Accelerating Oncology Drug Discovery With Quantum Computing

Johnathan D. Mercer
University of Washington

(Dated: December 8, 2020)

Cancer cells are as elusive as the quantum behavior from which they are governed, and we are now poised to 'fight fire with fire.' Cancer drug discovery is at the intersection of many fields, which now includes quantum computing. This paper intends to contribute to the interdisciplinary discussion by traversing these fields by way of the application of quantum annealing in ligand-based virtual screening. As such, it is written with introductions to new terms and resources for experts in one field, becoming familiar with another.

I. THE DRUG DISCOVERY PIPELINE

Our goal in precision medicine oncology is to outsmart evolution. Even though the biomedical and pharmaceutical industries have extremely passionate and highly trained scientists, the complexity of biology and the elusiveness of cellular proliferation has precluded the drug discovery process - from hypothesis to market - from overcoming the current orders of magnitude in the time dimension (decades) and cost dimension (Billions $).

The drug discovery process is an inventive one that's constantly evolving with new technology and methods. Although the word 'pipeline' is used and a natural mental model due to the staged filtering process, this doesn't imply it's a straight path. In fact, there has been effort to 'end the myth' of a pipeline view with systems based approaches [1].

A simplified diagram of the drug discovery process, Fig. 1, suites the purpose for our discussion. There are many analogies we can draw on here (software engineering: good design decisions early mitigates downstream risk; chaos theory: sensitive dependence on initial conditions) but obviously we want to make the best decisions early on in the discovery program to maximize efficiency and effectiveness. As shown, the discovery process has many stages, spanning from target identification all the way to post-market surveillance. Our discussion will focus on the early stages of drug discovery that entail virtual high-throughput screening (vHTS) *in silico* for which classical computing has enabled greater efficiencies in the search process [2–6], and quantum computing is hoped to augment modern computational pipelines with additional efficiencies and power to solve problems[7]. In some ways, however, the current state can be described as a method in search of a problem, in that quantum advantage is only expected for problems that can be cast as quantum algorithms. This reinforces the need for interdisciplinary expertise and collaborations that are capable of this translation process.

Screening methods are abundant and often fundamentally different, such as phenotypic drug discovery (PDD) vs. target-based drug discovery (TTD). Our focus is on TDD, but an excellent paper on the advances in discovery and comparison of strategies can be found in Holenz [8].

A. Target Identification

Cancer is a disease of the genome [9] in which the normal process of cell division is no longer controlled. There are many mechanisms by which proliferation can become excited, including damage to proto-oncogenes which promote cell growth and mitosis, to tumor suppressors (gene examples include p53 and BRCA1) that suppress mitosis and cell growth, or damage to genes involved in the DNA damage response pathway. Unfortunately, when something goes wrong and the cell division becomes uncontrolled, natural selection then drives even more aggressive progeny (which makes screening for early-stage detection vital for better prognosis).

A therapeutic target is defined as a "biological entity (usually a protein or gene) that interacts with, and whose activity is modulated by, a particular compound."[10] In cancer, attractive targets are known cancer pathway members (i.e., activity or inactivity is involved in cellular proliferation) whose function can be modulated by a compound, in which case the target is called 'druggable.' The entire research community (beyond cancer) continues to push the boundaries of the known druggable genome, such as with the "Illuminating the Druggable Genome (IDG)" project [11].

The standard of cancer care is currently founded on chemotherapy, radiation therapy and surgery. Precision medicine has ushered in primarily two additional modes of therapy: monoclonal antibodies (mAbs) and small-molecule ($\leq$ 500 Da) inhibitors (SMIs).

The former, mABs, harnesses the power of the immune system by combating cancer cells' ability to evade immune response. As a canonical example, pembrolizumab was approved by the FDA first in 2014 [12], and then notably in 2017 [13] because this approval was tissue agnostic and based on a computational biomarker called micro-satellite instability (MSI) which is a measure of genomic instability (another is called Tumor Mutation Bur-

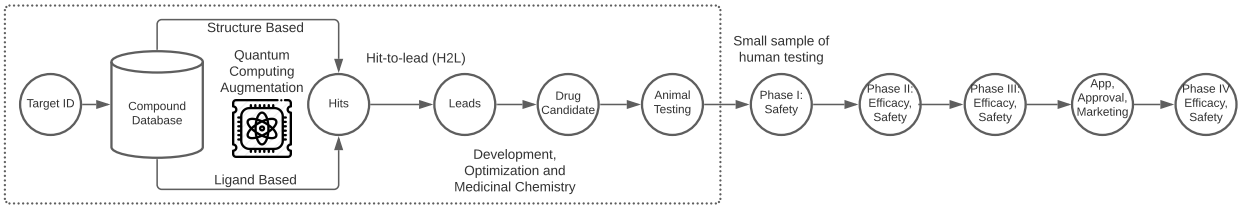

FIG. 1. Simplified diagram of the drug dicovery process (without feedback loops)

den (TMB)) which provides a signal that immunotherapy may be advantageous by precluding cancer cell ability to evade immune response (in this case via blocking the activity of a molecule called programmed death receptor-1 (PD-1) on the surface of the cell).

