

# Are mental properties supervenient on brain properties?

Joshua T. Vogelstein<sup>1\*</sup>, R. Jacob Vogelstein<sup>2</sup>, Carey E. Priebe<sup>1</sup>

<sup>1</sup>Department of Applied Mathematics & Statistics,  
Johns Hopkins University, Baltimore, MD, 21218,

<sup>2</sup>National Security Technology Department,  
Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723

The “mind-brain supervenience” conjecture suggests that all mental properties are derived from the physical properties of the brain. To address the question of whether the mind supervenes on the brain, we frame a supervenience hypothesis in rigorous statistical terms. Specifically, we propose a modified version of supervenience (called  $\epsilon$ -supervenience) that is amenable to experimental investigation and statistical analysis. To illustrate this approach, we perform a thought experiment that illustrates how the probabilistic theory of pattern recognition can be used to make a one-sided determination of  $\epsilon$ -supervenience. The physical property of the brain employed in this analysis is the graph describing brain connectivity (i.e., the brain-graph or connectome).  $\epsilon$ -supervenience allows us to determine whether a particular mental property can be inferred from one’s connectome to within any given misclassification rate  $> 0$ , regardless of the relationship between the two. This may provide motivation for cross-disciplinary research between neuroscientists and statisticians.

Questioning the relationship between the mind (our thoughts, beliefs, preferences, emotions, intelligences, 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 conjecture, 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. This work is an attempt to bridge the gap between these philosophical conjectures and experimentally testable hypotheses.

The primary contributions of this work flow from our introduction of a notion of supervenience amenable to empirical investigation. This renders the mind-brain dualism debate a hypothesis, rather than an assumption, both expanding the space of questions amenable to hypothesis testing, and placing limits on this space. Because hypothesis tests (implicitly sometimes) depend on a model, a very general model of brains and their associated mental properties is proposed. Fortunately, this formulation admits universally consistent classifiers, that is, classifiers guaranteed to find the relationship between minds and brains, if one exists, given sufficient data. Many previous investigations relating brains and mental properties can therefore be considered  $\varepsilon$ -supervenience hypothesis tests. This paradigm, therefore, generalizes previous approaches, embedding them in a rigorous statistical framework, and suggests avenues for future research.

## Results

### Statistical supervenience

Donald Davidson canonized the mind-brain supervenience relation in 1970 with the following quote: [3]

supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect.

This conjecture may be concisely and formally stated thusly:  $m \neq m' \implies b \neq b'$ , where  $(m, b), (m', b') \in \mathcal{M} \times \mathcal{B}$  are mind-brain pairs. This mind-brain supervenience relation does not imply an injective relation, a causal relation, or an identity relation (see Appendix 1 for more details and some examples). To facilitate both statistical analysis and empirical investigation, we convert this supervenience relation from a logical to a probabilistic relation.

Let  $b$  correspond to an agent's brain, which is a particular element from the set of all possible brains,  $\mathcal{B}$ . Similarly, let  $m$  correspond to an agent's mind, which is a particular element from the set of all possible minds,  $\mathcal{M}$ . Let  $\mathbb{P}[M, B]$  indicate a joint distribution of minds and brains. Statistical supervenience can thusly be defined:

**Definition 1.**  $\mathcal{M}$  is said to statistically supervene on  $\mathcal{B}$  for distribution  $\mathbb{P} = \mathbb{P}[M, B]$ , denoted  $\mathcal{M} \stackrel{S}{\sim}_{\mathbb{P}} \mathcal{B}$ , if and only if  $\mathbb{P}[m \neq m' | b = b'] = 0$ , or equivalently  $\mathbb{P}[m = m' | b = b'] = 1$ .

Statistical supervenience is therefore a probabilistic relation on sets (related to, but distinct from correlation; see Appendix 1 for details).

