $D_l = 3$. Because the length of a vessel defines the radius of a 3-dimensional volume of tissue supplied by that vessel, each successive vessel in the hierarchy also scales according to Eq. 1 with $D_l = 3$ [55, 3]. Similarly, the length of each successive wire on a 2-dimensional chip defines the area to which that wire delivers signals [39].

Thus, in the simplest networks that efficiently deliver resources homogeneously throughout a volume or area, D_l describes both the relative length of pipe between successive hierarchical levels and the physical dimension of the system. For example in Figure 1 C, where $\lambda = 2$ and $D_l = 2$, wires are $2^{1/2} = 1.41$ times longer when they connect to successively higher modules in the hierarchy.

Digital circuits scale in a third way beyond length and radius, which has no direct analog in mammalian cardiovascular networks. Digital circuits are partially *decentralized*, with networks that connect multiple sources and destinations, while vascular networks are centralized, with blood flowing from a single heart. In vascular networks, each pipe branches at each hierarchical level forming a tree structure (in the simplest case with $\lambda = 2$ forming a binary tree.) Chips, however, have many connections within each level of the network, and the number of these connections varies systematically with the hierarchical level. To account for this difference, we introduce a new equation, in which the communication (or number of wires) per module increases with the hierarchical level as:

$$w_i = w_0 \lambda^{\frac{i}{D_w}} \tag{3}$$

where D_w is the communication dimension and w_0 is the average number of wires per node. This hierarchical scaling of communication is a well-known pattern in circuit design called Rent's rule [15], where $p = \frac{1}{D_w}$ is the Rent's exponent.* This pattern is not unique to circuits and has been shown to occur in many biological networks [46, 7, 37, 50]. Vascular systems correspond to a special case where $w_i = 1$ for all i.

*Rent's rule is typically expressed as $C(n) = kn^p$, where C_n is the external communication of a module, n is the size of the module (number of nodes), k is the average external communication of a module with size 1, and p is Rent's exponent. For a hierarchy with branching factor of λ, the size of a module is given as $n = \lambda^i$, where i is the hierarchical level. Therefore, we can rewrite Rent's rule as $c_i = c_0 \times \lambda^{ip}$, where $c_0 = w_0$ and $p = \frac{1}{D_w}$.

2.1 Assumptions of the Unified Model

Before presenting the model and deriving scaling predictions, we state the model's assumptions and how they relate to earlier models, both in computation and biology:

1. **Time and energy are equally important constraints:** System designs seek to deliver the maximum quantity of resource per unit time for the minimum quantity of energy expended. In computer architecture this relationship is expressed as the *energy-delay product*, which formalizes the insight that a chip that is ten times faster or ten times more energy efficient is ten times better [27]. In synchronous systems, clock speed (delay between clock ticks) determines the maximum rate at which the system can compute.

2. **Steady state**: Resource supply matches processing demand [5, 3]. That is, the network supplies resources continually to the nodes and is always filled to capacity. This avoids network delays and the need to store resources in the system. Specifically,

   (a) System designs balance network delivery rates with node processing speeds, so that resources are delivered at exactly the same rate that they are processed.

   (b) Pipelining: A concept from computer architecture in which resources, e.g., computer instructions, leave the source at the same rate that they are delivered to the terminal nodes, and the network is always full. Consequently, resources (oxygen molecules or bits) flow through the network continually without bottlenecks, and they do not accumulate at source, sink, or intermediate locations.

3. **Terminal units and service volumes:** We follow previous scaling models of biology, which posit that the service volume (the volume of tissue that is supplied by a single terminal unit of the network) increases with system size and has a fixed metabolic rate [55, 3]. In contrast to [55] we do not assume that terminal branches of the vascular network

have fixed size. Following [3], we assume that the length (l_0) of the terminal branches of the network (e.g., capillaries) is proportional to the radius of the service volume. We also follow the assumptions in [3] that the capillaries have fixed radius, and that the speed of flow (u_0) through the service volume is proportional to its length, so that the rate of arrival of oxygen molecules to mitochondria in the service volume is constant across mammals. In chips, transistor size has shrunk over many orders of magnitude over the past 50 years. Similar to the length scaling of the service volume in mammals, the radius of the isochronic region (the service area) for chips scales proportionally with decreasing transistor size [39]. Thus, service regions are *smaller* in more powerful chips (which have more transistors), but they are *larger* in larger animals. We refer to the service volumes in mammals and the service regions on chips as *nodes*.

In addition to these general assumptions, we make the following refinements to accommodate salient differences between biology and computer architecture.

1. In biology, the energy processed by a node (E_{node}) is invariant with system size. That is, as the size of a service volume increases with body size, the total amount of energy it processes remains constant. We do not make this assumption for chips.

2. Component packing: In chips, we assume that total chip area is constant, and the number of transistors (N) is the square of the process size, i.e., the length of one side of a transistor.

In biology it is known that blood flow slows by several orders of magnitude as it travels from the aorta to the capillaries [55]. Earlier scaling models have generally not characterized this slowing [55, 3], but our equations include velocity as an explicit term, to highlight where it affects time and energy scaling. Here we model D_r as constant within an organism so that blood slows continuously from the heart to the capillaries. We also model D_w and D_l as constant.

Because rates of blood flow, oxygen delivery, and ATP synthesis can be converted one to another by a simple conversion constant, we treat them interchangeably in our scaling model.

3 Model Predictions for Mammals and Microprocessors

We define E_{net} and T_{net} respectively to be the energy dissipated and the time taken by the network to deliver a fundamental unit of resource to each node. For mammals the resource is oxygen (in mammals, carried by a unit volume of blood), and for computers the resource is a bit of information. Similarly, we define E_{node} and T_{node} as the energy dissipated and the time taken by the nodes to process that resource. For mammals, the node is the service volume corresponding to a region of tissue supplied by a single capillary [3], which corresponds to a volume of tissue containing a constant number of mitochondria [56], the organelles that process oxygen molecules to generate biologically useful energy in the form of ATP. A node is defined as having a constant rate of delivery of oxygen and processing of oxygen, but the volume of a node varies with organism size.

E_{net} is the energy required to deliver oxygen to the cells (as analyzed in [55]), and E_{node} is the energy dissipated by cells processing incoming oxygen. T_{net} is the time delay between delivering each oxygen molecule to the cell, and T_{node} is the time taken for the cell to process each oxygen molecule. From the steady-state assumption, $T_{net} = T_{node}$, i.e., supply matches demand as in [3].

In microprocessors, the nodes are transistors, and E_{net} and E_{node} represent the energy dissipated as bits are delivered to transistors and the energy required to process the bits at the node. T_{net} and T_{node} are the times required to deliver and process a bit at the node (i.e., network and transistor switching delay). In computers, the time taken to deliver and process bits is bounded by $\max(T_{net}, T_{node})$, i.e. a node cannot process another bit until it is delivered, and a node cannot process a new bit until it is finished processing the previous bit. For both mammals and

microprocessors we define the total energy as the sum of energy dissipated in the network plus the energy dissipated in the nodes: $E_{sys} = E_{net} + E_{node}$. [†]

In the following, we derive general scaling relationships between E_{net}, T_{net}, E_{node} and T_{node} and the number of nodes N, under the assumption that the energy-time product is minimized. N is our measure of system size (number of capillaries or number of transistors). In mammals, larger N implies larger organism volume and mass. For computer chips, N increases by shrinking components, and so increasing N does not imply increasing chip area, which we assume to be constant.

The hypothesis that mammals and computers minimize the energy-time product predicts that optimized system designs will achieve the highest performance per cost, where performance is given by flow and cost by energy expended. To show this mathematically, we express the optimal network design as a constraint optimization problem in which the whole system's energy-time product is minimized as:

$$\min_{D_r, D_w, D_l} (E_{sys} \times T_{sys}) \tag{4}$$

We derive expressions for E_{sys} and T_{sys} for mammals (Section 3.1) and microprocessors (Section 3.3) in terms of the dimensions D_r, D_w, and D_l where D_l is fixed by the external dimensions of the system.