Small-molecule inhibitors are able to interact with and change the state and function of a protein member of a cancer pathway in a variety of locations - extracellular, surface, and intracellular - and induce a change in cellular processes involved in proliferation. A pathway is a set of molecules that work together to conduct cellular activity (further information of cancer pathways can be found in the references [14–16]. For example, Gefitinib (alt. ZD1839, or Iressa) is a SMI of epidermal growth factor receptor (EGFR) tyrosine kinase that binds to its kinase ATP binding site, which in turn, interferes with the binding of adenosine triphosphate, thereby suppressing EGFR tyrosine kinase activity and its signal transduction of EGFR [17]. For more information of cancer drugs and their targets, an excellent resource is DrugBank [18].

B. Hit Search, Leads and Optimization

A 'hit' is a molecule with some evidence to interact with and induce the desired change in the target. The historic paradigm shift of this process has been the transition from expensive and costly *in vitro* experiments to either complementing or replacing this early stage with computational approaches to generate and/or narrow a large set of molecules down to a smaller set of candidates based on a set of criteria, the latter being commonly called computer-aided drug design (CADD). The methodology and process of CADD fall into two categories based on the knowledge and use of the three-dimensional (3-D) structure of the target. If the target structure is known and used, the pipeline is called - you may have guessed - *structure-based*. Otherwise, without 3-D information on the target, the two primary methods in *ligand-based* virtual screening are quantitative structure-activity relations (QSAR) and graph based methods. Both of these approaches use a set of known active ligands (i.e., known molecules that bind to the target) to rank the new candidates. The QSAR approach uses either conventional statistical methods or new machine learning methods to train a model based on the properties (it is convention in the literature to call them molecular descriptors) of the reference set to predict their experimentally derived activity and these trained models are then used for scoring candidates.

Graph based methods for ligand-based virtual screening also use the reference set, but the ranking of candidates is based on graph similarity (where these graphs may also use molecular descriptors) between candidates and the reference set: it is this category of problem which will be our focus in later sections.

The next phase of the discovery process is often referred to as hit-to-lead (HTL) because the hits undergo an additional battery of in silico and in vitro testing to further winnow the set down to *leads* that have greater likelihood of success.

II. MOLECULAR SIMILARITY AND GRAPH BASED METHODS

Molecular similarity, introduced in the 1990's [19], is a tenet of chemical informatics and medicinal chemistry. Ultimately we are interested in finding compounds that can induce a change in activity - first in the target and pathway of interest and thereby the cancer cell. However, similarity can be computed based on many characteristics and even dimensions, and furthermore, the crux is that similar molecular properties and structure does not always imply similar biological activity. The process of predicting activity from structure is known as structure-activity relation (SAR) for which the aforementioned QSAR is the computational counterpart to in-vitro methods. Cheminformatics is a vast and mature field and I will not attempt to survey it here but will list a few considerations of similarity elaborated on in Maggiora et al [20] so that, before we explore the computational problem, we fully appreciate the biological one:
a. Chemical vs Molecular vs Biological similarity Similarity of compounds based on physicochemical properties, structural features, or induced biological profiles.
b. 2D vs 3D similarity Using molecular graphs (or more efficient fingerprint method [21]) or additional con-

formational information in the 3D structure.

c. Local vs Global vs Medicinal Chemist's Perspective The consideration of compounds in their entirety (global) or a subset of atoms or groups (local), or additional context and intuition of the investigator such as differential activity under various physiological conditions.

Nevertheless, computational similarity searches can aid and complement human intuition and judgement, as is the case with other artificial intelligence. The graph-based methods alluded to above are attractive for many reasons, including the maturity of both the mathematics and the implementations of their algorithms (also other technology such as the advent of specialized graph databases enable harnessing real-world graphs). However in practice, as all stories go, things get messy and the mathematics of graph theory is often too rigid or the algorithms are intractable to solve for real-world problems and data set sizes. This is evidenced if we follow the progression of analysis when it is desired to use graph theory for modeling molecular similarity.

A graph G is defined as set of $G = (V, E)$ where V is a set of vertices (also called nodes) and E is a set of edges (also called links). The nodes in our case are atoms or groups of them and the edges encode atomic bonds.

1. Graph similarity starts with the concept of graph isomorphism. This definition is too rigid in that it doesn't account for partial similarity between, say graphs G_1 and G_2 - i.e. it's binary, they're isomorphic or not.

2. Variations of this definition are more flexible, such as the maximum common sub-graph (MCS) defined as the largest sub-graph of G_1 that is isomorphic to G_2

3. We want to account even further for the messiness and nuance of real-world systems and can extend MCS to labelled maximum common sub-graph (LMWCS) method which incorporates more information in the vertices and edges.

4. MCS and LMWCS are equivalent to yet a another metric between graphs called the maximal independent set (MIS) based on a third graph $G_3 = f(G_1, G_2)$ which introduces the idea of a conflict graph (here defined as the largest set of vertices such that there is no edge between all selected pairs).

