## Gerstein Lab experience in Quantum Computing

In the first phase of the PsychENCODE consortium, we successfully tapped into the modularity and data-agnosticism of machine learning methods like deep Boltzmann machines to build a predictive model of psychiatric phenotypes. This model was trained on genotyping, transcriptomic and epigenetic data, and involved learning parameters by searching across a vast energy landscape. While the second phase will expand upon published deep learning frameworks to incorporate single cell data and new functional genomics assays, we believe a parallel exploration of novel computational methodologies will be central to the success of such an endeavor. This parallel study is necessitated by the increasing size and complexity of the available data, which places significant burdens on the speed of model training and also substantially expands the parametric landscape to scales where it is hard to guarantee the discovery of optima, local or global. The scope of this project is, therefore, to evaluate quantum computing methods as an alternative to classical computing using tissue-level and single-cell functional genomics data collected as a part of the efforts of the PsychENCODE Data Analysis Core (DAC).

Concurrent with the evolution of “big data” efforts, quantum computing has emerged as a tantalizing new area of computational exploration. A large network of academic and private entities has arisen to tackle the development of scalable quantum hardware, and many incarnations of quantum computing systems have already been made accessible to the research community. An effort towards the implementation of quantum computing methods is especially timely given the recent passing of the U.S. National Quantum Initiative Act 2018, which calls for the implementation of a National Quantum Initiative directed towards the development of quantum information science and technology1, as well as the European Quantum Technologies Flagship, a ten-year effort with the advancement of quantum computing technology as one of the primary targets2, and the UK’s National Quantum Technologies Programme3. So-called “Noisy Intermediate-Scale Quantum” technology4, with 50-100 qubits, is due in the near future. These advances in hardware will potentially allow the implementation of a large cohort of quantum algorithms developed to address data analysis needs, such as quantum linear algebra, quantum deep learning, and, more broadly, quantum machine learning. These algorithms were designed to provide significant speed-ups relative to their classical counterparts on classical problems, and also to make tractable simulations of inherently quantum mechanical systems. While many such methods involve direct manipulation of qubits and the exploitation of the possible interference between qubit states, a small subset of these algorithms are, intriguingly, trainable on classical systems. This raises the prospect of testing these quantum algorithms on classical problems with a twofold purpose: (a) in the short term, these methods could possibly help to form an alternative model of multidimensional structure in complex datasets by training parameters using quantum Hamiltonians and optimization; and (b) in the long term, enable a seamless transition to eventual quantum hardware by casting existing questions in terms of the quantum computing paradigm.

The choice of running quantum computing algorithms directly on quantum computers or on classical machines through simulation depends on the availability of resources. We can run quantum algorithms based on quantum annealing as classical simulations, and to test models based on quantum logic gates directly on the publicly accessible IBM Q system5.

The parent grant for the PsychENCODE DAC addresses the goals of developing new methods for analyzing regulatory elements in the brain and integrating the data from PsychENCODE with those from other consortia to build a comprehensive picture of genomic regulation. The grant also included plans for the validation of prioritized regulatory elements using capture and whole-genome STARR-seq, and CRISPR knockout.

**Potential of Quantum Machine Learning for integrative modeling of psychiatric genomics data:**

Two main quantum computing methodologies have been implemented in practice: quantum annealing and gate model schemes. The first scheme is primed for the solution of ubiquitous optimization problems, where a cost function is mapped on to the Hamiltonian of a qubit architecture and the system is made to adiabatically transition to a final configuration through the steady modification of perturbing forces (say, magnetic fields). Quantum tunneling is, in general, a quintessential component of the transition process. The second scheme is premised on the application of quantum logic gates (single- or multi-qubit) to manipulate qubits, with the eventual promise of universal computation. Both these schemes are maturing fast, and have been the bases of numerous machine learning implementations6–9.

**Classical Boltzmann Machines with Simulated Quantum Boltzmann Machines** Characterization of the roles of enhancer elements in the brain can benefit greatly from embedding them in a gene regulatory network (GRN) modeling framework. GRNs model the regulatory relationships between transcription factors (TFs) and target genes, which may be mediated by TFs binding to promoters or enhancers of a target gene. Potential regulatory linkages may be found by analyzing the TF sequence motifs on promoters and enhancers, and analysis of the 3D conformation of the genome (Hi-C) for physical links between enhancers and genes. In our previous work10, we developed a classical Boltzmann Machine based model with embedded GRN structure, to provide a joint model of gene expression and enhancer activity levels within post-mortem prefrontal cortex tissue from psychiatric and control subjects. Our model also included dependencies between these variables and other molecular and high-level phenotypes, as well as their dependence on genetic variation, with the whole model referred to as a Deep Structured Phenotype Network (DSPN).

 We briefly summarize here a reduced form of the DSPN model, including only gene expression, enhancer activity and genetic variant variables, which we call an ‘imputation model’, since it may be used to impute genomic data from genotype observations alone. This reduced model has the form of a Conditional Boltzmann Machine11, and may be summarized by the following energy model:

,

where **,** and represent genomics, hidden and genetics variables respectively, where for convenience we assume and are binarized. Further, is split into sublayers, representing bulk-level enhancers activities (H3K27ac) and gene expression (). The model is parameterized by , which represent the dependencies between different variables as follows: are bias terms determined by eQTL and cQTL interactions (expression/chromatin Quantitative Trait Loci) with , are additional hidden bias terms, is a sparse matrix representing the GRN connectivity (i.e. only TF-target gene and enhancer-gene linkages are non-zero), and is aweight matrix connecting genomic and hidden variables. In ref. 10, we trained the model using Persistent Markov-Chain Monte-Carlo, and showed that the model improved prediction of gene expression and enhancer activities from genotype predictors compared to using independent additive polygenic predictors per gene/enhancer.

