Editorial Manager(tm) for PLoS ONE
Manuscript Draft

Manuscript Number: 10-PONE-RA-20839R1

Title: Are mental properties supervenient on brain properties?

Short Title: mind-brain supervenience

Article Type: Research Article

Section/Category: Other

Keywords: bayesian, statistical, supervenience, philosophy, hypothesis testing, machine learning, subspace identification, inference,

Corresponding Author: Joshua Vogelstein

Corresponding Author's Institution: Johns Hopkins University

First Author: Joshua Vogelstein

Order of Authors: Joshua Vogelstein;R. Jacob Vogelstein;Carey E Priebe

Abstract: The "mind-brain supervenience"' conjecture suggests that all mental properties (e.g. consciousness, intelligence, personality, etc.) are derived from the physical properties of the brain. The validity of this conjecture has been argued in philosophical terms for over 2,500 years. Alternative conjectures, including various non-physical causes of mental properties, seem rather implausible to many, but proving or disproving these alternatives has remained elusive.

To address the question of whether the mind supervenes on the brain, we here frame a supervenience hypothesis in rigorous mathematical 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.

To the philosopher, this work shows how philosophical conjectures can be transformed into statistical hypotheses that are amenable experimental investigation. This allows the philosopher to gain empirical support for her rational arguments. To the statistician, this work points out the limitations of hypothesis testing; and suggests that some of these limitations have not previously been fully appreciated. To the neuroscientist, this work indicates that much of contemporary research can be framed in terms of investigating supervenience. This should provide further motivation for cross-disciplinary research between neuroscientists and statistical graph-theoreticians.

Suggested Reviewers: Luc DeVroye
McGill
lucdevroye@gmail.com

In our opinion, his understanding of statistical pattern recognition is unmatched.

Patrick Suppes
Stanford
psuppes@stanford.edu
He has expertise in both neuroscience and philosophy.

Opposed Reviewers:

Response to Reviewers: Dear Prof Sporns and two helpful reviewers,

Thank you for your insightful comments. Based on your input, we have significantly revised the text of our manuscript entitled "Are Mental Properties Supervenient on Brain Properties?" Specifically, the manuscript is now re-organized to emphasize the flow of our thesis. Moreover, we have clarified the notion of supervenience with an appendix, and expanded upon the relationship of this work to both previous and future work in the discussion. We hope that you will find that our revisions adequate address your concerns. Below we provide detailed responses to specific comments:

Reviewer #1:

Thank you for your helpful comments. We agree that the original presentation of these ideas lacked some organization and clarity. The main theorem is that supervenience, which is currently known mostly by philosophers, actually makes a very strong claim about statistical classification accuracy. Only upon realizing this connection can hypothesis tests be created for supervenience. Given the possibility of such a test, we show that not only do such tests exist, but also that a quite general model of minds and brains admits a universally consistent classifier with power converging to unity. We believe that this idea is quite deep and interesting. In particular, the manuscript explains how many previously conducted investigations can be framed as epsilon-supervenience hypothesis tests, unifying previously disparate work. Moreover, these ideas further motivate the now burgeoning field of connectomics, which seems to operate on the assumption that a certain epsilon-supervenience holds for many mind-brain property pairs. Thus, it is our belief that this manuscript formally ties together many contemporary and future neurocognitive investigations in previously unrecognized ways.

Reviewer #2:

Thank you for your helpful comments. We address each separately.

First, regarding the very compact mathematical style, we have expanded to exposition of mathematical details in the main text with the intention that it will be more accessible to many interested philosophers and scientists, while recognizing that a certain level of proficiency with statistical inference and/or pattern recognition will likely be necessary to understand the mathematical details.

Second, regarding the definition of supervenience, we have expanded the text in the main document and added an appendix (Appendix A) to further expand upon the nature of supervenience relations, as well as compare and contrast it with other relations by providing examples, as you suggested. We have also now provided more description (final paragraph of Appendix B) and a reference to further expound upon the notion of isomorphism that we are employing in this manuscript.

Third, regarding the proof of universal consistency, we have modified the text to demonstrate that our "proof" is merely a realization that the composition of (i) representing a graph by its adjacency matrix

such that it lives in finite dimensional Euclidean space and (ii) a isomorphism-matching operator extends Stone's original proof to this domain. Therefore, we feel as though Stone's original proof is truly the workhorse here. Appendix B merely explains how one can extend the proof to this domain. Part of the beauty of this proof is that the the classifier works regardless of the structure of the brain-graph.

Finally, we fixed all the minor English flaws that you pointed out to us.

Please note that due to the significant restructuring of this manuscript, a red-lined version seemed inappropriate (in fact, upon making a red-lined version, essentially every line was either red or blue, which was effectively useless). As described above, in addition to re-organizing, Appendix A is wholly novel, and the discussion section has been greatly expanded. We hope the reviewers will be easily able to find these modifications.