5. Casting the problem as the conflict graph enables further relaxation of the definition of similarity and one such way is the maximum co-k-plex whereby we seek the largest set of vertices such that each vertex has at most k-1 edges (assuming no self-edges).

6. Finally, most graph algorithms are NP-hard (at least as hard as the hardest problems that require polynomial time) so in practice, we use heuristics, approximations methods and/or parallelism to make useful graph algorithms - but these hacks are at the expense of accuracy (optimal solutions).

7. If we don't want to sacrifice accuracy for practicality then we re-formulate the graph similarity problem into an equivalent form amenable to solving in practice.

In the first of a series of papers, Hernandez et al cast the graph similarity as as a Quadratic Unconstrained Binary Optimization (QUBO) problem [22]. Then followed this work with a comparison of classical QUBO optimization to quantum annealing[23], and finally extended the method to application in ligand-based virtual screening[24]. We will revisit and elaborate on the QUBO formulation in section IV.

III. QUANTUM ANNEALING

As alluded to in the introduction, there is great promise for the application of quantum algorithms in drug discovery. We are focused here on one approach, Quantum Annealing (QA)[25], for ligand-based and graph-based virtual-screening, but this is merely illustrative of the vast opportunity for QC in discovery. So before we focus exclusively on this application, let's first make a distinction between the two primary ways of harnessing quantum computing:

a. A more accurate model of nature. Modeling the behavior of quantum systems with actual quantum systems to overcome approximate behavior with simulations, such as simulating target-ligand docking structure (in structure-based target identification and optimization VS) to better understand how nature works. I will not elaborate on Quantum Chemistry but have provided several recent comprehensive surveys [26, 27]. Furthermore, quantum computing is hoped to impact fundamental physics such as the elucidation of the AdS/CFT duality framework between general relativity and quantum mechanics [28].

b. Harnessing quantum behavior to solve problems. Here we exploit the unique characteristics of quantum behavior: superposition, entanglement, Born interpretation and computational reversibility to create gate-based quantum circuits that act as 'supreme interferometers' if one is able to cast a classical problem in this framework. See Hidary [29] for my personal favorite introduction to quantum computing. There is one more quantum phenomenon, quantum tunneling, which is also exploited for problem solving and it is this behavior that enables quantum annealing-based algorithms for which we will explore here.

A. Introduction and Pseudo-code for QA

QA views an optimization problem through a physical lens of energy minimization. Closed physical systems

evolve towards a minimum energy state and remain there unless disturbed. We model the total energy of a physical system mathematically with an object called a Hamiltonian H.

1. Specify an initial *minimal energy* Hamiltonian $H_i(t)$
2. Specify a problem Hamiltonian $H_p(t)$ which encodes the energy landscape of the optimization problem. This is encoded via *biases* and *couplings* which will be elaborated on in IV.
3. Now define a new $H_s(t)$ which is a weighted combination of the initial and the problem Hamiltonian: $H_s(t) = T(t)*H_i(t)+L(t)*H_p(t)$ with weights $T(t)$ and $L(t)$ where the notation will become more intuitive later.
4. The *annealing* evolves $H_s(t)$ 'slow enough' by transferring the minimum energy state from $H_i(t)$ to $H_p(t)$, thereby finding the minimum energy (i.e., optimal solution) to the problem. In this process, $T(t) = 1-L(t)$ where most weight begins with $c_1(t)$ and the interpolation transfers the weight and ends $L(t) = 1$ so $H_s(t) = H_p(t)$ in minimal energy state (in principle, however in practice the runs are usually conducted many times).

But what does 'slow enough' mean exactly? I'll provide a summary of the formal definition in the next section but to preface will paraphrase a great and simple presentation from Griffiths [30]: suppose you have a perfect pendulum (no friction or air resistance) in a box, if you move the box slowly enough the pendulum will keep swinging with the same amplitude whereas if you move it quickly the plane of motion and the amplitude will change.

Visually, as the system evolves in time as show in Fig. 2, the system stays in the same energy level rather than jump to another level.

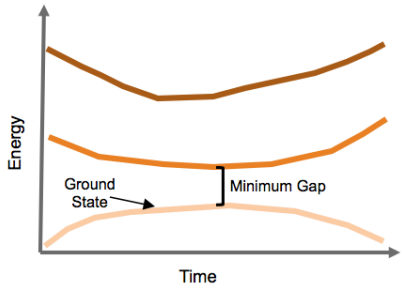

FIG. 2. A Simple Visual Illustrating the Adiabatic Evolution

B. The Adiabatic Theorem

The adiabatic theorem is the heart of QA. Without it, we wouldn't be able to transfer the minimal energy (i.e., *optimal solution*) state from $H_i(s)$ to $H_p(s)$.

