Project Description
The UMN-Mayo Computational Human Immuno Peptidome (CHIP) Center

1 The Grand Challenge

Perhaps more so than for any other discipline of science or engineering, the history of computer science has been one in which problems seem impossible until suddenly they are not. Translating languages, playing chess, recognizing objects, driving cars – all at first seemed so daunting that experts gave these as examples of tasks that computers might *never* be able to do as well as humans. Of course, none of these problems were solved "suddenly" with a single keystroke. In all cases, it was decades worth of concerted research, down blind alleys and with major paradigm shifts, that brought solutions. Raw computing power alone did not solve these problem; however, the tremendous increase in computing power that preceded these breakthroughs enabled them.

This proposal discusses a problem that computer science currently judges to be very difficult. It is a foundational problem in computational immunology which, if solved, could inform predictions of disease severity, enable treatments, and guide vaccine development. While such aspects of a pandemic response ultimately depend on experimental knowledge and trials, the computational results that we are proposing could play a critical role at two ends of the time spectrum:

- In the early phases, when identifying and characterizing the threat a new pathogen.
- In the long term, to develop a deep understanding of the molecular mechanisms of the infection.

Narrow Statement of Grand Challenge Problem

Stated succinctly, the computational problem that we will tackle is determining how strongly a given molecule binds to another. The given molecule is a peptide – a fragment of a protein derived from a pathogen, such as a virus. The other is a molecule called Major Histocompatibility Class I (MHC I) that is expressed on the surface of most of our cells. MHC I molecules have a cleft into which a peptide can bind. As illustrated in Figure 1, a peptide will only bind if it fits into the cleft like a key into a lock. The binding of a peptide to an MHC I molecule is a critical step in a critical component of the immune system, so-called *cellular immunity*.

Explained succinctly, cellular immunity allows circulating T-cells to kill off infected cells. When a cell is infected with a virus, it hijacks the host cell's machinery, forcing it to make viral proteins. Our cells have a defense mechanism: they chop up such proteins into fragments, called peptides, and transport them to the cell surface, bound to MHC I molecules. Presented this way on the cell surface, T-cells can identify a cell as being infected and can destroy it using toxins. If this mechanism succeeds, an infection is stopped in its tracks: T-cells kill off infected cells before they can do damage. If it fails, then infected cells become factories for reproducing copies of the virus and full-blown disease results.

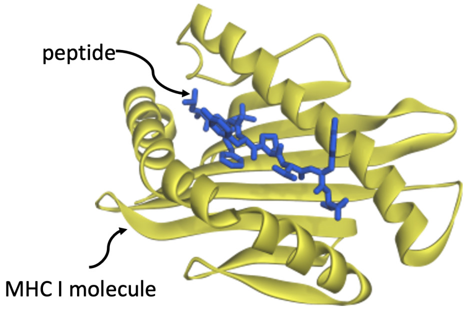

Figure 1: A peptide binds to an MHC I molecule if it fits into the cleft, like a key into a lock. There are perhaps 38,000 distinct peptides derived from proteins for a virus such as SARS-Cov-2. An individual has up to 6 variants of MHC I molecules. There are perhaps 21,000 variants of MHC I molecules in the human population.

Success or failure depends on whether peptides derived from the viral proteins bind to the MHC I molecules and so become targets. Binding depends on the biochemical affinity between the constituent building blocks of the pair of molecules. This, in turn, largely depends on molecular shape:

how well the metaphorical key fits into the metaphorical lock. There are many variants of MHC I molecules, coded for by a person's genes. These vary, sometimes subtly, in shape.

The set of all the peptides that can bind to person's MHC I molecules is called their *immunopeptidome*. This set is unique and determines the capacity of their immune system. Since the immune response of a person to, for instance, a viral infection like COVID-19 is dependent on whether their MHC I molecules present peptides derived from the virus, understanding and predicting the binding step is an important topic.

What we are proposing here appears to be a narrowly-defined problem: characterizing the binding strength between specific pairs of molecules. Most aspects of the molecular biology are well understood. And yet, we argue that the problem qualifies as a grand challenge, in terms of its difficulty and in terms of the impact of a solution. The difficulty lies with the *computational* requirements, stemming from the *combinatorial scale* of the problem. The impact of the solution will stem from an ability to precisely characterize, in advance and through purely computational means, how well a person's cellular immunity will cope with a novel pathogen.

Simulating molecular interactions has been a widely-studied and largely successful topic for perhaps five decades. Indeed, some of the earliest computers were applied to this problem [26]. Sophisticated software exists to simulate molecular binding events [7, 27]. The conventional approach with such tools is to simulate binding from first-principles, tracking the trajectories of all the atoms in all the molecules in three-dimensional space, numerically solving Newton's equations of motion. A variety of strategies are used to find low-energy configurations, including randomization, with so-called Monte Carlo methods. The problem is that simulating a *single* peptide-MHC I molecular binding takes *hours*, or even *days*, on a powerful computing cluster with this approach.

The SARS-CoV-2 virus, for example, has 29 distinct viral proteins. When chopped up, this translates into approximately 38,000 peptides. This is not an unmanageable number. However, the other side of the equation consists of the MHC I molecules. Every person has up to 6 variants, having inherited 3 from each parent. There are at least 21,000 variants in the human population [20]. Indeed, the genes that code for MHC I molecules are the most diverse in our genome. Evolution has ensured this, as humans and pathogens have been co-evolving together.

So, in the narrow formulation of the research problem, there are 38,000 peptides for a virus like SARS-Cov-2 each paired with 21,000 variants of MHC I molecules in the human population. This translates to three-quarters of a *billion* distinct pairings. If one is using existing simulation software, which requires hours or even days of computing time *per pairing*, one is confronted with billions of hours, or billions of days, of computing time – clearly an intractable proposition.

