Are mental properties supervenient on brain properties?

Joshua T. Vogelstein [*], **R. Jacob Vogelstein** [*] [†], **Carey E. Priebe** [*]

[*] Johns Hopkins University, Baltimore, MD, and [†] Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Submitted to Proceedings of the National Academy of Sciences of the United States of America

The "mind-brain supervenience" theorem suggests that all mental properties (e.g. consciousness, intelligence, personality, etc.) are derived from the physical properties of the brain. The validity of this theorem has been argued in philosphical terms for over 2,500 years, but it has not previously been approachable through experimental investigation. To address the question of whether the mind supervenes on the brain through empirical means, here we frame a supervenience hypothesis in rigorous mathematical terms and propose a modified version of supervenience (called ε-supervenience) that is amenable to scientific methods and statistical analysis. To elucidate this approach, we posit a thought experiment that illustrates how the probabilistic theory of pattern recognition can be used to make a one-sided determination of ε-supervenience. The physical property of the brain employed in this analysis is the graph describing brain connectivity (i.e., the *connectome*), and ε-supervenience allows us to determine whether a particular mental property can be inferred from one's connectome to within any given misclassification rate $\varepsilon > 0$, regardless of the relationship between the two. In addition to the theoretical results, we show via simulation that given reasonable assumptions about class conditional probabilities and the amount of data available, the thought experiment can actually be conducted on a simple organism, *Caenorhabditis elegans*, with currently available technology.

graph theory | universal consistency | neural circuit

Questioning the relationship between the mind (thoughts, beliefs, preferences, emotions, intelligence, etc.) and the brain (the physical structure inside our skulls) dates back at least as far as 400 BCE, when Plato wrote the dialogues, in which he posited immateriality of the soul [1]. Approximately two millennia passed before these ideas reached their canonical form through Descartes's discussion of mind-body dualism [2]. Then, in the 20th century, Donald Davidson stated and popularized the mind-brain supervenience theorem, which claims that an agent cannot alter in some mental property without altering in some physical property [3]. Contemporary fields of neural network theory and neuroscientific inquiry often assume mind-brain supervenience, or an even stronger assumption about mind-brain causality, but no previously proposed notion of supervenience seems amenable to empirical investigation. Here we define new versions of supervenience that formulate the theorem in rigorous mathematical terms and that can be experimentally tested.

Let $\mathcal{B}$ be the observation space for some physical property, such as brain connectivity structure (i.e., connectome; see [4, 5, 6]). Let $\mathcal{M}$ be the (finite) indicator space for some mental property, such as knowing calculus. Thus, for $b \in \mathcal{B}$ and $m \in \mathcal{M}$, the pair (b, m) represents a brain property/mind property pair.

Let $(B, M), (B_1, M_1), \cdots, (B_n, M_n)$ be random observation pairs taking their values in $\mathcal{B} \times \mathcal{M}$, independently and identically distributed according to some joint probability distribution $F = F_{BM}$. Abusing notation to conceptually identify the properties with their spaces, the statistical supervenience relation $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$ says that $M_i \neq M_j \implies B_i \neq B_j$ (almost surely; where $\implies$ does not suggest causation). That is, observing $B = b$ can allow us to assign m to M. While previously proposed notions of mind-brain-supervenience claim that all mental properties supervene on physical properties [7], here we consider empirically investigating only whether a particular mental property $\mathcal{M}$ statistically supervenes on a particular physical property $\mathcal{B}$.

Let $g : \mathcal{B} \to \mathcal{M}$ be a classifier, which takes as input an observed brain connectivity structure b and produces a classification $\widehat{m} = g(b)$ for the unobserved mental property m. The Bayes optimal classifier g^* minimizes $L_F(g)$ over all classifiers, where $L_F(g) = P_F[g(B) \neq M]$ denotes the probability of misclassification for classifier g under joint distribution $F = F_{BM}$. We can therefore rigorously define *statistical supervenience*:

Definition 1. *$\mathcal{M}$ is said to statistically supervene on $\mathcal{B}$ for distribution $F = F_{BM}$, denoted $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, if and only if $L_F(g^*) = 0$.*