**Original Cover Letter**

July 8, 2010

Dear Editorial Board:

On behalf of myself and my coauthors, I submit for your review a copy of our manuscript entitled, "Are mental properties supervenient on brain properties?" We hope that you will consider publishing this manuscript to the prestigious Public Library of Science (PLoS) ONE journal.

At the core of this paper is a thought experiment in which we mathematically prove that if the graph corresponding to brain connectivity (i.e., the *connectome*) is statistically related to a particular mental property (e.g., propensity for mathematics), one could build a classifier to determine whether any individual's brain exhibits that property with an arbitrarily small misclassification rate. The proof follows directly from the probabilistic theory of pattern recognition and is made possible by our novel exposition of universal consistency for a $k_n$-nearest neighbor classifer on graphs. In addition to these theoretical results, the manuscript describes a simulation that suggests how one could actually build such a classifier for the "brain" of *Caenorhabditis elegans* using today's technology.

Our results have many implications, some philosophical, some statistical, and some pragmatic. First, the philosophical implications: The relationship between the mind and brain has been investigated for millennia, using tools ranging from abstract arguments to microstimulation of individual neurons. However, until now, no one (to our knowledge) has proposed a framework that allows for empirical investigation of this relationship at the scale of the connectome. Such a framework will undoubtedly be of value as the multiple ongoing large-scale efforts to collect a connectome (cf. NIH Human Connectome Project) begin paying dividends.

In addition to the philosophical implications, our work has significant statistical implications. Whereas several groups have recently proposed algorithms for classifying graphs, there exists very little theoretical work on the limitations of these tools. In contrast, our work is founded on a constructive proof of a universally consistent classifier on graphs, the limitations of which are both known and well-understood. Furthermore, the proof is shown to be only one-sided, meaning that we can never gain confidence in the negation of the null hypothesis. Although we only apply this classifier to simulated brain-derived graphs, the methods are applicable to many other kinds of graph data. The recent emergence of networks and graphs as the preferred representations of structured relationships in multiple scientific disciplines suggests that the development of statistical tools for rigorous analysis of network data is a pressing need for the scientific community at large.

Finally, the results described in this manuscript have pragmatic implications. Studies relating brain structure to brain function are being published with increasing frequency due to the increasing availability of diffusion-weighted magnetic resonance imaging and other connectivity-imaging technologies. However, without sophisticated statistical tools for the analysis of such data, to-date the findings have (mostly) been limited to linear correlations between scalar measures of connectivity and the brain function under investigation. The graph-theoretic approach proposed here allows investigators to utilize the entirety of the information obtained from such measures, rather than using simple descriptive statistics.

In summary, we believe that the work described in this manuscript makes significant contributions in several contemporary fields of inquiry, and is therefore suited for a general-audience journal such as PLoS ONE. Please do not hesitate to contact us with any questions or comments.

Respectfully,

Joshua T. Vogelstein, PhD.
Johns Hopkins University
Dept. Applied Math. & Stat's
3400 N. Charles St.
Barton 320
Baltimore, MD 21210
(443) 858-9911

# Are mental properties supervenient on brain properties?

Joshua T. Vogelstein[1], R. Jacob Vogelstein[2], Carey E. Priebe[1]
[1]Department of Applied Mathematics & Statistics,
Johns Hopkins University, Baltimore, MD, 21218,
[2]National Security Technology Department,
Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723

February 18, 2011

**Abstract**

The "mind-brain supervenience" conjecture suggests that all mental properties (e.g. consciousness, intelligence, personality, etc.) are derived from the physical properties of the brain. The validity of this conjecture has been argued in philosophical terms for over 2,500 years. Alternative conjectures, including various non-physical causes of mental properties, seem rather implausible to many, but proving or disproving these alternatives has remained elusive.

To address the question of whether the mind supervenes on the brain, we here frame a supervenience hypothesis in rigorous mathematical terms. Specifically, we propose a modified version of supervenience (called $\varepsilon$-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 $\varepsilon$-supervenience. The physical property of the brain employed in this analysis is the graph describing brain connectivity (i.e., the brain-graph or connectome). $\varepsilon$-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.

To the philosopher, this work shows how philosophical conjectures can be transformed into statistical hypotheses that are amenable experimental investigation. This allows the philosopher to gain empirical support for her rational arguments. To the statistician, this work points out the limitations of hypothesis testing; and suggests that some of these limitations have not previously been fully appreciated. To the neuroscientist, this work indicates that much of contemporary research can be framed in terms of investigating supervenience. This should provide further motivation for cross-disciplinary research between neuroscientists and statistical graph-theoreticians.

## 1 Introduction

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 [**?**]. Approximately two millennia passed before these ideas reached their canonical form through Descartes's discussion of mind-body dualism [**?**]. 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 [**?**]. 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.