 The classical Boltzmann model above can be straightforwardly generalized to a quantum Boltzmann machine (QBM)6. In place of the classical energy model, we instead use a quantum Hamiltonian:

Here, the model is defined over q-bits, which are partitioned into subsets , representing genomics and hidden variables as above. Further, we use the notation:

where , , and similarly define in terms of the Pauli matrix . The parameters are defined similarly to the classical model, except that the weights may now take complex values (for tractability, we will use the GRN structure to impose sparsity on ). Also, the final term has no analogue in the classical energy; here the weights represent the strength of a local transverse field at q-bit , permitting the QBM to represent distributions of greater complexity than the classical BM through implicit higher-order interactions. The probabilistic model for the QBM is defined via the density matrix:

where is the partition function, . Inference may be performed by measuring the q-bits in the -basis, hence generating binary estimates of expression and enhancer activity variables. The probability of a joint (binary) configuration of can then be expressed via a partial trace: . As in ref. 6, the QBM can be trained by minimizing the negative log-likelihood . However, since evaluating gradient of can be intractable to compute for large-scale models, in ref. 6 an upper-bound is introduced whose gradient can be expressed in terms of the expectations of measured quantities, and thus approximated through sampling:

where is a ‘clamped’ Hamiltonian, whose q-bits are clamped to values corresponding to the classical data vector **.** Minimizing directly sets the transverse terms , while these must be treated as hyper-parameters when minimizing .

**QNN and SVM models on IBM Q systems** For the quantum computing classification problem, there are two recently published methods: a quantum variational classifier implemented on the IBM Q system8; and a quantum neural network (QNN) framework developed by Farhi and Neven7, that was evaluated through classical simulations. Both methods are based on the idea that a set of parametric one- and two-qubit unitary operators (or ‘gates’) acting on the input qubits can be optimized so as to yield a map that best predicts the output labels . We describe the general methodology of the variational classifier in the following, though the concepts underlying the QNN framework are similar. The input states are first transformed from their classical binary form to a quantum superposition of states through a series of phase-parameterized unitary one- and two-qubit gates. This is done to set up a scenario that is difficult to simulate classically. However, for the sake of the comparison attempted in this analysis, we could also directly use the classical binary encoding of the qubits. Subsequently, a sequence of single-qubit rotations and entangling gates are applied: the strengths of the rotations and entangling interactions form the parameters that can be optimized to yield the desired classifier map. Versions of stochastic gradient descent can be applied to search through parameter space. Finally, repeated iterations of the entire sequence of gates are necessary to allow sufficient measurements of the full superposition of qubit output states. This is because measurement collapses the state of the qubits, and so an acquisition of the statistics of the final superposition of qubit states requires many such measurement samples.

 Havlicek et al8 provide helpful Jupyter notebooks to run their classification analysis17.

**Simulated Quantum Annealing simulations:**

**Simulated Quantum Annealing for optimizing non-negative matrix factorization decompositions.** Simulated Quantum Annealing (SQA) sampling is a classical simulation of the evolution of a quantum mechanical system under a time-varying noise or temperature term. Analogous to classical simulated annealing, a system undergoes adiabatic evolution in the presence of a perturbation term that allows the system to stochastically explore different states. The energy function, or Hamiltonian, is initiated with a simpler functional form and gradually driven toward the target energy function that is to be minimized. The evolution is driven by a reduction in the noise term to zero. In the quantum mechanical case, there is the additional possibility of constructive and/or destructive interference of the system wavefunction, and tunneling through energy barriers. Indeed, it is the possibility of tunneling which offers potential speed-up advantages to quantum annealing, or even its classically simulated cousin, SQA. It is important to note that SQA is only an approximate method and has significant differences from true quantum annealing18,19. However, studies have shown that there are important similarities between the tunneling amplitude in the simulation, and that in the real system20,21.

Consider the framework of SQA, as described in recent work by Li et al22 for a linear decomposition of the form where the values represent the bulk expression of gene i, and the represent the expression of gene i in cell type j. The goal is to infer the cell fractions through regularized optimization by decomposing the gene expression matrix **B** into nonnegative matrices **V** and **H**, such that **B** ~ **VH**, where **V** consists of the latent feature vectors as its columns and **H** represents the corresponding contributions of those components to the bulk gene expression matrix. The goal of the decomposition is to minimize the Frobenius norm of the fitting errors ||**B**-**VH**||. In terms of the expression matrices, the idea is to decompose the bulk gene expression matrix into contributions from a set of latent feature vectors, each of which potentially represents a biological factor, say, a cell type.

Recent work by O’Malley et al9 applied a hybrid classical + quantum computing approach to the NMF problem. The latent feature matrix **V** was optimized in a classical fashion, while the contribution matrix **H** was binarized and optimized using D-Wave’s quantum annealer. For a fixed, classically determined matrix **V**, the columns of the matrix **H** can be determined as

where represents a column vector to be optimized, and the minimization is over a quadratic function of the difference between the bulk expression values and the decomposition. This optimization problem can be cast as an energy minimization process for a spin Hamiltonian with the components of the vector acting as the z-components of spins:

 where and are functions of the bulk expression and latent feature matrices.

Specifically, the implementation of SQA ca be carried out using path integral Monte Carlo12,13, which involves the mapping of a D-dimensional quantum problem on to an equivalent D+1-dimensional classical one and then taking stochastic steps while reducing the coupling term to 0. The coupling term allows the exploration of different spin configurations, while the reduction of this term to 0 ensures that the energy steadily decreases.

**NMF models on IBM Q system.** In previous work10, we showed that imputed cell-type fractions carry substantial predictive power for classifying case vs controls for psychiatric conditions, such as Schizophrenia, Bipolar disorder and Autism Spectrum disorder.