### Statistical supervenience is equivalent to perfect classification accuracy

If minds statistically supervene on brains,  $\mathcal{M} \stackrel{S}{\sim}_{\mathbb{P}} \mathcal{B}$ , then two different minds must supervene on two different brains. This means that there exists a deterministic function mapping each brain to its supervening mind,  $g(\cdot) : \mathcal{B} \mapsto \mathcal{M}$ ; therefore, one could in principle construct this function. The optimal such classifier,  $g^*$ , has the smallest expected misclassification rate,  $L_{\mathbb{P}}(g^*)$ . The minimum misclassification rate is called Bayes error (see Methods for details). The primary result of casting supervenience as a statistical framework is the following theorem:

**Theorem 1.**  $\mathcal{M}$  is said to statistically supervene on  $\mathcal{B}$  for distribution  $\mathbb{P} = \mathbb{P}[M, B]$ , denoted  $\mathcal{M} \stackrel{S}{\sim}_{\mathbb{P}} \mathcal{B}$ , if and only if  $L_{\mathbb{P}}(g^*) = 0$ . Formally,  $\mathcal{M} \stackrel{S}{\sim}_{\mathbb{P}} \mathcal{B} \Leftrightarrow L_{\mathbb{P}}(g^*) = 0$ .

If minds supervene on brains, then, by the definition of supervenience, there exists a function that maps each brain deterministically to a particular mind. This means that one could draw a decision boundary between all equivalence classes of brains, each class corresponding to a different mind, and no mind will reside within two different equivalence classes. Thus, the optimal classifier would correctly find these decision boundaries, and therefore have no opportunity to err.  $\square$

The above argument shows (for the first time to our knowledge) that statistical supervenience and zero Bayes error are equivalent. Statistical supervenience can therefore be thought of as a constraint on the possible distributions on minds and brains. Specifically, let  $\mathcal{P}$  indicate the set of all possible joint distributions on minds and brains, and let  $\mathcal{P}_s$  be subset of distributions for which supervenience holds. Theorem 1 implies that  $\mathcal{P}_s = \{\mathbb{P}[M, B] : L_{\mathbb{P}}(g^*) = 0\} \subseteq \mathcal{P}$ .

### $\varepsilon$ -supervenience admits hypothesis testing

Although the above theorem is of potential theoretical interest, the arguments rely on knowing the typically unknown  $\mathbb{P}[M, B]$  and  $g^*$ , rendering them useless pragmatically. However, both  $\mathbb{P}[M, B]$  and  $g^*$  could be estimated from data. Before explicitly considering the problem of testing for statistical supervenience, we define a relaxed notion of supervenience:

**Definition 2.** Given  $\varepsilon > 0$ ,  $\mathcal{M}$  is said to  $\varepsilon$ -supervene on  $\mathcal{B}$  for distribution  $\mathbb{P} = \mathbb{P}[M, B]$ , denoted  $\mathcal{M} \stackrel{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ , if and only if  $L_{\mathbb{P}}(g^*) < \varepsilon$ .

Given this relaxation, consider the problem of testing for  $\varepsilon$ -supervenience. Let the null hypothesis be  $H_0: L_{\mathbb{P}}(g_n) \geq \varepsilon$ , and the alternative hypothesis be  $H_A: L_{\mathbb{P}}(g_n) < \varepsilon$ . Hold-out error,  $\hat{L}_{\mathbb{P}}^{n'}(g_{\bar{n}})$ , as defined in Methods, is a test statistic for this hypothesis. We reject for values of the test statistic lower than the critical value, that is, we reject if and only if  $\hat{L}_{\mathbb{P}}^{n'}(g_{\bar{n}}) < c_{\alpha}(n', \varepsilon)$ . The critical value is available under the least favorable distribution  $n' \hat{L}_{\mathbb{P}}^{n'}(g_{\bar{n}}) \sim \text{Binomial}(n', \varepsilon)$ . Thus, rejection implies that we are  $100(1-\alpha)\%$  confident that  $\mathcal{M} \stackrel{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ . The definition of  $\varepsilon$ -supervenience therefore admits, for the first time to our knowledge, a statistical test of supervenience, given a specified  $\varepsilon$  and  $\alpha$ .