## 2 Statistical supervenience

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

> 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: 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, $\mathcal{M}$. The supervenience conjecture implies that: $m \neq m' \implies b \neq b'$, 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 A 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 $\mathbb{P}[M, B]$ indicate a joint distribution of minds and brains.

**Definition 1.** *$\mathcal{M}$ is said to statistically supervene on $\mathcal{B}$ for distribution $\mathbb{P} = \mathbb{P}[M, B]$, denoted $\mathcal{M} \overset{S}{\sim}_{\mathbb{P}} \mathcal{B}$, 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 A for details).

## 3 Statistical supervenience is equivalent to perfect classification accuracy

If minds statistically supervene on brains, $\mathcal{M} \overset{S}{\sim}_{\mathbb{P}} \mathcal{B}$, 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. It may be the case that subsets of brains form equivalence classes, such that any brain in that subset is mapped to the same mind. 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 $\widehat{m}$ denote the output of a classifier, $g(b) = \widehat{m}$. 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: $g^* = \operatorname{argmin}_g L_{\mathbb{P}}(g)$. Thus, the *Bayes error*, or Bayes risk, $L_{\mathbb{P}}(g^*)$ is the minimum possible misclassification rate. 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 $\mathcal{M} \overset{S}{\sim}_{\mathbb{P}} \mathcal{B}$, if and only if $L_{\mathbb{P}}(g^*) = 0$. Formally, $\mathcal{M} \overset{S}{\sim}_{\mathbb{P}} \mathcal{B} \Leftrightarrow L_{\mathbb{P}}(g^*) = 0$.*

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. □

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 $\mathcal{P}_s = \{\mathbb{P}[M, B] : L_{\mathbb{P}}(g^*) = 0\} \subseteq \mathcal{P}$.

# 4 A hypothesis test for supervenience

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. Let $\mathcal{T}_n = \{(m_1, b_1), (m_2, b_2), \ldots, (m_n, b_n)\}$ 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: $g_n(b; \mathcal{T}_n) = \widehat{m}$. 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: $\widehat{L}_{\mathbb{P}}^{n'}(g_{\widetilde{n}}) = \mathbb{P}[g_{\widetilde{n}}(B) = M | \mathcal{T}_{\widetilde{n}}]$, where $\widetilde{n} = n - n'$, and $n' < n$ is the number of held-out training samples (samples not used to obtain $g_{\widetilde{n}}(\cdot)$). The approximate number of misclassified minds therefore has a binomial distribution: $n' \widehat{L}_{\mathbb{P}}^{n'}(g_{\widetilde{n}}) \sim \text{Binomial}(n', L_{\mathbb{P}}(g_{\widetilde{n}}))$.

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 $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$, 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 $H_A$: $L_{\mathbb{P}}(g_n) < \varepsilon$. We reject for values of the test statistic lower than the critical value, that is, we reject if and only if $n' \widehat{L}_{\mathbb{P}}^{n'}(g_{\widetilde{n}}) < c_\alpha(n', \varepsilon)$. The critical value is available under the least favorable distribution Binomial$(n', \varepsilon)$. Thus, rejection implies that we are $100(1 - \alpha)$% confident that $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$. 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 $n, n' \to \infty$. 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 $\mathbb{E}[L_{\mathbb{P}}(g_n)] \to L_{\mathbb{P}}(g^*)$ as $n \to \infty$. 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 $\mathbb{P} = \mathbb{P}[M, B]$ [**?**]. 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 $\beta > \alpha$ [**?**]. 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 $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ holds, but we can never be confident in its negation; rather, it may be the case that the evidence in favor of $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ 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 [**?**].

# 5 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 B for proof). Therefore, the existence of a universally consistent classifier guarantees that eventually (in $n, n'$) we will be able to conclude $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ 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 B also extends the proof to deal with other matrix norms (which might speed up convergence), and the regression scenario, where $|\mathcal{M}|$ is infinite.*

# 6 Discussion

## 6.1 Summary

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.

## 6.2 Practical issues

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 $\widehat{L}$) 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 C 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.

## 6.3 A unified quest

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 [**?**]. In the vertebrate literature, vertices often correspond to types of neurons in particular regions, and edges correspond to tendencies of connections [**?**]. For primates [**?**] and humans [**?**] vertices frequently represent functionally distinct neuroanatomical regions, and edges represent regional interconnectivity. Furthermore, this practice also transcends analytical background, including anatomists [**?**], philosophers [**?**], statisticians [**?**], and physicists [**?**]. 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 neurcognitive investigations.