Unfortunately, it is in general impossible to determine whether $L_F(g^*) = 0$ without knowing F. Therefore, we relax the above statistical supervenience to define *ε-supervenience*:

Definition 2. *Given $\varepsilon > 0$, $\mathcal{M}$ is said to ε-supervene on $\mathcal{B}$ for distribution $F = F_{BM}$, denoted $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$, if and only if $L_F(g^*) < \varepsilon$.*

Now, generalizing the concept of a classifier g to allow consideration of training data, consider $g_n : \mathcal{B} \times (\mathcal{B} \times \mathcal{M})^n \to \mathcal{M}$ which takes as input an observed brain connectivity structure b and n training pairs $(b_1, m_1), \cdots, (b_n, m_n)$ and produces a classification $\widehat{m} = g_n(b; (b_1, m_1), \cdots, (b_n, m_n))$. Let $L_F(g_n) = E[P_F[g_n(B; (B_1, M_1), \cdots, (B_n, M_n)) \neq M | (B_1, M_1), \cdots, (B_n, M_n)]]$.

Consider the problem of testing for ε-supervenience. Let the null hypothesis be given by $H_0 : L_F(g_n) \geq \varepsilon$ so that if we reject at level $\alpha > 0$ in favor of the alternative hypothesis $H_A : L_F(g_n) < \varepsilon$ then we can conclude, with $100(1-\alpha)\%$ confidence, that $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$. Letting $\widehat{L}_F^{n'}(g_n)$ denote the hold-out estimate of misclassification performance based on n' test observations, we note that $\widehat{L}_F^{n'}(g_n)$ is distributed $Binomial(n', L_F(g_n))$. The test rejects for small $\widehat{L}_F^{n'}(g_n)$. The level α critical value $c_\alpha(n', \varepsilon)$ is available under the least favorable distribution $Binomial(n', \varepsilon)$. Furthermore, $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ implies $L_F(g^*) < \varepsilon$, and thus if g_n is a *consistent* classifier for $F = F_{BM}$ — that is, if $\lim_n L_F(g_n) = L_F(g^*)$ — then the power of this test (the probability of rejecting when in fact the alternative is true) goes to unity as $n, n' \to \infty$. Thus we have an inference procedure:

Reserved for Publication Footnotes

Theorem 1. *Given $\alpha > 0$, we can test $\mathcal{M}\overset{\varepsilon}{\sim}_F\mathcal{B}$ so that rejection implies $\mathcal{M}\overset{\varepsilon}{\sim}_F\mathcal{B}$ holds with probability greater than or equal to $1-\alpha$. Furthermore, given a consistent classifier the power of the test converges to unity.*

Since the joint distribution $F = F_{BM}$ is unknown, the utility of Theorem 1 requires that g_n be a *universally consistent* classifier — that is, $\lim_n L_F(g_n) = L_F(g^*)$ for all distributions $F = F_{BM}$. Unfortunately, the rate of convergence of $L_F(g_n)$ to $L_F(g^*)$ depends on the (unknown) distribution $F = F_{BM}$ [8]. Furthermore, arbitrarily slow convergence theorems regarding the rate of convergence of $L_F(g_n)$ to $L_F(g^*)$ demonstrate that there is no universal n, n' which will guarantee that the test has power greater than any specified target $\beta > \alpha$ [9]. For this reason, the test outlined above can provide only a one-sided conclusion: if we reject we can be $100(1-\alpha)\%$ confident that $\mathcal{M}\overset{\varepsilon}{\sim}_F\mathcal{B}$ holds, but we can never be confident in its negation. Thus, without restrictions on F_{BM}, arbitrarily slow convergence theorems imply that our theorem of ε-supervenience does not satisfy Popper's *falsifiability* requirement [10]. Given these caveats, consider the following thought experiment:

