Fall 2017

Accelerated many-body protein side-chain repacking using gpus: application to proteins implicated in hearing loss

Mallory RaNae Tollefson
University of Iowa

Recommended Citation

Tollefson, Mallory RaNae. "Accelerated many-body protein side-chain repacking using gpus: application to proteins implicated in hearing loss." MS (Master of Science) thesis, University of Iowa, 2017.
https://doi.org/10.17077/etd.dljipt1k

Follow this and additional works at: https://ir.uiowa.edu/etd
Part of the Biomedical Engineering and Bioengineering Commons

ACCELERATED MANY-BODY PROTEIN SIDE-CHAIN REPACKING USING GPUS: APPLICATION TO PROTEINS IMPLICATED IN HEARING LOSS

by

Mallory RaNae Tollefson

A thesis submitted in partial fulfillment
of the requirements for the Master of Science
degree in Biomedical Engineering in the
Graduate College of
The University of Iowa

December 2017

Thesis Supervisors: Assistant Professor Michael J. Schnieders
Professor Richard J. H. Smith

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

MASTER’S THESIS

This is to certify that the Master's thesis of

Mallory RaNae Tollefson

has been approved by the Examining Committee for
the thesis requirement for the Master of Science degree
in Biomedical Engineering at the December 2017 graduation.

Thesis Committee:

Michael J. Schnieders, Thesis Supervisor

Richard J. H. Smith

Terry Braun

Michael Mackey

Thomas Casavant

To my husband, William Tollefson, for encouraging me and supporting my goals; and to my mother, Korbi Muñoz, for teaching me to try again when I fail and to stay positive when life gets hard.

ACKNOWLEDGEMENTS

I would like to thank my advisors, Dr. Michael Schnieders and Dr. Richard Smith, for providing me with valuable research opportunities. Their guidance and expertise have been instrumental in my growth as a scientist. Dr. Schnieders, thank you for introducing me to computational biophysics and always challenging me to keep learning. I can't thank you enough for the countless times you stayed late in the lab to guide me through problems or worked weekends to review my work. Dr. Smith, thank you for providing a clinical perspective to enhance my research projects; your influence has helped to contextualize and make my work more impactful. Your positive disposition is so uplifting.

I would also like to thank the members of my examining committee, Dr. Michael Mackey, Dr. Thomas Casavant, and Dr. Terry Braun. Dr. Mackey provided opportunities for me to practice teaching in a hands-on assistant position, where I learned teaching skills that will serve me throughout my career. Both Dr. Casavant and Dr. Braun have always been willing to offer new perspective on my research with insightful feedback.

Thank you to all of my peers in the laboratory; Jacob Litman, Armin Avdic, Hernan Bernabe, Jill Hauer, and Stephen LuCore have been particularly helpful in providing background knowledge, advice, and friendship. I would also like to thank the members of the Molecular Otolaryngology and Renal Research Laboratories (MORL) for providing an encouraging work environment. Dr. Hela Azaiez and Kevin Booth have offered constant support of my work through insightful feedback and encouragement.

Finally, I would like to thank my husband, William Tollefson, and my mother, Korbi Muñoz, for supporting and inspiring me as I pursue my goals.

ABSTRACT

With recent advances and cost reductions in next generation sequencing (NGS), the amount of genetic sequence data is increasing rapidly. However, before patient specific genetic information reaches its full potential to advance clinical diagnostics, the immense degree of genetic heterogeneity that contributes to human disease must be more fully understood. For example, although large numbers of genetic variations are discovered during clinical use of NGS, annotating and understanding the impact of such coding variations on protein phenotype remains a bottleneck (i.e. what is the molecular mechanism behind deafness phenotypes). Fortunately, computational methods are emerging that can be used to efficiently study protein coding variants, and thereby overcome the bottleneck brought on by rapid adoption of clinical sequencing.

To study proteins *via* physics-based computational algorithms, high-quality 3D structural models are essential. These protein models can be obtained using a variety of numerical optimization methods that operate on physics-based potential energy functions. Accurate protein structures serve as input to downstream variation analysis algorithms. In this work, we applied a novel amino acid side-chain optimization algorithm, which operated on an advanced model of atomic interactions (i.e. the AMOEBA polarizable force field), to a set of 164 protein structural models implicated in deafness. The resulting models were evaluated with the MolProbity structure validation tool. MolProbity "scores" were originally calibrated to predict the quality of X-ray diffraction data used to generate a given protein model (i.e. a 1.0 Å or lower MolProbity score indicates a protein model from high quality data, while a score of 4.0 Å or higher reflects relatively poor data). In this work, the side-chain optimization algorithm improved mean MolProbity score from 2.65 Å (42nd percentile) to nearly atomic resolution at 1.41 Å (95th percentile). However, side-chain optimization with the AMOEBA many-body potential function is computationally expensive. Thus, a second contribution of this work is a parallelization scheme that utilizes nVidia graphical processing units (GPUs) to accelerate the side-chain repacking algorithm. With the use of one GPU, our side-chain optimization algorithm achieved a 25 times speed-up compared to using two Intel Xeon E5-2680v4 central processing units (CPUs). We expect the GPU acceleration scheme to lessen demand on computing resources dedicated to protein structure optimization efforts and thereby dramatically expand the number of protein structures available to aid in interpretation of missense variations associated with deafness.

PUBLIC ABSTRACT

Protein structural modeling plays an important role in elucidating the function of proteins, studying the changes in function caused by genetic variations that may underlie disease, and aiding in the development of new disease therapeutics. However, modeling proteins at atomic resolution is computationally expensive; to mitigate this, nearly all publicly available protein structures result from highly simplified versions of atomic interactions. Unfortunately, models produced via such simplifications, although less expensive, often contain structural errors that can influence downstream interpretations. Although structural errors can be corrected by application of an advanced model of atomic interactions, there is an associated increase in computational resources. Therefore, there is a need for new protein modeling algorithms that incorporate the accuracy of advanced molecular interactions while maintaining efficiency. The resulting protein models can then be used to aid in interpreting missense variations discovered clinically in the context of non-syndromic hearing loss.

In this work, we apply an advanced model of atomic interactions to correct and improve existing protein structures with a novel optimization algorithm that maintains efficiency. We developed a parallelization scheme – a method that can perform multiple calculations simultaneously – that allows our algorithm to use multiple sources of computing power concurrently, reducing the time needed to analyze each protein structure. We also designed a method that utilizes the most recent advances in computer hardware to increase the speed of simulations. Using cutting-edge hardware and our parallelized algorithm, we achieved a 25-fold speed-up compared to previous attempts at correcting protein structures with an advanced model of atomic interactions. This increase in efficiency will allow, for the first time, high-quality protein structures to be produced for a majority of the genes associated with non-syndromic hearing loss.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

As the most common human sensory deficit, deafness impacts an estimated 360 million people worldwide. Recent advances in targeted genetic sequencing technology such as the development of the OtoSCOPE[1,2] (Otologic Sequence and Capture of Pathogenic Exons) genetic testing platform have popularized the use of genetic testing in clinical diagnostics. OtoSCOPE identifies an average of 545 variants per patient, and all variants sequenced by OtoSCOPE are curated in the publicly available Deafness Variation Database[3,4] (DVD http://deafnessvariationdatabase.org). With hundreds-of-thousands of documented genetic variants available from sequencing platforms (nearly 800,000 variants in the DVD), interpretation and evaluation of variants is crucial to understanding the pathogenic mechanisms that cause deafness. With this large number of sequenced genetic variations and limited experimental resources, computational biophysical simulations can be used to help elucidate the structural and biochemical impacts that missense variants have on inner ear function. Specifically, using massively parallelized algorithms for advanced hardware makes physics-based protein modeling an effective and efficient method for studying genetic variants. Prior to studying variants using physics-based thermodynamic simulations, we must first construct high-quality models for each protein implicated in deafness.

Presently, most protein structures found in either the Protein Databank (PDB)[5] or homology modeling databases[6,7] are based on structure refinement with pairwise potential energy functions (*i.e.* force fields) such as the fixed charge Amber[8, 9], CHARMM[10, 11] and OPLS-AA[12, 13] models. There is currently a trend to augment central processing units (CPUs) with the massively parallel architecture provided by graphical processing units

(GPUs). The efficient, parallel capacity of GPUs has made protein simulation with a many-body potential tractable. Many-body potential energy functions for proteins include polarizable force fields, such as the AMOEBA[14, 15] and CHARMM Drude[16] models, and continuum representations of solvation[17, 18]. Previous studies suggest that the resolution of many protein structures in the PDB can be improved by these more advanced potential energy functions[19]. Many-body potentials provide an alternative to pairwise models that promise to improve protein modeling accuracy. To assess protein structure quality, the MolProbity[20,21] assessment tool is widely used. MolProbity was calibrated to use knowledge-based heuristics to estimate the X-ray diffraction resolution used to produce protein structures. For example, MolProbity evaluates all-atom contacts and scores side-chain rotamers based on empirical distributions. Unfavorable overlap of atoms is measured with a clash score, representing the number of unphysical and unfavorable atom overlaps per one thousand atoms. Side-chain conformations (*i.e.* rotamers) are flagged as outliers if they fall within the least probable 1 percent of rotamers. By assigning a clash score and identifying rotamer outliers, MolProbity is able to pinpoint structural errors that could be addressed during later refinement procedures.