3.1 Mammallian Cardiovascular Network

In this section, we derive general energy and time scaling relations for the cardiovascular network and nodes, and then use them to minimize Eq. 4. We first define scaling relationships for

[†]For computers it is intuitive that these quantities can be treated independently. In biology, this is less obvious because the heart which powers the vascular network is itself composed of cells (nodes) that require oxygen delivery, an apparent circularity. However, the metabolic power of the heart (E_{net}) is supplied by oxygen delivered directly to the heart by the coronary artery, bypassing the rest of the vascular network. Thus we treat E_{net} independently from E_{node}.

the four key quantities E_{net}, T_{net}, E_{node} and T_{node}, and then show how they scale with N when Eq. 4 is minimized. In contrast to computer scaling, several theoretical scaling models have been proposed for animals over the last century, e.g., [52, 55, 6, 18, 3]. The influential West et al. model [55] predicted scaling relationships by minimizing energy dissipation, whereas an alternative model [3] maximized metabolic rate by minimizing the time to deliver oxygen. Not surprisingly, scaling models that assume different optimization principles make different predictions [42]. Our model combines both energy and time constraints into a single framework.

E_{net}: From basic principles of hydraulics, the energy dissipated to transport a constant volume of blood through the network is given by the loss in pressure from the aorta to the capillaries multiplied by the volume being transported. The loss in pressure is the product between hydraulic resistance (R) and flow (Q), so $\Delta P = RQ$. Thus $E_{net} \propto \Delta P = RQ$.

E_{node}: Following [55] and [39], we assume that the quantity of energy dissipated to metabolize a fixed quantity of oxygen in each node is constant. Therefore, the energy summed over all nodes is $E_{node} \propto N$.

T_{net}: The time to deliver a fixed number of oxygen molecules to the nodes is given by the volume of blood being transported divided by the flow (Q). Since a constant volume is delivered to each node in parallel, we consider the volume being distributed per unit time to all nodes, giving $T_{net} \propto N/Q$.

There is no distance term in the T_{net} equation. This is because T_{net} is defined as the time to deliver the 'next' oxygen molecule from a capillary, consistent with the steady-state assumption. It is not the time it takes a single molecule to traverse the network (i.e. it is not τ in [3]), but rather the inverse of the rate at which oxygen molecules are delivered to the nodes, analogous to the inverse of clock speed in computer chips.

T_{node}: From the steady-state assumption, $T_{node} \propto T_{net} \propto N/Q$.

Substituting these relationships into Eq. 4 (where $E_{sys} = RQ + N$, and $T_{sys} \propto N/Q$) gives:

$$\min(E_{sys} \times T_{sys}) = \min_{D_r, D_w, D_l} (RN + \frac{N^2}{Q}) \tag{5}$$

where N is the number of terminal units.

We now show how R and Q scale with N. The resistance of a pipe is given by the well-known Hagen-Poiseuille's equation, where R at hierarchical level i is $R_i = \frac{8\mu l_i}{\pi r_i^4}$ and μ is the viscosity constant. The total network resistance R is given by [55]:

$$R = \sum_{i=0}^{H} \frac{8\mu l_i}{\pi r_i^4} \frac{1}{n_i} = \frac{8\mu l_0}{\pi r_0^4} \lambda^{-H} \sum_{i=0}^{H} \lambda^{i\left(\frac{1}{D_l} - \frac{4}{D_r} + 1\right)} \tag{6}$$

where there are $H + 1$ hierarchical levels, and $n_i = \lambda^{H-i}$ is the total number of pipes at hierarchical level i.

Next, we consider upper and lower bounds for D_r given the objective of minimizing the energy-time product (Equation 5). Recalling that $\lambda^{-H} = N^{-1}$, in the case where $D_r \leq \frac{4D_l}{1+D_l}$, the summation in Eq. 6 converges to a constant ($\log(N)$ in the case of equality), and

$$R \propto l_0 N^{-1} \tag{7}$$

As D_r increases above $\frac{4D_l}{1+D_l}$, R increases from $\propto l_0 N^{-1}$ to $\propto l_0 N^{\frac{1}{D_l} - \frac{4}{D_r}}$. See Section 6 for details of the calculation.

Flow through a pipe is defined as $Q = u\pi r^2$, where u is the fluid velocity. Therefore, flow through the aorta equals $Q = u_H \pi r_H^2$, and substituting from Eq. 2, $Q = u_0 \pi r_0^2 \lambda^{\frac{2H}{D_r}} = u_0 \pi r_0^2 N^{\frac{2}{D_r}}$. Since we do not assume that u_H is independent of N, u_0 appears in the equations. If Q is equal at all levels of the network (steady state assumption) then:

$$Q \propto u_0 N^{\frac{2}{D_r}}. \tag{8}$$

With R and Q in hand, we now substitute these relationships into the equations for E_{net},

E_{node}, T_{net}, and T_{node}, obtaining the scaling predictions shown in the first column of Table 1. It it is evident that the scaling behavior or E_{net} depends on the value of D_r:

Case 1: $D_r \leq \frac{4D_l}{1+D_l}$: $E_{net} \propto l_0 u_0 N^{\frac{2}{D_r}-1}$

Case 2: $D_r > \frac{4D_l}{1+D_l}$: $E_{net} \propto l_0 u_0 N^{\frac{1}{D_l}-\frac{2}{D_r}}$

Given that $D_l = 3$ for 3 dimensional animals, and that D_r must be greater than 2 to accommodate the necessary slowing of blood as it flows toward the capillaries then [55], Case 1 applies for $2 \leq D_r \leq 3$, and Case 2 applies for $D_r \geq 3$.

Section 6 gives the derivations for E_{net} for all values of D_r. Here we show the case ($D_r \leq 3$) that minimizes the scaling of the energy-time product (Eq. 5):

$$\min_{D_r}(RN + \frac{N^2}{Q}) \propto l_0 + u_0^{-1} N^{2-\frac{2}{D_r}} \tag{9}$$

The energy-time product is dominated by the second term in Eq. 9, which is minimized by setting D_r to its minimum possible value. Thus, minimizing the energy time-product requires $D_r = 2$ (Case 1), and:

$$E_{net} \propto l_0 u_0 N^{\frac{2}{D_r}-1} \propto l_0 u_0 \tag{10}$$

		$\mathbf{D_r \leq \frac{4D_l}{1+D_l}}$	$\mathbf{D_l = 3, D_r = \frac{24}{11}}$
Mammals	E_{net}	$l_0 u_0 N^{\frac{2}{D_r}-1}$	$N^{\frac{1}{12}}$
	E_{node}	N	N
	T_{net}	$u_0^{-1} N^{1-\frac{2}{D_r}}$	N^0
	T_{node}	$u_0^{-1} N^{1-\frac{2}{D_r}}$	N^0
	$E_{sys} \times T_{sys}$	$l_0 + u_0^{-1} N^{2-\frac{2}{D_r}}$	$N^{\frac{1}{12}} + N$
		$\mathbf{D_w \geq \frac{D_l}{D_l - 1}}$	$\mathbf{D_l = 2, D_w = 2}$
Computers	E_{net}	$N^{1-\frac{1}{D_l}}$	$N^{\frac{1}{2}}$
	E_{node}	$N^{1-\frac{1}{D_l}}$	$N^{\frac{1}{2}}$
	T_{net}	N^0	N^0
	T_{node}	$N^{-\frac{1}{D_l}}$	$N^{-\frac{1}{2}}$
	$E_{sys} \times T_{sys}$	$N^{1-\frac{1}{D_l}} + N^{1-\frac{1}{D_l}}$	$N^{\frac{1}{2}} + N^{\frac{1}{2}}$

Table 1: Predicted scaling relationships for mammals and computer chips. The first column shows the general scaling equation under conditions that minimize the energy-time product. The second column shows how each quantity scales with N given the values of the dimensional parameters that minimize the energy-time product, $D_r = \frac{24}{11}$ for mammals and $D_w = 2$ for chips.

3.2 Biological scaling predictions from the energy-time minimization model

Earlier scaling models showed that area-preserving branching ($D_r = 2$) leads to the 3/4 power scaling of metabolic rate with body size known as Kleiber's law (e.g., [55, 3]). However, in animal circulatory networks blood must slow before reaching capillaries in order to reduce pressure on the walls of small vessels and to allow oxygen to be dissociated from hemoglobin in the capillaries. Under this circumstance, perfect area-preserving branching is not feasible, and D_r must be greater than 2.