Thought experiment 1. *Let the physical property under consideration be brain connectivity structure ("connectome"), so b is a graph with vertices representing neurons (or neuroanatomical regions) and edges representing connections between neurons (or white matter tracts). Further let $\mathcal{B}$, the observation space, be the collection of all graphs on a finite number of vertices, and let $|\mathcal{B}|$ be countable. Now, imagine collecting very large amounts of very accurate independent and identically distributed brain-graph data and the associated mental property indicators. A k_n-nearest neighbor classifier using an isomorphism-matching Frobenius norm is universally consistent (see Appendix 1 for proof). Therefore, Theorem 1 applies and the existence of a universally consistent classifier guarantees that eventually (in n, n') we will be able to conclude $\mathcal{M}\overset{\varepsilon}{\sim}_F\mathcal{B}$ for this mental/brain property pair, if indeed ε-supervenience holds. This logic holds for directed graphs or multigraphs or hypergraphs with discrete edge weights and vertex attributes.*

While the above thought experiment addresses the question of ε-supervenience, it does not address causality. Assuming we have confirmed $\mathcal{M}\overset{\varepsilon}{\sim}_F\mathcal{B}$ for a particular mental/brain property pair with confidence level α, then morphing the brain (by altering edges) could be used to determine whether the relation is in fact causal.

Practical issues regarding actually conducting the above thought experiment include: (1) as stated, we must consider the space $\mathcal{B}$ to be the quotient space of graphs mod graph isomorphism, unless the vertices are *labeled*; (2) a more informative and tractable distance on $\mathcal{B}$ may be desired, as the k_n-nearest neighbor classifier under our Frobenius norm may have a rate of convergence so slow and a computational demand so high as to be impractical; and (3) collecting enough sufficiently accurate independent and identically distributed brain-graph data and the associated mental property indicators may be beyond current technological capabilities. Regardless, related experimental work includes collecting various types of brain graph data [11, 12, 13] and various approaches to inference on brain graphs [14, 15, 16], suggesting feasibility of such an experiment in the near future (see Appendix 2 for a simulated example of a feasible experiment). Nevertheless, our thought experiment suggests that we can hope to determine that a given mental property under consideration ε-supervenes on a brain's connectivity structure.

Appendix 1: k_n-nearest neighbor universal consistency for graphs

Assume first that all graphs are simple, on the same set of vertices, and that the graphs are labeled so that we know which vertex in one graph corresponds to which vertex in another. Then the Frobenius distance function $d(b_1, b_2)$ can be written in terms of the associated adjacency matrices A_1 and A_2: $d(b_1, b_2) = ||A_1 - A_2||_F$. If the graphs are identical, then $d(b_1, b_2) = 0$, and if the graphs are different, then $d(b_1, b_2) \geq 1$. Since the space $\mathcal{B}$ is finite, n large enough guarantees that with probability approaching unity at least k_n training samples coincide with each atom, so long as $k_n/n \to 0$. Then $k_n \to \infty$ guarantees that the nearest neighbor vote-winner for each atom will eventually coincide with Bayes' choice, yielding universal consistency.

In the foregoing argument, there exists a smallest nonzero atomic probability p_{min}, and "n large enough" is driven by this probability. Generalizing to countable $\mathcal{B}$ with discrete weights, we see that given $\delta > 0$, there is a finite set S with $P[S] > 1 - \delta$ and smallest atomic probability p_{min}, so that $L_F(g_n) \to c \leq L_F(g^*) + \delta$, yielding universal consistency.

If the graphs may have different numbers of vertices, and are unlabeled, we consider the isomorphism-matching Frobenius norm. Assume without loss of generality that b_1 has at least as many vertices as b_2, and write A_2^P for the adjacency matrix associated with b_2 "padded" to include extra isolated vertices so that A_2^P is the same size as A_1. Then $d(b_1, b_2) = \min_Q ||QA_1Q^T - A_2^P||_F$ where the minimum is taken over all permutation matrices [17]. Under the equivalence relation induced by this isomorphism-matching, the foregoing universal consistency argument holds.

Several points of note: Isolated vertices are ignored in our equivalence relation; the class-conditional signal is entirely encompassed by the connectivity structure; the graph isomorphism problem is computationally hard [18, 19]; and the argument employed here does not capture the concept of "nearness implies likelihood of similar class"—we simply rely on atomic behavior.