Importantly, the utility of any statistical test depends both on the p-value, the probability of obtaining a test statistic at least as extreme as the observed value (under the assumed model), and its power, the probability that the test will reject a false null hypothesis. Ideally, the power of this test would go to unity, as  $n, n' \rightarrow \infty$ . A sufficient condition for power to approach unity is that  $g_n$  is a *consistent* classifier. A classifier is consistent if and only if its expected misclassification rate converges to the Bayes optimal limit with sufficient data, that is  $\mathbb{E}[L_{\mathbb{P}}(g_n)] \rightarrow L_{\mathbb{P}}(g^*)$  as  $n \rightarrow \infty$ . Below, we show that under a very general mind-brain model, one can construct a consistent classifier whose power approaches unity with sufficient data.

Unfortunately, the rate of convergence of  $L_{\mathbb{P}}(g_n)$  to  $L_{\mathbb{P}}(g^*)$  depends on the (unknown) distribution  $\mathbb{P} = \mathbb{P}[M, B]$  [4]. Furthermore, arbitrarily slow convergence theorems regarding the rate of convergence of  $L_{\mathbb{P}}(g_n)$  to  $L_{\mathbb{P}}(g^*)$  demonstrate that there is no universal  $n, n'$  which will guarantee that the test has power greater than any specified target  $\beta > \alpha$  [5]. 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  $\mathcal{M} \stackrel{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$  holds, but we can never be confident in its negation; rather, it may be the case that the evidence in favor of  $\mathcal{M} \stackrel{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$  is insufficient for any number of reasons, including that we simply have not yet collected enough data. Thus, without restrictions on  $\mathbb{P}[M, B]$ , arbitrarily slow convergence theorems imply that our theorem of  $\varepsilon$ -supervenience does not strictly satisfy Popper's *falsifiability* requirement [6].

### A Gedankenexperiment demonstrating consistency and unity power

To ensure consistency and therefore unity power, the classifier  $g_n(\cdot)$  must be able to converge to the truth, regardless of the true distribution,  $\mathbb{P}$ . We therefore make explicit a model for brains, and show that under this very general model, universally consistent classifiers are available.

**Gedankenexperiment 1.** Let the physical property under consideration be brain connectivity structure (“connectome”), so  $b$  is a brain-graph (or, network) with vertices representing neurons (or neuroanatomical regions) and edges representing synapses (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 2 for proof). Therefore, the existence of a universally consistent classifier guarantees that eventually (in  $n, n'$ ) we will be able to conclude  $\mathcal{M} \tilde{\sim}_{\mathbb{P}} \mathcal{B}$  for this mind/brain property pair, if indeed  $\varepsilon$ -supervenience holds. This logic holds for directed graphs or multigraphs or hypergraphs with discrete edge weights and vertex attributes. Furthermore, Appendix 2 also extends the proof to deal with other matrix norms (which might speed up convergence), and the regression scenario, where  $|\mathcal{M}|$  is infinite.

## Discussion

We have introduced the notion of  $\varepsilon$ -supervenience, which states that the Bayes optimal misclassification rate for any mind/brain property pair is less than  $\varepsilon$ . Furthermore, when we restrict the space of minds and brains to the setting of *Gedankenexperiment 1*, we have shown that  $k_n$ -NN classifiers are universally consistent, such that one can derive a hypothesis test, with confidence level  $\alpha$ , that is guaranteed to converge to the Bayes optimal misclassification rate, given sufficient data, no matter the true (but unknown) distribution of mind/brain pair properties. Alas, this is a one-sided test, so although power converges to unity, one can never determine whether (i) more data is necessary to get a lower p-value, or (ii) that the particular  $\varepsilon$ -supervenience does not hold.

Importantly, we are *not* claiming that actually determining  $\varepsilon$ -supervenience in humans is practically possible for any particular mental property at this time (at least when the vertices represent individual neurons). Rather, the claim is that *if* one had sufficient data (a very large number of exchangeable mind/brain property pairs), then *in theory*, some level  $\varepsilon$ -supervenience hypothesis test could be performed. Even in such a scenario, a universally consistent classifier is just one of many possible kinds of classifiers, and not necessarily the best one (in terms of  $\hat{L}$ ) for any particular dataset. Further, even if a universally consistent classifier is used, a more informative and tractable distance on  $\mathcal{B}$  may be desired, as the  $k_n$ -nearest neighbor classifier under a Frobenius norm may have a rate of convergence so slow and a computational demand so high as to be impractical (but see Appendix 3 for a simulated example in which convergence is relatively fast). Whichever classifier is used, it is likely to benefit from a large amount of domain-specific knowledge, which the proposed classifier completely neglects.