We make a specific prediction for the value of D_r that minimizes the energy-time product while both slowing the flow of blood to the capillaries and matching the supply and demand for oxygen in the nodes. By our definition of a node as the volume of tissue that processes oxygen at a fixed rate, T_{node} must be invariant. Table 1 shows the model prediction $T_{node} \propto u_0^{-1} N^{1-\frac{2}{D_r}}$.

Following [3], in the optimal case u_0 increases with organism mass, and therefore with N. See Section 6.1 for the derivation that $u_0 \propto l_0 \propto N^{\frac{2}{3D_r}-\frac{2}{9}}$. Substituting this equation for u_0 into the equation for T_{node} in Table 1, we find that T_{node} is invariant with respect to N when $D_r = \frac{24}{11} = 2.18$. The last column of Table 1 lists the scaling predictions given this value of D_r.

We test the prediction that $D_r = 24/11$ using data from [31]. This influential Kolokotrones et al. paper showed that metabolic rate is elevated in both small and very large mammals, indicating systematic deviations from a simple power-law relationship between metabolism and mass. Although the deviation appears only as a slight curvature in the canonical log-log plots, as shown in Figure 2, it is important because it calls into question prior scaling models that purport to explain a universal scaling exponent.

We derive the equation relating metabolism (B) to mass (M), following the approach used in [3], but we relax the assumption that $D_r = 2$ giving[‡] $M \propto N^{\frac{2}{d_r}+\frac{1}{3}}$ and

[‡]These expressions are consistent with those in [3], specifically when $D_r = 2$, $N \propto M^{\frac{3}{4}}$ and $l_0 \propto M^{\frac{1}{12}} \propto N^{\frac{1}{9}}$ and when $D_r = 3$, $N \propto M$ and $l_0 \propto M^0 \propto N^0$.

$$B \propto M^{\frac{18-8D_r}{6+Dr}} + M^{\frac{24-2D_r}{18-3D_r}} \tag{11}$$

See subsection 6.1 for details of the calculations.

Although this prediction for B is not as simple as the $\frac{3}{4}$ scaling predicted by West et al. [55] or the alternative models proposed by Kolokotrones et al. [31], the exponents in Equation 11 arise naturally by combining two scaling relationships: that of the metabolic rate of the nodes and the metabolic power required to drive the network.

By considering blood slowing through the network due to $D_r > 2$ and by including energy dissipated in both the network and the nodes, each with different scaling exponents, the model naturally generates the curvature observed in the data. Intuitively, in smaller animals a greater fraction of energy is consumed by E_{node}, a term that is linear in the number of nodes.

We tested the predicted value of $D_r = \frac{24}{11}$, which minimizes the energy-time product, and find a marginally better fit (solid blue line in Figure 2), than alternative models in [31]. The Mean Squared Error (MSE) for our model is 0.0271 vs 0.0287 for the extended West et al. model (red dotted line in Figure 2). The alternative models in [31] that were specifically designed to account for curvature have MSE 0.274 and 0.0277. We also calculated a value of D_r that is the best statistical fit to the data. Following [31] we use least squares regression, eliminate the orca which is an outlier, and choose scaling constants to best fit the data. We find that $D_r = 2.50$ gives the best statistical fit (dashed green line in Figure 2). Alternative fitting methods and inclusion of the outlier have negligible effect on the best-fit value of D_r.

The energy-time minimization model is the only model that naturally generates curvature accounting for the elevated metabolic rate of the largest mammals as well as the smallest. The predicted value of D_r between 2 and 3 is also consistent with the idea that the upper region of the network is area preserving with $D_r = 2$, while $D_r = 3$ in the lower region as proposed by [55], and it is consistent with the empirical radius scaling reported in [42].

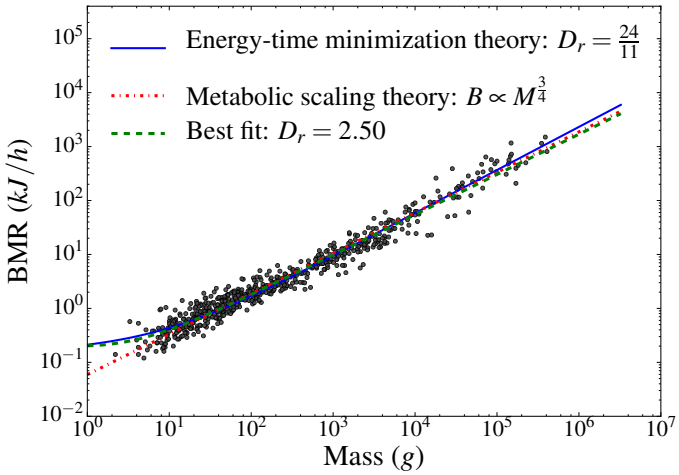

Figure 2: The energy-time minimization model predicts metabolic scaling in mammals. Data from [31] show slight, but theoretically important, curvature in the scaling of metabolic rate vs. mass of mammals. The theoretical optimum predicted by Eq. 11 with $D_r = \frac{24}{11}$ is shown as a solid blue line. The West et al. $3/4$ scaling prediction [55] is shown as a dashed red line, and the best empirical fit of Eq. 11 to the data is shown as a dashed green line ($D_r = 2.50$).

3.3 Microprocessor Model

We now apply the same reasoning to computer chips. In computers, unlike biology, nodes (transistors) are not constant size but have shrunk by many orders of magnitude over 40 years of microarchitecture evolution. During this time, total chip area has grown much more slowly, and we assume it to be constant for our calculations. In addition, the total area of all transistors on the chip is a fixed fraction of the area of the chip [39]. Putting these two constraints together, the linear dimensions of transistors decrease with transistor count as $N^{-1/2}$ (more generally, N^{-1/D_l}). The width of the smallest wires is $r_0 \propto N^{\frac{-1}{D_l}}$ because minimum transistor size and wire width are both determined by the process size. Similarly $l_0 \propto N^{\frac{-1}{D_l}}$ because transistor linear density is increasing as $N^{\frac{1}{2}}$. Intuitively, this means that the number of nodes increases as smaller transistors are placed closer together and connected with smaller and shorter wires. In the following, we assume that all wires carry the same flow and that information is transferred

synchronously. We now calculate how E_{net}, T_{net}, E_{node} and T_{node} scale with the number of transistors, N, and the three scaling dimensions, D_l, D_r and D_w.

E_{net} can be calculated from basic principles of electronics as the energy dissipated to transmit a bit over a wire: $\frac{CV^2}{2}$, where C is capacitance and V is voltage. Because V has remained approximately constant over the last four decades (decreasing only by a factor of five while transistor count increased by six orders of magnitude [43]), we estimate that the total energy to transmit all bits over the network scales as C [11]. Ignoring fringe effects and for an aspect ratio of 1, wire capacitance is proportional to wire length, $C = \epsilon l$ [58], where ϵ is the dielectric constant. Thus, the network capacitance is the sum of the capacitances of all wires, which is proportional to the total wire length of the network [19]:

$$E_{net} \propto C \propto \sum_{i=0}^{H} l_i w_i n_i \propto l_0 w_0 \lambda^H \sum_{i=0}^{H} \lambda^{i\left(\frac{1}{D_l}+\frac{1}{D_w}-1\right)} \tag{12}$$

where at all levels i, l_i is the length of wire, w_i is the number of wires per module, and n_i is the number of modules. Recalling that $l_0 \propto N^{-1/D_l}$ and $\lambda^H \propto N$ gives:

$$E_{net} \propto N^{(1-\frac{1}{D_l})} \sum_{i=0}^{H} \lambda^{i\left(\frac{1}{D_l}+\frac{1}{D_w}-1\right)}. \tag{13}$$

Note that the scaling of E_{net} with N depends on D_l and D_w, but not on D_r. Similarly to energy scaling in mammals, how E_{net} scales depends on whether the exponent $\frac{1}{D_l} + \frac{1}{D_w} - 1$ in Eq. 13 is positive or negative. If $D_w \geq \frac{D_l}{D_l - 1}$ the exponent is negative and the summand converges to a constant ($\log(N)$ in the case of exact equality), leaving $E_{net} \propto N^{1-\frac{1}{D_l}}$. When $D_w < \frac{D_l}{D_l-1}$, $C \propto N^{\frac{1}{D_w}}$. Given $D_l = 2$ for two-dimensional chips, E_{net} is minimized when $D_w \geq 2$. See Section 7 for details.