## 6.4 Human applications

In recent years, with the advent of the field of "connectomics" [**?**, **?**], neuroimaging has driven an explosion of studies investigating the human connectome and relating connectomes to cognitive properties [**?**]. 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 [**?**], one can nearly perfectly differentiate between schizophrenic individuals and normal (control) individuals [**?**]. 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 [**?**], Q-Ball Imaging [**?**], and diffusion spectrum imaging [**?**], similar results could be obtained with other, more subtle cognitive properties. Furthermore, the utilization of other imaging technologies, such as polarized light imaging [**?**] and high-throughput electron microscopy [**?**, **?**], will continue to improve the effective resolution of these inferred connectomes from human brains. While determining from a brain scan whether a particular individual 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.

## 6.5 Alternative explanations

Although hypothesis testing for a particular $\epsilon$-supervenience appears to be possible based on the *Gedanken-experiment* in Section 5, 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 [**?**, **?**]. 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 [**?**]. 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 [**?**]. 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 $\varepsilon$.

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

## 6.6 Dynamics vs. statics

The *Gedankenexperiment* in Section 5 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 [**?**, **?**, **?**, **?**, **?**], 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).

## 6.7 Concluding thoughts

This *Gedankenexperiment*, together with (i) the formal definition of $\varepsilon$-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.

# A 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). For instance, 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 $b'$; or that a single additional synapse is formed between a pair of neruons. 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. Consider, for example, acceleration and velocity. Clearly, acceleration supervenes on velocity, as acceleration cannot change without velocity changing (assuming one does not consider gravity as acceleration). Similarly, velocity supervenes on position, as velocity cannot change without position changing. Therefore, acceleration supervenes on position, by the transitive property of supervenience, but it is not the case that a change in acceleration is equal to a change in position. Rather, position can change with constant velocity, meaning without acceleration changing. Thus, to claim that something supervenes on another is not equivalent to claim that the two are identical.

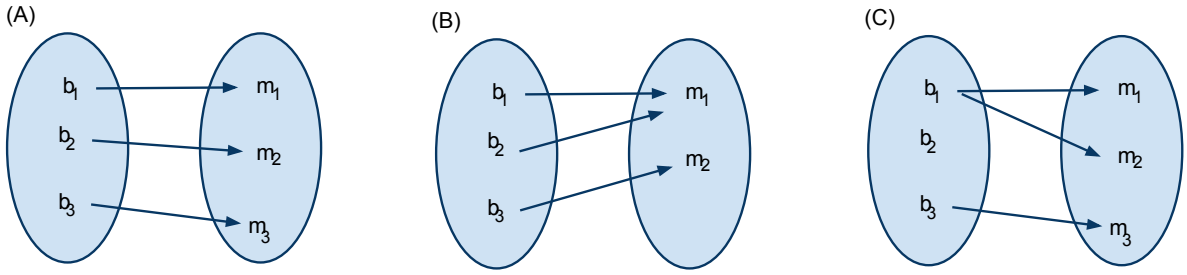


Figure 1: Possible relations between minds and brains. (A) Minds supervene on brains, and it so happens that there is a bijective relation from brains to minds. (B) Minds supervene on brains, and it so happens that there is a surjective (a.k.a., onto) relation from brains to minds. (C) Minds are *not*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

What supervenience does imply, however, is the following. Imagine finding two different minds. If $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$, 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 $\mu_X$ and $\sigma_X$ are the mean and variance of $X$, and $\mathbb{E}[X]$ is the expected value of $X$. If $\rho_{MB} = 1$, then both $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ and $\mathcal{B} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{M}$. Thus, perfect correlation implies supervenience, but supervenience does not imply correlation.

## B $k_n$ nearest neighbor algorithm

Consider the following problem setup. We have a collection of training data, $\mathcal{T}_n = \{(m_i, b_i)\}_{i=1}^n$, each sampled exchangably from some unknown joint distribution, $(m_i, b_i) \overset{iid}{\sim} \mathbb{P}[M, B]$, where $m_i$ and $b_i$ are the observed mental and brain properties of experiment $i$, respectively. A new brain, $b$, called the "test brain", is then observed, and one desires to find the most likely class of the new brain, $m$. It is further assumed that the test mind/brain pair is sampled from the same distribution as the training data, $(m, b) \sim \mathbb{P}[M, B]$, and $m$ is unobserved. Further assume that $m$ can take one of a finite number of possible values, that is, $|\mathcal{M}| < \infty$.

The 1-nearest neighbor (1-NN) classifier works as follows. Compute the distance between the test brain and all the training brains, $d_i = d(b, b_i)$ for all $i \in [n]$, where $[n] = 1, 2, \ldots, n$. Then, sort them, $d_{(1)} < d_{(2)} < \ldots < d_{(n)}$, and their corresponding mental properties, , $m_{(1)}, m_{(2)}, \ldots, m_{(n)}$, where parenthetical indices indicate rank order. The 1-NN algorithm predicts that the unobserved mind is of the same class as the closest brain's class: $\widehat{m} = m_{(1)}$. The $k_n$ nearest neighbor is a straightforward generalization of this approach. It says that the test mind is in the same class as which ever class is the plurality class of the $k_n$ nearest neighbors, $\widehat{m} = \operatorname{argmax}_{m'} \mathbb{I}\{\sum_{i=1}^{k_n} m_{(i)} = m'\}$. Given a particular choice of $k_n$ (the number of nearest neighbors to consider), and a choice of $d(\cdot, \cdot)$ (the distance metric used to compare the test datum and training data), one then has a relatively simple and intuitive algorithm.