A central (perhaps *the* central) quest in much of neuroscience, psychology, and cognitive science is to discover the brain properties that subvene under various mental properties, although questions are rarely cast within a supervenience formalism. Moreover, the particular brain properties that are often believed to subvene under these mental properties are neural circuits, or brain-subgraphs. To this end, many investigations in these fields include schematic diagrams showing a particular brain-subgraph subvening under a particular mental phenotype. This practice transcends the evolutionary hierarchy of neuroscientific research. For instance, in the invertebrate literature, vertices correspond to particular labeled neurons, and edges correspond to synapses [7]. In the vertebrate literature, vertices often correspond to types of neurons in particular regions, and edges correspond to tendencies of connections [8]. For primates [9] and humans [10] vertices frequently represent functionally distinct neuroanatomical regions, and edges represent regional interconnectivity. Furthermore, this practice also transcends analytical background, including anatomists [11], philosophers [12], statisticians [13], and physicists [14]. The near ubiquity of this practice suggests that a fundamental quest is to determine which brain-subgraphs subvene under which mental properties (although perhaps causality, not supervenience, is the true desideratum). Perhaps supervenience is therefore a framework that can fruitfully be applied to myriad and varied neurocognitive investigations.

In recent years, with the advent of the field of “connectomics” [15, 16], neuroimaging has driven an explosion of studies investigating the human connectome and relating connectomes to cognitive properties [17]. Many of these studies can be framed as  $\varepsilon$ -supervenience hypothesis tests. For instance, a recent study showed that using data from diffusion tensor imaging [18], one can nearly perfectly differentiate between schizophrenic individuals and normal (control) individuals [19]. As the resolution and signal-to-noise ratio of magnetic resonance imaging continue to improve, especially with more advanced techniques such as High Angular Diffusion Imaging [20], Q-Ball Imaging [21], and diffusion spectrum imaging [22], similar results could be obtained with other, more subtle cognitive properties. Furthermore, the utilization of other imaging technologies, such as polarized light imaging [23] and high-throughput electron microscopy [24, 25], will continue to improve the effective resolution of these inferred connectomes from human brains. While determining from a brain scan whether a particular indi-

vidual knows calculus might be quite distant, many other cognitive and psychological supervenience hypotheses have already been tested, and the gap between testing for calculus and testing for schizophrenia seems to be diminishing.

Although hypothesis testing for a particular  $\epsilon$ -supervenience appears to be possible based on the *Gedankenexperiment* in Results, a natural question to raise is: what are the conceivable alternative hypotheses? We consider three such alternatives.

First, perhaps brains are more accurately characterized as quantum networks over classical networks. Several authors have suggested that brains have certain hypercomputational properties that classical computers could not achieve [26, 27]. However, assuming that the computer can be represented as a network, the above results hold regardless of whether computations in the brain are quantum or classical. This follows because quantum networks merely speed up computation for certain classes of problems; they cannot, however, solve problems that classical computers cannot [28]. This means that if the above analysis failed to reject the null hypothesis at level  $\alpha$ , it will fail regardless of whether one assumes quantum or classical computations.

Second, perhaps minds stochastically supervene on brains [29]. While perhaps difficult to imagine, much like a non-deterministic world was difficult to imagine prior to modern physics, it is not inconceivable that mental properties are only stochastically determined by physical ones.

A third alternative hypothesis is supernatural causal effects. The above analysis could be considered an empirical test for whether we have souls, or, perhaps whether souls play a causal role in our mental properties over and above the physical role played by the brain, or whether the data we have suggests that the probability that our souls play a measurable causal role over and above the physical is less than  $\epsilon$ .

It therefore seems that failing to reject the null hypothesis that a particular mental property  $\epsilon$ -supervenes on a particular brain property could potentially be explained by stochastic or supernatural forces, but not a quantum network brain model.