Appendix 2: Simulation

As an example of a feasible experiment, one may consider a species whose nervous system consists of the same (small) number of labeled neurons for each organism. *Caenorhabditis elegans* is believed to be such a species [20]. The hermaphroditic C. elegans' somatic nervous system consists of 279 interconnected neurons. While the graph with these neurons as vertices and edges defined by chemical synapses between neurons is not identical across individuals, it is reasonably consistent [20]. Furthermore, these animals exhibit a rich behavioral repertoire that depends on circuit properties [21]. Thus, one may design an experiment by describing the joint distribution F_{BM} via class-conditional distributions $F_{B|M=m_j}$ for the C. elegans brain-graph for two mental properties of interest, m_0 and m_1, along with the prior probability of class membership $P[M = m_1]$. Here the mental property corresponds to the C. elegans exhibiting or not exhibiting a particular behavior (e.g., response to an odor).

Simulations suggest that one may build a classifier, practically and with a manageable training sample size n, that demonstrates ε-supervenience with reasonable choices for ε and α and a plausible joint distribution F_{BM} (Figure 1). To generate the data, we let the class-conditional random variable $E_{ij}|M = m_0$ be distributed Poisson$(A_{ij} + \eta)$, where A_{ij} is the number of chemical synapses between neuron i and neuron j according to [22], with noise parameter $0 < \eta \ll 1$. The

class-conditional random variable $E_{ij}|M = m_1$ is distributed Poisson$(A_{ij} + k_{ij})$ for neurons $i, j \in \mathcal{D}$, where $\mathcal{D}$ is the set of edges deemed responsible for odor-evoked behavior according to [23], with signal parameter k_{ij} uniformly sampled from $[-5, 5]$. We consider k_n-nearest neighbor classification of labeled multigraphs (directed, with loops) on 279 vertices, under Frobenius norm. The k_n-nearest neighbor classifier used here satisfies $k_n \to \infty$ as $n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$, ensuring universal consistency. (Better classifiers can be constructed for the joint distribution F_{BM} used here; however, we demand universal consistency.)

Importantly, conducting this experiment *in actu* is not beyond current technological limitations. 3D superresolution imaging [24] combined with neurite tracing algorithms [25, 15, 16] allow the collection of a brain-graph within a day. Genetic manipulations, laser ablations, and training paradigms can each be used to obtain a non-wild type population for use as $M = m_1$ [21], and the class of each organism (m_0 vs. m_1) can also be determined automatically [26].

ACKNOWLEDGMENTS. The authors would like to acknowledge helpful discussions with J Lande and B Vogelstein. This work was supported in part by the NSA Research Program in Applied Neuroscience.

1. Plato. Plato: complete works. Hackett Pub Co, (1997).
2. Descartes, R. Meditationes de prima philosophia. (1641).
3. Davidson, D. Experience and Theory, chapter Mental Events. Duckworth (1970).
4. Sporns, O., Tononi, G., and Kotter, R. PLoS Computational Biology 1(4), e42 (2005).
5. Lichtman, J. W., Livet, J., and Sanes, J. R. Nat Rev Neurosci 9(6), 417–422 Jun (2008).
6. Seung, H. Neuron 62(1), 17–29 (2009).
7. Kim, J. Phiilosophy of Mind. Westview Press, second edition, (2005).
8. Devroye, L., Györfi, L., and Lugosi, G. A Probabilistic Theory of Pattern Recognition. Springer, (1996).
9. Devroye, L. Probability Theory and Related Fields 62(4), 475–483 (1983).
10. Popper, K. (1959).
11. White, J., Southgate, E., Thomson, J. N., and Brenner, S. Philosophical Transactions of Royal Society London. Series B, Biological Sciences 314(1165), 1–340 (1986).
12. W. Denk, W. and Horstmann, H. PLOS Biol. 2, e329 (2004).
13. Briggman, K. and Denk, W. Current opinion in neurobiology 16(5), 562–570 (2006).
14. Macke, J. H., Maack, N., Gupta, R., Denk, W., Schlkopf, B., and Borst, A. J Neurosci Methods 167(2), 349–357 Jan (2008).
15. Mishchenko, Y. J Neurosci Methods 176(2), 276–289 Jan (2009).
16. Lu, J., Fiala, J. C., and Lichtman, J. W. PLoS ONE 4(5), e5655 05 (2009).
17. Horn, R. and Johnson, C. Matrix analysis. Cambridge Univ Pr, (1990).
18. Conroy, J. M., Kratzer, S. G., and Podrazik, L. J. In Society for Industrial and Applied Mathematics, (1997).
19. Zaslavskiy, M., Bach, F., and Vert, J. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 2227–2242 (2008).
20. Durbin, R. M. Studies on the Development and Organisation of the Nervous System of Caenorhabditis elegans. PhD thesis, University of Cambridge, (1987).
21. de Bono, M. and Maricq, A. V. Annu Rev Neurosci 28, 451–501 (2005).
22. Varshney, L., Chen, B., Paniagua, E., Hall, D., and Chklovskii, D. ArXiv (2009).
23. Chalasani, S. H., Chronis, N., Tsunozaki, M., Gray, J. M., Ramot, D., Goodman, M. B., and Bargmann, C. I. Nature 450(7166), 63–70 Nov (2007).
24. Vaziri, A., Tang, J., Shroff, H., and Shank, C. V. Proc Natl Acad Sci U S A 105(51), 20221–20226 Dec (2008).
25. Helmstaedter, M., Briggman, K. L., and Denk, W. Curr Opin Neurobiol 18(6), 633–641 Dec (2008).
26. Buckingham, S. D. and Sattelle, D. B. Invert Neurosci 8(3), 121–131 Sep (2008).