In this work, we applied a global amino acid side-chain optimization algorithm based on an advanced model of atomic interactions (*i.e.* the AMOEBA polarizable force field) to a set of 164 protein homology structural models implicated in deafness. All structural models were obtained from the SwissProt[22] homology database, based on the requirement that the homology template had at least 30% sequence identity to the target protein. We assessed the results of our algorithm by analyzing pre- and post-optimized structures with the MolProbity protein structure validation tool. Side-chain optimization improved the mean MolProbity score from 2.65 Å (42nd percentile compared to all structures in the PDB) to nearly atomic resolution at 1.41 Å (95th percentile). However, side-chain optimization with the AMOEBA many-body potential function increases the demand for computing resources relative to simple pairwise potential energy functions. We recently designed a parallelization method that uses the Parallel Java[23] (PJ) message passing interface (MPI) library to parallelize across compute nodes and the OpenMM library[24] to perform side-chain optimization using nVidia GPUs *via* the CUDA language. With the addition of only one GPU to two Intel Xeon E5-2680v4 central processing units (CPUs), our side-chain optimization algorithm achieved a 25 times speed-up compared to using just the two CPUs. Our parallel algorithm utilizes 94% of the GPU capacity according to the nVidia device monitor for a ~100 residue protein domain. Our massively parallelized, many-body, side-chain optimization technique improves protein structure quality while efficiently using computing resources. We expect the more efficient GPU acceleration scheme to lessen demand on computing resources dedicated to protein structure optimization efforts.

CHAPTER 2: BACKGROUND

2.1: Overview of NSHL and OtoSCOPE

Approximately 70 percent[2] of individuals with congenital deafness are diagnosed with non-syndromic hearing loss (NSHL)–hearing loss in the absence of additional symptoms or clinical phenotypes. Nearly a million variants spanning over 80 genes have been identified to cause NSHL, which makes NSHL heterogenic and the task of pinpointing the genetic cause of many NSHL cases difficult. However, targeted genomic enrichment with massive parallel sequencing (TGE+MPS) is transforming the diagnostics of heterogenic disease by laying the foundation for personalized health care. The TGE+MPS platform, OtoSCOPE, has been validated as a clinical tool that simultaneously interrogates the coding exons and flanking intronic sequence of all genes implicated in NSHL[25]. First developed in 2010, the current version, v8, tests NSHL-causing genes, non-syndromic mimics like the Usher syndrome genes, mitochondrial genes, and genes that have been implicated in the more common forms of syndromic hearing loss. OtoSCOPE discovers many variants in each patient sample, and ultimately provides a definitive genetic diagnosis for 42 percent of cases[1]. However, the unannotated novel variants discovered during sequencing result in inconclusive results; as such, variant interpretation remains a bottleneck for genetic diagnostics in the context of heterogenic disease.

With the larger number of variants sequenced by OtoSCOPE and a limited amount of resources available for variant interpretation, computational biophysical simulations show promise in helping to elucidate the structural and biochemical impacts that missense variations have on inner ear protein function. This, in turn, will help to alleviate the bottleneck of variant interpretation. One limitation of using protein structural simulations for variant interpretation is the necessity for a high-quality input models for the protein of interest. Only a small fraction of the human proteome has been structurally solved by experimental approaches, however, comparative protein modeling can be used to increase the structural coverage. Comparative models created from a homologous protein with 30 percent or greater sequence identity provides a high likelihood that the protein backbone fold has been evolutionarily conserved[26]. To date, approximately 40 percent of the human proteome has been comparatively modeled[6] against homologous structures with 30 percent sequence identity or greater. Analogously, 43 percent of the genes examined by OtoSCOPE have corresponding homology models based on 30 percent sequence identity or greater. Fortunately, structural coverage of the human proteome has significantly increased in recent years and advances in electron cryo-electron microscopy will likely advance structural coverage of the human reference proteome in coming years.

As clinical tools, validated TGE+MPS platforms such as OtoSCOPE have changed the clinical evaluation of heterogenic diseases like hearing loss. Establishing a genetic diagnosis is no longer an exclusionary process relegated to the end of an extensive and expensive diagnostic algorithm. Rather, because of its comprehensive design, genetic testing now has the highest yield of any test used in the evaluation of hearing loss and as such, has moved to the beginning of the diagnostic algorithm. A positive genetic diagnosis can obviate downstream tests, guide subsequent care, and save healthcare dollars. In addition, as alternative and personalized therapies for hearing loss are developed, modeling the proteins and variants sequenced during comprehensive genetic testing will be foundational to therapy implementation.

2.2: Force Fields

As a mathematical description of molecular energetics, a force field can be used to simulate the properties of biomolecules. Beginning in the 1940's, mathematical models were developed for diphenyl derivatives and other small, organic compounds[27]. In contrast, present-day simulations often involve whole proteins consisting of thousands of atoms. An accurate description of molecular energetics is essential for atomic resolution biomolecular simulations, but increased accuracy often introduces increased computational cost. Thus, modeling large systems with a quantum mechanical level of detail, though in principle most accurate, is too computationally expensive given currently available computer hardware. Alternatively, classical molecular mechanics decreases computational costs associated with simulation, but may also introduce inaccuracies. For example, some molecular mechanics force fields are based on fixed atomic partial charges, which is often too simplistic for accurate protein energetics or structural optimization. However, force fields with explicit polarization and high order permanent multipoles stand at the intersection between accuracy and computational cost. Consequently, in this work we use the AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) polarizable force field for accurate protein structure optimization. We use parallelization techniques and recent advances in graphical processing units to decrease the time costs associated with use of a polarizable force field.

The AMOEBA force field incorporates induced dipoles and fixed atomic multipoles up to quadrupoles, which in principle makes the model more accurate than fixed-charge force field counterparts. First published to define water[28] in 2003, the

AMOEBA model is transferable as it can represent molecules across aqueous and vapor environments. The functional form of the AMOEBA force field is described in equation 1.

$$U_{AMOEBA} = U_{\mathrm{bond}} + U_{\mathrm{angle}} + U_{\mathrm{b\theta}} + U_{\mathrm{oop}} + U_{\mathrm{torsion}} + U_{\mathrm{tor-tor(GLY)}} + U_{\mathrm{vdW}} + U_{ele}^{perm} + U_{ele}^{ind}$$
$$+U_{\mathrm{vdW}} + U_{ele}^{perm} + U_{ele}^{ind}$$

Equation 1

The first six terms describe bonded interactions, including bond stretching, angle bending, bond-angle cross term, out-of-plane bending, and torsional rotation. The final three terms describe non-bonded interactions including van der Waals, permanent electrostatics, and polarization energy. Deviations from the ideal angle and bond length, θ_0 and b_0 respectively, are accounted for in the bond and angle energy contributions (the first three terms of equation 1).

Bond stretching is described by equation 2,

$$U_{bond} = K_{\mathrm{b}}(b-b_0)^2[1-2.55(b-b_0)+3.793125(b-b_0)^2]$$

Equation 2

and a bond angle's energy is described by equation 3.

$$U_{angle} = K_{\theta}(\theta-\theta_0)^2\,[1-0.014(\theta-\theta_0)+5.6\times10^{-5}(\theta-\theta_0)^2$$
$$-7.0\times10^{-7}(\theta-\theta_0)^3+2.2\times10^{-8}(\theta-\theta_0)^4]$$

Equation 3

The bond-angle cross term, which represents the coupling of bond stretching and angle bending, is shown in equation 4.

$$U_{b\theta} = K_{\mathrm{b\theta}}[(b-b_0)+(b'-b'_{\theta})](\theta-\theta_0)$$

Equation 4

Equation 5 describes out-of-plane bending as an angle, χ, between a vector and plane.

$$U_{oop} = K_{\chi}\chi^2$$

Equation 5

Four linearly bonded atoms will have rotational favorability represented by the torsion energy in equation 6. A Fourier expansion with *n* terms describes the dihedral angle, ϕ**.** K_n and δ_n represent the magnitude and phase of the *n*th Fourier term, respectively.

$$U_{torsion} = \sum_{n} K_n[1 + cos(1 + n\phi \pm \delta_n)]$$

Equation 6

The van der Waals interactions are described by equation 7, where ε_{ij} represents well depth, and ρ_{ij} represents the quotient of separation distance between two atoms (R_{ij}) and separation distance given by the lowest energy (R^0_{ij}).

$$U_{vdW} = \varepsilon_{ij}\left(\frac{1 + \delta}{\rho_{ij} + \delta}\right)^{n-m}\left(\frac{1 + \gamma}{\rho_{ij}^m + \gamma} - 2\right)$$

Equation 7

Using a buffered 14-7 potential, the van der Waals interactions are specifically described by equation 8.

$$U_{vdW} = \varepsilon_{ij}\left(\frac{1.07}{\rho_{ij} + 0.07}\right)^{7}\left(\frac{1.12}{\rho_{ij}^7 + 0.12} - 2\right)$$

Equation 7

When combining heterogeneous atom pairs, $R_{ij}^{0} = \frac{\left(R_{ii}^{0}\right)^{3}+\left(R_{jj}^{0}\right)^{3}}{\left(R_{ii}^{0}\right)^{2}+\left(R_{jj}^{0}\right)^{2}}$ defines the minimum energy distance and $\varepsilon_{ij} = \frac{4\varepsilon_{ii}\varepsilon_{jj}}{\left(\varepsilon_{ii}^{1/2}+\varepsilon_{jj}^{1/2}\right)^{2}}$ defines well depth.