In Quantum Mechanics(QM), we model a particle (or system of particles) as a wave function $\Psi(\vec{r}, t)$ which is found by solving the time-dependent Schrödinger equation:

$$H\Psi = i\hbar\frac{\partial\Psi}{\partial t} \tag{1}$$

for which the solution:

$$\Psi(\vec{r}, t) = \psi(\vec{r})e^{-iEt/\hbar} \tag{2}$$

where $\psi(\vec{r})$ is the solution of the time-independent Schrödinger equation $H\psi = E\psi$. But this is quantum statics, because the potential V, which is buried in H operator, *is not* a function of time:

$$\hat{\mathrm{H}} = -\frac{\hbar^2}{2m}\nabla^2 + V \tag{3}$$

In QM, *perturbation theory* handles *small* changes in H, but from the pseudo-code we see that we're wholesale evolving from one H to another! (i.e. $H_i(s) \rightarrow H_p(s)$)

The adiabatic[1] theorem comes to the rescue, which states that a physical system will remain in its *instantaneous eigenstate* if the perturbation acting on it is 'slow enough' and if there is a gap between the eigenvalue and the rest of the Hamiltonian's spectrum [31].

The *instantaneous eigenstate* is defined by $\psi(t)$ and the proof starts with a ansantz: $H(t)\psi(t) = E(t)\psi(t)$. It can be shown [30, 32] that $\Psi(t)$ can be approximated:

$$\Psi(t) \approx e^{i\theta(t)}e^{i\gamma(t)}\psi(t) \tag{4}$$

by the instantaneous eigenstate if the evolution of the system is slow enough.

IV. APPLIED QUANTUM ANNEALING WITH DWAVE

The quantum computing space is flourishing with an ecosystem of both major tech companies and start-ups in both quantum hardware and software - many of which are trying to innovate in the field of drug discovery [33]. QC and QA applications are being tested and applied in genomics research, and as a complementary example, Li et al [34] casts the computational biology problem of the classification of binding affinities of transcription factors and their gene targets into one amenable to QA.

[1] Adiabatic is defined as a process by which *heat* does not enter or leave the system of interest.

A. Interdisciplinary Nature of Science: From Compounds to QA

The interface between, and cross-pollination of ideas across, disciplines is where magic happens so let's briefly summarize where we're at in the journey:

1. Cancer cells have one or more changes in DNA than make the proliferate out of control.
2. We identify targets in known pathways that, if we can change their behavior with a compound, could reduce or stop the cell from the pathological behavior.
3. To search for compounds with the desired effect we want to compare how similar they are to a reference set.
4. To define 'similar' we can model the compounds as a graph $G = (V, E)$ where nodes can be atoms or groups of atoms and edges can be atomic bonds or some other relationship between them.
5. Graph similarity measures -with flexibility- can be cast as a QUBO problem.
6. QUBOs can be cast to a physical system, as we'll see shortly, and solved using quantum annealing.

This process of casting and recasting problems within and across domains illustrates the beauty and art of the interconnections of science and hopefully the brief outline reinforces the importance of interdisciplinary skills and even ambassadors or translators between fields.

B. DWave Computers and QA Setup

DWave has a somewhat controversial history in the field because of the opinion of the academic community of their early marketing liberties taken regarding performance and technical achievements. I'm not aware of a definitive analysis showing quantum speedup, but our analysis is illustrative so we will not let that stop us. After studying Dwave capabilities, using their Ocean platform and learning from their resources my impressions are positive and they seem to be a modern and mature SaaS company for QA software. For example, I've provided a great overview series in the references for their QA YouTube videos as an introduction [35].

DWave's architecture design uses superconducting flux qubits and have shown to exhibit the properties of quantum systems such as entanglement of qubits [36]. Superconducting qubits are a form of macroscopic quantum behavior in that the system is orders of magnitudes larger than quantum objects and the two state superposition of the physical system is bidirectional current achieved via a Josephson junction that introduces the non-linearities needed to turn a superconducting circuit into a qubit [37].

DWave's processor implements the aforementioned structure of the QA Hamiltonion using these superconducting qubits with the following model of $i = 1...N$ spins:

$$H_s(s) = -\frac{1}{2}T(s)\sum_{i=1}^{N}\sigma_i^x + L(s)H_p(s) \tag{5}$$

where H_p is:

$$H_p(s) = -\sum_{i=1}^{N}h_i\sigma_i^z + \sum_{i<j}J_{i,j}\sigma_i^z\sigma_j^z \tag{6}$$

There's a lot to unpack here. Let's take them one-by-one:

1. You may immediately notice that this differs from the pseudo-code in that the system evolves in s, this is merely because in the literature you'll find that time is scaled so that $s = \frac{t}{t_f}$ where t_f is the total run time.
2. The $T(t)$ and $L(t)$ parameters here are the same as in the previous pseudo-code and represent the Transverse and Longitudinal energies of the spins. Hopefully you find my notation more intuitive than found elsewhere!
3. Notice that both H_i and H_p are comprised of all qubits. This was a turning point for me in understanding because my blush impression after first looking at this was that there was two sets of qubits and the the ground state was being transferred - somehow - from one set to another, which now of course, is nonsensical. Rather, the system of qubits are initialized in the ground state, i.e., $\otimes_{i=1}^{N}\frac{1}{\sqrt{2}}(|\uparrow\rangle + |\downarrow\rangle)$ and what we are transitioning slowly is the biases and the couplings which encode the optimization problem in H_p.
4. The first term of H_s represents the energy of the initial ground state $\otimes_{i=1}^{N}\frac{1}{\sqrt{2}}(|\uparrow\rangle + |\downarrow\rangle)$ of the system shown above. Here, the σ_i^x objects for the respective qubits are called Pauli X matrices corresponding to the energy of the initial system when initialized with a transverse (recall the notation for the $T(s)$ weight) magnetic field along the x-axis.
5. The second term, H_p is of the form originally used in statistical mechanics called an Ising Model [38]. Conventionally this models a system of magnetic dipoles that couple with their neighbors (the emergent properties and behaviors of the system from local interactions is the theme of this field).
6. Let's dissect H_p. Before, I mentioned biases and the couplings which are used to encode the optimization problem in this model. Biases are the h_i on the σ_i^z matrices (which are called Pauli Z matrices) and couplings by $J_{i,j}$. These are linear and