We now calculate the scaling of E_{node} ignoring leakage power.[§] For a single node, compu-

[§]Transistors and other devices conduct a small amount of current even when they are not being used. This energy loss is referred to as leakage power and is a significant issue in modern microprocessor design not explicitly addressed by our model.

tation energy is given by the transistor's (dynamic) energy dissipation as $\frac{CV^2}{2}$. Again assuming constant V and the capacitance of a transistor proportional to its length (l_0), E_{node} is obtained by summing the capacitance across all N nodes giving $E_{node} \propto Nl_0 \propto N^{1-\frac{1}{D_l}}$.

We calculate T_{net} as the time to transmit a bit over the last wire in the network that connects to each transistor. This assumes perfect pipelining so there is no delay in signal arriving at the last wire (Appendix 7 shows that perfect pipelining requires $D_r = 2$). Thus, T_{net} is equivalent to the wire latency which equals resistance multiplied by the capacitance of the wire (RC). For wires with aspect ratio 1, $R_i = \rho l_i / r_i^2$, where ρ is the resistivity of the material, and $C_i \propto l_i$ as above. Thus,

$$T_{net} \propto R_0 C_0 \propto \frac{{l_0}^2}{{r_0}^2} \propto N^0. \tag{14}$$

$\frac{{l_0}^2}{{r_0}^2}$ is constant because in chips $l_0 \propto r_0$ and both are determined by process size.

Computation time for each node, T_{node}, is calculated as the transistor delay, $\frac{CV}{I}$ [2], where again V is constant and C is proportional to transistor length: $T_{node} \propto C_0 \frac{V}{I} \propto l_0 \propto N^{-\frac{1}{D_l}}$.

Before calculating the energy-time product, we observe that T_{net} is the only term that depends on D_r, so we set $D_r = 2$ to minimize T_{net}. Similarly, E_{net} is the only term that depends on D_w, and we set D_w to minimize E_{net}. In summary, given $D_l = 2$, the terms of the energy-time product are minimized when $D_r = 2$ and $D_w > 2$. Although the energy-time product is minimized for values of D_w greater than 2, this would entail greater communication locality, which is challenging to engineer and doesn't improve the energy-time product. Thus, the model predicts that $D_w = 2$, which is consistent with observed Rent's exponents that approach 1/2 [59, 50]. The scaling relations for various quantities are summarized in Table 1.

3.4 Predictions for Microprocessors

Summarizing the results from the previous section, the energy-time product for chips is minimized when $D_l = D_r = 2 = D_w$. This result corresponds to ideal scaling, as suggested by Dennard [17], where the linear dimensions of transistors and wires scale at the same rate, wire delay is constant, and the Rent's exponent is 1/2.

The final energy-time product scales as $N^{1/2}$ (Table 1), showing that, unlike mammals, as size increases, the energy-delay product per node decreases systematically. Thus, chips have become faster and they consume less energy per transistor as more transistors are packed onto a chip. Of course, this trend arises from the remarkable miniaturization of transistors and wires described by Moore's law. It is not surprising that transistors are faster (T_{node}) and require less energy (E_{node}) as they become smaller. It also makes sense that E_{net} grows sub-linearly with the number of transistors, because as N increases the distance between nodes is reduced. Additionally, $D_w = 2$, means that most bits move locally, so the distance between nearest nodes affects the average distance that bits are transmitted. The only term in the energy-time product that does not decrease with increased N and decreased process size is T_{net} which remains constant under Dennard scaling where wire radius and length scale proportionally to each other.

These scaling models make two testable predictions. First, power consumption (P) in chips (total energy dissipated per unit of time) scales as

$$P = \frac{E_{sys}}{T_{sys}} \propto N^{1/2}. \tag{15}$$

Second, performance, measured as computations executed per unit of time, or throughput (Tp), is predicted to scale linearly with N, i.e.

$$Tp \propto \frac{N}{T_{sys}} \propto N. \tag{16}$$

We compared our theoretical predictions for active power consumption (ignoring leakage power as discussed above) with data obtained for 523 different microprocessors over a range of approximately 6 orders of magnitude in transistor count (See section 7.3 for details of the data collection). The data are shown in Figure 3, where the measured exponent was 0.495 (95% confidence interval = 0.46 to 0.53) which agrees closely with our prediction of 0.5. Consistent data on performance across many technology generations is difficult to obtain because reporting standards have changed over the years and their adoption by different vendors is not uniform. We obtained normalized performance data for 100 different Intel chips, measured with Dhrystone Millions of Instructions per Second (DMIPS), from a variety of sources (see Section 7.3). These sources included a variety of published third-party performance comparisons from different generations over a range of 6 orders of magnitude in transistor count. The best-fit exponent for these data is 1.11 (95% confidence interval =1.07 to 1.15), as shown in Figure 4. This is close to our predicted exponent of 1, suggesting that engineered designs slightly outperform the theoretical optimum defined by the model. Performance and throughput were fit using least squares regression, assuming that there are no significant errors in the reported count of the number of transistors [36].

It is somewhat counter-intuitive that performance increases only linearly with the number of transistors. Given that transistor switching times have decreased dramatically as size has decreased, one might expect performance to increase as the product of clock speed and transistor number (N). However, this is not the case, and we show the expected performance if time were actually the inverse of clock speed in the red dotted line in Figure 4. Some performance increases are achieved by increasing clock speed for a given manufacturing process, which may account for the higher than predicted scaling exponent[¶]. This analysis confirms that the network is indeed the bottleneck. The network delivers bits to transistors at a constant rate per

[¶]Additionally, higher end chips are more likely to be benchmarked, potentially leading to a bias in the data towards higher performing chips.

transistor (Eq. 14), so performance has increased only linearly with transistor number even though, in principle, smaller transistors could process information more quickly. As in biology, performance cannot be understood without considering the constraints of the network.

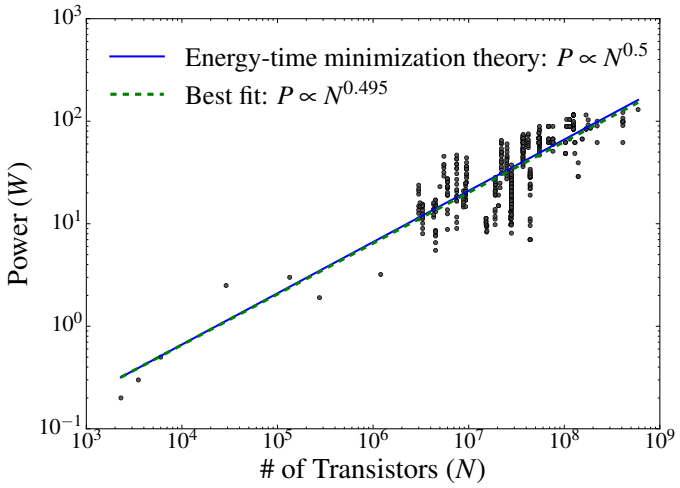

Figure 3: The energy-time minimization model predicts power scaling in chips. Each data point represents a microprocessor chip, with active power and number of transistors per chip from [39]. The energy-time minimization model prediction (Eq. 15) is shown in blue, and the best fit line is shown in green.

Our model provides a simple theoretical explanation for the scaling of power and performance in computers over 40 years of microprocessor technology improvements. The excellent agreement between the theoretical optimum and experimental data suggests that through successive generations of trial-and-error, innovation and optimization, engineered designs are highly successful, approaching and sometimes exceeding the theoretical optimum predicted by the model.

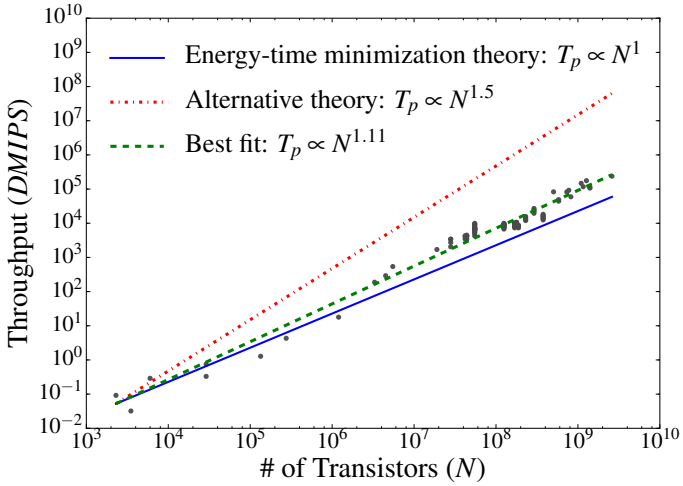

Figure 4: The energy-time minimization model predicts how throughput scales with the number of transistors. The raw data and their sources are included as supplementary material. The model prediction (Eq. 16) is shown as a solid blue line. The dotted red line shows an alternative prediction if throughput were bound by the nodes (switching speed) rather than the network. The green dashed line is the best fit to the data.

