

Supplementary Materials for: Are mental properties supervenient on brain properties?

Joshua T. Vogelstein^{1*}, R. Jacob Vogelstein², Carey E. Priebe¹

¹Department of Applied Mathematics & Statistics,
Johns Hopkins University, Baltimore, MD, 21218,

²National Security Technology Department,
Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723

1 Relations between sets

In this appendix we aim to provide more intuition regarding supervenience by discussing the limitations and extent of its implications.

First, a supervenient relation does not imply an injective relation. An injective relation is any relation that preserves distinctness. Thus if minds are injective on brains, then $b \neq b' \implies m \neq m'$ (note that the directionality of the implication has been switched relative to supervenience). However, it might be the case that a brain could change without the mind changing. Consider the case that a single subatomic particle shifts its position by a Plank length, changing brain state from b to b' . It is possible (likely?) that the mental state supervening on brain state b remains m , even after b changes to b' . In such a scenario, the mind might still supervene on the brain, but the relation from brains to minds is not injective. This argument also shows that supervenience is not necessarily a *symmetric* relation. Minds supervening on brains does not imply that brains supervene on minds.

Second, supervenience does not imply causality. For instance, consider an analogy where M and B correspond to two coins being flipped, each possibly landing on heads or tails. Further assume that every time one lands on heads so does the other, and every time one lands on tails, so do the other. This implies that M supervenes on B , but assumes nothing about whether M causes B , or B causes M , or some exogenous force causes both.

Third, supervenience does not imply identity. The above example with the two coins demonstrates this, as the two coins are not the same thing, even if one has perfect information about one from the other.

What supervenience does imply, however, is the following. Imagine finding two different minds. If $\mathcal{M} \stackrel{S}{\sim}_F \mathcal{B}$, then the brains subvening under those two minds must be different. In other words, there cannot be two different minds, either of which could supervene on a single brain. Figure 1 shows several possible relations between the sets of minds and brains.

Note that statistical supervenience is distinct from statistical correlation. *Statistical correlation* between brain states and mental states is defined as $\rho_{MB} = \mathbb{E}[(B - \mu_B)(M - \mu_M)] / (\sigma_B \sigma_M)$, where μ_X and σ_X are the mean and variance of X , and $\mathbb{E}[X]$ is the expected value of X . If $\rho_{MB} = 1$, then both $\mathcal{M} \stackrel{S}{\sim}_F \mathcal{B}$ and $\mathcal{B} \stackrel{S}{\sim}_F \mathcal{M}$. Thus, perfect correlation implies supervenience, but supervenience does not imply correlation. In fact, supervenience may be thought of as a generalization of correlation which can be applied to arbitrary valued random variables (such as mental or brain properties), unlike correlation which is only defined for scalar valued random variables.

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 [4]. The hermaphroditic *C. elegans*' somatic nervous system consists of 279 interconnected neurons. Although the graph with these neurons as vertices and edges defined by chemical synapses between neurons is likely not identical across individuals, it appears to be reasonably consistent [4]. Furthermore, these animals exhibit a rich

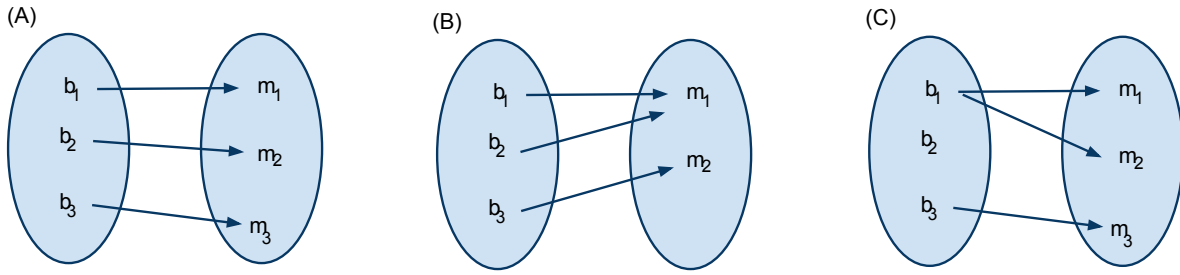


Figure 1: Possible relations between minds and brains. (A) Minds supervene on brains, and it so happens that there is a bijective relation from brains to minds. (B) Minds supervene on brains, and it so happens that there is a surjective (a.k.a., onto) relation from brains to minds. (C) Minds are *not* supervenient on brains, because two different minds supervene on the same brain.

behavioral repertoire that seemingly depends on circuit properties [5]. Thus, one may design an experiment by describing the joint distribution $F_{M,B}$ 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 $F_{M=m_j}$. 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 $F_{M,B}$ (Figure 2). To generate the data, let E_{ij} be an integer-valued random variable whose value indicates the number of synapses (edges) between neurons (vertices) i and j . Let the class-conditional random variable $E_{ij}|M = m_0$ be distributed $\text{Poisson}(A_{ij} + \eta)$, where A_{ij} is the number of chemical synapses between neuron i and neuron j according to [6], with noise parameter $\eta = 0.05$. Let \mathcal{E} be the set of edges deemed responsible for odor-evoked behavior according to [7]. Therefore, the distribution of these edges must differ between the two classes. The class-conditional random variable $E_{ij}|M = m_1$ is distributed $\text{Poisson}(A_{ij} + z_{ij})$, where the signal parameter $z_{ij} = \eta$ for all edges not in \mathcal{E} , and z_{ij} is uniformly sampled from $[-5, 5]$ for all edges within \mathcal{E} .