The permanent electrostatic energy between two atoms *i* and *j*, is represented as $U_{ele}^{perm}(r_{ij}) = M_i^T T_{ij} M_j$, where r_{ij} is the separation distance between the two atoms. T_{ij} is given by equation 9, and M_i is a permanent multipole with a charge (*q*), dipole (*μ*), and quadrupole (*Q*) as shown in equation 10.

$$T_{ij} = \begin{bmatrix} 1 & \frac{\partial}{\partial_{x_j}} & \frac{\partial}{\partial_{y_j}} & \frac{\partial}{\partial_{z_j}} & \cdots \\ \frac{\partial}{\partial} & \frac{\partial^2}{\partial_{x_i}\partial_{x_j}} & \frac{\partial^2}{\partial_{x_i}\partial_{y_j}} & \frac{\partial^2}{\partial_{x_i}\partial_{z_j}} & \cdots \\ \frac{\partial}{\partial} & \frac{\partial^2}{\partial_{y_i}\partial_{x_j}} & \frac{\partial^2}{\partial_{y_i}\partial_{y_j}} & \frac{\partial^2}{\partial_{y_i}\partial_{z_j}} & \cdots \\ \frac{\partial}{\partial} & \frac{\partial^2}{\partial_{z_i}\partial_{x_j}} & \frac{\partial^2}{\partial_{z_i}\partial_{y_j}} & \frac{\partial^2}{\partial_{z_i}\partial_{z_j}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \frac{1}{r_{ij}}$$

Equation 9

$$M_i = \left[q_i, \mu_{i,x}, \mu_{i,y}, \mu_{i,z}, Q_{i,xx}, Q_{i,xy}, Q_{i,xz}, \dots, Q_{i,zz}\right]^T$$

Equation 10

An induced dipole at each atomic center can be used to describe polarization shown in equation 11, where α_i is the polarizability of atom i and $E_{i,\alpha}$ is the electric field along any axis, α.

$$\mu_{i,\alpha}^{ind} = \alpha_i E_{i,\alpha}$$

Equation 11

Permanent multipoles and induced dipoles at all other atomic centers (*i.e.* not *i*) generate the electric field described by equation 12.

$$\mu_{i,\alpha}^{ind} = \alpha_i \left(\sum_{\{j\}} T_{\alpha}^{ij} M_j + \sum_{\{j'\}} T_{\alpha\beta}^{ij'} \mu_{j',\beta}^{ind} \right)$$

Equation 12

2.3: Side-Chain Optimization with a Many-Body Potential

Under the AMOEBA many-body potential, the total energy of a protein, $E(\mathbf{r})$, can be defined to arbitrary precision using the expansion in equation 13,

$$E(\mathrm{r}) = E_{\mathrm{env}} + \sum_{i} E_{\mathrm{self}}(r_i) + \sum_{i} \sum_{j>i} E_2(r_i, r_j) + \sum_{i} \sum_{j>i} \sum_{k>j} E_3(r_i, r_j, r_k) + \cdots$$

Equation 13

where E_{env} is the energy of the environment (*i.e.* the protein backbone and residues that are not being optimized), $E_{\mathrm{self}}(r_i)$ is the self-energy of residue i including its intra-molecular bonded energy terms and non-bonded interactions with the backbone, and $E_2(r_i, r_j)$ is the two-body non-bonded interaction energy between residues *i* and *j* excluding other residues. The self, two-body, and three-body energy terms are calculated as shown in equations 14, 15 and 16, respectively.

$$E_{\text{self}}(r_i) = E_{BB/SC}(r_i) - E_{\text{env}}$$

Equation 14

$$\boldsymbol{E}_2(\boldsymbol{r_i}, \boldsymbol{r_j}) = \boldsymbol{E_{BB/SC}}(\boldsymbol{r_i}, \boldsymbol{r_j}) - \boldsymbol{E}_{\text{self}}(\boldsymbol{r_i}) - \boldsymbol{E}_{\text{self}}(\boldsymbol{r_j}) - \boldsymbol{E}_{\text{env}}$$

Equation 15

$$\begin{aligned}\boldsymbol{E}_3(\boldsymbol{r_i}, \boldsymbol{r_j}, \boldsymbol{r_k}) = \boldsymbol{E_{BB/SC}}(\boldsymbol{r_i}, \boldsymbol{r_j}, \boldsymbol{r_k}) - \boldsymbol{E}_{\text{self}}(\boldsymbol{r_i}) - \boldsymbol{E}_{\text{self}}(\boldsymbol{r_j}) - \boldsymbol{E}_{\text{self}}(\boldsymbol{r_k}) - \boldsymbol{E}_2(\boldsymbol{r_i}, \boldsymbol{r_j}) \\ - \boldsymbol{E}_2(\boldsymbol{r_i}, \boldsymbol{r_k}) - \boldsymbol{E}_2(\boldsymbol{r_j}, \boldsymbol{r_k}) - \boldsymbol{E}_{\text{env}}\end{aligned}$$

Equation 16

A visual description of equations 14, 15 and 16 are shown in figures 1, 2 and 3 respectively.

$$E_{self}(R_i^{\alpha})$$

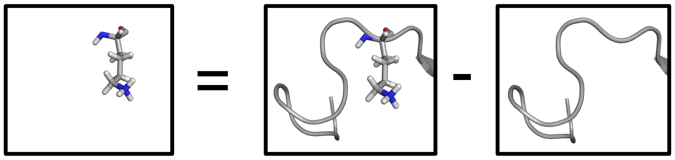

Figure 1. A visual description of equation 14. The self-energy of residue *i*, with side-chain rotamer conformation α, is equal to the energy of the backbone plus residue combination, less the environment (i.e. the backbone and all residues that are not being optimized).

$$E_2\left(R_i^{\alpha}, R_j^{\beta}\right)$$

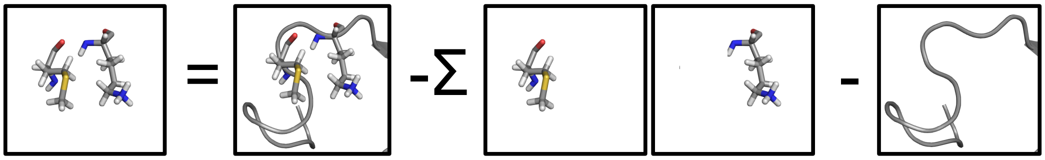

Figure 2. A visual description of equation 15. The pair-energy for residues *i* and *j*, with side-chain rotamer conformations α and β, respectively, is equal to the energy of the combination of the backbone plus residues *i* and *j*, less both self-energies and less the environment.

$$E_3\left(R_i^{\alpha}, R_j^{\beta}, R_k^{\gamma}\right)$$

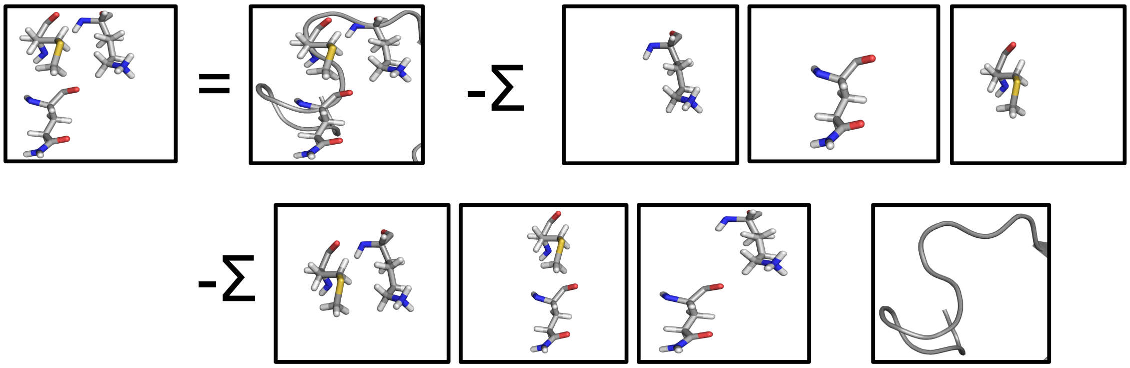

Figure 3. A visual description of equation 16. The three-body energy for residues *i*, *j*, and *k*, with side-chain rotamer conformations α, β, and γ, respectively, is equal to the energy of the backbone plus residues *i*, *j*, and *k*, less three self-energies, three pair-energies and the environment.

CHAPTER 3: GPU ACCELERATION OF SIDE-CHAIN OPTIMIZATION

3.1: Massive Parallelization of the Many-Body Expansion

Computing the self, two-body and three-body energy terms as a function of rotamer conformation is computationally expensive. To address this challenge, our Force Field X (FFX) program[29,30] (http://ffx.biochem.uiowa.edu) utilizes two complementary parallelization approaches, including 1) use of the Parallel Java (PJ) message passing interface (MPI) library to distribute terms among multiple processes, and 2) the OpenMM library to perform force field energy evaluations using nVidia GPUs *via* the CUDA language. FFX uses PJ to sub-divide each shared memory node of a multiple node compute cluster into one or more processes. Energy terms are then assigned to processes, evaluated and globally communicated using MPI, with synchronization steps between calculation of the self, two-body, and three-body terms (*i.e.* two-body terms depend on self-terms as shown in Equation 15, and thus must be calculated after self-energies are completed). For nodes with one or more nVidia GPU coprocessors, the FFX-OpenMM interface (based on Java Native Access wrappers to the OpenMM C API) can be used to offload energy evaluations from FFX running on CPUs to OpenMM on GPUs (figures 1 and 2). Once all energy terms are calculated, side-chain rotamers and rotamer pairs are eliminated by low-energy alternatives based on rigorous mathematical inequalities that have been described for pairwise force fields (i.e. dead-end elimination[31] and/or Goldstein elimination[32]) and more recently for many-body force fields[19] such as the polarizable atomic multipole AMOEBA model.

Prior to exercising dead-end or Goldstein elimination, we apply a pruning technique to reduce the number of high-energy side-chain rotamers in a computationally inexpensive method. Side-chain rotamers with a self-energy that is 30 kcal/mol larger than the side-chain's lowest energy rotamer (*i.e.* often corresponding to unphysical clashes with the backbone) are pruned from the simulation to avoid superfluous, expensive pair and triple energy evaluations involving that high-energy rotamer. Pruning greatly diminishes the number of pair and triple energy evaluations necessary prior to dead-end or Goldstein elimination.

In order to perform a global optimization, the energy of all possible rotamer pairs, triples, etc., must be evaluated to find the rotamer combination with the lowest energy. The effect that any side-chain has on another side-chain acts as a function of distance, where in general the further apart the rotamers, the smaller the effect. The advantage of calculating the energy for a set of rotamers with a large distance between them is small relative to the cost of computing. To avoid expensive energy evaluations on a set of rotamers separated by a large distance, we assign a cut-off distance where the effect that any two or more side-chains have on each other is assumed infinitesimal if those side-chains are separated by a distance larger than the cut-off.

The combination of possible side-chain conformations grows exponentially with protein size, making large proteins a particularly difficult optimization problem. In order to simplify the optimization of large proteins, we used a "box optimization" method. In this method, a box containing several amino acids is centered at the alpha carbon of the N-terminus. The box is treated as an independent system where amino acids outside of the box are assumed to have no effect inside of the box. The combinations of all rotamers that lie in the box are evaluated until the lowest energy for the box is solved. The box is then incrementally moved along the protein and solved until all side-chains are in their lowest energy conformation. Box-optimization avoids the unnecessary evaluation of side-chains separated by a large distance through only evaluating rotamers lying within the relevant boxed area.