4 Discussion

4.1 Summary of scaling predictions

Scaling analyses provide a framework for understanding critical parameters and constraints on the design of both biological and computational systems spanning an enormous range of sizes. We have presented a unified model which predicts scaling relationships for both mammals and microprocessors by simultaneously minimizing energy dissipation and delivery time. The energy-time minimization model highlights the similarities and differences between biological networks that deliver oxygen and computational networks that deliver information. Earlier scaling models focus either on minimizing energy dissipation or on minimizing delivery time, e.g., [55, 3]. Here we extend that work by considering minimization of energy and time simultaneously and investigating the tradeoffs between them.

This theoretical model makes testable scaling predictions for biological metabolism and for

the power and performance of computers. In biology, Section 3.2 extends previous predictions to explain the observed curvature in metabolic scaling of mammals. Other studies have interpreted the deviation from linear scaling as indicating that there is no single unified metabolic scaling theory, for example as imperfect matching of supply and demand [5]. The framework presented here accounts for curvature in the optimization model by including both time and energy minimization in both the network and the nodes. In computation, the unified model accurately predicts Rent's exponents, active power consumption and chip performance in over 40 years of chip design. Thus, the model provides evidence of strong convergence between natural and engineered designs due to physical constraints despite the obvious differences between them.

The model presented here is, of course, a simplification of the more complex reality. For example, our analysis assumes that D_l, D_r, and D_w are fixed constants throughout the network both within and across systems. In reality, each of these may vary. For example, [42] did not find evidence for a constant $D_l = 3$ in mouse vasculature, suggesting that the network does not deliver resources uniformly throughout the body volume. This is not surprising given that different tissues and organs have different metabolic requirements. D_r may vary within the vascular network with area-preserving branching closer to the heart and area-increasing branching slowing blood velocity in smaller vessels, but [42] find values for D_r consistent with our predictions. Similarly, there is evidence that D_w varies across hierarchical levels in computer chips [45]. Including these factors in the model would allow more accurate predictions, but they are unlikely to substantively change the order-of-magnitude predictions of our simple unified model.

Our model makes novel predictions both for mammals and microprocessors. For mammals we give the first quantitative prediction for D_r that accounts both for blood slowing through the network and for the empirically observed curvature in scaling relations that cause small and very

large mammals to deviate from $3/4$ scaling predictions. Additionally, this prediction ($D_r = \frac{24}{11}$) gives an energy-time product that is approximately linear with N ($E_{sys} \propto N^{1/12} + N^1$, Table 1). Highlighting the inherent tradeoff between energy dissipation and delivery times has important implications for understanding the energetic basis of fitness. Some have proposed that biological fitness maximizes metabolic power (energy/time) [34, 44], whereas others have proposed that it minimizes biological times (e.g., generation times, which is equivalent to maximizing vital rates) [33, 49]. The invariance of the energy-time product on a per-node basis is consistent with the idea that organism fitness is largely independent of body mass. Mammals of all sizes, from small, fast, low-power mice to large, slow, powerful elephants, coexist and, therefore, are likely nearly equally fit. This implies a direct trade-off between maximizing metabolic power and minimizing generation times, which holds over the many orders-of-magnitude variation in body mass. The energy-time product reflects powerful geometric, physical and biological constraints on the evolution of organism designs.

In computation, the model accurately predicts power consumption and performance of computer chips as simple functions of the number of transistors. These order-of-magnitude performance predictions highlight that delivery of bits through the network, rather than processing bits at the transistors, is the rate limiting step that constrains performance. More precise predictions may be obtained by incorporating additional factors, for example leakage power, which comprises an increasing fraction of the power budget of computer chips [26].

4.2 Implications for Evolutionary Transitions

The similarities between biological and computational scaling suggest future trajectories in computing based on how the fundamental structural and functional properties of organisms from bacteria to mammals have changed over evolutionary time. Work by Delong et al. [16] demonstrated that the slopes and intercepts of metabolic scaling relations change at the evo-

lutionary transitions: prokaryote (bacteria) metabolic rate varies *superlinearly* with size, unicellular protist rate varies *linearly*, and whole-organism metabolic rate of multicellular animals scales *sublinearly*, converging to the canonical $3/4$ exponent that approximates the mammalian scaling described above. They hypothesize that these discontinuous scaling shifts arise from body plans overcoming pre-existing constraints, and then accommodating to new constraints, as body size and complexity increase.

Delong et al. hypothesize the following: Larger bacteria have higher metabolic rates because their larger genomes allow increased use of metabolic substrates, but eventually cell surface area limits metabolic processing. Unicellular protists overcome this constraint by internalizing the metabolic machinery into respiratory organelles (i.e., mitochondria that convert oxygen into ATP). The number of mitochondria increases linearly with cell size until intracellular transport constraints begin to limit the rate of metabolic processing. Next, multi-cellular animals have effectively invariant cell size and intracellular transport, but as as body size and number of cells increased, vascular networks evolved to rapidly and efficient deliver metabolites. However, vascular networks introduce the sublinear network scaling effects characterized above.

Delong et al. highlight the importance of both time and energy constraints, and these change at each evolutionary transition, with the consequence that the absolute time and quantity of energy required to deliver each molecule of oxygen increase across the major evolutionary transitions. This suggests that the energy-time minimization framework which we have used to predict the curvature in metabolic scaling in mammals, might be also be applicable across the full range of living organisms, with different constraints on time and energy emerging at each evolutionary transition. The explanations that they hypothesize are also directly relevant to understanding how energy-time minimization affects the ongoing evolution of computer hardware.

4.2.1 Innovations in chip design mimic innovation in the evolution of bacteria

The chip scaling described above shows how time and energy dissipation have decreased while performance increased as larger numbers of smaller transistors have been packed onto each chip. During this era, technological innovations in chips have emerged that optimize against physical constraints. Just as bacteria have evolved larger genomes and used the new genes to exploit new metabolic niches, new materials, switching methods, etching processes and cooling technologies have pushed physical boundaries, allowing transistors to shrink and more of them to be packed onto each chip. Like bacteria, however, there are limits to this process. There are no elephant-sized bacteria, and there will be no silicon-based single core chips with quadrillions of transistors.

4.2.2 Single core chip scaling mimics unicellular protists

Historical chip scaling mimics the linear relationship between performance and size (Figure 4) seen in protists. Unicellular protists show linear increases in metabolic rate with size (Figure 1 of reference [16]) as more energy processing nodes (mitochondria) are packed into larger cells. As size continues to increase however, this design strategy also reaches physical limits. Our analysis suggests that the internal transport network already constrains processing speeds (T_{net} constrains T_{sys}). Further, the requirement to dissipate heat over a fixed surface area constrains both cells and chips.

4.2.3 Multi-core chips echo the transition to multicellularity

Computer chips are currently undergoing the evolutionary transition to multicore, resembling the biological transition to multicellularity. Our unified scaling framework suggests some future scenarios. As the era of transistor minimization wanes, additional transistors will require increased physical area, and therefore networks that span greater distances. Similar to

multicellular organisms, we expect that as the number of cores grows, an increasing fraction of chip power will be devoted to these ever larger Networks-on-Chip (NoC) connecting more cores. Larger networks will consume more power, take more time to traverse, and ultimately the time-energy minimization will be increasingly difficult to sustain as chips increase in size. Clock speeds have already leveled off as power, footprint and cooling requirements dominate chip design considerations [53]. If chips follow biology, we can expect that the most important future advances in chip design will increase network efficiency, for example by using optical networks.