Unfortunately, no such algorithm is universally consistent. Let $g_n$ be the $k_n$ nearest neighbor classifier when there are $n$ training points. Then, a collection of such algorithms, $\{g_n\}$, with $k_n$ increasing with $n$, can be universally consistent under certain constraints. In particular, as $n$ increases, $k_n$ must also increase, but not quite as quickly. Formally, $k_n$ must satisfy: (i) $k_n \to \infty$ as $n \to \infty$ and (ii) $k_n/n \to 0$ as $n \to \infty$. In Stone's original proof, $b$ was assumed to be a $d$-dimensional vector, and the $L_2$ norm ($d(b, b') = \sum_{j=1}^d (b_j - b'_j)^2$, where $j$ indexes elements of the $d$-dimensional vector) was shown to satisfy the constraints on a distance metric for this collection of classifiers to be universally consistent. Later, others extended these results to apply to any $L_p$ norm [**?**]. When brain-graphs are represented by their adjacency matrices, one can stack the columns of the adjacency matrices, effectively embedding graphs into finite Euclidean space, in which case Stone's theorem applies. Stone's original proof applied to the scenario when $|\mathcal{M}|$ was infinite, resulting in a universally consistent regression algorithm as well.

Note that the above extension of Stone's original theorem to the graph domain implicitly assumed that vertices were labeled, such that elements of the adjacency matrices could easily be compared across graphs. In theory, when vertices are unlabeled, one could first map each graph to a quotient space invariant to isomorphisms, and then proceed as before. Unfortunately, solving the graph-matching problem is currently NP-Incomplete (meaning it is not known to be either P or NP) [**?**], so in practice, dealing with unlabeled vertices will likely be computationally challenging.

## C 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 [**?**]. 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 [**?**]. Furthermore, these animals exhibit a rich behavioral repertoire that seemingly depends on circuit properties [**?**]. Thus, one may design an experiment by describing the joint distribution $\mathbb{P}[M, B]$ via class-conditional distributions $\mathbb{P}[B|M = m_j]$ 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 $\mathbb{P}[M = m_1]$. 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 $\varepsilon$-supervenience with reasonable choices for $\varepsilon$ and $\alpha$, and a plausible joint distribution $\mathbb{P}[M, B]$ (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 Poisson$(A_{ij} + \eta)$, where $A_{ij}$ is the number of chemical synapses between neuron $i$ and neuron $j$ according to [**?**], with noise parameter $\eta = 0.05$. Let $\mathcal{E}$ be the set of edges deemed responsible for odor-evoked

behavior according to [**?**]. Therefore, the distribution of these edges must differ between the two classes. The class-conditional random variable $E_{ij}|M = m_1$ is distributed 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 $k_n \to \infty$ 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 $\mathbb{P}[M, B]$ 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.

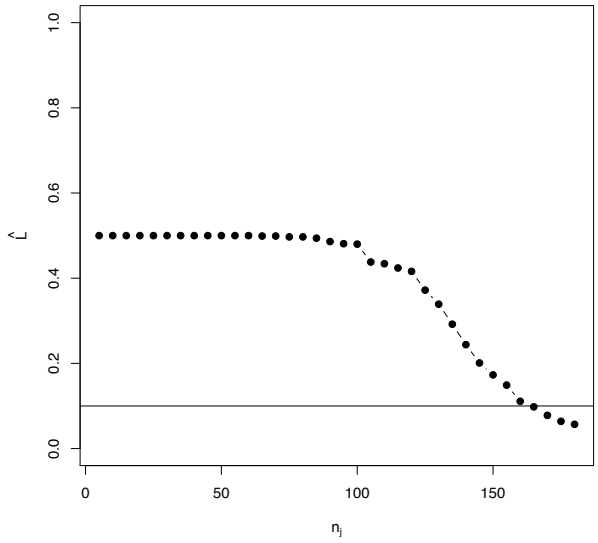


Figure 2: C. elegans graph classification simulation results. $\widehat{L}$ (the misclassification rate estimated upon with $1000$ testing samples) is plotted as a function of class-conditional training sample size $n_j = \widetilde{n}/2$, suggesting that for $\varepsilon = 0.1$ we can determine that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

Importantly, conducting this experiment *in actu* is not beyond current technological limitations. 3D super-resolution imaging [**?**] combined with neurite tracing algorithms [**?**, **?**, **?**] 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 $M = m_1$ [**?**], and the class of each organism ($m_0$ vs. $m_1$) can also be determined automatically [**?**].