Figures 4 and 5 describe the parallelization methods used during the calculation of self and pair energies, respectively. This parallelization technique is extended to three-body terms in our global protein side-chain optimization algorithm as well.

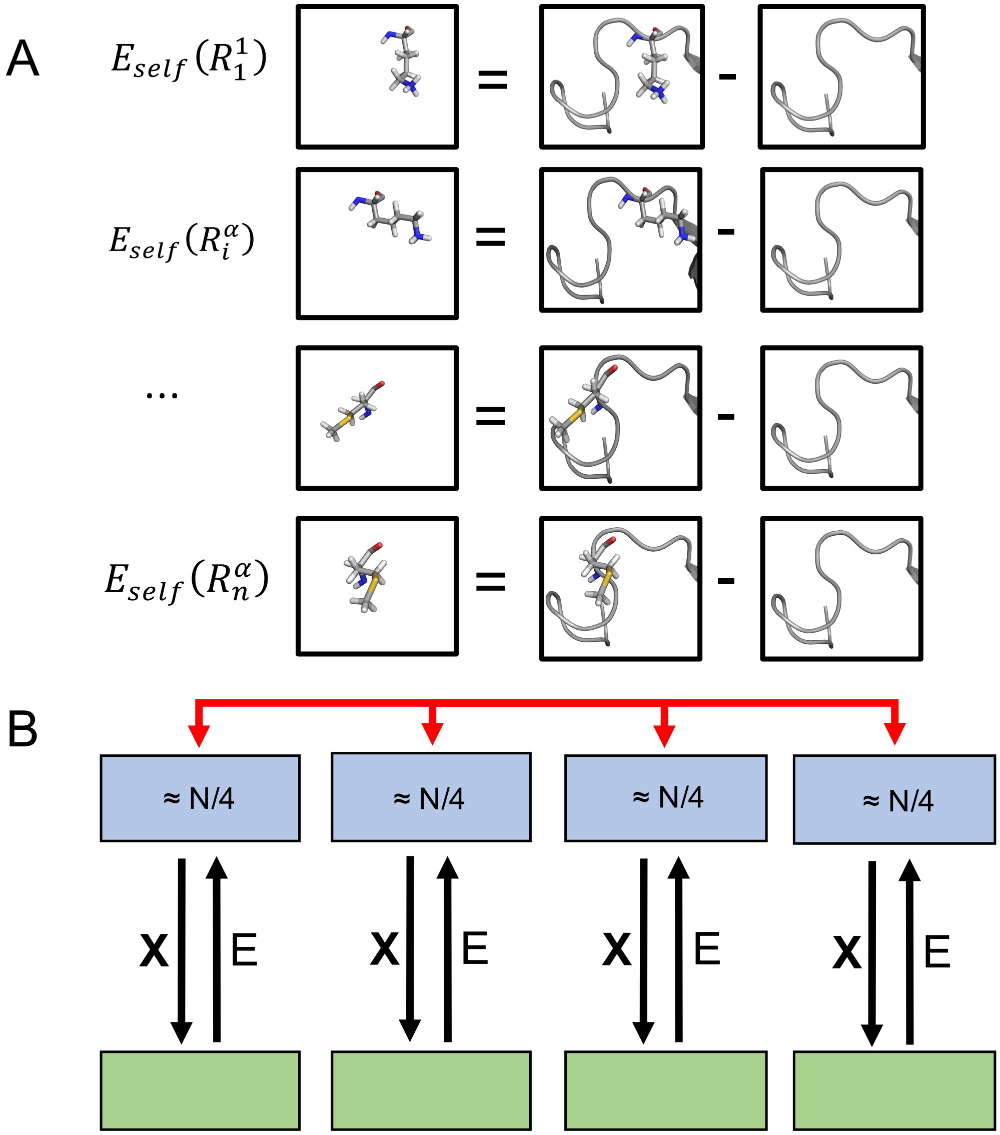

Figure 4. A) $E_{self}(R_i^{\alpha})$ represents the self-energy for residue i in rotamer conformation α. The total number of rotamer self-energies to be calculated is $N_{\mathrm{self}}=\sum_{i=1}^{n} \mathrm{rot}_i$ (where rot_i is the number of rotamers for residue i). B) A scheduler dynamically assigns calculation of each self-energy to a node (blue box). In this example, each of the four nodes will evaluate approximately $N_{\mathrm{self}}/4$ self-energies. Java MPI (red arrows) is used to synchronize and communicate among nodes. Each node sends the current rotamer-dependent coordinates (X) to a GPU (green box), which returns the self-energy (E) for that rotamer. Use of GPUs accelerates the calculation of energy terms

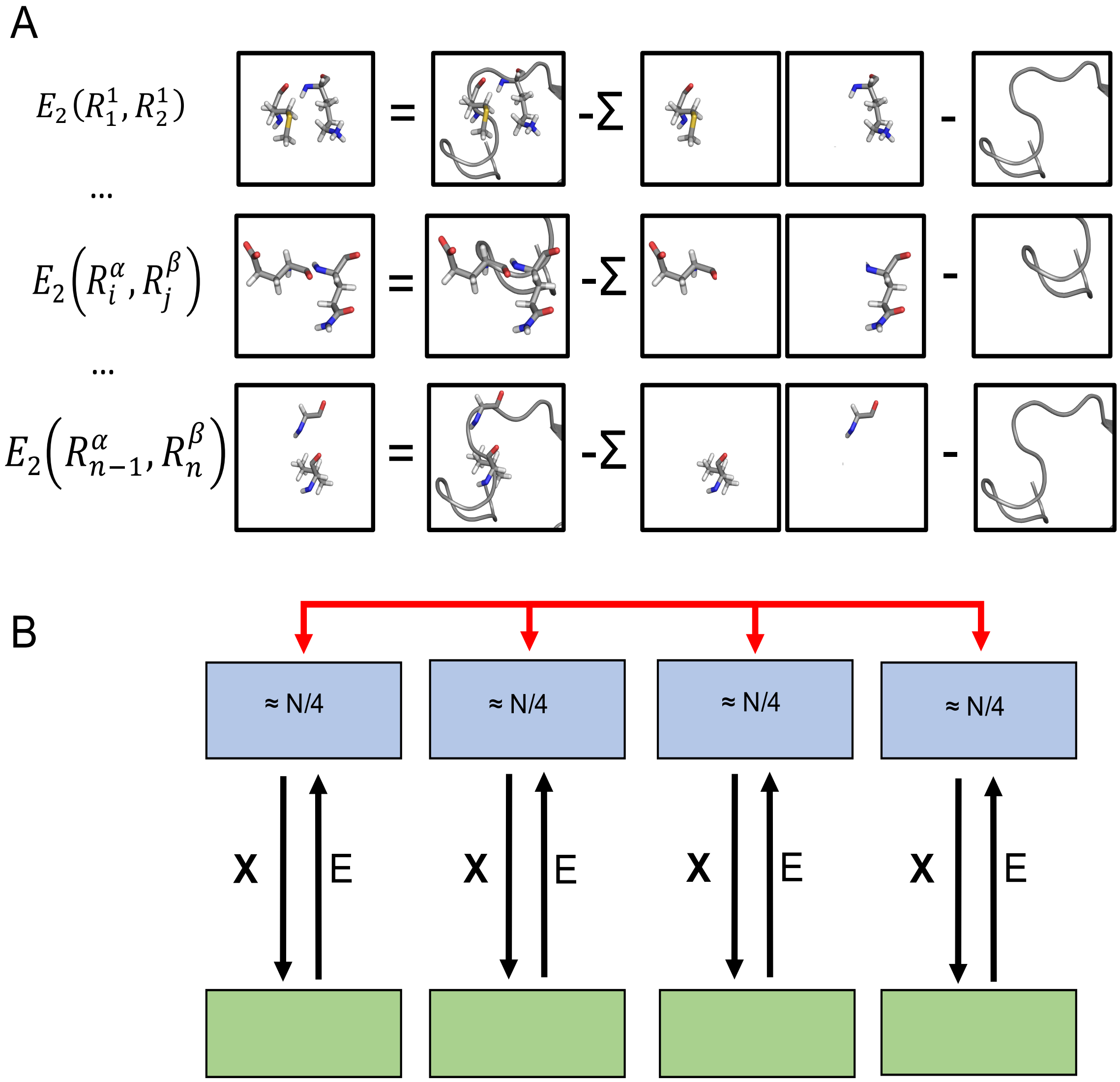
A
$E_2(R_1^1, R_2^1)$
=
-Σ
-
...
$E_2\left(R_i^\alpha, R_j^\beta\right)$
=
-Σ
-
...
$E_2\left(R_{n-1}^\alpha, R_n^\beta\right)$
=
-Σ
-
B
≈ N/4
≈ N/4
≈ N/4
≈ N/4
X
E
X
E
X
E
X
E

Figure 5. A) $E_2\left(R_i^\alpha, R_i^\beta\right)$ represents the pair-energy of residues i and j with rotamer conformations α and β, respectively. The total number of rotamer pair-energies to be calculated is $N_{\text{pair}}=\sum_{i=1}^{n}\left(\text{rot}_i \sum_{j=i+1}^{n} \text{rot}_j\right)$ B) A scheduler dynamically assigns calculation of each pair-energy to a node (blue box). In this example, each of the four nodes will evaluate approximately $N_{\text{pair}}/4$ pair-energies. Java MPI (red arrows) is used to synchronize and communicate among nodes. Each node sends the current rotamer-dependent coordinates (X) to a GPU (green box), which returns the energy (E) for that rotamer pair. Use of GPUs accelerates the calculation of energy terms

3.2: Acceleration Results

We tested our Java MPI parallelization scheme by collecting the timings for a global side-chain optimization of a domain in the USH1C protein, which is commonly implicated in hearing loss. We used nodes consisting of two Intel Xeon E5-2680v4 CPUs and tested the algorithm using one, two, three and four total nodes. Results from the energy evaluation timings for two-body interactions are shown in Table 1, and the respective results for three-body interactions are shown in Table 2. Using multiple nodes achieved a linear speed-up for both two-body and three-body interaction simulations. Side-chain optimization visibly improved the USH1C domain by extending secondary structure and increasing hydrogen bonding (Figures 6 and 7).