4.2.4 Computer scaling deviates from biological scaling in important ways

There are also important differences between scaling of oxygen delivery in biology and information delivery in computation, which play an important role in evolutionary transitions. In particular, on-chip computer networks have two advantages not available to cardiovascular networks. First, the shrinking of 'process' size (smaller transistors and wires) reduces both energy and delay in the nodes as the number of nodes increases. This reduction in process size will ultimately end as physical limits are reached [53]. Second, the locality of network traffic, characterized by the Rent's exponent and D_w, reduces long distance communication over computer networks. As shown above, this effect reduces E_{net} and leads to a smaller wire footprint as N increases on single core chips. This advantage will likely continue for multi-core chips where communication, and therefore network bandwidth, footprint and energy consumption of NOCs can be reduced by keeping communication primarily local [10, 60]. Communication locality has the potential to produce more favorable scaling in multi-core computation than is achievable in multicellular biology.

4.2.5 Decentralized designs in the transition to sociality

We now consider how the lessons learned from computer architecture may lend insights into an important biological evolutionary transition, the transition to social animal societies. Understanding and improving the flow of energy, materials and information through human societies is one of the greatest challenges facing science and engineering, and scaling analyses lend an important perspective on this problem [41]. Sociality is an important evolutionary transition, reflected in the ecological dominance of humans and ants whose networked systems transport both energy and information. These social species have experienced great success, dispersing over vast territories across the globe and capturing a large fraction of available energy [23, 25]. Recent evidence suggests that ant colonies and human societies follow similar scaling relationships as individual organisms [40, 9, 14, 28, 54].

In social animal systems and networked computer systems, networks are at least partially decentralized, for example traffic flow within cities [47] and among ant nests [22]. Tainter et. al. [51] argue that complex human and ant societies are able to exploit "low-gain" energy systems: those that provide low concentrations of dispersed energy, but that are ubiquitous and therefore can be exploited by complex systems capable of processing and storing vast quantities of energy. Understanding the forces that have driven the tremendous power and performance scaling in computing may lend insights into how other technologies exploit similar scaling relationships [13]. In particular, communication locality in computation suggests an important strategy in the transition to sociality: animal societies can escape the constraints of the centralized distribution network by evolving systems for decentralized transportation and modular communication. Indeed, the transition to solar energy is capitalizing on the same kind of dramatic technological performance improvements that computer technology experienced as Moore's Law [21]. The history of computing suggests large gains in the efficiency of energy delivery if increasingly powerful solar cells utilize dispersed solar energy locally to escape the centralized distribution overhead of the fossil fuel based economy.

Moreover, power laws as a function of size are not unique to organisms and computers but are observed across a wide variety of complex systems in nature, society and technology. The scaling of white and grey matter [61] and communication modularity [37] in the brain, of flow through river networks that minimize transportation costs [4], of energy use and GDP in countries [12], and the pace of life and population in cities [8] are all additional examples that a unifying scaling theory might explain. Because cost and performance, i.e., energy and time, impose universal constraints, we suggest that a common design principle may govern the scaling of many evolved and engineered complex systems that process energy, materials and information.

5 Conclusion

Our analysis provides a unifying explanation for the origin of scaling laws in biology and computing. Despite obvious differences in form and function, the scaling of organisms and computers is governed by the same simple principle: minimizing the energy and time to deliver and process resources. Both natural selection and human engineering have evolved designs that manage the trade-off between cost and performance to minimize energy dissipation and time to deliver resources, resulting in general scaling laws that predict metabolic rate, and microprocessor power and performance over several orders-of-magnitude variation in system size.

Engineering ingenuity and economic pressures have created increasingly fast and powerful computers through a series of innovations, including integrated circuits, innovations in materials and other technological tricks, synchronizing clock trees, multi-core chips and networked and distributed computation. Today, technology is undergoing another major evolutionary transition as distributed computing changes the metabolic landscape of technology that is becoming more tightly coupled with the environment. As computers are embedded in more physical de-

vices, physical proximity and energy concerns for low-power devices may drive computational scaling to more closely resemble biological scaling. In computation, dramatic changes have emerged over the last 35 years, but to a surprising extent, their trajectories mimic the biological transitions that took billions of years to evolve simple unicellular bacteria into the largest and most powerful animals and societies on earth.

6 Appendix A: Details of Scaling in Organisms

This appendix gives the full derivation of the scaling equations. We begin with the total network resistance discussed in Section 3.1, and its subsequent effect on the energy time product scaling. Assume that $D_l = 3$ for 3 dimensional organisms and that $\lambda^{-H} = N^{-1}$. Using these values and simplifying, equation 6 is transformed.

$$R = \frac{8\mu l_0}{\pi r_0^4} N^{-1} \sum_{i=0}^{H} \lambda^{i(\frac{4}{3} - \frac{4}{D_r})} \tag{17}$$

Let the summand $S = \sum_{i=0}^{H} \lambda^{i(\frac{4}{3} - \frac{4}{D_r})}$. $R \propto N^{-1}S$. How S scales with N is dependent on the exponent $\frac{4}{3} - \frac{4}{D_r}$, and reduces to three different cases:

Case 1: $D_r = 3$: In this case the exponent is equal to 0, and the $S = H + 1 \propto \log(N)$, and $R \propto \frac{\log(N)}{N}$, because $\log(N)$ in this case grows much more slowly than N, it is reasonable to conclude that $R \propto N^{-1}$. In this case the energy time product scales as $l_0 + u_0^{-1} N^{2 - \frac{2}{D_r}}$.

Case 2: $D_r < 3$: Here (and in subsequent cases) we can use the geometric series to calculate the exact value of S. In particular

$$\begin{aligned} S &= \frac{(1 - \lambda^{(\frac{4}{3} - \frac{4}{D_r})(H+1)})}{1 - \lambda^{\frac{4}{3} - \frac{4}{D_r}}} \\ &= \frac{1 - \left(\lambda^H\right)^{(\frac{4}{3} - \frac{4}{D_r})} \lambda^{(\frac{4}{3} - \frac{4}{D_r}))}}{1 - \lambda^{\frac{4}{3} - \frac{4}{D_r}}} \\ &= \frac{1 - N^{(\frac{4}{3} - \frac{4}{D_r})} \lambda^{(\frac{4}{3} - \frac{4}{D_r}))}}{1 - \lambda^{\frac{4}{3} - \frac{4}{D_r}}} \end{aligned}$$

If we let $c = \lambda^{(\frac{4}{3} - \frac{4}{D_r})}$ we see that

$$S = \frac{1 - cN^{(\frac{4}{3} - \frac{4}{D_r})}}{1 - c} \tag{18}$$

Because $\frac{4}{3} - \frac{4}{D_r} < 0$ is negative in this case and N is large in practice, $cN^{(\frac{4}{3}-\frac{4}{D_r})}$ is small, and S is proportional to a constant ($S \approx \frac{1}{1-c}$). This implies that $R \propto N^{-1}$. Once again, equation 4 scales as $\propto l_0 + u_0^{-1} N^{2-\frac{2}{D_r}}$.

Case 3: $D_r > 3$: In this case the exponent in S is positive, meaning that S scales directly with N. Note that $c > 1$ in this case and we can write

$$S = \frac{cN^{(\frac{4}{3}-\frac{4}{D_r})} - 1}{c - 1} \tag{19}$$

This means that $S \propto N^{(\frac{4}{3}-\frac{4}{D_r})}$. This implies that $R \propto N^{-1}S \propto N^{(\frac{1}{3}-\frac{4}{D_r})}$. This means that resistance still scales inversely with size, but at a faster rate than if $D_r \leq 3$. This implies the energy time produce scales as $\propto N^{\frac{4}{3}-\frac{4}{D_r}} + N^{2-\frac{2}{D_r}}$. Note that this results in a positive scaling of R with N if $D_r > 12$.