## Acknowledgments

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

# Are mental properties supervenient on brain properties?

Joshua T. Vogelstein[1,*], R. Jacob Vogelstein[1,2], Carey E. Priebe[1]
**1 Department of Applied Mathematics and Sciences, Johns Hopkins University, Baltimore, MD, USA**
**2 National Security Technology Department, Johns Hopkins University Applied Physics Laboratory, Baltimore, MD, USA**
∗ **E-mail: joshuav@jhu.edu**

## Abstract

The "mind-brain supervenience" conjecture suggests that all mental properties (e.g. consciousness, intelligence, personality, etc.) are derived from the physical properties of the brain. The validity of this conjecture has been argued in philosophical terms for over 2,500 years. Alternative conjectures, including various non-physical causes of mental properties, seem rather implausible to many, but proving or disproving these alternatives has remained elusive.

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 $\varepsilon$-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 $\varepsilon$-supervenience. The physical property of the brain employed in this analysis is the graph describing brain connectivity (i.e., the *connectome*), and $\varepsilon$-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.

The potential significance of this work can be divided into distinct disciplines. To the philosopher, this work demonstrates that philosophical conjectures can be morphed into statistical hypotheses, amenable to experimental investigations, allowing the philosopher to add empirical support to their rational arguments. To the statistician, herein lies the first proof to our knowledge of the existence of a universally consistent classifier on graphs, and a constructivist one at that. To the neuroscientist, a theoretically possible experiment is proposed to garnish support for a hypothesis that is widely believed: that mental properties supervene on brain properties.

## Introduction

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 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. Here we define new versions of supervenience that formulate the conjecture in rigorous mathematical terms and that can be experimentally tested as a hypothesis.

The primary contributions of this work are as follows. First, a notion of supervenience amenable to empirical investigation is formally introduced. This renders the mind-brain dualism debate a hypothesis, rather than an assumption. Second, in addition to expanding the space of questions amenable to hypothesis testing, we also demonstrate the the limits of hypothesis testing. Third we posit a very general model of brains and their associated mental properties that admits statistical analysis in a graph theoretical and statistical framework. Fourth, we prove that this formulation admits a universally consistent classifier that is guaranteed to find the relationship between minds and brains, if one exists. Fifth we demonstrate through simulation that the proposed universally consistent classifier has reasonable convergence properties on simulated brain-graph data.

# Results

Let $\mathcal{B}$ be the observation space for some physical property, such as brain connectivity structure (i.e., connectome; see [4–6]). 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 $L_F(g^*) = 0$.*

(Note that this definition does not imply a one-to-one mapping.) To allow for the possibility of only *partial* supervenience, we relax the above statistical supervenience to define *$\varepsilon$-supervenience*:

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

Unfortunately, in general $F$ is unknown, but can be estimated from the data. Therefore, 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 $\vec{d}_n = (b_1, m_1), \cdots, (b_n, m_n)$ and produces a classification $\widehat{m} = g_n(b; \vec{d}_n)$. Let $L_F(g_n) = E[P_F[g_n(B; \vec{D}_n) \neq M | \vec{D}_n]]$.

Consider the problem of testing for $\varepsilon$-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 $\alpha$ 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:

**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 $\varepsilon$-supervenience does not strictly 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 (or, network) 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 $\varepsilon$-supervenience holds. This logic holds for directed graphs or multigraphs or hypergraphs with discrete edge weights and vertex attributes. Furthermore, Appendix 1 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

While the above thought experiment addresses the question of $\varepsilon$-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 $\alpha$, 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 (or other norm; see Appendix 1 for details) 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 [11–13] and various approaches to inference on brain graphs [14–16], 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 $\varepsilon$-supervenes on a brain's connectivity structure. This thought experiment, together with (i) the formal definition of $\varepsilon$-supervenience, (ii) the brain-graph model, and (iii) the universal consistency proof on graphs, is the first *proof* (to our knowledge) that empirically investigating supervenience is at least theoretically possible.

## Appendix 1: $k_n$-nearest neighbor universal consistency for graphs

Assume first that all graphs are simple (meaning undirected with no loops and binary binary edges), 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 non-zero 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.

Finally, the above proof can be straightforwardly generalized to utilize any matrix norm, $\|b_1 - b_2\|_A = \|A(b_1 - b_2)\|$, assuming that $AA^\mathsf{T}$ is positive definite (see [8] pg. 455 for proof when $b_i$'s are vectors). Furthermore, we can relax the constraint that mental properties finite, that is, we can allow any $|\mathcal{M}| = \infty$ (see [20] for proof when $b_i$'s are vectors). The proofs for the case when $b$ is an adjacency matrix follows immediately upon first defining a bijective embedding of the matrix into a vector space, and then embedding each adjacency matrix in that space.