quadratic coefficients, respectively and will be determined by the optimization problem you're encoding.

7. Finally, the aforementioned quadratic unconstrained optimization (QUBO) problem at the end of section II is defined as min/max of $\sum_{i<=j} x_i Q_{i,j} x_j$. There are a few variants of this formulation and a tutorial for formulating these problems can be found in Glover [39]. This can be mapped to an Ising Model if you set $q = 2x - 1$ where q is the value of the qubit above and x are the binary variables from the classical QUBO, so the QUBO formulation works in binary variables (0,1) whereas Ising variables take (-1,1). Now it's easier to understand the biases h_i and couplings $J_{i,j}$ because they map to the QUBO matrices by $H_i \rightarrow Q_{i,i}$ and $J_{i,j} \rightarrow Q_{i,j}$. When thinking about these variables and relationships as networks, the H_i are node properties and the $J_{i,j}$ edge weights encoding relationships between nodes.

V. LIGAND-BASED VIRTUAL SCREENING WITH QA

The original intent of this section was the culmination of the building blocks we developed in previous sections to apply QA to ligand-based VS for graph-based molecular similarity. I reduced the scope of this work to meet the deadline, but rather than fully transitioning to a 'toy example,' I tried to keep the kernel of most of the learning and coding work in various areas, including:

1. Familiarity and manipulation of .mol files [1] with an emphasis on databases used for virtual-screening.
2. Molecular data analysis and visualization and modeling them as graphs with graph libraries.
3. Application of DWave QA based optimization, their Quantum Processing Units (QPU), the Ocean platform and API when running code locally to send/receive results back from their QPUs
4. A comparison of quantum and classical computing with respect to a graph problem. Here, instead of similarity I'm computing the Maximal Independent Set (MIS) aforementioned in section II. The comparison includes both run-time and results comparison for computing MIS with DWave vs. standard classical libraries.

To that end, I used the Directory of Useful Decoys (DUD) database [40] which includes targets and their respective sets of ligands and decoys and created a Python pipeline that ingests, analyzes and does graph computations using both DWave and classical computing methods.

DUD v2 includes 40 targets, including for example, EGFR mentioned at the end of section IA. The pipeline then prepares a final data set (based on customizable parameters) of 11,194 molecules comprised of 688,621 atoms and 721,401 bonds. The final analysis however was only on 2,218 molecules because I exceeded my free trial plan limits for QPU computation time. Explanation of the procurement and preparation of the data can be found in appendix A.

With respect to run time analysis, it's difficult to compare quantum speedup with the size of these networks and parsing the DWave run time locally because the time includes network latency of submitting the job as well as multiple layers of timing for the QPU, including: initial programming, anneal, readout and thermalization time [2]. Furthermore, annealing is typically conducted many times per run and the final solution is aggregated so this time will ultimately depend on the number of samples defined per computation. I set the *num_reads* parameter to 10 for all computations which isn't terribly high and the average additional time DWave took to complete the MIS was surprisingly only 1 second (with network and all other QA times taken into account above), with the distribution of the difference between classical and quantum computation shown in Fig. 3. Note, this shows that QA took more time, not less, than the classical approach for this problem. We expect that as the networks grow this will reverse, but conversely, QA is limited in scalability by the number of qubits because you need a qubit for each variable in your constrained optimization problem.

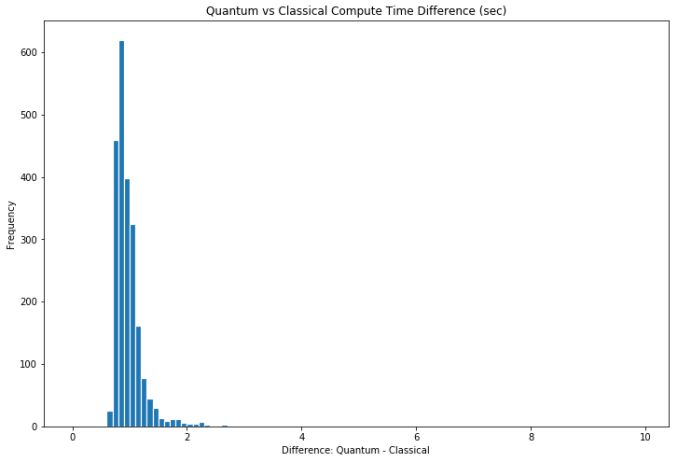

FIG. 3. Quantum vs Classical MIS Compute Time Difference

The MIS comparison results of this section are purely

[1] MOL files are plain text files that have molecular data: atom, bonds, coordinates and structural information.