With this grant, we will develop new, highly targeted algorithms to make such computation tractable. While existing software for these sorts of atomic simulations is sophisticated, it is general-purpose, written in FORTRAN decades ago [7]. The most widely-used software packages have been written to simulate molecular interactions of nearly any type, from crystals, to proteins, to polymer chemistry [27]. Others specifically simulate protein binding [8, 12, 32]. Observing simulation trials for such general-purpose software, most of the computational time is spent moving molecules randomly in space, looking for energetically-favorable states; most of the random movement is wasted for this particular problem. Only the final steps, as the peptide settles into an optimal configuration in the cleft of the MHC I molecule, matter. There is domain-specific knowledge here than can really help.

Pilot Project

A pilot project for this Phase I Development Grant is to develop modern, efficient, custom software for the specific problem of peptide-MHC I binding. With heuristics, and with a hierarchical approach, we believe that we can turn a billion hours of supercomputing time into one million minutes. In Phase II, as we assemble a Center, we will deploy the software at scale with cloud computing resources and on a custom-built cluster of graphical-processing units (GPUs). We will seek industry support for this. (We have had conversations with both Google and Amazon Web Services about the possibility of them donating cloud-computing time, and they have been receptive.)

Broader Statement of Grand Challenge Problem

The grand challenge that were are positing is much broader than simply writing better algorithms and deploying the code on supercomputing clusters, although this will be a significant aspect of the effort. The problem statement is broader in two main respects:

1. Not all aspects of the biochemistry of binding are understood or have been well characharacterized from a computational perspective.

2. Structural models do not exist for novel peptides, nor for most variants of MHC I molecules that one encounters in the human population

Needless to say, the biochemistry of the immune system is a complex and vast topic of study by a large community of experimentalists. We can point to the work of Prof. Mark Davis at Stanford. His lab has been striving to understand the structural and biochemical underpinnings of peptide binding for decades [1].

Our focus is applied, translational, and computational. We aim to incorporate knowledge from structural and molecular biology into efficient computational models and deliver useful predictions, at scale. The goal of this development grant – and that of the eventual CHIP center – is not to develop new, experimental or biochemical knowledge of the immune system, but rather to synthesize and apply knowledge as it evolves.

We have relevant expertise from the participants at University of Minnesota, the Mayo Clinic and Iowa State University. We point to collaborative work that we have recently done on an important aspect of the binding problem, hydrophobicity, as an example [28]. Hydrophobicity plays an important role in peptide:MHC I binding, yet has not been explicitly considered in computational models. We have shown how to incorporate it and how this improves binding predictions.

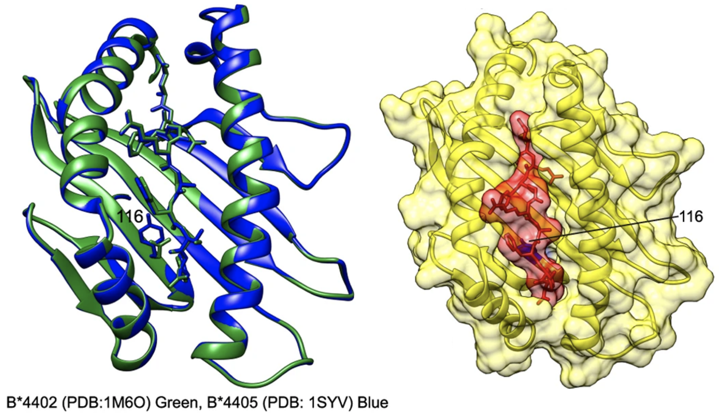

Figure 2: Inferring the structure of an MHC I molecule by homology and through protein-folding. **On Left**: Comparison of MHC I molecule B*4402 (in green) and MHC I molecule B*4405 (in blue). The root-mean squared distanced between these two structures is 0.3 Å. The structure of the green molecule was inferred by homology from the structure of the blue molecule.
On Right: Combined ribbon and surface representation of the green MHC I molecule, after folding, with the peptide 116 from SARS-Cov-2 bound in the cleft.

Beyond detailed and accurate biochemical models of binding, a significant challenge for this research is the availability of models. On the one hand, viral proteins are readily characterized. In the case of SARS-Cov-2, the virus was first identified in Dec. 2019, By February 2020, the amino acid sequence of its proteins had already been published [31]. Viral proteins get chopped into fragments called peptides by intracellular mechanisms, each 8 to 15 amino acids in length. Given a novel pathogen, most of these peptides will be new to science, having never been encountered before. (Similarly, such peptides will be new to an individual's immune system!)

The first challenge is to construct structural models for novel peptides. We discuss this in Section 2.3. The second, more significant challenge is to construct structural models for MHC I molecules. As noted above, there is tremendous diversity in these molecules, with perhaps 21,000 variants in the human population. Only a small fraction of these have been characterized experimentally. Doing so entails significant effort, with experimental techniques such as crystallography and mass spectrometry. As we discuss in Section 5, this has led to an unfortunate social consequence: nearly all of the variants of MHC I molecules that have been characterized experimentally are those found predominately in people of European descent. Addressing this inequity will be a significant focus of our Broader Impacts activities.

This project will pursue a novel strategy for constructing structural models of variants of MHC I molecules that are not available. Instead of waiting for them to be characterized experimentally, we will construct the models *de novo*. We will do so first by *homology*: beginning with a structural model for an MHC I molecule that it most closely resembles, we will construct a model for a new MHC I molecule by substituting amino acids. This is illustrated in Figure 2. Here we will apply domain-specific expertise, provided by Prof. James Cornette from Iowa State University. (He is not a PI on this development grant, but he is a close collaborator. He will join the team of an eventual center in Phase II. Please see "Other Personnel" for his biography.) While all MHC I variants have different shapes, their overall structure is similar; differences are primarily in the location of side chains. These structural differences can be inferred.