## 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 [21]. 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 [21]. Furthermore, these animals exhibit a rich behavioral repertoire that depends on circuit properties [22]. 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 $\varepsilon$-supervenience with reasonable choices for $\varepsilon$ and $\alpha$ 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 [23], with noise parameter $0 < \eta \ll 1$. The class-conditional random variable $E_{ij}|M = m_1$ is distributed Poisson$(A_{ij} + z_{ij})$ for neurons $i, j \in \mathcal{D}$, where $\mathcal{D}$ is the set of edges deemed

responsible for odor-evoked behavior according to [24], with signal parameter $z_{ij}$ 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.)

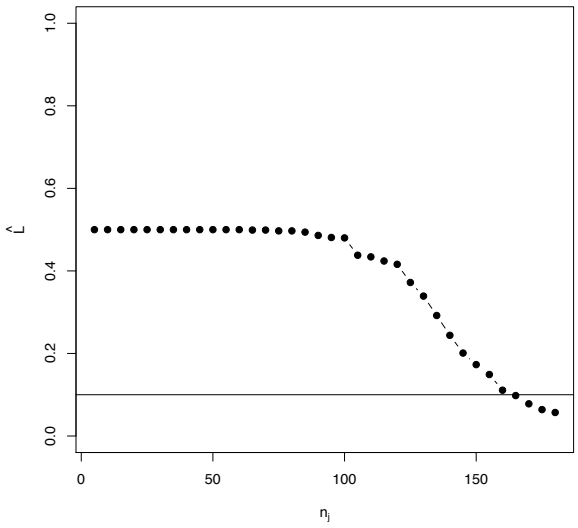


**Figure 1.** C. elegans graph classification simulation results. $\widehat{L}_F^{1000}(g_n)$ is plotted as a function of class-conditional training sample size $n_j$, suggesting that for $\varepsilon = 0.1$ we can determine that $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ holds with 99% 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.

Importantly, conducting this experiment *in actu* is not beyond current technological limitations. 3D superresolution imaging [25] combined with neurite tracing algorithms [15, 16, 26] allow the collection of a brain-graph in a day. Genetic manipulations, laser ablations, and training paradigms can each be used to obtain a non-wild type population for use as $M = m_1$ [22], and the class of each organism ($m_0$ vs. $m_1$) can also be determined automatically [27].

## Acknowledgments

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

## References

1. Plato (1997) Plato: complete works. Hackett Pub Co.

2. Descartes R (1641) Meditationes de prima philosophia.

3. Davidson D (1970) Experience and Theory, Duckworth, chapter Mental Events.

4. Sporns O, Tononi G, Kotter R (2005) The human connectome: A structural description of the human brain. PLoS Computational Biology 1: e42.

5. Lichtman JW, Livet J, Sanes JR (2008) A technicolour approach to the connectome. Nat Rev Neurosci 9: 417–422.

6. Seung H (2009) Reading the Book of Memory: Sparse Sampling versus Dense Mapping of Connectomes. Neuron 62: 17–29.

7. Kim J (2005) Phiilosophy of Mind. Westview Press, second edition, 352 pp.

8. Devroye L, Györfi L, Lugosi G (1996) A Probabilistic Theory of Pattern Recognition. Springer.

9. Devroye L (1983) On arbitrarily slow rates of global convergence in density estimation. Probability Theory and Related Fields 62: 475–483.

10. Popper K (1959) The logic of scientific discovery .

11. White J, Southgate E, Thomson JN, Brenner S (1986) The structure of the nervous system of the nematode caenorhabditis elegans. Philosophical Transactions of Royal Society London Series B, Biological Sciences 314: 1-340.

12. W Denk W, Horstmann H (2004) Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. PLOS Biol 2: e329.

13. Briggman K, Denk W (2006) Towards neural circuit reconstruction with volume electron microscopy techniques. Current opinion in neurobiology 16: 562–570.

14. Macke JH, Maack N, Gupta R, Denk W, Schlkopf B, et al. (2008) Contour-propagation algorithms for semi-automated reconstruction of neural processes. J Neurosci Methods 167: 349–357.

15. Mishchenko Y (2009) Automation of 3d reconstruction of neural tissue from large volume of conventional serial section transmission electron micrographs. J Neurosci Methods 176: 276–289.

16. Lu J, Fiala JC, Lichtman JW (2009) Semi-automated reconstruction of neural processes from large numbers of fluorescence images. PLoS ONE 4: e5655.

17. Horn R, Johnson C (1990) Matrix analysis. Cambridge Univ Pr.

18. Conroy JM, Kratzer SG, Podrazik LJ (1997) A continuous method for the quadratic assignment problem. In: Society for Industrial and Applied Mathematics.