Table 1. Timings for energy evaluations of a global rotamer optimization through two-body interactions (617 self, 80358 pair) for an USH1C domain using a varying number of nodes (each node has two Intel Xeon E5-2680v4 CPUs).

# Nodes	Seconds	Speed-Up
1	13760	1x
2	6608	2x
3	4436	3x
4	3341	4x

Table 2. Timings for energy evaluations of a global rotamer optimization through three-body interactions (617 self, 80358 pairs, 4123053 triples) for an USH1C domain using a varying number of nodes (each node has two Intel Xeon E5-2680v4 CPUs).

# Nodes	Seconds	Speed-Up
1	729200	1x
2	349700	2x
3	233600	3x
4	174600	4x

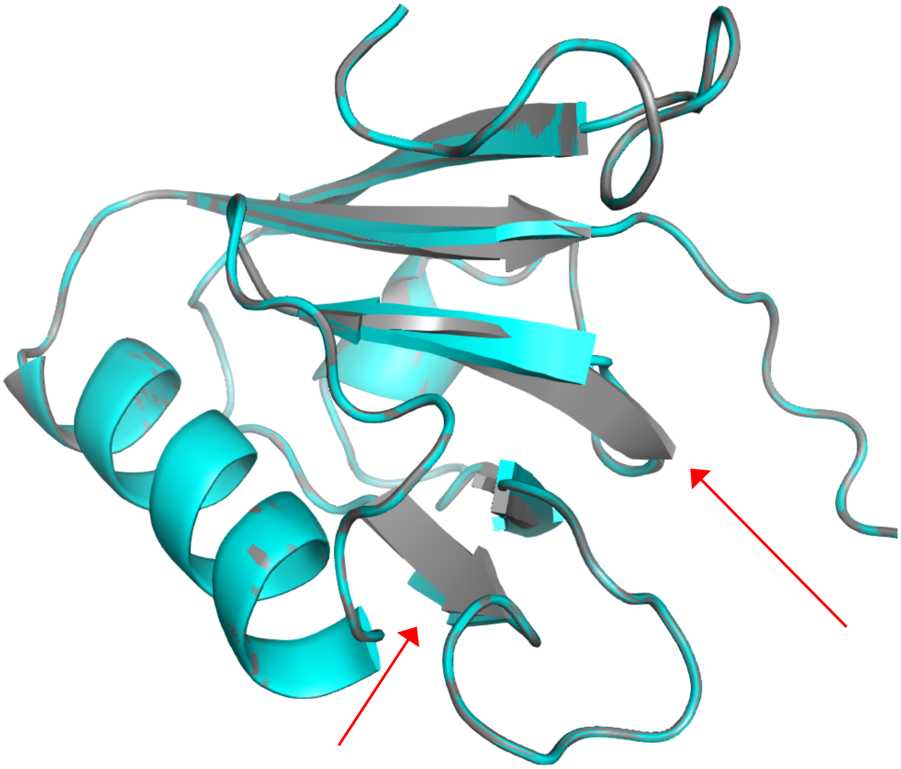

Figure 6. USH1C domain before (blue) and after (grey) side-chain optimization. The superimposed domains show that the optimization procedure extended the protein's secondary structure (red arrows).

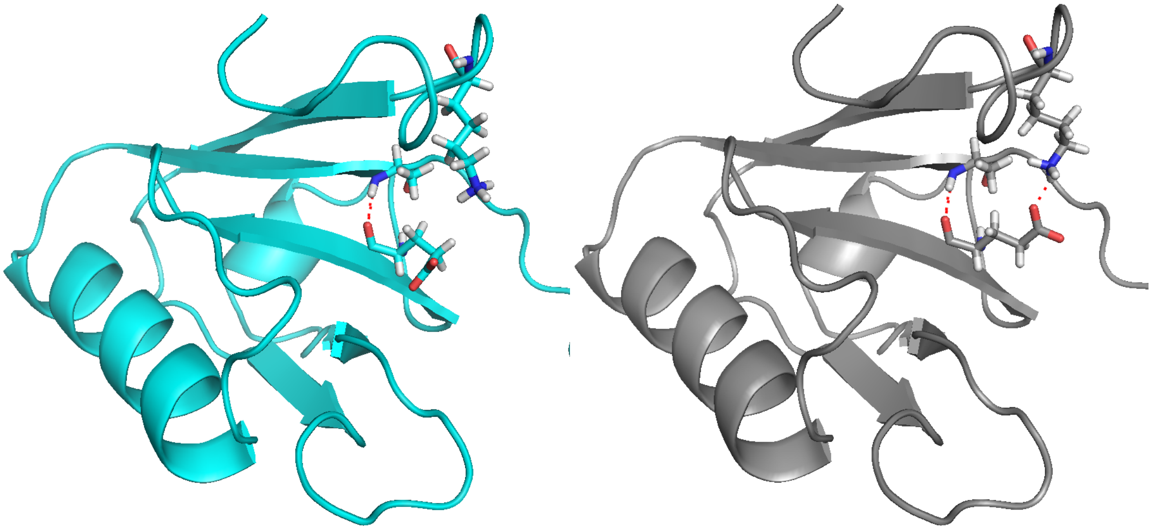

Figure 7. USH1C domain before (blue, left) and after (grey, right) two-body side-chain optimization. Finding low-energy rotamers increased the hydrogen bonding networks of the domain as evidenced by the additional hydrogen bond between glutamic acid and lysine (on right).

We then tested our GPU parallelization on the USH1C domain using a varying number of nodes and GPUs. Offloading energy calculations to OpenMM on a single node equipped with a GPU (two Intel Xeon E5-2680v4 CPUs and one nVidia GPU) resulted in a 24-fold speed-up for both two-body and three-body interaction simulations compared to using one node (two Intel Xeon E5-2680v4 CPUs) with no GPU. According to the nVidia device-monitoring tool, our algorithm efficiently used the GPUs at 94% utilization. Results from the two-body and three-body GPU accelerated side-chain optimization are shown in Tables 3 and 4, respectively.

Table 3. Energy evaluation timings for two-body (617 self, 80358 pair) global side-chain optimization for an USH1C domain using a varying number of nodes (each node with two Intel Xeon E5-2680v4 CPUs and one nVidia Tesla p100 GPU).

Computing Unit (# Nodes / # GPUs)	Time (sec)	Speed-Up Relative to 1 Node (without using GPU)
1 Node / 1 GPU	561.8	24.5x
2 Nodes / 2 GPUs	296.5	46.4x
3 Nodes / 3 GPUs	204.3	67.4x
4 Nodes / 4 GPUs	157.9	87.1x

Table 4. Energy evaluation timings for three-body global side-chain optimization for USH1C domain using a varying number of nodes (each with two Intel Xeon E5-2680v4 CPUs and one nVidia Tesla p100GPU).

Computing Unit (# Nodes / # GPUs)	Time (sec)	Speed-Up Relative to 1 Node (without using GPU)
1 Node / 1 GPU	29970	23.3x
2 Nodes / 2 GPUs	16390	42.7x
3 Nodes / 3 GPUs	11230	62.3x
4 Nodes / 4 GPUs	8647	80.9x

Pruning of high-energy rotamers resulted in an additional 2.28-fold speed-up of a global side-chain optimization on the USH1C domain. Table 5 shows the timings for USH1C domain optimization with and without pruning of high self-energy rotamers.

Table 5. Energy evaluation timings for two-body (617 self, 80358 pair) global side-chain optimization for an USH1C domain using a varying number of nodes (each with two Intel Xeon E5-2680v4 CPUs and one nVidia Tesla p100 GPU) both with and without pruning of high-energy rotamers.

Computing Unit (# Nodes / # GPUs)	Time Without Pruning (sec)	Time With Pruning (sec)	Relative Speed-Up
1 Node / 1 GPU	2460	1078	2.28x

Figure 8 shows a graph of optimization acceleration from the parallel methods described as timed on the University of Iowa High Performance Computing Clusters (HPC) Neon and Argon. Using Neon nodes consisting of two Intel Xeon E5-2609 CPUs, a global two-body side-chain optimization of the 110 residue USH1C domain required 400 minutes of simulation. Parallelizing over nodes and across nVidia GPUs allows our algorithm to utilize an nVidia GPU equipped Argon nodes consisting of two Intel Xeon E5-2680 CPUs simultaneously, resulting in a 40-fold speed-up compared to previous generation Neon hardware. Parallelizing across four Argon GPU nodes reduced the simulation time to less than three minutes.