Given a *de novo* model of an MHC I molecule that is accurate in terms of its atomic configuration, we will apply computing power to fold it into its actual shape. Here we will make use of the latest breakthrough, AlphaFold, an AI-based solution recently announced by Google's Deepmind project [12].

Goals of the UMN-Mayo Computational Human Immuno Peptidome (CHIP) Center

Perhaps the most ambitious aspect of this project, and the aspect that qualifies it as a Grand Challenge, is translating the computational modelling into practice. Of course, we emphasize that the computational challenge is significant; it will require the deployment of supercomputing power in Phase II to be realized. Characterizing the immuno-peptidome will not be a separate activity from applications. Rather, there will be tight synergy between the computing team and the practitioners. Here we point to some of the expected applications: what will be possible if the CHIP Center delivers accurate binding predictions for peptides of novel pathogens? Details of the anticipated organizational structure of the CHIP center are provided in Section 3.

- ***Predict disease severity for a new pathogen for different individuals*.**

  Applications: Early Response & Triaging. Translating the computing results to practice will generally entail a focus on the individual. As explained above, every person inherits up to 6 distinct variants of MHC I molecules, three from each parent. A form of genetic testing called human leukocyte antigen (HLA) typing can be performed to establish which variants a person has. This type of testing is convenient, widely available, and inexpensive, as it used for paternity testing. Performing such tests on a large group, say everyone at risk in a pandemic, is feasible.

  With population-wide typing, our computational tools could predict which individuals are most likely to mount a strong antiviral immune response to a novel pathogen, given their MHC I variants; these individuals would be at the lowest risk for severe disease. Conversely, our tools could predict which individuals are least likely to mount a strong antiviral response; these individuals would be at the highest risk.

  Consider the implications for early response and biodefense. With the computational ability that we will deliver, when faced with a novel biological threat, an early-response team could predict which of its personnel are likely to have immunity and which might be most vulnerable. This could be assessed based on prior HLA typing of the personnel. With the computing ability described above, all that would be required is a proteome profile of the novel virus or bacteria. Such profiles are usually easy to obtain, often available within weeks when a new pathogen is identified.

- ***Disease severity for different variants of a virus for different individuals.***

  Applications: Resource Allocation & Population Monitoring. As we have seen with SARS-Cov-2, viral mutations are perhaps the single greatest confounding factor to a pandemic response. The more widespread a pandemic, the more hosts a virus infects. With more hosts, there are more opportunities for it to mutate. Given the extent of the COVID19 pandemic, some virologists have hypothesized that nearly all mutations favorable to its spread will be discovered by the virus before the pandemic abates [24].

  Mutations confound a response because vaccines and treatments may be less effective against new variants. Here our prediction tools could transform both planning and resource allocation. Once the protein sequence of a new variant is identified, the differences from the original strain can be analyzed. Differences in the proteins expressed will translate to a different set of peptides. With population-wide HLA typing, a distribution of MHC I variants can be constructed for sub-groups – perhaps different demographic groups in different geographic regions, or perhaps even a fine-grained map tagging all individuals in the group with their specific MHC I variants. These MHC variants can be paired up with the novel peptides from the viral variant to assess the risk of severe disease for the individual. If the analysis is done at the level of a group, then a statistical analysis of the risk can be performed against the distribution of MHC I molecules in the group.

- ***Effectiveness of different vaccines for different variants of a virus for different individuals.***

  Applications: Tailoring Vaccines to Individuals. It is likely that future historians will point to this pandemic as an inflection point for society, not due to the damage that was inflicted, as great as this as been, but due to the progress made in science as a consequence of it. The development of mRNA vaccines, in particular, is a startling success story [22]. They have been deployed in record time, and on an unprecedented scale. Significantly, mRNA vaccines are readily customizable. Once the infrastructure is developed, different mRNA vaccines could be administered to different groups at different times, with little extra production cost; all that is required is swapping out the RNA sequence in the vaccine doses.

  This flexibility offers the possibility to administer mRNA vaccines that elicit the best immune response for each individual, in response to the specific viral variants that pose a threat at that time. Recall that mRNA codes for proteins, such as the infamous spike protein of the SARS-Cov-2 virus. Dividing proteins into units 8 to 15 amino acids long yields the requisite peptides to target in mRNA vaccine production. So our computational tools will allow screening of the peptides of viral variants, matching of those against MHC I molecular variants, and then choosing which peptides to target in the vaccine production.

2 Research Agenda

2.1 Background

There are many steps in the activation of the mechanisms of cellular immunity. Firstly, intracellular proteins are degraded by the proteasome, which breaks them into peptides – protein fragments that are

8 to 15 amino acids in length. These peptides are transported to sites where they can bind to MHC I molecules. MHC I molecules select and present a limited set of peptides from a broad repertoire. How MHC I molecules make this selection is unclear. Evidence suggest the they initially bind many peptides because of highly flexible binding pockets. Peptide binding is followed by a selection step wherein a large fraction of these peptides are released. Only peptides with very high binding affinity remain bound [9]. Once tightly bound, a peptide-MHC I complex is transported to the surface of the cell where it is presented to killer CD8+ T-cells. Of all these steps, the binding of the peptide to the MHC I molecule is the most selective and consequently, of highest interest in predicting cellular immunity. It is this step that will be the focus of our computational efforts.