[2] Information regarding these times can be found by clicking: DWave Timing

illustrative because I realized, after exceeding my maximum DWave credits, that I used a classical algorithm of *Maximal* Independent Set rather than Maximum Independent Set - the Maximum is the largest of the potential Maximal Independent Sets if more than one exists and the results of these two functions may be different if more than one maximal set exists and they are of different sizes. I'm unable to rerun the analysis, but we can compare the classical Maximal Independent Set and the DWave Maximum Independent Set. For similarity between results, I used the Jaccard (J) Similarity [3] which helps normalize across molecule sizes. The average similarity $\bar{J} = 0.56$ and the distribution of Jaccards across all molecules is provided in Fig. 4.

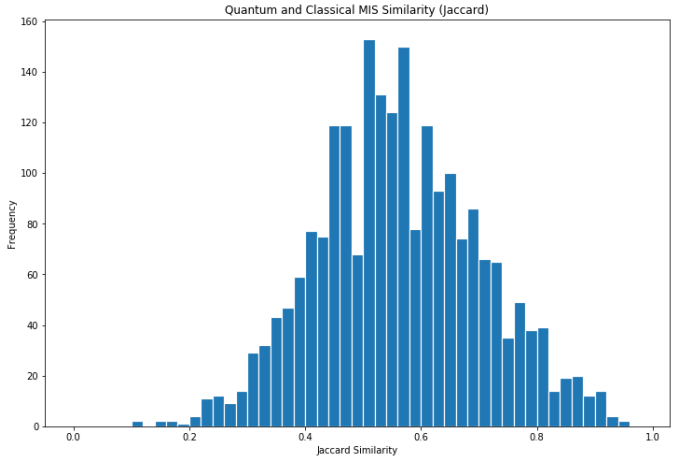

FIG. 4. Jaccard Similarity Between MIS Methods

As the distribution suggests, there is often large differences between the classical and quantum results, and as an example, ZINC00175819 is shown in Fig. 5 below with the MIS nodes encoded red.

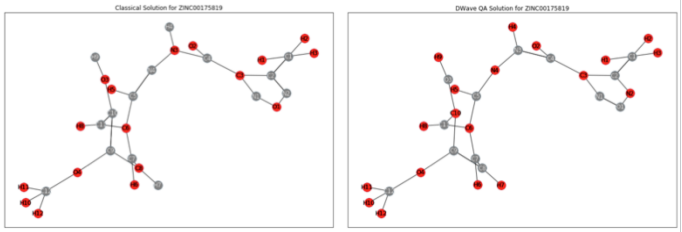

FIG. 5. Comparison of MIS Nodes On the Atomic Bond Graph Representation for ZINC00175819: Classical (Left) and DWave(Right)

[3] For sets A and B, the Jaccard coefficient $J(A,B) = |A \cap B|/|A \cup B|$. The Jaccard is 1 if the sets are equal and 0 if disjoint.

VI. DISCUSSION

At first blush, it's easy to classify this paper as an investigation of the application of quantum computing in drug discovery. Rather, I hope it contributes to reinforcing the importance and power in interdisciplinary research collaboration and the translation process observed as we journeyed from the problems in the biology of cancer, the mathematics of graph theory and optimization, to the physics of quantum objects to help us understand, harness and even one day outsmart nature.

ACKNOWLEDGMENTS

I would like to thank Dr. Boris Blinov for the great quantum computing course at the University of Washington and for putting me into an indefinite state of happy confusion!

$$|me\rangle = \alpha\,|\text{  }\rangle + \beta\,|\text{  }\rangle \tag{7}$$

Appendix A: Data Procurement and Preparation

I downloaded the complete database (which is actually a set .tar.gz files) which saves to the *./in-data* sub-folder in the repository. Then I wrote a script to programmatically unzip individual targets and their ligands/decoys as well as the .mol2 files which are prepared and easily iterable unzipped files are collected into *./out-data* subfolder in the repository. The .mol2 files (n=128) then go through a preparation script where I extract both the atom information and the atomic bonds from each file in separate data frames [4]. Although the pipeline can run on the entire set of files, I have a parameter for downsampling. For the purpose of this analysis, I cap at the .mol file level, after the algorithm exceeds n=10,000 (in this case it stopped at 11,194 molecules (which takes 32 minutes to prepare the data frames). These atom and bond data frames have molecule identifiers and they contain data on all targets, ligands and decoys which can be persisted as DB tables in the future. This preparation enables me to iterate and subset the data easily to quantum and classically compute MIS for each molecule and compile a summary data set of both the run time and consistency of the results.

[4] The biopandas package does not yet have functionality to extract the bonds so I wrote a custom function to extract them into the needed data structure.

Appendix B: Python Pipeline

The analysis pipeline has five components and a brief summary is provided here.