6.1 Length, Velocity and Mass Scaling

We determine the scaling of l_0, u_0 and M following the method presented in Banavar et al. [3]. Specifically, we assume that $M \propto V \propto V_{net}$, where V is the volume of the organism, V_{net} is the volume of the network transporting oxygen molecules, and

$$u_0 \propto l_0 \propto \left(\frac{V}{N}\right)^{\frac{1}{3}} \tag{20}$$

We calculate V_{net} as:

$$
\begin{aligned}
V_{net} &= \sum_{i=0}^{H} V_i = \sum s^{-1} l_i r_i^2 n_i \\
&= \sum_{i=0}^{H} {l_0}^{-1} l_0 \lambda^{\frac{i}{D_l}} r_0^2 \lambda^{\frac{2i}{D_r}} \lambda^{H-i} \\
&= \lambda^H r_0^2 \sum_{i=0}^{H} \lambda^{i\left(\frac{i}{D_l}+\frac{2}{D_r}-1\right)} \\
&\propto N^{\frac{1}{D_l}+\frac{2}{D_r}}
\end{aligned}
$$

Note that $s \propto l_0$ is the linear distance between oxygen molecules that will be delivered to the same capillary, allowing narrower vessels when oxygen travels at higher velocity (see [3] Figure 3 for further explanation). The calculation assumes that $\frac{1}{D_l} + \frac{2}{D_r} - 1 > 0$ which will always be the case with $2 < D_r < 3$, and $D_l = 3$.

Therefore $M \propto N^{\frac{1}{D_l}+\frac{2}{D_r}}$, and $N \propto M^{\frac{1}{\frac{1}{3}+\frac{2}{D_r}}}$, where $D_l = 3$. Additionally

$$
\begin{aligned}
u_0 \propto l_0 &\propto \left(\frac{V}{N}\right)^{\frac{1}{3}} \\
&\propto \left(N^{\frac{1}{3}+\frac{2}{D_r}-1}\right)^{\frac{1}{3}} \\
&\propto N^{\frac{2}{3D_r}-\frac{2}{9}}
\end{aligned}
$$

Combining these results with E_{sys} and T_{sys} in Table 1, we derive how metabolic rate, B (measured in power), scales with the mass of an organism.

$$
\begin{aligned}
B = \frac{E_{sys}}{T_{sys}} =& \frac{RQ^2}{N} + Q \\
& \propto u_0^2 l_0 N^{-2} N^{\frac{4}{D_r}} + u_0 N^{\frac{2}{D_r}} \\
& \propto N^{\frac{6}{D_r} - \frac{8}{3}} + N^{\frac{8}{3D_r} - \frac{2}{9}} \\
& \propto M^{\frac{\frac{6}{D_r} - \frac{8}{3}}{\frac{2}{D_r} + \frac{1}{3}}} + M^{\frac{\frac{8}{3D_r} - \frac{2}{9}}{\frac{2}{D_r} + \frac{1}{3}}} \\
& \propto M^{\frac{18 - 8D_r}{6 + Dr}} + M^{\frac{24 - 2D_r}{18 - 3D_r}}
\end{aligned}
$$

7 Appendix B: Details of Scaling in Electronics

In this section we give a detailed analysis of the derivation of the scaling of the network capacitance and network latency discussed in Section 3.3.

7.1 Capacitance

Recall that $D_l = 2$ for 2 dimensional computer chips and that $\lambda^{-H} = N^{-1}$. We can then calculate capacitance as:

$$C \propto N^{(1-\frac{1}{D_l}} \sum_{i=0}^{H} \lambda^{i\left(\frac{1}{D_l}+\frac{1}{D_w}-1\right)} \tag{21}$$

Similar to how we handled organsims we are interested in whether the exponent $\frac{1}{D_l}+\frac{1}{D_w}-1$ is positive or negative.

Let the summand $S = \sum_{i=0}^{H} \lambda^{i(\frac{1}{D_l}+\frac{1}{D_w}-1)}$. $C \propto N^{1-\frac{1}{D_l}} S$.

Case 1: $D_r = \frac{D_l}{D_l-1}$: In this case the exponent is equal to 0, and the $S = H + 1 \propto \log(N)$, and $C \propto \log(N) N^{1-\frac{1}{D_l}}$, because $\log(N)$ in this case grows much more slowly than $N^{1-\frac{1}{D_l}}$ and we know $D_l = 2$ for 2 dimensional chips, it is reasonable to conclude that $C \propto N^{\frac{1}{2}}$

Case 2: $D_r > \frac{D_l}{D_l-1}$: Here (and in subsequent cases) we can use the geometric series to calculate the exact value of S, using a similar approach to 6. In this case the exponent is negative and S is a small constant, leaving $C \propto N^{\frac{1}{2}}$

Case 3: $D_r < \frac{D_l}{D_l-1}$: In this case the exponent in S is positive, meaning that S scales directly with N. Now the summand contributes an $N^{\frac{1}{D_l}+\frac{1}{D_w}-1}$ and $C \propto N^{\frac{1}{D_w}}$.

7.2 Network Delay

Recall that we wish to determine the network latency L which is defined as:

$$T_{net} \propto \max_i L_i \tag{22}$$

with

$$L_i \propto RC = \frac{\rho \epsilon l i^2}{r_i^2} = \frac{\rho \epsilon l_0^2}{r_0^2} \lambda^{i\left(\frac{2}{D_l} - \frac{2}{D_r}\right)} \tag{23}$$

L_i will scale differently depending on the relative values of D_r and D_l.

Case 1: $D_r > D_l$: In this case the fraction in the exponent is greater than 0 and the latency will be highest when $i = H$, resulting in $L \propto N^{\frac{2}{D_l} - \frac{2}{D_r}}$.

Case 2: $D_r < D_l$: In this case the exponent is negative and the highest latency occurs at the bottom of the network $i = 0$, leaving $L \propto \frac{l_0^2}{r_0^2} \propto N^0$

Case 3: $D_r = D_l$: In this case the exponent is 0 and there is equal latency at all levels and $L \propto N^0$.

7.3 Chip Data

We obtain data on chip power usage and throughput from several third party sources. For power output we consulted two web archives of over 523 chips listing power consumption and transistor count [48, 30]. When possible the figures were cross checked with data directly from the manufacturer. For throughput data, we consulted a combination of third party sources, starting with a list published on wikipedia [57]. Each source for the data was consulted independently and verified before inclusion in the dataset. Additional throughput data was obtained from benchmarks performed by online technology publication Tom's Hardware [24]. The data can be found in the supplementary information.

References and Notes

[1] O. Azizi, A. Mahesri, B. C. Lee, S. J. Patel, and M. Horowitz. Energy-performance tradeoffs in processor architecture and circuit design: a marginal cost analysis. In *ACM SIGARCH Computer Architecture News*, volume 38, pages 26–36. ACM, 2010.

[2] H. Bakoglu. *Circuits, Interconnections, and Packaging for VLSI*. 1990.

[3] J. Banavar, M. Moses, J. Brown, J. Damuth, A. Rinaldo, R. Sibly, and A. Maritan. A general basis for quarter-power scaling in animals. *Proceedings of the National Academy of Sciences*, 107(36):15816–15820, 2010.

[4] J. R. Banavar, F. Colaiori, A. Flammini, A. Maritan, and A. Rinaldo. Topology of the fittest transportation network. *Physical Review Letters*, 84(20):4745, 2000.

[5] J. R. Banavar, J. Damuth, A. Maritan, and A. Rinaldo. Supply–demand balance and metabolic scaling. *Proceedings of the National Academy of Sciences*, 99(16):10506–10509, 2002.

[6] J. R. Banavar, A. Maritan, and A. Rinaldo. Size and form in efficient transportation networks. *Nature*, 399(6732):130–132, 1999.

[7] D. Bassett, D. Greenfield, A. Meyer-Lindenberg, D. Weinberger, S. Moore, and E. Bullmore. Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS computational biology*, 6(4):e1000748, 2010.

[8] L. Bettencourt, J. Lobo, D. Helbing, C. Kuhnert, and G. West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301, 2007.

[9] L. M. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, and G. B. West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences*, 104(17):7301–7306, 2007.

[10] G. B. Bezerra, S. Forrest, M. Forrest, A. Davis, and P. Zarkesh-Ha. Modeling noc traffic locality and energy consumption with rent's communication probability distribution. In *Proceedings of the 12th ACM/IEEE international workshop on System level interconnect prediction*, pages 3–8. ACM, 2010.

[11] B. Bingham and M. Greenstreet. Computation with energy-time trade-offs: Models, algorithms and lower-bounds. In *Parallel and Distributed Processing with Applications, 2008. ISPA'08. International Symposium on*, pages 143–152. IEEE, 2008.

[12] J. Brown, W. Burnside, A. Davidson, J. Delong, W. Dunn, M. Hamilton, N. Mercado-Silva, J. Nekola, J. Okie, W. Woodruff, et al. Energetic limits to economic growth. *BioScience*, 61(1):19–26, 2011.