MHC I molecules are proteins that contain a groove on their surfaces that is flanked by two binding pockets, as shown in Figure 3. The ends of a peptide bind in the pockets; the rest of it fits neatly in the groove or else bulges out slightly above the surface. MHC I molecules are, in fact, complex folded structures; their binding pockets are formed by various amino acid residues, many of which would not be adjacent if the molecule were stretched out into a linear conformation. The composition of these residues determines which amino acids from a peptide can binding in that pocket. A variety of biochemical factors play a role in deciding whether binding is energetically favored. For example, the binding pockets of the MHC 1 molecule labelled HLA-A*02:01 favor hydrophobic amino acids, leading to a preference for amino acids such as Leucine, Methionine, Isoleucine, and Valine at the ends. In contrast, the MHC 1 molecule labeled HLA-B*40:01 favors acidic residues such as Glutamic Acid in one of its pockets.

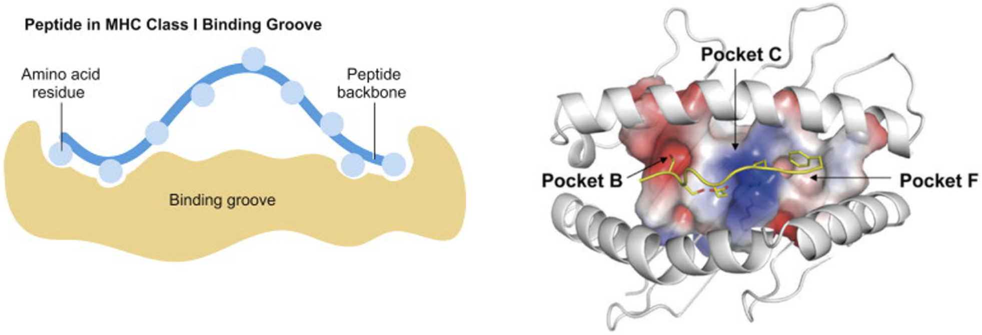

Figure 3: **On Left**: Illustration of the binding pockets of MHC I molecules. Generally there are two binding pockets with a groove in between them. **On Right**: some MHC I molecules have more complicated topologies, with multiple binding pockets. Evolution has ensured diversity.

Let us assume a sequence length of 275 amino acids per MHC I molecule [2] and a binding peptide consisting of 9 amino acids. Let us also assume each amino acid has a mean of 19 atoms [3]. This equates to 5,396 atoms per peptide-MHC I complex. Now, each atom has 3 spatial coordinates equating to 16,188 distinct coordinate values per peptide:MHC I complex. Even though most aspects of the physical chemistry are understood, tracking all these coordinate values through space in order to search for a minimum energy condition is a taxing problem, even when programmed on a modern computer.

2.2 Existing Tools

To our knowledge, no one has attempted to simulate peptide binding at the level of physical chemistry, at scale, pairings tens of thousands of peptides with tens of thousands of variants of MHC I molecules. Rather, people have turned to neural networks.

We have investigate the use of these tools, through a preliminary "EAGER" grant from the NSF.[1]

[1] #2036064 Computationally Predicting and Characterizing the Immune Response to Viral Infections

We point to three packages that perform exactly the predictions that we are discussing: NetMHC [4], PickPocket [35], and SYFPETHI [17]. These tools have been used to study cancer immune escape mechanisms [19], checkpoint blockade immunotherapy for tumors [16], and identifying T-cell response targets [10]. All are efficient, returning binding predictions in a matter of seconds for queries. So, in might seem that the grand challenge that we are positing has already been solved.

Unfortunately, it has not. These tools are trained with one-dimensional data: text *labels* for MHC I molecules, paired with amino acid *letter* sequences of peptides, scored according to the experimentally observed binding strength. So these neural networks are trained on *textual* data, as illustrated in Figure 4. The network predicts how strongly a novel peptide will bind to a given MHC I molecule, according to the similarity of the amino acid sequence only. No information regarding molecular shape or binding chemistry is used.

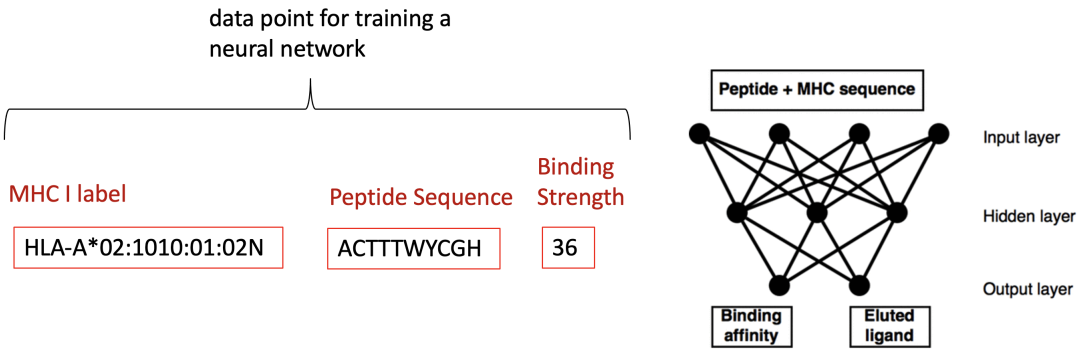

Figure 4: Neural Networks for Binding Prediction. These are trained on data points such as the one on illustrated here. It consists of a *label* for an MHC I molecule, paired with an amino acid *letter sequence* for the peptide, scored accorded to an experimentally observed binding strength. The network, illustrated on the right, predicts how strongly a novel peptide will bind to a given MHC I molecule, according to the similarity of the amino acid letter sequences.