1. Setup: Installation of necessary Python libraries.
2. Raw Data Ingestion: programmatically unzipping all .tar.gz files and then all .mol2 files within the target sub-directories.
3. DWave and Classical Graph Library Experiments: experimentation with both DWave and graph libraries, computing MIS, and visualizing the graphs with MIS encodings.
4. Create the Full Atom and Bond Data Sets: Transforming the .mol2 text files into atom and atomic bond data frames.
5. Conduct Full DWave vs. Classical Testing on Full Data: Iterating through the atomic bonds data frame and computing MIS with both DWave and classical library, consolidating the performance results, visualizing the MIS nodes on network visualizations, and conducting the distributional analysis of the Jaccard Similarity and run times.

All of the code is available on request via Github.

[1] K. Baxter *et al.*, An end to the myth: There is no drug development pipeline, Science Translational Medicine **5**, 171cm1 (2013).

[2] P. Ripphausen, B. Nisius, L. Peltason, and J. Bajorath, Quo vadis, virtual screening? a comprehensive survey of prospective applications, Journal of Medicinal Chemistry **53**, 8461 (2010), pMID: 20929257, https://doi.org/10.1021/jm101020z.

[3] A. Gimeno, M. J. Ojeda-Montes, S. Tomás-Hernández, A. Cereto-Massagué, R. Beltrán-Debón, M. Mulero, G. Pujadas, and S. Garcia-Vallvé, The light and dark sides of virtual screening: What is there to know?, International journal of molecular sciences **20**, 1375 (2019), https://doi.org/10.3390/ijms20061375.

[4] E. Glaab, Building a virtual ligand screening pipeline using free software: a survey., Briefings in bioinformatics **17**, 352–366 (2016), https://doi.org/10.1093/bib/bbv037.

[5] X. L. Xiaoqian Lin, Xiu Li, A review on applications of computational methods in drug screening and design, Molecules (2020), https://doi.org/10.3390/molecules25061375.

[6] A. Rosales, J. Wahlers, E. Limé, *et al.*, Rapid virtual screening of enantioselective catalysts using catvs, Nature Catalysis **2**, 41–45 (2019), https://doi.org/10.1038/s41929-018-0193-3.

[7] Y. Cao, J. Romero, and A. Aspuru-Guzik, Potential of quantum computing for drug discovery, IBM Journal of Research and Development **62**, 6:1 (2018).

[8] J. Holenz and P. Stoy, Advances in lead generation, Bioorganic Medicinal Chemistry Letters **29**, 517 (2019).

[9] F. Collins, Cancer: A disease of the genome, Cancer Research **67**, PL01 (2007), https://cancerres.aacrjournals.org/content.

[10] A. Smith, A. Smith, D. Bender, B. Bender, S. Datta, O. U. Press, E. Datta, G. Smith, E. Campbell, P. Campbell, *et al.*, *Oxford Dictionary of Biochemistry and Molecular Biology* (Oxford University Press, 1997).

[11] D.-T. Nguyen, S. Mathias, C. Bologa, S. Brunak, N. Fernandez, A. Gaulton, A. Hersey, J. Holmes, L. J. Jensen, A. Karlsson, G. Liu, A. Ma'ayan, G. Mandava, S. Mani, S. Mehta, J. Overington, J. Patel, A. D. Rouillard, S. Schürer, T. Sheils, A. Simeonov, L. A. Sklar, N. Southall, O. Ursu, D. Vidovic, A. Waller, J. Yang, A. Jadhav, T. I. Oprea, and R. Guha, Pharos: Collating protein information to shed light on the druggable genome, Nucleic Acids Research **45**, D995 (2016), https://academic.oup.com/nar/article-pdf/45/D1/D995/8846748/gkw1072.pdf.

[12] C. Robert, A. Ribas, J. Wolchok, F. Hodi, O. Hamid, R. Kefford, J. Weber, A. Joshua, W. Hwu, T. Gangadhar, A. Patnaik, R. Dronca, H. Zarour, R. Joseph, P. Boasberg, B. Chmielowski, C. Mateus, M. Postow, K. Gergich, J. Elassaiss-Schaap, X. Li, R. Iannone, S. Ebbinghaus, S. Kang, and A. Daud, English (US)Anti-programmed-death-receptor-1 treatment with pembrolizumab in ipilimumab-refractory advanced melanoma: A randomised dose-comparison cohort of a phase 1 trial, The Lancet **384**, 1109 (2014).

[13] L. Marcus, S. J. Lemery, P. Keegan, and R. Pazdur, Fda approval summary: Pembrolizumab for the treatment of microsatellite instability-high solid tumors, Clinical Cancer Research **25**, 3753 (2019).

[14] F. Sanchez-Vega *et al.*, Oncogenic signaling pathways in the cancer genome atlas., Cell **173**, 321 (2018).

[15] B. Vogelstein and K. Kinzler, Cancer genes and the pathways they control, Nature Medicine **10**, 789–799 (2004), https://doi.org/10.1038/nm1087.