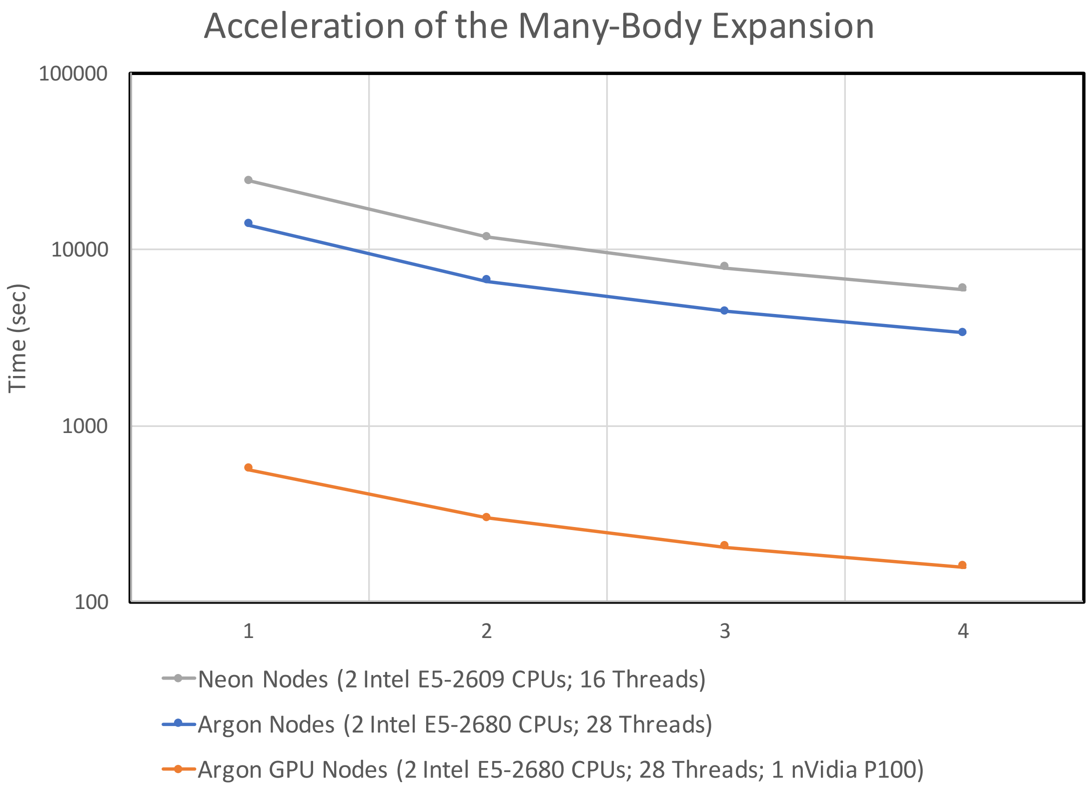

Figure 8. Parallel performance for computing many-body energy terms for a 110 residue protein. After parallelization using a single nVidia P100 GPU, the walk-clock time was reduced from more than 400 minutes on a Neon node down to less than 10 minutes on an Argon GPU node (*i.e.* a 40x speed-up). An additional 4x speed-up is demonstrated via parallelization across multiple nodes (*i.e.* a 160x speed-up overall).

CHAPTER 4: OPTIMIZATION OF PROTEIN STRUCTURES IMPLICATED IN DEAFNESS

4.1: Side-Chain Optimization of Deafness Proteins

Comparative protein models for 104 genes (164 total protein structural models) included in the OtoSCOPE genetic testing platform were acquired from SwissProt[6] and ModBase[7], which are both exhaustive databases of protein homology models derived from alignments to sequences whose structures have been determined using experimental methods such as x-ray crystallography. Comparative protein modeling provides a means by which researchers can predict the structure of a protein whose atomic coordinates have not been solved experimentally by crystallography, NMR or analogous methods.[33] Many human genes implicated in hearing loss have not been studied experimentally, so computational approaches are necessary to generate plausible protein structures. Comparative protein modeling begins from an experimental structure for an evolutionarily related protein, which is used as a template for the target sequence.[7, 34] The percent sequence identity between the homologues provides an estimate on the reliability of the model.[35]

Starting from the sequence alignment to an experimentally solved homolog, structural refinement algorithms based on pairwise fixed partial charge potential energy functions are used to locally optimize the new amino acid backbone and side-chain coordinates. Although both SwissProt and ModBase provide structural coverage for a large portion of the human exome, the breadth of these databases limits the compute resources spent on any individual protein structure. Comparative protein models from leading databases can usually be further refined so that their structure is more consistent with what is known about molecular conformations and side-chain packing. Thus, use of advanced potential energy functions, such as the AMOEBA force field, in tandem with global optimization of amino acid side-chains[19] can greatly improve the quality of SwissProt or ModBase structures as assessed by tools such as MolProbity. MolProbity is widely used by x-ray crystallographers to aid refinement of models by reporting structural features that are known to be unphysical and by recommending low-energy, favorable conformational changes to amino acid backbones and/or side-chains. Lower MolProbity scores indicate a structure that is consistent with higher quality x-ray diffraction data (*i.e.* a MolProbity score of 1.0 reflects the resolution of data at 1.0 Å). All homology models used in this study have sequence identity of 30% or greater to their template structures, which correspond to a high likelihood that the protein backbone fold has been evolutionarily[26] conserved. High quality protein structural models, in turn, provide optimal starting models for downstream molecular dynamics algorithms that can be used to compute thermodynamic changes in protein folding or binding stability.

Homology models were refined using the AMOEBA force field as the potential energy function. The input homology models were subject to a minimization algorithm, followed by global side-chain repacking, a second minimization, and finally an iterative local side-chain optimization and minimization. The initial minimization step was accomplished by means of a quasi-Newton optimization algorithm presented by Broyden[36], Fletcher[37], Goldfarb[38], and Shanno[39] (BFGS) in which the Hessian matrix is approximated using a series of gradients to improve algorithm convergence. The global side-chain optimization[40] assumes a rigid backbone and is based on an energy expansion that includes the energy of the environment, self-energy terms for each residue from a discretized side-chain conformation or rotamer, and pairwise terms for the interaction between two side-chains. For smaller systems, the energy due to three-body interactions (interaction of three residues) was also calculated. Using the Goldstein elimination criteria for elimination of single rotamers and rotamer pairs[31,32,40], the relative energy of rotamers were compared to eliminate conformations not found in the global minimum energy conformation (GMEC). Upon determining the lowest energy state of the side chain conformations, the model was subject to a second BFGS minimization.

The purpose of the first minimization was to eliminate obvious steric clashes in the starting structure. The RMS gradient convergence criterion was conservatively set to 1.0. The global rotamer optimization algorithm was then used to place side-chains into an optimal set of rotamers based on the energy of each set of conformations. During this procedure, to reduce the permutation space for large models, the sliding box optimization method was used to partition the model into a user-specified number of separate sections in three-dimensional space or boxes. In each box, the GMEC was found given a certain combination of rotamers, which were then incorporated into the coordinates for the final, optimized structure. Additionally, the procedure allows for specification of box overlap, limiting the amount of bias resulting from semi-global rotamer optimization compared to global rotamer optimization.

The second minimization was done in order to let the rigid conformations of the rotamers relax and establish bonding networks as might be seen *in vivo*. Since rotamers are structurally stiff, they only give an approximate location of side-chain atoms; consequently, this last minimization is a necessary step toward their correct placement. Finally, local rotamer optimization and minimization were performed to ensure that prior optimization steps (primarily minimization) did not result in rotamers moving too far from their lowest energy conformation. Iterative rotamer optimization and minimization allows large energy barriers to be overcome, helping to place the protein structure coordinates in a low energy, favorable conformation.

4.2: Optimization Results

For both starting homology models and refined structures, integrity was measured using the MolProbity scoring metric, a heuristic algorithm that assesses bad steric clashes, poor rotamers and placement in the Ramachandran Plot. Among crystallographers, the MolProbity score is widely accepted as an approximation of the equivalent x-ray resolution of a model (*e.g.* a MolProbity score of 1.0 approximately corresponds to an X-ray resolution of 1.0 Å). On average, by using our AMOEBA and the side-chain repacking algorithm, we reduced the number of steric clashes above 0.4Å per 1000 atoms from 43.2 to 1.6 and decreased the percentage of poor rotamers from 2.70% to 0.21% (a poor rotamer is one in which the side chain position is inconsistent with the majority of proteins in the protein data bank). The MolProbity statistics for all of the optimized comparative models are shown in Table 6. The complete list of all statistics for each model is shown in Table A1.

Table 6. Average MolProbity refinement statistics for all OtoSCOPE proteins considered in this work.

Sequence		Clash		Poor	MolProbity	
Identity	Model	Score	%tile	Rotamers	Score	%tile
	Original	43.2	31	2.70%	2.65	42
63.4%	FFX	1.6	99	0.21%	1.41	95

CHAPTER 5: CONCLUSIONS

5.1: Summary of GPU Accelerated Optimization

This work describes the massive parallelization of a global amino acid side-chain protein optimization algorithm, followed by application to understanding missense variants implicated in non-syndromic hearing loss. This optimization algorithm uses a polarizable force field to dramatically improve the structural quality of protein models as accessed by the MolProbity tool, which ultimately prepares the structures for use in downstream free energy simulation algorithms. Using the Parallel Java message-passing interface (MPI), we parallelized the optimization algorithm across compute nodes and achieved a linear speed-up as a function of node count. In addition to MPI parallelization, amino acid side-chain energy evaluations were evaluated using nVidia GPUs via the OpenMM library. Compared to using 1 to 4 nodes without GPUs, inclusion of 1 GPU per node achieved a 25-fold speed-up (i.e. 4 GPU nodes was 100x faster than one node with no GPU). We also showed that pruning unphysical, high self-energy rotamers provides an additional 2.3-fold speed-up. Our accelerated approach opens the door to routine use of advanced polarizable force fields during protein optimization for the first time.

We applied the algorithm to a set of SwissProt and ModBase comparative protein models that are implicated in non-syndromic hearing loss. According to MolProbity, the average quality of the protein set before optimization ranked in the 42nd percentile compared to structures in the PDB. After optimization, the average quality of the protein set ranked in the 95th percentile, demonstrating that our algorithm effectively optimizes the protein structures and prepares the models for use in free energy simulations studies.

5.2: Future Directions and Alternative Applications

Advances in genetic sequencing platforms, such as the clinical tool OtoSCOPE used to understand non-syndromic hearing loss, have exposed genetic heterogeneity in many human diseases. Fully understanding and annotating the thousands of genetic variants sequenced by OtoSCOPE remains a challenge; however, simulation techniques can be used to increase our understanding of missense variations at a larger scale than is possible experimentally. For example, beginning from the models described in this work, its possible calculate thermodynamic changes in protein folding stability or protein-protein binding affinity caused by genetic variants. Such protein phenotypes can help to explain patient phenotypes and support clinical diagnostics. Overall, the optimized structures in this work lay a foundation for using high-quality protein models in thermodynamic variant analysis simulations. Future work will include 1) updating and optimizing the collection of proteins implicated in deafness as new structures are solved, comparatively modeled, or created from *ab initio* techniques and 2) using thermodynamic simulations with a polarizable force field to study variations discovered by OtoSCOPE.