These tools are valuable; we have used them extensively in our research [28]. However, the predictions that they provide are coarse:

1. The neural networks are trained without any data on molecular shape, and without any reference to the underlying physical chemistry. The inferences that they provide are based on peptide amino acid sequence only. However, the peptides have three-dimensional shapes. Small differences in amino acid sequence can translate to very different shapes and very different binding affinities.

2. The neural networks are trained on experimental data that comes from a wide variety of domains. For instance, a large fraction of the data for NetMHC comes from studies of proteins derived from the HIV virus [4]. However, peptides from a novel pathogen such as SARS-Cov-2 might bear little similarity to these. Most will be new to science. Neural networks perform statistical inference, interpolating to produce answers. Inferring from data that is too dissimilar from the target generally yields poor results.

Indeed, acknowledgement that neural networks provide **spurious inferences** for peptide binding strength is widely acknowledged. In particular, the tools seem to deliver many **false positives**. It is possible that machine learning and neural networks are the right way to attack this problem. (We are considering such techniques in our approach.) However, training on the letter sequence of amino acids simply cannot provide reliable answers to complex questions pertaining to physical chemistry, such as this one. One must incorporate molecular shape and biochemical aspects of binding into the modelling.

2.3 Technical Approach

As the Pilot Project in this Phase I Development grant, we will develop and apply a new **mechanistic model** for predicting peptide-MHC I binding. In contrast to general-purpose software for molecular simulations, such as CHARMM [7] and Amber [27], ours will be specifically optimized for this problem.

The starting point is a three-dimensional molecular model of both the peptide and the MHC I molecule. We will make use of structural models of MHC I variants that have been characterized experimentally, through crystallography. When such a structural model is not available, we will infer it by homology: we will construct it starting from the structural model that it most closely resembles, substituting structural models of amino acids where it differs. Then we will make use of existing software for simulating protein folding [12]. So the starting point for our calculations will be optimal folded structures for the MHC I molecules. We will follow the same strategy for generating structural models for the peptides. (These are much shorter, with fairly simple molecular structures, so this is a much easier task.)

Next, in each simulation, we will position the peptide roughly aligned in the binding pockets. This is likely the most difficult step, as the optimal position of the peptide might not be known. Here, we will endeavor to incorporate as much domain-specific knowledge as possible. We anticipate that intelligent placement of the peptide is the single most important factor in reducing computation time.

For binding strength, we will implement energy calculations based a variety of biochemical factors: electrostatic interaction, acidic/basic pH, hydrogen bonds, van der Waals forces, shape complementarity, hydrophobicity, π-interaction, steric effects, and solvation energy.

Finally, we will perform a rigorous search for a minimum-energy binding configuration. Compared to existing methods in general-purpose software, we will reduce the dimensional space by fixing bond lengths and amino acids sidechains in advance. Thus the only variables that we will manipulate will be the dihedral angles along the backbones, as well as the dihedral angles of the amino acid sidechains. If these moves are not sufficient, we will also flex the backbone of the MHC I molecule.

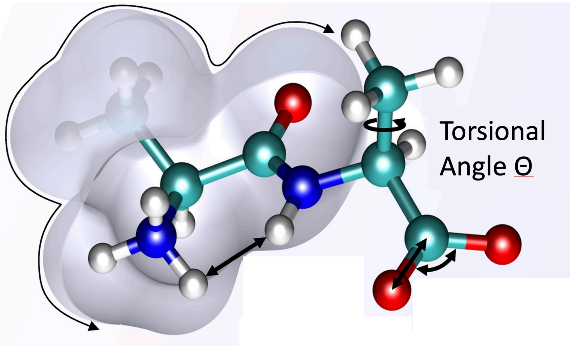

Figure 5: We will explore moves in the torsional space to find minimum-energy configurations. Such moves are much more efficient than moves in the general cartesian space.

Instead of carrying out the search in a space with cartesian coordinates, we will do most of the molecular maneuvering in the **torsional space**. That is to say, we will rotate sidechains instead of randomly displacing and flexing them. Moving in the cartesian space necessitates tracking 3 x 19 = 57 variables per amino acid; however, moving in the torsional space necessitates tracking only 3 dihedral angles per amino acid. Furthermore, we can restrict moves to just the residues present in the binding pockets of the MHC I molecule, reducing the 275 amino acids per MHC I molecule to no more than 70.

We can only justify this claim by developing and deploying the code, but we anticipate that by devising efficient, custom algorithms, we can turn one billion days of simulation time into one million minutes for our grand challenge problem, where we are looking at SARS-Cov-2 binding predictions, with 38,000 peptides paired with 21,000 MHC I variants.

2.4 Applying the Computational Toolset

As a follow up Pilot Project, using the data and algorithms discussed in the prior sections, we will identify commonly occurring haplotypes in the U.S. population that may make individuals vulnerable to COVID-19. We will apply the algorithms that we have developed to predict viral peptide binding to variants of MHC I molecules across populations. We will reduce the combinatorial space by concentrating on the low-mutational rate regions of the viral proteins as well as the most frequent MHC I alleles found in a given population. If we set cutoff thresholds, the numbers become much more manageable. For example, there are only 374 variants of MHC I molecules that occur with a frequency above one-in-one-thousand in the U.S. population; only 77 are above 2% frequency in the U.S. population. These are considerably smaller sets than the 21,000 possible variants discussed above.

We acknowledge here that this specific strategy seems to run contrary to the concern discussed in our Broader Impacts section, namely that genetic research in this and other fields has focused too much on populations that are primarily of European descent. When working with genetic data from the U.S. population, we will always prioritize analyzes of data from underrepresented groups: African Americans, Hispanics, and Native Americans.