[13] M. Buchanan. Generalizing moore. *Nature Physics*, 12(3):200–200, 2016.

[14] W. R. Burnside, J. H. Brown, O. Burger, M. J. Hamilton, M. Moses, and L. Bettencourt. Human macroecology: linking pattern and process in big-picture human ecology. *Biological Reviews*, 87(1):194–208, 2012.

[15] P. Christie and D. Stroobandt. The interpretation and application of rent's rule. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 8(6):639–648, 2000.

[16] J. P. DeLong, J. G. Okie, M. E. Moses, R. M. Sibly, and J. H. Brown. Shifts in metabolic scaling, production, and efficiency across major evolutionary transitions of life. *Proceedings of the National Academy of Sciences*, 107(29):12941–12945, 2010.

[17] R. Dennard, F. Gaensslen, V. Rideout, E. Bassous, and A. LeBlanc. Design of ion-implanted mosfet's with very small physical dimensions. *Solid-State Circuits, IEEE Journal of*, 9(5):256–268, 1974.

[18] P. S. Dodds. Optimal form of branching supply and collection networks. *Physical review letters*, 104(4):048702, 2010.

[19] W. Donath. Placement and average interconnection lengths of computer logic. *Circuits and Systems, IEEE Transactions on*, 26(4):272–277, 1979.

[20] W. E. Donath. Wire length distribution for placements on computer logic. *IBM J. Res. and Development*, 25:152–155, 1981.

[21] J. D. Farmer and F. Lafond. How predictable is technological progress? *Research Policy*, 45(3):647–665, 2016.

[22] T. P. Flanagan, N. M. Pinter-Wollman, M. E. Moses, and D. M. Gordon. Fast and flexible: Argentine ants recruit from nearby trails. *PloS one*, 8(8):e70888, 2013.

[23] H. Haberl, K. H. Erb, F. Krausmann, V. Gaube, A. Bondeau, C. Plutzar, S. Gingrich, W. Lucht, and M. Fischer-Kowalski. Quantifying and mapping the human appropriation of net primary production in earth's terrestrial ecosystems. *Proceedings of the National Academy of Sciences*, 104(31):12942–12947, 2007.

[24] T. Hardware. Performance charts. `http://www.tomshardware.com/charts/`, Nov. 2015.

[25] B. Hölldobler and E. O. Wilson. *The ants*. Harvard University Press, 1990.

[26] M. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, and K. Bernstein. Scaling, power, and the future of cmos. In *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pages 7–pp. IEEE, 2005.

[27] M. Horowitz, T. Indermaur, and R. Gonzalez. Low-power digital design. In *Low Power Electronics, 1994. Digest of Technical Papers., IEEE Symposium*, pages 8–11. IEEE, 1994.

[28] C. Hou, M. Kaspari, H. B. Vander Zanden, and J. F. Gillooly. Energetic basis of colonial living in social insects. *Proceedings of the National Academy of Sciences*, 107(8):3634–3638, 2010.

[29] M. Kleiber. Body size and metabolic rate. *Physiological Reviews*, 27(4):511, 1947.

[30] I. Knowledge. History of the ic. `https://web.archive.org/web/20120207092719/http://www.icknowledge.com/history/history.html`, Feb. 2012.

[31] T. Kolokotrones, V. Savage, E. J. Deeds, and W. Fontana. Curvature in metabolic scaling. *Nature*, 464(7289):753–756, 2010.

[32] J. Koomey, S. Berard, M. Sanchez, and H. Wong. Implications of historical trends in the electrical efficiency of computing. *Annals of the History of Computing, IEEE*, 33(3):46–54, 2011.

[33] S. Lindstedt and W. Calder III. Body size, physiological time, and longevity of homeothermic animals. *Quarterly Review of Biology*, pages 1–16, 1981.

[34] A. Lotka. *Elements of mathematical biology*. Dover Publications New York, 1956.

[35] B. Mandelbrot. *The fractal geometry of nature*. Macmillan, 1983.

[36] B. McArdle. The structural relationship: regression in biology. *Canadian Journal of Zoology*, 66(11):2329–2339, 1988.

[37] D. Meunier, R. Lambiotte, and E. T. Bullmore. Modular and hierarchically modular organization of brain networks. *Frontiers in neuroscience*, 4, 2010.

[38] G. Moore et al. Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85, 1998.

[39] M. Moses, S. Forrest, A. Davis, M. Lodder, and J. Brown. Scaling theory for information networks. *Journal of the Royal Society Interface*, 5(29):1469, 2008.

[40] M. E. Moses and J. H. Brown. Allometry of human fertility and energy use. *Ecology Letters*, 6(4):295–300, 2003.

[41] M. E. Moses and S. Forrest. Beyond biology. *Metabolic Ecology: A Scaling Approach*, page 293, 2012.

[42] M. G. Newberry, D. B. Ennis, and V. M. Savage. Testing foundations of biological scaling theory using automated measurements of vascular networks. *PLoS Comput Biol*, 11(8):e1004455, 2015.

[43] T. H. Ning. A perspective on the theory of MOSFET scaling and its impact. *IEEE Solid State Circuits Newsletter*, 12(1):27–30, 2007.

[44] H. Odum et al. *Environment, power and society.* New York, USA, Wiley-Interscience, 1971.

[45] H. M. Ozaktas. Information flow and interconnections in computing: extensions and applications of rent's rule. *Journal of Parallel and Distributed Computing*, 64(12):1360–1370, 2004.

[46] S. Reda. Using circuit structural analysis techniques for networks in systems biology. In *Proceedings of the 11th international workshop on System level interconnect prediction*, pages 37–44. ACM, 2009.

[47] H. Samaniego and M. E. Moses. Cities as organisms: Allometric scaling of urban road networks. *Journal of Transport and Land use*, 1(1), 2008.

[48] T. B. Server. Chipset benchmarks. `https://web.archive.org/web/20070809235011/http://balusc.xs4all.nl/srv/har-chi.html`, Aug. 2007.

[49] R. Sibly. The life-history approach to physiological ecology. *Functional Ecology*, 5(2):184–191, 1991.

[50] R. V. Solée, S. Valverde, M. R. Casals, S. A. Kauffman, D. Farmer, and N. Eldredge. The evolutionary ecology of technological innovations. *Complexity*, 18(4):15–27, 2013.

[51] J. A. Tainter, T. Allen, A. Little, and T. W. Hoekstra. Resource transitions and energy gain: contexts of organization. *Conservation Ecology*, 7(3):4, 2003.

[52] D. Thompson. Darcy w (1917) on growth and form, 1942.

[53] M. M. Waldrop. The chips are down for moores law. *Nature News*, 530(7589):144, 2016.

[54] J. S. Waters, C. T. Holbrook, J. H. Fewell, and J. F. Harrison. Allometric scaling of metabolism, growth, and activity in whole colonies of the seed-harvester ant pogonomyrmex californicus. *The American Naturalist*, 176(4):501–510, 2010.

[55] G. West, J. Brown, and B. Enquist. A general model for the origin of allometric scaling laws in biology. *Science*, 276(5309):122, 1997.

[56] G. B. West, W. H. Woodruff, and J. H. Brown. Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2473–2478, 2002.

[57] Wikipedia. Millions of instructions per second. `https://en.wikipedia.org/wiki/Instructions_per_second#Millions_of_instructions_per_second`, Nov. 2015.

[58] N. Wilhelm. *Why Wire Delays Will No Longer Scale for VLSI Chips*. Sun Microsystems Laboratories, 1995.

[59] X. Yang, E. Bozorgzadeh, and M. Sarrafzadeh. Wirelength estimation based on rent exponents of partitioning and placement. In *Proceedings of the 2001 international workshop on System-level interconnect prediction*, pages 25–31. ACM, 2001.

[60] P. Zarkesh-Ha, G. B. Bezerra, S. Forrest, and M. Moses. Hybrid network on chip (hnoc): local buses with a global mesh architecture. In *Proceedings of the 12th ACM/IEEE international workshop on System level interconnect prediction*, pages 9–14. ACM, 2010.

[61] K. Zhang and T. Sejnowski. A universal scaling law between gray matter and white matter of cerebral cortex. *Proceedings of the National Academy of Sciences*, 97(10):5621, 2000.