Title: A Protein Structure Prediction and Enzyme Classifier Tool: A (PROSPECT) for the Future

Relevance Statement: Protein structure is at the core of nearly every biological phenomenon and disease and recent scientific advances have enabled an unprecedented growth in computational solutions to previously unanswered questions. Our proposal, PROSPECT, is an ambitious new tool that will allow massive parallelization of protein structure prediction (PSP) for both medical and scientific pursuits that seeks to combine both quantum computing and cutting-edge neural network (NN) frameworks**.** The success of this project could inform scientific discoveries and provide further intuition that could help solve many diseases.

Abstract: The purpose of this proposal is to propose an ambitious framework for an innovative combination of breaking NN techniques, meta-data analysis, and quantum computing in order to drive discovery and innovation in the field of PSP. Typically speaking there are a few simple ways in which the field of computational PSP evolves in its efficiency, but most salient among them are through the creation of new algorithms and methodologies that more-efficiently predict the outcome of the structures with the same computational resources. Here we propose two distinct aims. Aim 1 of this proposal will be to improve our tried and true I-TASSER model to incorporate powerful NN technologies. While the Zhang server was consistently dominant at CASP competitions, the most recent CASP showed the immense value and potential of NNs at predicting dimensional structure with limited information [1]. We intend to incorporate this value, and expand on it, by creating a series of two cooperative NNs. The first NN will be an autoencoder which will compress the structural and protein sequence data, preserving only the important features, adaptively reducing the amount of information needed to provide our second NN, as well as homogenizing the length of the input (which is a key necessity for NN function). The first NN will analyze and compress ab-initio modeling data which will be used as the basis of training for the second NN. The second NN will take train using the simplified data from the autoencoder and backpropagate results form the simplified solutions and train against the wealth of structural information present in the Protein Data Base. Aim 2 will be an ambitious attempt at projecting the future value of quantum computers and quantum algorithms in the future of PSP. While the first aim focuses on improving computational efficiency through honing NN applications, the second aim is directed harnessing an incoming technology. Recent studies have shown that there are many applications in which quantum algorithms, which have been shown to be applicable to the PSP problem [2], can be adapted to be first operable on binary-based machines [3]. We propose that the second aim be dedicated to the pursuit of these adaptations, as well as true quantum implementation as the quantum computers become increasingly powerful in the upcoming years. Through the diversified approach of these two aims, our balance in improving the our I-TASSER online system will allow us to continue, and improve upon the design which has helped hundreds of thousands of researchers over the past decade.

Research Strategy:

A. Significance: The significance of this proposal is that it combines the proven, outstanding, capabilities of NNs with an outlook on the potential of quantum algorithms for computational speedup.

PSP is one of the fastest growing fields in all of modern computation. Each year petabytes of new genomic information are being generated, thus providing an immense wealth of data to explore. With ever-increasingly large repertoires of data, as well as increasingly powerful computers there is immense value in investing in research strategies that take advantage of these exponentially growing resources. PSP is one such area of research.

The value of PSP hardly needs to be belabored. An intrinsic understanding of the relationship between one-dimensional amino acid sequence and three/four-dimensional structure cannot be understated. Protein function, and by extension protein folding, is at the centrum of nearly every biological phenomenon. Ergo, scientific progress in this field is paramount to uncovering many potential avenues for medical and scientific experimentation. As a central focus in this proposal will be NNs it is important to explore their significance.

The utility of NNs is that they are incredibly good at function approximation. They act as black boxes that, with the proper training, learn to give a certain output given a certain input, in a very computationally efficient manner. The reason they are computationally efficient is that they have the ability to only focus on aspects of given data that matter, massively increasing the speed and efficiency of computational analysis. This is one of the reasons the usage of NNs in protein prediction is growing on popularity in the field of PSP. Instead of having to calculate individual interactions between hundreds to hundreds of thousands of atoms, a NN is able to generalize what effect a group of atoms may have on a single atom, or another group of atoms, without having to compute these effects atom by atom. In the training process, it also is able to learn that some atoms will be highly unlikely to be in close proximity (bonded), and will, therefore, waste less computation.

Finally, to briefly mention the significance of quantum computers, it has already been shown that some of the very basic amino acid chains can be solved through quantum algorithms, and as time progresses, and the strength of quantum computers rises, it will become more and more prudent to solve the NP complex problem of protein folding on hardware designed to tackle those problems [2,5].

B. Innovation:

Already there have been over one-hundred thousand human genomes sequenced [5], and with even greater advances in genomic sequencing just on the horizon there is no question to the immense value in being able to efficiently and dynamically predict proper structural outcomes for individualized genomes. Similarly novel have been recent discoveries in the field of computational protein prediction. NNs like AlphaFold have recently dethroned our I-TASSER framework in the recent CASP competition, proving their tenacity and potential as cornerstones to structural prediction. Paralleling the innovation in NNs have been recent innovations in quantum computing, where the promise of quantum supremacy looms on the horizon.