APPENDIX

Table A1. The complete list of MolProbity statistics for all OtoSCOPE proteins studied in this work. The gene and residue ranges being modeled are shown in the two left-most columns, followed by the sequence identity to the template experimental structure the model was based off of, and then output from the MolProbity scoring algorithm.

	Residue	**Sequence**		**Clash**		**Poor**	**MolProbity**	
Gene	**Range**	**Identity**	**Model**	**Score**	**%tile**	**Rotamers**	**Score**	**%tile**
ACTG1	6-375	100%	Original	7.78	83	2.88%	2.2	64
			FFX	2.44	99	0.00%	1.48	96
CDH23	24-233	98%	Original	4.03	96	0.00%	1.19	99
			FFX	0.93	99	0.00%	0.99	100
	418-537	39%	Original	161.21	0	0.92%	3.24	15
			FFX	1.05	99	0.00%	1.55	94
	586-817	33%	Original	65.85	1	3.43%	3.53	8
			FFX	0.86	99	0.49%	1.55	94
	934-1312	32%	Original	117.75	0	2.74%	3.49	9
			FFX	2.75	98	0.00%	1.79	86
	1358-1567	34%	Original	64.38	1	2.81%	3.41	10
			FFX	1.59	99	0.56%	1.66	90
	1628-1742	32%	Original	39.36	8	1.96%	3.04	21
			FFX	2.28	99	0.00%	1.81	85
	1845-1935	38%	Original	35.23	10	0.00%	2.77	33
			FFX	0.72	99	0.00%	1.45	96
	2063-2172	40%	Original	50.96	3	0.00%	2.93	25
			FFX	3.6	97	1.05%	1.77	87
	2231-2340	34%	Original	31.06	14	4.17%	3.03	22
			FFX	1.19	99	0.00%	1.55	94
	2396-2503	39%	Original	128.54	0	2.13%	3.44	10
			FFX	1.84	99	0.00%	1.66	90
	2503-2605	39%	Original	37.45	9	3.33%	3.2	16

Table A1 – Continued.

	Residue	Sequence		Clash		Poor	MolProbity	
Gene	Range	Identity	Model	Score	%tile	Rotamers	Score	%tile
COCH	165-281	33%	Original	45.01	5	1.05%	2.51	47
			FFX	2.2	99	0.00%	1.3	98
	365-516	32%	Original	37.7	9	0.80%	2.62	40
			FFX	0.88	99	0.00%	1.45	96
DFNB31	132-226	96%	Original	25.1	21	11.11%	3.47	9
			FFX	0.7	99	0.00%	1.45	96
	264-378	99%	Original	5.64	92	13.83%	2.95	25
			FFX	0.56	99	0.00%	1.41	97
	813-904	98%	Original	11.1	66	2.70%	2.44	51
			FFX	0	100	0.00%	0.71	100
DIAPH1	92-452	91%	Original	16.09	45	5.49%	2.73	35
			FFX	2.05	99	0.30%	1.35	98
	762-1215	92%	Original	5.91	91	7.28%	2.49	48
			FFX	1.21	99	0.73%	1.3	98
ESPN	9-336	31%	Original	67.8	1	1.20%	3.08	20
			FFX	1.66	99	0.40%	1.55	94
ESRRB	97-186	99%	Original	3.48	97	2.67%	2.05	73
			FFX	1.39	99	0.00%	1.56	93
	211-433	80%	Original	9.02	77	0.51%	1.48	96
			FFX	0.55	99	0.00%	0.74	100
EYA4	369-639	77%	Original	18.96	35	3.88%	2.64	39
			FFX	2.11	99	0.00%	1.5	95
GPIC3	108-196	61%	Original	13.59	56	0.00%	1.96	78
			FFX	0.72	99	0.00%	1.34	98
GJB2	2-217	93%	Original	31.11	14	5.13%	3.22	16
			FFX	1.12	99	0.51%	1.39	97

Table A1 – Continued.

Gene	Residue Range	Sequence Identity	Model	Clash Score	Clash %tile	Poor Rotamers	MolProbity Score	MolProbity %tile
GJB3	2-210	56%	Original	52.92	3	2.69%	3.21	16
			FFX	3.19	97	0.00%	1.71	89
GJB6	2-216	74%	Original	46.2	5	4.10%	3.31	13
			FFX	2.25	99	0.00%	1.58	93
GPSM2	20-381	98%	Original	5.05	94	1.05%	1.28	99
			FFX	2.34	99	0.00%	1.01	100
	594-648	95%	Original	15.66	47	6.00%	2.72	35
			FFX	0.58	99	0.00%	1.16	99
HGF	34-289	99%	Original	5.86	91	1.72%	1.95	78
			FFX	1.95	99	0.43%	1.47	96
	305-470	45%	Original	50.06	4	0.00%	2.79	32
			FFX	4.32	96	0.67%	1.85	83
	495-721	100%	Original	8.21	80	0.00%	1.68	90
			FFX	2.83	98	0.52%	1.26	99
HOMER2	3-111	91%	Original	10.34	70	0.00%	2.11	69
			FFX	2.3	99	0.00%	1.44	96
KCNQ4	174-332	31%	Original	68.91	1	0.00%	3.02	22
			FFX	1.29	99	0.00%	1.39	97
LRTOMT	78-143	42%	Original	60.32	2	0.00%	2.89	27
			FFX	0	100	0.00%	1.24	99
MARVELD2	441-548	31%	Original	40.02	8	0.00%	2.08	71
			FFX	1.08	99	0.00%	0.81	100
MSRB3	49-166	62%	Original	21.17	30	0.00%	2.12	69
			FFX	0.56	99	0.98%	1.25	99
MYH9	7-806	82%	Original	83.62	0	15.06%	4.08	2
			FFX	3.56	97	0.14%	1.69	90

Table A1 – Continued.

	Residue	Sequence		Clash		Poor	MolProbity	
Gene	Range	Identity	Model	Score	%tile	Rotamers	Score	%tile
MYH14	49-799	98%	Original	20.87	30	2.95%	2.62	40
			FFX	1.41	99	0.16%	1.44	96
MYO15A	1222-1838	46%	Original	65.34	1	0.19%	2.74	34
			FFX	1.51	99	0.00%	1.43	97
	2871-2956	38%	Original	130.34	0	1.41%	3.41	10
			FFX	2.9	98	0.00%	1.81	85
MYO3A	16-287	53%	Original	33.82	12	0.00%	2.48	48
			FFX	2.07	99	0.00%	1.49	95
	323-996	36%	Original	71.37	1	0.17%	2.78	32
			FFX	1.39	99	0.33%	1.42	97
MYO6	2-825	98%	Original	6.2	90	4.38%	2.2	65
			FFX	1.59	99	0.55%	1.21	99
	840-992	93%	Original	20.25	31	10.26%	2.67	38
			FFX	1.45	99	0.00%	1.39	97
	1175-1277	98%	Original	2.34	99	2.22%	1.28	99
			FFX	2.92	98	0.00%	1.08	100
MYO7A	60-686	44%	Original	71.86	1	0.18%	2.71	36
			FFX	1.19	99	0.90%	1.4	97
	817-935	33%	Original	37.32	9	3.03%	3	23
			FFX	0.96	99	0.00%	1.51	95
	993-1686	84%	Original	29.16	16	2.16%	2.6	41
			FFX	1	99	0.17%	1.28	98
OTOF	1-124	91%	Original	8.52	79	1.77%	1.93	79
			FFX	0.5	99	0.00%	0.88	100
	1494-1574	34%	Original	135.01	0	7.14%	4.05	2
			FFX	1.57	99	0.00%	1.70	89

Table A1 – Continued.

Gene	Residue Range	Sequence Identity	Model	Clash Score	 %tile	Poor Rotamers	MolProbity Score	 %tile
PCDH15	27-255	95%	Original	10.89	67	0.00%	1.55	94
			FFX	0.56	99	0.00%	0.92	100
POU3F4	189-338	77%	Original	128.08	0	7.52%	3.81	4
			FFX	1.23	99	0.00%	1.28	99
POU4F3	186-332	49%	Original	171.95	0	4.03%	3.65	6
			FFX	1.25	99	0.00%	1.5	95
PRPS1	3-317	98%	Original	12.94	58	3.00%	2.32	58
			FFX	1.64	99	0.00%	1.45	96
RDX	1-325	90%	Original	14.65	51	5.42%	2.75	34
			FFX	1.28	99	0.00%	1.35	98
	494-583	72%	Original	7.62	83	1.28%	1.72	88
			FFX	0.69	99	0.00%	0.94	100
SLC26A4	516-577	35%	Original	25.46	21	0.00%	2.11	70
			FFX	1.02	99	0.00%	1.49	95
	620-727	34%	Original	54.7	3	0.00%	2.71	36
			FFX	1.71	99	0.00%	1.46	96
SLC26A5	505-718	48%	Original	38.99	8	1.67%	2.65	39
			FFX	1.8	99	0.00%	1.50	95
TMPRSS3	217-449	46%	Original	34.13	11	0.52%	2.48	48
			FFX	1.14	99	0.52%	1.41	97
USH1C	1-192	99%	Original	3.2	97	1.72%	1.29	98
			FFX	0.64	99	0.57%	0.71	100
	441-552	95%	Original	22.25	27	10.11%	3.42	10
			FFX	1.76	99	0.00%	1.61	92

Table A1 – Continued.