Using the algorithms developed, we will identify the variants of MHC I molecules that commonly occur in U.S. populations, as reported by [17]. We will then identify those that may confer more, and those that may confer less, protection that average against COVID-19. According to our hypothesis, individuals with variants of MHC I molecules to which some viral peptides bind strongly are likely to mount an effective immune response. Those with variants to which all viral peptides bind only weakly are not. With this data, we can make predictions across sample populations. These will be validated against any clinical data that is available. (No new clinical data will be generated.)

3 Center Development

3.1 Multi-Organizational Team

The proposed research is targeted and focused on delivering outcomes that will aid in pandemic preparedness. However, our approach is ambitious, aiming to develop a very general capability in computational immunology. It builds upon expertise and prior research in the Vasmatzis lab in genomics and computational techniques for cancer immunotherapy; work in the the Block lab on patient immune response monitoring for SARS-CoV-2; as well as expertise and prior research by Julia Udell in graft-versus-host predictions at "Be-the-Match."

A premise is the deployment of high-performance computing. Prof Riedel's experience with circuit design and molecular dynamics will be brought to bear on this aspect. This project will provide the computational infrastructure to predict whether peptides derived from viral proteins will bind to allelic variants of MHC I molecules. The same computational infrastructure could be deployed for other pathogens. It could also be transformative in other contexts, for instance for treatments of cancer via immunotherapy as well as for the treatment of autoimmune diseases. Capitalizing on our past experience in modeling MHC-peptide complexes [15, 30, 34]; monitoring of immune responses to peptide vaccines [5, 6, 13, 14, 33]; our ability to develop fast computational algorithms and pipelines [11, 18, 29]; and our access to clinical data, we are poised to develop transformative tools and deliver critical information for pandemic preparedness, adhering to a tight timeline.

We have assembled a strong multidisciplinary team that, combined, has both breadth and depth of expertise in computational biology, genomics/bioinformatics, immunology, graphics, computer science and high-performance computing.

- **Dr. Vasmatzis** from the Mayo Clinic and **Prof. Jim Cornette**[2] from Iowa State University have been involved in pioneering work related to peptide-MHC binding predictions [23], TCR and antibody structure prediction [25], and modeling statistical potentials of atomic interactions [26]. The Vasmatzis lab at Mayo Clinic also has expertise in developing highly sophisticated genomics pipelines. For example, the group has developed a pipeline that allows the accurate determination of tumor-specific neoantigens based on tumor-specific DNA junctions that are often the source of neopeptides in tumors. Their technique, called MPseq, detects with high sensitivity, specificity and cost-effectiveness many complex rearrangement events such as chromoplexis and chromothripsis. These truncate highly expressed genes and result in altered protein sequence juxtaposed on normal truncated proteins. Their group at Mayo has used this technique for mesothelioma cases that exhibit a higher potential of rearrangements to produce neoantigens compared to single nucleotide variants. The results were reported in JTO [34]. Also, the group at the Mayo Clinic has world-class expertise in genomics and sequencing, as well as access to patient data sets needed to validate the computational results.

- **Professor Riedel** from the University of Minnesota has extensive experience with molecular computing that can be brought to bear on the research. Funded by seven major NSF grants, he has spearheaded the development of novel computing constructs with DNA. He currently has a DARPA grant to develop DNA storage systems. This research is predicated on algorithmic expertise in molecular simulation and design, adapted by Prof. Riedel from the realm of electronic circuit design. The circuit-design community has unique expertise that can be brought to bear on the challenging computational problems encountered in molecular simulation. Applications in molecular biology, in turn, offer a wealth of interesting problems in modeling and algorithmic development. With its cross-disciplinary emphasis, this project will bring new perspectives to both fields. This project will leverage the infrastructural support of the Minnesota Supercomputing Institute, applying advanced algorithmic techniques such as stochastic simulation and machine learning.

- Prof. Riedel and Dr. Vasmatzis co-advise **Julia Udell**, a Ph.D. student in the Bioinformatics and Computational Biology program at the University of Minnesota, who will play a significant role in this project. She has both computational and immunological expertise, having worked as a biostatistician for Stanford's HLA typing lab prior to beginning her doctorate. She is the author of the neoantigen-ranking algorithm validated in the JTO study, and will take the lead in applying this algorithm to the SARS-CoV-2 proteome.

- **Dr. Matthew Block**[3] is an immunologist and a medical oncologist with an interest in understanding the mechanisms that influence anti-tumor immunity in patients with melanoma and ovarian cancer. His research efforts have focused on preclinical translational studies and therapeutic clinical trials testing novel cancer vaccines and immunotherapies. As part of his efforts to identify novel immunotherapy approaches to cancer, his laboratory has developed T cell and antibody-based assays to measure changes in antigen-specific immune responses using patient samples. With the onset of the COVID-19 pandemic, he has developed methods to measure immune responses to SARS-CoV-2.

Our "grand challenge" problem is ambitious both from the standpoint of computation as well as in terms of the requisite expertise in immunology. Our team is uniquely positioned for this challenge: the researchers at Mayo have expertise in molecular modelling and immunology, as well as access to clinical

[2] Prof. Jim Cornette is not on this Phase I Development Grant. However, he is a close collaborator. He will be a part of an eventual Phase II Center. See "Other Personnel."

[3] The same is true for Dr. Matt Block. He too will part of an eventual Phase II Center.

data; the researchers at UMN have complementary strengths in molecular simulation, as well as experience with developing and deploying large-scale computation projects.

3.2 Scaling to a Center

While we have emphasized the pilot projects and the conceptual challenges in this proposal, much of the effort funded by this Phase I grant will be in assembling a full proposal for a center: the UMN-Mayo Computational Human Immuno Peptidome (CHIP) Center. The center will be physical, occupying offices on both campus, with computing equipment on the University of Minnesota Campus, and experimental equipment at Mayo. It will also have a virtual setup, as researchers from both campuses interact without the 90 minute drive between Minneapolis and Rochester, Minnesota.