The *Gedankenexperiment* in Section did not require simulating any dynamics; rather, the dynamics are necessarily a function of the model parameters (statics). Similarly, for the question of mind-brain supervenience in humans, one need not ever observe any activity of the brain, one must merely observe the model that determines the activity (in a potentially stochastic process). Thus, this approach to understanding the relationship between mind and brain is distinct from the standard systems neuroscience paradigm, in which the goal is typically to understand the neural activity “code.” In contrast, if mind-brain supervenience holds, it motivates a search for the neural *connectivity* “code,” a so-called “engram” for memories [30, 31, 32, 33, 34], or more generally a *mengram*, the neural signature of any mental property, be it cognitive, psychological, or otherwise (note that supervenience allows for the particular mengram of a mental property to vary both across individuals and time).

This *Gedankenexperiment*, together with (i) the formal definition of  $\epsilon$ -supervenience as a constraint on distributions, (ii) the brain-graph model, and (iii) the universal consistency proof on graphs, is the first demonstration (to our knowledge) that empirically investigating supervenience is at least theoretically possible. The above discussion suggests that many previously conducted investigations either assume supervenience, or test it. Further, new technologies facilitate testing supervenience of mental properties on brain-graphs more easily.

## Methods

Assuming for the moment that the space of all possible minds is finite, that is,  $|\mathcal{M}| < \infty$ , then we call any such function a *classifier* (this assumption will later be relaxed). Let  $\hat{m}$  denote the output of a classifier,  $g(b) = \hat{m}$ . Define misclassification rate as  $L_{\mathbb{P}}(g) = \mathbb{P}[g(B) \neq M]$  which denotes the probability that  $g$  misclassifies  $b$ . The Bayes optimal classifier  $g^*$  minimizes  $L_{\mathbb{P}}(g)$  over all classifiers, that is:  $g^* = \arg\min_g L_{\mathbb{P}}(g)$ . Thus, the *Bayes error*, or *Bayes risk*,  $L_{\mathbb{P}}(g^*)$  is the minimum possible misclassification rate.

Let  $\mathcal{T}_n = \{(m_1, b_1), (m_2, b_2), \dots, (m_n, b_n)\}$  be a set of random samples taking their values in  $\mathcal{M} \times \mathcal{B}$ , each independently and identically distributed according to  $\mathbb{P}[M, B]$ . Generalizing the concept of a classifier  $g$  to allow incorporation of training data, consider  $g_n : \mathcal{B} \times (\mathcal{M} \times \mathcal{B})^n \mapsto \mathcal{M}$  which takes as input an observed brain  $b$  and training data  $\mathcal{T}_n$ , and produces a classification:  $g_n(b; \mathcal{T}_n) = \hat{m}$ . Misclassification rate for this classifier will be a random variable, because the training data  $\mathcal{T}_n$  are random samples. The expected misclassification rate for this classifier is therefore approximated by “hold-out” error:  $\hat{L}_{\mathbb{P}}^{n'}(g_n) = \mathbb{P}[g_n(B) = M | \mathcal{T}_{\tilde{n}}]$ , where  $\tilde{n} = n - n'$ , and  $n' < n$  is the number of held-out training samples (samples not used to obtain  $g_n(\cdot)$ ). The approximate number of misclassified minds therefore has a binomial distribution:  $n' \hat{L}_{\mathbb{P}}^{n'}(g_n) \sim \text{Binomial}(n', L_{\mathbb{P}}(g_n))$ .