[16] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences **102**, 15545 (2005), https://www.pnas.org/content/102/43/15545.full.pdf.

[17] M. Ranson and S. Wardell, Gefitinib, a novel, orally administered agent for the treatment of cancer, Journal of Clinical Pharmacy and Therapeutics **29**, 95 (2004), https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2710.2004.00543.x.

[18] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Research **34**, D668 (2006), https://academic.oup.com/nar/article-pdf/34/suppl_1/D668/3924741/gkj067.pdf.

[19] G. Klopmand, Concepts and applications of molecular

similarity, by mark a. johnson and gerald m. maggiora, eds., john wiley & sons, new york, 1990, 393 pp. price: $65.00, Journal of Computational Chemistry **13**, 539 (1992).

[20] G. Maggiora, M. Vogt, D. Stumpfe, and J. Bajorath, Molecular similarity in medicinal chemistry, Journal of Medicinal Chemistry **57**, 3186 (2014), pMID: 24151987, https://doi.org/10.1021/jm401411z.

[21] P. Willett, Similarity searching using 2d structural fingerprints, in *Chemoinformatics and Computational Chemical Biology*, edited by J. Bajorath (Humana Press, Totowa, NJ, 2011) pp. 133–158.

[22] M. Hernandez, A. Zaribafiyan, M. Aramon, and M. Naghibi, A novel graph-based approach for determining molecular similarity (2016), arXiv:1601.06693 [cs.DS].

[23] M. Hernandez and M. Aramon, Enhancing quantum annealing performance for the molecular similarity problem, Quantum Information Processing **16**, 10.1007/s11128-017-1586-y (2017).

[24] M. Hernandez, G. Liang Gan, K. Linvill, C. Dukatz, J. Feng, and G. Bhisetti, A quantum-inspired method for three-dimensional ligand-based virtual screening, Journal of Chemical Information and Modeling **59**, 4475 (2019), pMID: 31625746, https://doi.org/10.1021/acs.jcim.9b00195.

[25] D.-W. Systems, How the quantum annealing process works.

[26] Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kieferová, I. D. Kivlichan, T. Menke, B. Peropadre, N. P. D. Sawaya, S. Sim, L. Veis, and A. Aspuru-Guzik, Quantum chemistry in the age of quantum computing, Chemical Reviews **119**, 10856 (2019), pMID: 31469277, https://doi.org/10.1021/acs.chemrev.8b00803.

[27] B. Bauer, S. Bravyi, M. Motta, and G. Kin-Lic Chan, Quantum algorithms for quantum chemistry and quantum materials science, Chemical Reviews **0**, null (0), pMID: 33090772, https://doi.org/10.1021/acs.chemrev.9b00829.

[28] L. Susskind, Dear qubitzers, gr=qm (2017), arXiv:1708.03040 [hep-th].

[29] J. Hidary, *Quantum Computing: An Applied Approach* (Springer, 2019).

[30] D. J. Griffiths, *Introduction to Quantum Mechanics (2nd Edition)*, 2nd ed. (Pearson Prentice Hall, 2004).

[31] M. Born and V. Fock, Proof of the adiabat theorem, Physik **51**, 165 (1928), ttps://doi.org/10.1007/BF01343193.

[32] M. OpenCourseWare, Mit 8.06 quantum physics iii, spring 2018, [YouTube video] (2019), accessed Nov. 1, 2020.

[33] A. Buvailo, 18 startups using quantum theory to accelerate drug discovery (2020).

[34] R. Li, R. D. Felice, R. Rohs, *et al.*, Quantum annealing versus classical machine learning applied to a simplified computational biology problem, npj Quantum Information **4** (2018), https://doi.org/10.1038/s41534-018-0060-8.

[35] D-Wave Systems, Quantum annealing explained (playlist), [YouTube video] (2016), accessed nov. 1, 2020.

[36] T. Lanting, A. J. Przybysz, A. Y. Smirnov, F. M. Spedalieri, M. H. Amin, A. J. Berkley, R. Harris, F. Altomare, S. Boixo, P. Bunyk, N. Dickson, C. Enderud, J. P. Hilton, E. Hoskinson, M. W. Johnson, E. Ladizinsky, N. Ladizinsky, R. Neufeld, T. Oh, I. Perminov, C. Rich, M. C. Thom, E. Tolkacheva, S. Uchaikin, A. B. Wilson, and G. Rose, Entanglement in a quantum annealing processor, Phys. Rev. X **4**, 021041 (2014).

[37] M. L. Bellac, *A Short Introduction to Quantum Information and Quantum Computation* (Cambridge University Press, USA, 2006).

[38] E. Ising, Contribution to the Theory of Ferromagnetism, Z. Phys. **31**, 253 (1925).

[39] F. W. Glover and G. A. Kochenberger, A tutorial on formulating QUBO models, CoRR **abs/1811.11538** (2018), arXiv:1811.11538.

[40] N. Huang, B. K. Shoichet, and J. J. Irwin, Benchmarking sets for molecular docking, Journal of Medicinal Chemistry **49**, 6789 (2006), pMID: 17154509, ttps://doi.org/10.1021/jm0608356.