Our model uses a specialized NN, referred to as an autoencoder. Typically, protein structure is interpreted with specific information on every atom involved in the structure, as well as the myriad of intermolecular interactions it may have. The autoencoder works as a data compressor reduces both storage and computational work needed for each prediction by preserving an abstract interpretation of the original input. Autoencoders can be precisely tuned to preserve a certain level of accuracy, so the potential loss in accuracy with corresponding storage saving will be known *a-priori*. This method is frequently used in many NN studies, but has yet to make its way into the lexicon of techniques explored in the field of PSP. Additionally, the shortened interpretation of the protein structure has added value of being of the same dimensions every time. This will provide an input of consistent length for the second NN, as NNs must have consistent inputs. The novelty of using the method described is that the predictions per each atom will be based not only on the atoms that are next in the sequence, but also on other atoms that may be close, even if they are very far away in terms of the amino acid sequence. This will help to account not only for the atoms that are covalently bonded, but also those that form due to hydrogen bonding and dispersion forces, and this speedup is seldom used in contemporary NN implementations for PSP.

The reaching goal of the implementation of quantum computers, and adapting quantum algorithms to a classical system is that the field of computational PSP has only rudimentary approaches to these methodologies so far. We propose the first, aggressive plan to begin the first real attempt at implementing these methodologies, carving out new territory for the field of PSP. By adopting the quantum algorithms developed to classical systems we also ensure that there is tractable progress now, rather than a lofty ideal, making concrete this objective as a significant step forward.

C. Approach:

Aim 1: Develop a workflow that utilizes, and expands upon, the recent advances in NN architecture in order to quickly and efficiently identify likely candidates for medical treatment.

The first aim of our approach is to further optimize the efficiency and accuracy of the I-TASSER modeling. This will allow for: larger protein predictions, more predictions by independent investigators, and increased prediction accuracy. To achieve this, we propose using a NN architecture for the purpose of data compression and prediction, in conjunction with training data supplied via *ab-initio* simulations.

Our proposed architecture contains a series of two collaborative NNs. The first NN is an autoencoder that reduces the complexities of *ab-initio* modeling (e.g. atomic position, and molecular interactions) into a vector of consistent size. This simplified vector will be fed into a secondary NN which will used these simplified interpretations of simulated proteins to train itself. In the end, the second NN will receive a linear sequence of amino acids and make inferred structural predictions based on its training data. These output predictions will be compared to structures which have been solved, and accuracy metrics will be drawn. This process will be iterated over until an optimized series of neural nets are created.

Furthermore, expanding on past I-TASSER data, we will utilize these neural nets to be the new gold standard of publicly-submitted protein sequences. Since our service is uniquely primed with an understanding of important contemporary protein structures, we will be able to create a queue of similarly sequenced known genes. One of the main focuses of this proposal is to optimize and enhance the scientific community's ability to innovate by proposing PSPs, and our researchers will use the meta-data gathered through analyzing the contemporarily submitted proteins to use server down-time to predict similar structures that may have medical or scientific import. By proactively computing these structures in server down-time we can optimize each dollar.

There are, however, some limitations to this nested NN approach. The first limitation is that ab-initio simulations, which will be used to create training data, are computationally expensive. This will be one of the main resource costs of this proposal, however the end product has immense value. The library of ab-initio training data could be used many times over and could provide an invaluable resource to the scientific community in the pursuit to create their own custom neural networks for PSP. Another limitation of the approach is that autoencoders intrinsically have poor predictive power toward the N and C terminus of proteins due to the linear method in which predictions are created. This, however, can be rectified when the average predicted position of all the atoms in the list is taken. A final limitation we consider is that, once trained, our NNs are static, and will interpret large and small proteins the same. This can be overcome if we batch the length of the protein structure to be predicted into categories and then have individually trains NNs for proteins of that same relative size. These disparate models should have greater fidelity, with less generalizability.

The primary method this aim will be testable and validated is through comparison of the predicted protein structures to the solved references in the PDB. These online structures will be invaluable in the generation of our NNs, and have been proven both through I-TASSER and AlphaFold to be exceptionally useful when training computational methods for computational PSP. We can see how our NNs perform when blind tested against select structures in the PDB, and these accuracy metrics will inform how to rank our NNs against common standards like AlphaFold and I-TASSER.

Aim 2: Creation of quantum-inspired classical algorithms that can offer discovery frameworks for PSP

The second aim of this proposal focuses on the utilization of quantum computing, and related algorithms, as a method for increased computational speed up. In recent studies it has been shown that many types of quantum algorithms can be adapted to be implemented in classic computers [3]. Additionally, even

prototypic quantum annealers [2] which have significantly less versatility than their true quantum computer brethren. Have already begun to solve some initial, small, proteins like Chignolin at 10 residues. These initial successes show the immense promise and lay the groundwork for the methodology of our proposal.