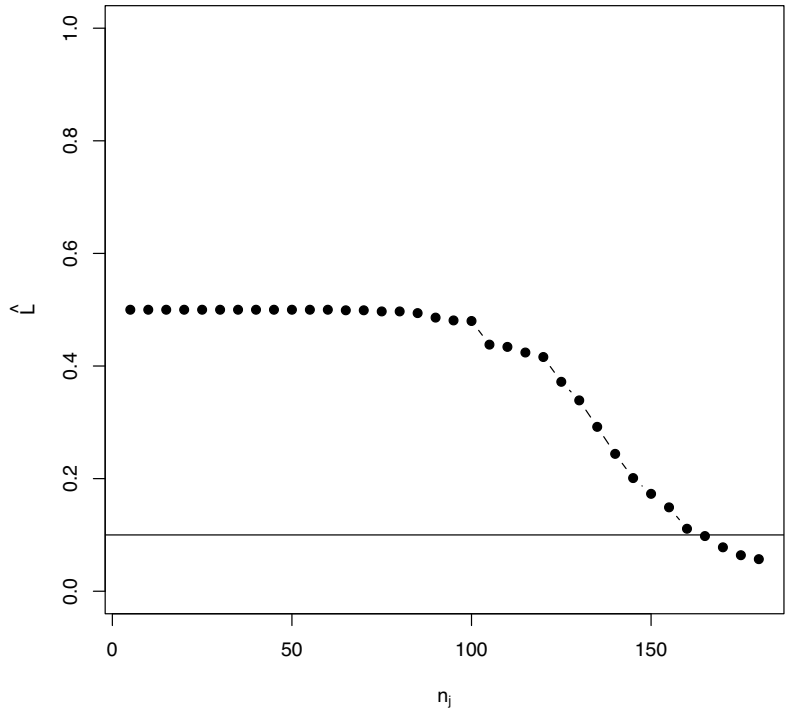

Figure 1: C. elegans graph classification simulation results. $\widehat{L}_F^{1000}(g_n)$ is plotted as a function of class-conditional training sample size n_j, suggesting that for $\varepsilon = 0.1$ we can determine that $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ holds with 99% confidence with just a few hundred training samples generated from F_{BM}. Each dot depicts an estimate for $L_F(g_n)$; standard errors are $(L_F(g_n)(1 - L_F(g_n))/1000)^{1/2}$. E.g., $n_j = 180$; $k_n = 53$; $\widehat{L}_F^{1000}(g_n) = 0.057$; standard error less than 0.01. We reject $H_0 : L_F(g^*) \geq 0.10$ at $\alpha = 0.01$. $L_F(g^*) \approx 0$ for this simulation.