Given the combinatorics and scale of the problem, we will employ modern software engineering techniques for this effort, with rapid prototyping of algorithmic changes. We will run simulations first on graphical processing units (GPUs) on local computers. Once the algorithms are validated, we will deploy them at scale on supercomputing clusters or on cloud-computing resources. As part of the center development, we will seek support from industry. We have had discussions with both Google and Amazon Web Services. Both companies have expressed interest in donating cloud computing time to the project. (However, we were unable to secure a formal commitment that we can include with this proposal. Our Phase II proposal is contingent on this.)

We have institutional support to hire a *Project Coordinator* for a Phase II Center proposal. This person would become the *Center Administrator* if we receive a Phase II Center grant. His/her responsibilities will include: (1) handling administrative logistics with regard to sub-grants and contracts; (2) outreach to industry and work on tech transfer details with our technology innovation offices; (3) scheduling and logistics for large-scale meetings; (4) managing the participants for undergraduate research experiences; (5) publicizing our activities; and (6) working with external evaluators regarding the effectiveness of our Broader Impacts and Broader Participation in Computing efforts.

Both PIs are Graduate Faculty in the joint **UMN/Mayo Biomedical Informatics and Computational Biology** program. Building on the success of twice-yearly workshops through this program, the UMN-Mayo CHIP Center will hold workshops, inviting a diverse set of both internal and external participants. The students participating in our graduate programs as well as those on undergraduate internships are natural candidates for these workshops. In addition to scientific topics, these workshops will offer training in areas such as inclusive mentoring skills, conflict management, and proposal preparation. This training will be offered by the senior personnel; we will also solicit external keynote speakers, not all of them speaking on technical subjects. An important goal of the workshop is to provide training opportunities for under-represented students.

4 Risk and Mitigation Plan

The "grand challenge" problem that we are proposing is ambitious. The reviewer might ask how the problem of predicting the binding strength of peptides compares to the more general problem of protein folding. Indeed we make use of protein folding software in our approach. However, protein folding is still, to an extent, an *unsolved* problem – perhaps the most notorious unsolved problem of this type. Early discoveries in proteomics started the race to develop methods for determining the three dimensional structure of proteins, but the road to get there proved far more difficult than scientists initially imagined. The combinatorial space is simply enormous. Only with painstakingly crafted heuristics and, more recently, with sophisticated artificial intelligence has there been real progress [12].

Our problem entails the folding of proteins to obtain the initial conformations of our molecules, and then entails detailed binding predictions. Simulating physical chemistry at this level itself qualifies as a "grand challenge." As we have emphasized, existing software tools require inordinate computing time; they simply cannot be scaled to provide the millions of binding predictions that we require, each for large, complex proteins.

Accordingly, the reviewer might ask if we are biting off more than we can chew, even for a Research Center proposal. We are confident that, while the research is ambitious, it is doable. The key difference is that we are tackling a very specific molecular interaction: we are simulating the binding of peptides to MHC I molecules. Peptides are short, consisting of simple protein fragments without complex structure. MHC I molecules are large with very complex structure, but the structure is not all that dissimilar between variants. To construct an MHC I molecule from its amino acid sequence, one can start with a similar molecule as a template, already folded into the correct shape, and make simple substitutions of side chains and other structures. Folding the changes into an optimal configuration requires computing power, but not on the scale of predicting a three-dimensional structure from scratch.

In order to characterize the immunopeptidome, we will develop a hierarchical simulation framework, performing coarse estimates to eliminate many of the peptide candidates early, and then deploying heuristics to perform more accurate estimates of binding strength on the most promising candidates. This approach might not succeed. We might either get too few peptide candidates – so false negatives – or too many peptide candidates – so false positives. If so, we will adapt the search.

To reduce the false positives, we will first return to the preliminary results of the neural network tools. We will refine and customize the training data, selecting from the largest possible pool of relevant peptides, but not over-training. We will seek to smooth and normalize the training samples. To reduce false negatives, we will refine the biochemical models, deploying a wide variety of heuristics. To make the computation of binding energies more efficient, we will only explore moves in the torsional space, as opposed to the three-dimensional cartesian space. We will try constraining moves to the space obtained by interpolating between known binding configurations. With these heuristics, we are confident that we can make the simulation tractable. Over time, we will strive to improve its accuracy.

5 Broader Impacts

Intrinsically, the research in the proposal aims to characterize cellular immunity at the level of the *individual*. The computational engine that we will develop will predict how well peptides bind to specific variants of MHC I molecules, coded for by an individual's genes. As discussed in Section 1, such understanding could inform the early response to a pandemic, rapidly identifying individuals at acute risk as well as those at low risk from an emerging pathogen. Given a genetic profile of a group or population, information can be used to infer risk, inform treatment, and guide vaccine development tailored to groups or populations.

This focus on the genetic differences between individuals is a opportunity but also a significant moral hazard, one that will be the primary focus of our Broader Impacts efforts in this proposal. As with nearly all genetics research, there is a significant bias towards individuals of European descent in the available data on MHC I molecules. Without rectifying this, the benefits to pandemic preparedness discussed in Section 1 might only accrue to this demographic, further exacerbating the health disparities in our society.

Since the human genome was first sequenced in April 2003, the vast majority of genetic studies have been performed on people of European descent. According to a 2009 analysis, 96% of participants in genome-wide association studies were from this demographic [21]. By some measures, the percentage of people from other demographics has improved since, but only with respect to the inclusion of people of Asian descent [25]. The percentage of people of African and Latin American descent as barely shifted. People from different indigenous groups – in the Americas, Asia, the Arctic, Austronesia – are similarly underrepresented.