## References

- [1] Plato. *Plato: complete works* (Hackett Pub Co, 1997).
- [2] Descartes, R. *Meditationes de prima philosophia* (1641).
- [3] Davidson, D. *Experience and Theory*, chap. Mental Events (Duckworth, 1970).
- [4] Devroye, L., Györfi, L. & Lugosi, G. *A Probabilistic Theory of Pattern Recognition* (Springer, 1996).
- [5] Devroye, L. On arbitrarily slow rates of global convergence in density estimation. *Probability Theory and Related Fields* **62**, 475–483 (1983).
- [6] Popper, K. *The logic of scientific discovery* (1959).
- [7] Geoffrey North, R. J. G. (ed.) *Invertebrate neurobiology* (CSHL Press, 2007).
- [8] Shepherd, G. *The synaptic organization of the brain* (Oxford University Press New York, 2004).
- [9] Felleman, D. & Van Essen, D. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex* **1**, 1 (1991).
- [10] S. Mori, P. v. Z. L. N.-P., S. Wakana. *MRI Atlas of Human White Matter* (Elsevier Science, 2005).
- [11] Abeles, M. *Corticonics* (Cambridge University Press, 1991).
- [12] Koch, C. & Davis, J. *Large-scale neuronal theories of the brain* (The MIT Press, 1994).
- [13] Rao, R., Olshausen, B. & Lewicki, M. *Probabilistic models of the brain: Perception and neural function* (The MIT Press, 2002).
- [14] Chow, C., Gutkin, B., Hansel, D., Meunier, C. & Dalibard, J. (eds.) *Methods and Models in Neurophysics* (Elsevier, 2003).
- [15] Sporns, O., Tononi, G. & Kotter, R. The human connectome: A structural description of the human brain. *PLoS Computational Biology* **1**, e42 (2005).
- [16] Hagmann, P. *From diffusion MRI to brain connectomics*. Ph.D. thesis, Institut de traitement des signaux (2005).
- [17] Sporns, O. *Networks of the Brain* (MIT Press, 2010).
- [18] Basser, P. J., Mattiello, J. & LeBihan, D. Mr diffusion tensor spectroscopy and imaging. *Biophys J* **66**, 259–267 (1994). URL [http://dx.doi.org/10.1016/S0006-3495\(94\)80775-1](http://dx.doi.org/10.1016/S0006-3495(94)80775-1).
- [19] Ardekani, B. A. *et al.* Diffusion tensor imaging reliably differentiates patients with schizophrenia from healthy volunteers. *Hum Brain Mapp* (2010). URL <http://dx.doi.org/10.1002/hbm.20995>.
- [20] Tuch, D. *et al.* High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. *Magnetic Resonance in Medicine* **48**, 577–582 (2002).
- [21] Tuch, D. Q-ball imaging. *Magnetic Resonance in Medicine* **52**, 1358–1372 (2004).
- [22] Wedeen, V., Hagmann, P., Tseng, W., Reese, T. & Weisskoff, R. Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging. *Magnetic Resonance in Medicine* **54**, 1377–1386 (2005).
- [23] Palm, C. *et al.* Towards ultra-high resolution fibre tract mapping of the human brain—registration of polarised light images and reorientation of fibre vectors. *Frontiers in Human Neuroscience* **4** (2010).
- [24] W. Denk, W. & Horstmann, H. Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. *PLOS Biol.* **2**, e329 (2004).

- [25] Hayworth, K., Kasthuri, N., Schalek, R. & Lichtman, J. Automating the collection of ultrathin serial sections for large volume TEM reconstructions. *Microscopy and Microanalysis* **12**, 86–87 (2006).
- [26] Penrose, R. & Gardner, M. *The emperor's new mind: concerning computers, minds, and the laws of physics* (Oxford University Press, USA, 1999).
- [27] Satinover, J. *The quantum brain: the search for freedom and the next generation of man* (Wiley, 2002).
- [28] Nielson, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge University Press, 2000).
- [29] Craver, C. F. Stochastic supervenience (2009).
- [30] Semon, R. W. *The Mneme* (G. Allen & Unwin Ltd., 1921).
- [31] Lashley, K. In search of the engram. *Symposia of the society for experimental biology* **4**, 30 (1950).
- [32] Zhang, W. & Linden, D. The other side of the engram: experience-driven changes in neuronal intrinsic excitability. *Nature Reviews Neuroscience* **4**, 885–900 (2003).
- [33] Shema, R., Sacktor, T. & Dudai, Y. Rapid erasure of long-term memory associations in the cortex by an inhibitor of PKM {zeta}. *Science* **317**, 951 (2007).
- [34] Berry, J., Krause, W. & Davis, R. Olfactory memory traces in *Drosophila*. *Progress in brain research* **169**, 293–304 (2008).

## Acknowledgments

The authors would like to acknowledge helpful discussions with J. Lande, B. Vogelstein, and S. Seung.

## Author Contributions

JTV, RJV, and CEP conceived of the manuscript. JTV and CEP wrote it. CEP ran the experiment.

## Additional Information

The authors have no competing financial interests to declare.