Gene	Residue Range	Sequence Identity	Model	Clash Score	Clash %tile	Poor Rotamers	MolProbity Score	MolProbity %tile
USH1G	7-149	38%	Original	56.9	2	0.85%	2.75	34
			FFX	1.79	99	1.69%	1.59	93
	388-461	99%	Original	2.51	98	1.54%	1.18	99
			FFX	1.67	99	0.00%	0.92	100
USH2A	326-728	34%	Original	83.63	0	0.83%	3.13	18
			FFX	3.27	97	0.83%	1.79	85
	768-901	34%	Original	57.42	2	0.85%	2.85	29
			FFX	2.52	98	1.69%	1.93	79
	922-1052	33%	Original	51.43	3	1.80%	2.93	26
			FFX	1.61	99	0.00%	1.64	91
	1716-1871	31%	Original	78.58	0	3.05%	3.38	11
			FFX	2.89	98	0.00%	1.75	87

REFERENCES

[1] Shearer, A. E., Black-Ziegelbein, E. A., Hildebrand, M. S., Eppsteiner, R. W., Ravi, H., Joshi, S., Guiffre, A. C., Sloan, C. M., Happe, S., Howard, S. D., Novak, B., DeLuca, A. P., Taylor, K. R., Scheetz, T. E., Braun, T. A., Casavant, T. L., Kimberling, W. J., LeProust, E. M., and Smith, R. J. H. (2013) Advancing genetic testing for deafness with genomic technology, *J. Med. Genet. 50*, 627-634.

[2] Shearer, A. E., and Smith, R. J. H. (2012) Genetics: advances in genetic testing for deafness, *Curr. Opin. Pediatr. 24*, 679-686.

[3] Shearer, A. E., Eppsteiner, Robert W., Booth, Kevin T., Ephraim, Sean S., Gurrola, J., Simpson, A., Black-Ziegelbein, E. A., Joshi, S., Ravi, H., Giuffre, Angelica C., Happe, S., Hildebrand, Michael S., Azaiez, H., Bayazit, Yildirim A., Erdal, Mehmet E., Lopez-Escamez, Jose A., Gazquez, I., Tamayo, Marta L., Gelvez, Nancy Y., Leal, Greizy L., Jalas, C., Ekstein, J., Yang, T., Usami, S.-i., Kahrizi, K., Bazazzadegan, N., Najmabadi, H., Scheetz, Todd E., Braun, Terry A., Casavant, Thomas L., LeProust, Emily M., and Smith, Richard J. H. (2014) Utilizing Ethnic-Specific Differences in Minor Allele Frequency to Recategorize Reported Pathogenic Deafness Variants, *The American Journal of Human Genetics 95*, 445-453.

[4] Ephraim, S. S., Anand, N., DeLuca, A. P., Taylor, K. R., Kolbe, D. L., Simpson, A. C., Azaiez, H., Sloan, C. M., Shearer, A. E., Hallier, A. R., Casavant, T. L., Scheetz, T. E., Smith, R. J. H., and Braun, T. A. (2014) Cordova: Web-based management of genetic variation data, *Bioinformatics 30*, 3438-3439.

[5] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The protein data bank, *Nucleic Acids Res. 28*, 235-242.

[6] Bienert, S., Waterhouse, A., de Beer, Tjaart A. P., Tauriello, G., Studer, G., Bordoli, L., and Schwede, T. (2017) The SWISS-MODEL Repository—new features and functionality, *Nucleic Acids Res. 45*, D313-D319.

[7] Pieper, U., Webb, B. M., Dong, G. Q., Schneidman-Duhovny, D., Fan, H., Kim, S. J., Khuri, N., Spill, Y. G., Weinkam, P., Hammel, M., Tainer, J. A., Nilges, M., and Sali, A. (2014) ModBase, a database of annotated comparative protein structure models and associated resources, *Nucleic Acids Res. 42*, D336-D346.

[8] Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters, *Proteins: Struct., Funct., Bioinf. 65*, 712-725.

[9] Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005) The AMBER biomolecular simulation programs, *J. Comput. Chem. 26*, 1668-1688.

[10] Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009) CHARMM: The biomolecular simulation program, *J. Comput. Chem. 30*, 1545-1614.

[11] MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins, *J. Phys. Chem. B 102*, 3586-3616.

[12] Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides, *J. Phys. Chem. B 105*, 6474-6487.

[13] Jorgensen, W. L., and Tirado-Rives, J. (1988) The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin, *J. Am. Chem. Soc. 110*, 1657-1666.

[14] Shi, Y., Xia, Z., Zhang, J., Best, R., Wu, C., Ponder, J. W., and Ren, P. (2013) Polarizable atomic multipole-based AMOEBA force field for proteins, *J. Chem. Theory Comput. 9*, 4046-4063.

[15] Ponder, J. W., Wu, C., Ren, P., Pande, V. S., Chodera, J. D., Schnieders, M. J., Haque, I., Mobley, D. L., Lambrecht, D. S., DiStasio, R. A., Head-Gordon, M., Clark, G. N. I., Johnson, M. E., and Head-Gordon, T. (2010) Current status of the AMOEBA polarizable force field, *The Journal of Physical Chemistry B 114*, 2549-2564.

[16] Lemkul, J. A., Huang, J., Roux, B., and MacKerell, A. D. (2016) An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications, *Chem. Rev. 116*, 4983-5013.

[17] Schnieders, M. J., and Ponder, J. W. (2007) Polarizable atomic multipole solutes in a generalized Kirkwood continuum, *J. Chem. Theory Comput. 3*, 2083-2097.

[18] Schnieders, M. J., Baker, N. A., Ren, P. Y., and Ponder, J. W. (2007) Polarizable atomic multipole solutes in a Poisson-Boltzmann continuum, *J. Chem. Phys. 126*, 124114.

[19] LuCore, Stephen D., Litman, Jacob M., Powers, Kyle T., Gao, S., Lynn, Ava M., Tollefson, William T. A., Fenn, Timothy D., Washington, M. T., and Schnieders, Michael J. (2015) Dead-End Elimination with a Polarizable Force Field Repacks PCNA Structures, *Biophys. J. 109*, 816-826.

[20] Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2009) MolProbity: all-atom structure validation for macromolecular crystallography, *Acta Crystallogr. D 66*, 12-21.

[21] Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S., and Richardson, D. C. (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids, *Nucleic Acids Res. 35*, W375-W383.

[22] Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., and Schwede, T. (2009) The SWISS-MODEL Repository and associated resources, *Nucleic Acids Res. 37*, D387-D392.

[23] Kaminsky, A. (2007) Parallel Java: a unified API for shared memory and cluster parallel programming in 100% Java, In *2007 IEEE International Parallel and Distributed Processing Symposium*, IEEE, Long Beach, CA, USA.

[24] Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y. T., Beauchamp, K. A., Wang, L. P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., and Pande, V. S. (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics, *PLoS Comput. Biol. 13*, 17.

[25] Shearer A. E., DeLuca A. P., Hildebrand M. S., Taylor K. R., Gurrola J., 2nd, Scherer S., Scheetz T. E., Smith R. J. Comprehensive genetic testing for hereditary hearing loss using massively parallel sequencing. Proc Natl Acad Sci U S A. 2010; 107(49):21104–9.

[26] Balaji, S., & Srinivasan, N. (2007). Comparison of sequence-based and structure-based phylogenetic trees of homologous proteins: Inferences on protein evolution. *Journal of Biosciences,* 32(1):83-96.

[27] Westheimer, F. H., & Mayer, J. E. (1946). The Theory of the Racemization of Optically Active Derivatives of Diphenyl. *The Journal of Chemical Physics,* 14(12), 733-738. doi: http://dx.doi.org/10.1063/1.1724095

[28] Ren, P. Y., & Ponder, J. W. (2003). Polarizable atomic multipole water model for molecular mechanics simulation. *Journal of Physical Chemistry B, 107*(24), 5933-5947. doi:10.1021/jp027815+

[29] Fenn, T. D., and Schnieders, M. J. (2011) Polarizable atomic multipole X-ray refinement: Weighting schemes for macromolecular diffraction, *Acta Crystallogr. D 67*, 957-965.

[30] LuCore, Stephen D., Litman, Jacob M., Powers, Kyle T., Gao, S., Lynn, Ava M., Tollefson, William T. A., Fenn, Timothy D., Washington, M. T., and Schnieders, Michael J. (2015) Dead-End Elimination with a Polarizable Force Field Repacks PCNA Structures, *Biophys. J. 109*, 816-826.

[31] Desmet, J., Maeyer, M. D., Hazes, B., and Lasters, I. (1992) The dead-end elimination theorem and its use in protein side-chain positioning, *Nature 356*, 539-542.

[32] Goldstein, R. F. (1994) Efficient rotamer elimination applied to protein side-chains and related spin glasses, *Biophys. J. 66*, 1335-1340.

[33] Fiser, A., Feig, M., Brooks, C. L., and Sali, A. (2002) Evolution and physics in comparative protein structure modeling, *Acc. Chem. Res. 35*, 413-421.

[34] Arnold, K., Kiefer, F., Kopp, J., Battey, J. N. D., Podvinec, M., Westbrook, J. D., Berman, H. M., Bordoli, L., and Schwede, T. (2009) The Protein Model Portal, *Journal of Structural and Functional Genomics 10*, 1-8.

[35] Eramian, D., Eswar, N., Shen, M.-Y., and Sali, A. (2008) How well can the accuracy of comparative protein structure models be predicted?, *Protein Sci. 17*, 1881-1893.

[36] Broyden, C. G. (1970) The convergence of a class of double-rank minimization algorithms 1. General considerations, *IMA J. Appl. Math 6*, 76-90.

[37] Fletcher, R. (1970) A new approach to variable metric algorithms, *Computer J. 13*, 317-322.

[38] Goldfarb, D. (1970) A family of variable-metric methods derived by variational means, *Math. Comput. 24*, 23-26.

[39] Shanno, D. F. (1970) Conditioning of quasi-newton methods for function minimization, *Math. Comput. 24*, 647-656.

[40] LuCore, S. D., Litman, J. M., Powers, K. T., Lynn, A. M., Gao, S., Tollefson, W. T. A., Fenn, T. D., Washington, M. T., and Schnieders, M. J. (2015) Dead-End Elimination for Many-Body Potentials Repacks Low-Resolution X-ray Diffraction Models into Atomic Resolution Structures, *Proc. Natl. Acad. Sci. U.S.A. (submitted)*.