Data indicate that inequalities in health care are being exacerbated by the bias in genetic data. Patients of African and Native American ancestry are currently more likely than those of European ancestry to receive ambiguous genetic test results after sequencing, or be told that they have variants of unknown significance [23]. This is not only unfair; it is a missed opportunity scientifically, since the genetic diversity in these groups would enrich these studies. Similarities aren't that interesting when it comes to how genetics impacts diseases; differences are.

The bias in genetic data is evident in the structural models of MHC I molecules that are available for our research. Overwhelmingly, the detailed models are for those variants coded by genes that are prevalent in populations of European descent [20]. As explained in Section 2, a significant facet of our Grand Challenge problem is to construct *de novo* structural models of MHC I variants that have not been characterized experimentally.

We commit to the following tenets for this NSF Development Grant as well as for the eventual Minnesota-Mayo Center on the Computational Human Immuno-Peptidome (CHIP). At all times, we will prioritize using genetic data, generating data, and performing analyzes **on the immunopeptidome of underrepresented groups**:

- As examples when developing algorithms and implementing code;
- As examples in figures and tables, in all reports and publications, including internal ones; and
- When deploying computation at scale on cloud computing infrastructure or on graphical-processing unit (GPU) clusters.

As was explained in Section 2.3, a thrust of our research is to create structural models for MHC I molecules for variants that have not been characterized experimentally. The paucity of data is largely from variants found in underrepresented groups. Delivering these models, and simulating peptide binding for these, is a deliverable for this grant.

5.1 Broadening Participation

The PIs have an excellent track record of working with women and other underrepresented groups in computer science, computer engineering, and computational immunology. PI Riedel and Vasmatzis both currently have female PhD students and have graduated a much higher number of female students than the average in their departments.

The PIs will work with the University of Minnesota's College of Science and Engineering Diversity and Outreach program to involve underrepresented students in this research. This program manages the NSF-funded **North Star STEM Alliance–Minnesota's Louis Stokes Alliance for Minority Participation** (LSAMP). One of the core principles of the Diversity and Outreach program is that introducing students to research opportunities early in their undergraduate career is the best practice for retention. The program places students from underrepresented groups in research labs and provides funding, for instance for them to travel to conferences.

6 Results from Prior NSF-funded Research

1. NSF Grant 2036064 "EAGER: Computationally Predicting and Characterizing the Immune Response to Viral Infections," start date 08/01/2020, 24 Months, $200,000.
    - Intellectual Merit: This project will apply neural networks and machine learning to predict which peptides derived from SARS- CoV-2 will bind to each allelic variant of MHC-I molecule commonly found in the U.S. population.

- Broader Impacts: The project has funded 2 Ph.D. students. It has delivered software for predicting peptide:MHC binding.

2. NSF Grant 1408123: “SHF: Medium: Back to the Future with Printed, Flexible Electronics Design in a Post-CMOS Era when Transistor Counts Matter Again,” start date 07/31/2014, 36 Months, $808,000.
    - Intellectual Merit: This project developed and applied the paradigm of stochastic bit stream computation to design challenges in emerging areas such as printed electronics.
    - Broader Impacts: The project has funded 6 Ph.D. students, and resulted in 18 published papers and three patents.
3. NSF grant 1423407, “SHF: Small: Advanced Digital Signal Processing with DNA,” start date 07/31/2014, 36 Months, $408,000.
    - Intellectual Merit: This project demonstrated the synthesis of Markov chains, and polynomials with a fractional representation of molecular values
    - Broader Impacts: The project has funded 2 Ph.D. students, and resulted in 8 published papers.
4. NSF Grants 1241987, 1338382: “EAGER: Digital Yet Deliberately Random – Synthesizing Logical Computation on Stochastic Bit Streams,”
start date 06/30/2012, 24 Months, $328, 423.
    - Intellectual Merit: The project introduced the concept of stochastic computing with state machines; this significantly reduces hardware cost and allows for trade-offs between accuracy and resource usage.
    - Broader Impacts: The project has funded 3 Ph.D. students and resulted in 11 published papers.
5. NSF grant 1117168, “SHF: Small: Digital Signal Processing with Biomolecular Reactions,” start date 07/31/2011, 36 Months, $400,000.
    - Intellectual Merit: This project demonstrated filtering and other signal processing operations with molecular constructs.
    - Broader Impacts: The project has funded 2 Ph.D. students and resulted in 7 published papers.
6. NSF Grant 0845650, “CAREER Award: Computing with Things Small, Wet, and Random—Design Automation for Digital Computation with Nanoscale Technologies and Biological Processes”, start date 09/14/2009, 60 Months, $500,000.
    - Intellectual Merit: This award established novel and transformative approaches to design automation guided by physical views of computation. A broad theme was application of expertise from an established field, digital circuit design, to new fields, such as nanotechnology and synthetic biology.
    - Broader Impacts: With its cross-disciplinary emphasis, this project brought new perspectives to both the field of circuit design and the field of synthetic biology. The project has funded 3 Ph.D. students and a Postdoc. It resulted in 13 published papers.
7. NSF grant 0946601, “EAGER: Synthesizing Signal Processing Functions with Biochemical Reactions,” start date, 07/31/2009, 24 Months, $200,000.
    - Intellectual Merit: This project demonstrated iterative computation with molecular reactions. It only introduced the paradigm of rate-independent computation, which has transformed the field.
    - Broader Impacts: The project has funded 2 Ph.D. students and resulted in 7 published papers.