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 $k_n \rightarrow \infty$ as $n \rightarrow \infty$ and $k_n/n \rightarrow 0$ as $n \rightarrow \infty$, ensuring universal consistency. (Better classifiers can be constructed for the joint distribution $F_{M,B}$ used here; however, we demand universal consistency.) Figure 2 shows that for this simulation, rejecting $\varepsilon = 0.1$ -supervenience at $\alpha = 0.01$ only requires a few hundred training samples.

Importantly, conducting this experiment *in actu* is not beyond current technological limitations. 3D super-resolution imaging [8] combined with neurite tracing algorithms [9, 10, 11] allow the collection of a *C. elegans* 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 $M = m_1$ [5], and the class of each organism (m_0 vs. m_1) can also be determined automatically [12].

References

- [1] Stone, C. J. *The Annals of Statistics* **5**(4), 595–620 July (1977).
- [2] Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. Springer, (1996).
- [3] Garey, M. R. and Johnson, D. S. *Computers and intractability. A guide to the theory of NP-completeness. A Series of Books in the Mathematical Sciences*. WH Freeman and Company, San Francisco, Calif, (1979).
- [4] Durbin, R. M. *Studies on the Development and Organisation of the Nervous System of Caenorhabditis elegans*. PhD thesis, University of Cambridge, (1987).
- [5] de Bono, M. and Maricq, A. V. *Annu Rev Neurosci* **28**, 451–501 (2005).

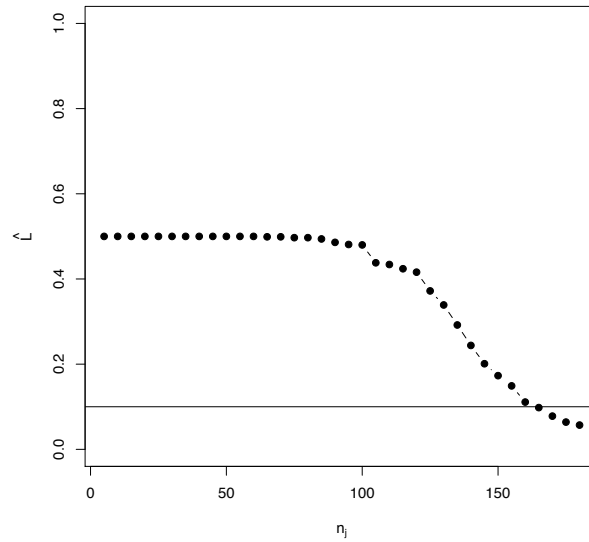


Figure 2: *C. elegans* graph classification simulation results. $\hat{L}_F^{n'}(g_n)$ (the misclassification rate estimated with $n' = 1000$ testing samples) is plotted as a function of class-conditional training sample size $n_j = n/2$, suggesting that for $\varepsilon = 0.1$ we can determine that $\mathcal{M}_{F,B}^{\varepsilon}$ holds with 99% confidence with just a few hundred training samples generated from $F_{M,B}$. Each dot depicts an estimate for $\hat{L}_F^{n'}(g_n)$; standard errors are $(\hat{L}_F^{n'}(g_n)(1 - \hat{L}_F^{n'}(g_n))/n')^{1/2}$; e.g., $n_j = 180$; $k_n = 53$; $\hat{L}_F^{n'}(g_n) = 0.057$; 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.

- [6] Varshney, L. R., Chen, B. L., Paniagua, E., Hall, D. H., Chklovskii, D. B., Spring, C., and Farm, J. *World Wide Web Internet And Web Information Systems*, 1–41.
- [7] Chalasani, S. H., Chronis, N., Tsunozaki, M., Gray, J. M., Ramot, D., Goodman, M. B., and Bargmann, C. I. *Nature* **450**(7166), 63–70 November (2007).
- [8] Vaziri, A., Tang, J., Shroff, H., and Shank, C. V. *Proceedings of the National Academy of Sciences of the United States of America* **105**(51), 20221–6 December (2008).
- [9] Helmstaedter, M., Briggman, K. L., and Denk, W. *Current opinion in neurobiology* **18**(6), 633–41 December (2008).
- [10] Mishchenko, Y. *J Neurosci Methods* **176**(2), 276–289 January (2009).
- [11] Lu, J., Fiala, J. C., and Lichtman, J. W. *PLoS ONE* **4**(5), e5655 (2009).
- [12] Buckingham, S. D. and Sattelle, D. B. *Invertebrate neuroscience : IN* **8**(3), 121–31 September (2008).