19. Zaslavskiy M, Bach F, Vert J (2008) A path following algorithm for the graph matching problem. IEEE Transactions on Pattern Analysis and Machine Intelligence 31: 2227-2242.

20. Stone C (1977) Consistent nonparametric regression. The annals of statistics : 595–620.

21. Durbin RM (1987) Studies on the Development and Organisation of the Nervous System of Caenorhabditis elegans. Ph.D. thesis, University of Cambridge.

22. de Bono M, Maricq AV (2005) Neuronal substrates of complex behaviors in c. elegans. Annu Rev Neurosci 28: 451–501.

23. Varshney L, Chen B, Paniagua E, Hall D, Chklovskii D (2009) Structural Properties of the Caenorhabditis elegans Neuronal Network. ArXiv .

24. Chalasani SH, Chronis N, Tsunozaki M, Gray JM, Ramot D, et al. (2007) Dissecting a circuit for olfactory behaviour in caenorhabditis elegans. Nature 450: 63–70.

25. Vaziri A, Tang J, Shroff H, Shank CV (2008) Multilayer three-dimensional super resolution imaging of thick biological samples. Proc Natl Acad Sci U S A 105: 20221–20226.

26. Helmstaedter M, Briggman KL, Denk W (2008) 3d structural imaging of the brain with photons and electrons. Curr Opin Neurobiol 18: 633–641.

27. Buckingham SD, Sattelle DB (2008) Strategies for automated analysis of c. elegans locomotion. Invert Neurosci 8: 121–131.

# Figure Legends

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.

**(this is just copied and pasted from response to reviewers)**

Dear Prof Sporns and two helpful reviewers,

Thank you for your insightful comments. Based on your input, we have significantly revised the text of our manuscript entitled "Are Mental Properties Supervenient on Brain Properties?" Specifically, the manuscript is now re-organized to emphasize the flow of our thesis. Moreover, we have clarified the notion of supervenience with an appendix, and expanded upon the relationship of this work to both previous and future work in the discussion. We hope that you will find that our revisions adequate address your concerns. Below we provide detailed responses to specific comments:

Reviewer #1:

Thank you for your helpful comments. We agree that the original presentation of these ideas lacked some organization and clarity. The main theorem is that supervenience, which is currently known mostly by philosophers, actually makes a very strong claim about statistical classification accuracy. Only upon realizing this connection can hypothesis tests be created for supervenience. Given the possibility of such a test, we show that not only do such tests exist, but also that a quite general model of minds and brains admits a universally consistent classifier with power converging to unity. We believe that this idea is quite deep and interesting. In particular, the manuscript explains how many previously conducted investigations can be framed as epsilon-supervenience hypothesis tests, unifying previously disparate work. Moreover, these ideas further motivate the now burgeoning field of connectomics, which seems to operate on the assumption that a certain epsilon-supervenience holds for many mind-brain property pairs. Thus, it is our belief that this manuscript formally ties together many contemporary and future neurocognitive investigations in previously unrecognized ways.

Reviewer #2:

Thank you for your helpful comments. We address each separately.

First, regarding the very compact mathematical style, we have expanded to exposition of mathematical details in the main text with the intention that it will be more accessible to many interested philosophers and scientists, while recognizing that a certain level of proficiency with statistical inference and/or pattern recognition will likely be necessary to understand the mathematical details.

Second, regarding the definition of supervenience, we have expanded the text in the main document and added an appendix (Appendix A) to further expand upon the nature of supervenience relations, as well as compare and contrast it with other relations by providing examples, as you suggested. We have also now provided more description (final paragraph of Appendix B) and a reference to further expound upon the notion of isomorphism that we are employing in this manuscript.

Third, regarding the proof of universal consistency, we have modified the text to demonstrate that our "proof" is merely a realization that the composition of (i) representing a graph by its adjacency matrix such that it lives in finite dimensional Euclidean space and (ii) a isomorphism-matching operator extends Stone's original proof to this domain. Therefore, we feel as though Stone's original proof is truly the workhorse here. Appendix B merely explains how one can extend the proof to this domain. Part of the beauty of this proof is that the the classifier works regardless of the structure of the brain-graph.

Finally, we fixed all the minor English flaws that you pointed out to us.

Please note that due to the significant restructuring of this manuscript, a red-lined version seemed inappropriate (in fact, upon making a red-lined version, essentially every line was either red or blue,

which was effectively useless). As described above, in addition to re-organizing, Appendix A is wholly novel, and the discussion section has been greatly expanded. We hope the reviewers will be easily able to find these modifications.

**LaTeX Source File of Revised Manuscript (TEX file)**

Click here to download LaTeX Source File (TEX file): supervenience.tex

**LaTeX Bibliography (BIB file)**

Click here to download LaTeX Bibliography (BIB file): supervenience.bbl