An important question to guide the discussion of this approach is: how is it currently proposed that quantum computers solve protein structures? Quantum computers operate by entangling qubits into an array. These qubits, unlike traditional binary bits, do not necessarily exist in 0 or 1 positions. Instead, they can exist in a *quantum* state that must be sampled multiple times. The aggregate percentage of these outcomes, which collapse either into a 0 or 1, will result in the characteristic profile of the qubit (e.g. a qubit that has 40% 0 characteristic, and 60% 1 characteristic will, average that distribution upon sampling.) The inherit stochasticity of this system is reduced by sampling many times over. The computational power of qubits becomes evident when multiple qubits are entangled or disentangled adaptively to create logic gates. Entanglement occurs by forcing two qubits to be either similar or dissimilar in nature, hence creating AND, NOR, etc. logic gates.

A central focus of this aim will be in the classical implementation of quantum algorithms to yield exponential speedup of PSP. While there is no explicit procedure in discovering new mathematical implementations of different algorithms, each is an entirely unique pursuit, we will be taking direction from contemporary work in the literature that suggests that the folding problems are NP complete.

This said, we are aware that initial attempts at creating quantum algorithms have indeed been shown to yield quantum speedup and give accurate solutions to protein folding problems. Current implementations that we will be expanding upon focus on representing the bond angles of these proteins as a series of entangled qubits. As seen in Fig 1 simplified interpretations of residues can be plotted into two or three dimensions and then have their interactions ranked.

The most direct pitfall of this aim is its boldness and its relative unfoundedness in the scientific literature. What we aim to do is created a concentrated effort in solving an unsolved problem, and hence intended implementations of solutions may be flawed in their initial approach. Additionally, the ideal full-blown quantum computers that are needed for true quantum algorithms do not yet exist, which may impinge testing speed.

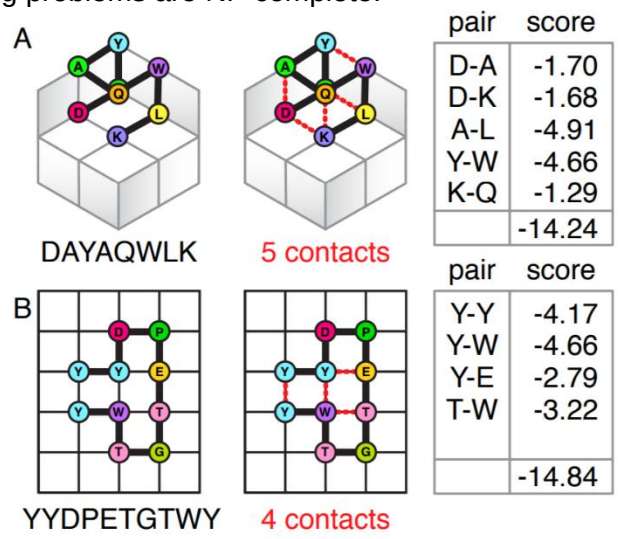

FIG. 1. Visualization of the ground state lattice folds for A) the 8 residue Trp-Cage snippet (DAYAQWLK) on a cubic lattice and B) the 10 residue Chignolin (YYDPETGTWY) on a square lattice. Both lattice folds were obtained with the D-Wave 2000Q quantum annealer. The tables on the right show the MJ interaction strengths for the interacting amino acid pairs in the depicted lattice folds and the obtained total ground state energy

Similar to the testing methods of the previous aim, the primary method for validating the experimental accuracy of these quantum algorithm will be comparison against the PDB, and, by extension, the performance of the leading classical algorithms. Ideally these quantum methods will provide models with higher accuracy, but it is important to note that the major focus of these algorithms are not necessarily in improvement of accuracy, but instead in the exponential speed-up of the protein folding solutions. Hence, we will judge the quality of our algorithms and quantum computer implementation not on the perceived increase in accuracy, but in the decrease in computational time needed to achieve the same accuracy. Ideally we can implement many of our classical algorithms into quantum algorithms, and vice versa, to see increase in speed.

References: [1] "De novo structure prediction with deep-learning based scoring R.Evans", J.Jumper, A.W.Senior **[2]** "Coarse-grained lattice protein folding on a quantum annealer Tomas Babej" and Christopher Ing **[3]** "A quantum-inspired classical algorithm for recommendation systems" Ewin Tang **[4]** "Resource-Efficient Quantum Algorithm for Protein Folding" Anton Robert, Ivano Tavernelli **[5]** "The Protein Data Bank" H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne