Ab initio ground state energy calculations for protein design with continuous flexibility

Hunter Stephens

April 22, 2021

1 Introduction

Proteins preform a myriad of functions within the body from providing structural support to defending against foreign bodies. One of the major factors in determining the function is the three dimensional structure of the folded protein. Protein design and engineering seeks to find a protein sequence that results in a desired structure and function [1]. This is a well defined computational problem that can be solved with a conformational space model and energy function [1]. Initial attempts at solving the problem involved placing two constraints on the formulations. The first is that the conformational space be discrete, being compromised of a selected number of residues and conformations called rotamers [2]. The second is that the energy function be residue pairwise in that it is a sum of terms that depend on the amino acid type and conformation at no more than two residue sites. Several algorithms have been developed to solve this problem including the DEE algorithm [3], the A* algorithm [4], the DEE/A* algorithm [5, 6], and the COMETS algorithm [7].

In reality, proteins are generally flexible in both side chains [8] and backbone [9]. That flexibility can provide significantly lower energies than those provided by a discrete set of rotamers. It is common that an ideal rotamer be physically unrealizable due to nuclear repulsion, but for small perturbations to lead to an optimal confirmation [8, 10, 6]. The optimal sequence is often significantly different and more realizable when the protein is considered to be flexible [8, 10]. One of the most recent approaches to including continuous flexibility and non-pairwise energy functions into the protein design problem incorporates machine learning and reduces the continuous design problem to a discrete one by fitting the energy to a from suitable for discrete methods. The EPIC algorithm [11] learns a polynomial expansion of the continuous energy surface and the LUTE algorithm [12] learns a pairwise energy matrix. Both of these algorithms reduce the energy function call bottleneck by reducing the number of function calls allowing for more accurate *ab initio* energy calculations to be incorporated. In theory, any energy formulation could be learned and fitted by both EPIC and LUTE. In addition to this fit, knowing bounds on the conformational energy is needed for the design algorithm and in the case for use of the iMinDEE [13] algorithm a pruning interval is needed in addition to a pair-wise lower bound matrix.

Both the EPIC and LUTE algorithms rely on energy function calls, $E(\mathbf{x})$, where $\mathbf{x}$ represents the internal continuous coordinates of the protein system, namely the positional and spin coordinates $\{\mathbf{r}, \sigma\}$. We will now write,

$$\min E(\mathbf{x}) = \langle \Psi_o | \, H \, | \Psi_o \rangle = \xi_o(\mathbf{x}) \tag{1}$$

We have replaced the energy function with an energy functional where the energy of a conformation is the ground state energy of the molecular Hamiltonian. Using the Born-Oppenheimer (BO) approximation we can consider the eigenfunctions as the product of a nuclear and electronic wave function allowing us to neglect the nuclear kinetic energies. This is possible due to the relative masses of electrons to nuclei ($m_e/M \approx 10^{-3} - 10^{-5}$). With this we can write, H as

$$H = -\frac{1}{2}\sum_i \nabla_i^2 - \sum_i \sum_\alpha \frac{Z_\alpha}{|\mathbf{r}_i - \mathbf{R}_\alpha|} + \frac{1}{2}\sum_i \sum_{j \neq i} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{2}$$

where $\mathbf{r}_i$ represents electron coordinates, $\mathbf{R}_\alpha$ are the nuclear coordinates, and Z_α are the nuclear charges. We have set $e = m_e = \hbar = 4\pi\epsilon_o = 1$. The ground state energy is an eigenvalue with corresponding eigenfunction for the following relation,

$$H\left|\Psi_o\right\rangle = \xi_o\left|\Psi_o\right\rangle \tag{3}$$

For the given H in (2), the eigenvalues and eigenfunctions cannot be calculated exactly and numerical techniques are needed.

One of the most powerful ideas behind the numerical calculations of the ground state energy is the variational theorem. Given some trial-wavefunction, $\left|\tilde{\Psi}\right\rangle$, the variational theorem states,

$$\xi_o \leq \frac{\left\langle\tilde{\Psi}\right| H \left|\tilde{\Psi}\right\rangle}{\left\langle\tilde{\Psi}\middle|\tilde{\Psi}\right\rangle} \tag{4}$$

where ξ_o is the true ground state energy of the system. This allows for a couple of things. The first is that we can now approximate the ground state energy by making an *ansatz* for the wavefunction and then minimizing. Most methods use a linear superposition of orbital basis functions where $\left|\tilde{\Psi}\right\rangle$ takes the form of,

$$\left|\tilde{\Psi}\right\rangle = \sum_{i=0}^{N} c_i\left|\psi_i\right\rangle \tag{5}$$

The second revelation from the variational theorem is that since our minimized ground state energy is always equal to or greater than the true energy we know that any method or basis producing a lower energy is a better approximation. Thus, the Hartree-Fock limit will be an upper-bound for calculations of energies using post-Hartree-Fock methods. In this project, we will introduce the implementation of *ab initio* energy calculations and explore the idea of bounding these energy calculations as well reducing the computational cost for calculations on whole proteins. By no means is what presented a comprehensive study, a lot of time and effort was put into understanding the *de novo* protein design problem as well as the chemical physics of *ab initio* calculations. What this does provide, is a starting point and possible road-map for future work and development. This will be highlighted more in the discussion section.

2 Methods

2.1 Side-chain flexibility

We considered a single flexible residue, cysteine, with two degrees of freedom. The degrees of freedom were the two χ angles depicted in Figure 1. The modelling and transformation of the residue were performed using the Biotite library [14]. Energy calculations were performed on voxels within the conformational space of the residue using the Psi4 package [15]. We focused on three methods, namely Hartree-Fock (HF) [16], Møller-Plesset perturbation theory (MPPT)[17], and Density Functional Theory (DFT)[18]. For MPPT we focused only on second order corrections (MP2) as seemed to be the norm in the literature. A custom script was written to convert the Biotite structure to a Psi4 input file. This is easily extended to take as input a pdb file for larger protein systems. Figure 2 shows the energy profiles of the first χ angle for the three methods. Each profile is similar in shape as seen by the energy minimum translated plot, but with differing magnitudes. HF as the simplest and less accurate acts as an upper bound on more accurate methods. This is know as the Hartree-Fock limit. While not as accurate, HF provides a very fast estimate of the energy upper bound of the system. Careful consideration must be placed on this however. Methods like DFT are not based on the HF idea of minimizing a linear combination of orbitals. Thus, the HF method can only provide a provable upper bound for post-HF methods, like MP2. In practice, DFT is more accurate than HF, but may not be able to be provably bounded above by the HF limit. It can be seen from Figures 2 and 3 that the energy landscape is quite smooth with definite extrema.

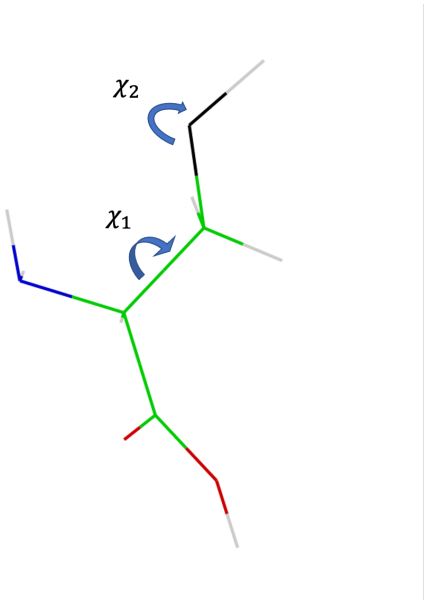

Figure 1: Cysteine with two internal degrees of freedom as rotations around two bonds of the side-chain.

2.2 Nuclear perturbations

These methods are extremely expensive for large atomic systems, like proteins. Transforming the energy into a residue pairwise form with methods like LUTE could quickly become intractable. With this in mind, we began to investigate ways to reduce the computational cost while maintaining good accuracy on the energy or energy bounds. This could be moot if available computational power allows for LUTE or EPIC to completely handle the learning of the energy space. To summarize, we will describe the flexibility of the residues as perturbations of the atomic structure. In essence, we will perturb a set of nuclei, freeze the remaining orbitals, and then find the energy minimum of the new perturbed potential. It must be noted that presented here is only the derivation. More work is needed for this to implemented.

First, consider a set of perturbations on a subset of the nuclei given by $\{\delta R_A\}$, where $R_a = R_{A,o} + \delta R_A$. In our electronic Hamiltonian, only the nuclear-electron coulomb potential would be affected by the perturbation. First, let us write out the full potential

$$V_{eN} = -\sum_A V_{eN}^A = -\sum_A \sum_i \frac{Z_A}{|\mathbf{r}_i - \mathbf{R}_A|} \tag{6}$$

Without a loss of generality, we will consider only one of the perturbed nuclei in the following derivation and drop all subscripts and superscripts. First, expand the potential around $R_{A,o}$, $V(R_A) = V(R_{A,o}) + \delta R_A \cdot \nabla V(R_{A,o}) + O(\delta R_A^2)$. The, the gradient term is simply the electric field of the surrounding electrons on the original nuclear position. This is very interesting as the first order correction to the potential is the work needed to move the nucleus in the external electric field of the electrons. This is seen by the perturbation dotted into the electric field term. So we can write as our new Hamiltonian,

$$\tilde{H} = H - \sum_A \delta R_A E(R_{A,o}) = H_0 + H_1 \tag{7}$$

Writing our Hamiltonian in this form allows for the use of time-independent perturbation theory where $\xi_o = \xi_o^{(0)} + \xi_o^{(1)} + ...$ and

$$\xi_o^{(1)} = \langle \Psi_o | H_1 | \Psi_o \rangle \leq \frac{\left\langle \tilde{\Psi} \middle| H_1 \middle| \tilde{\Psi} \right\rangle}{\left\langle \tilde{\Psi} \middle| \tilde{\Psi} \right\rangle} \tag{8}$$

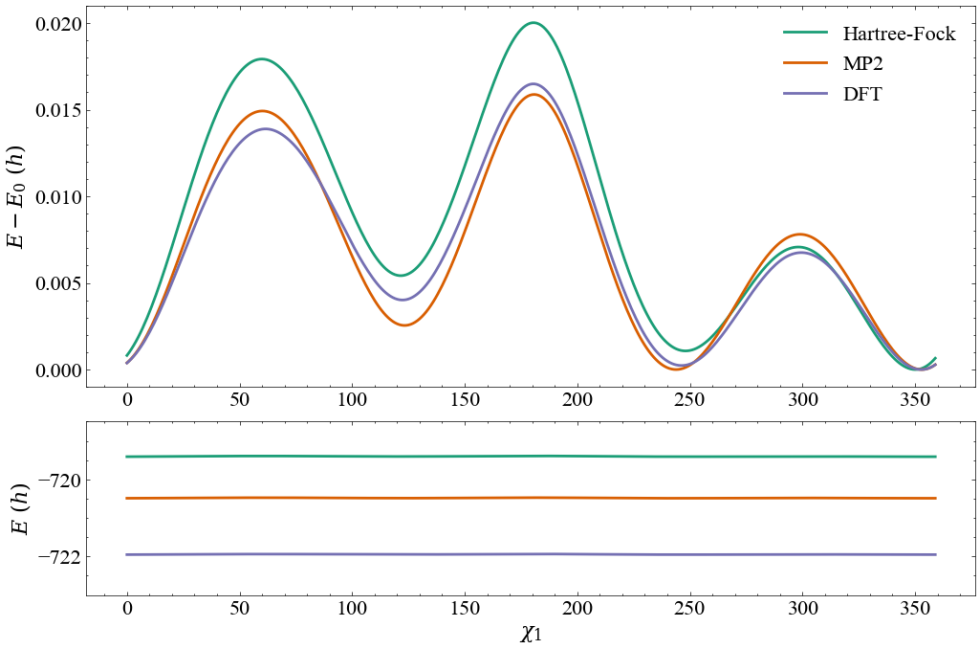

Figure 2: Comparison of the calculated energy versus the first internal degree of freedom for cysteine. (top) Energy profiles translated to the minimum energy. (bottom) absolute energies for the various methods demonstrating the idea of less accurate methods bounding more accurate methods.

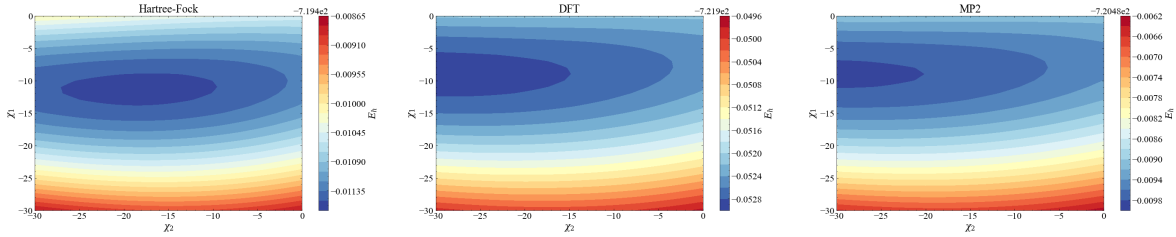

Figure 3: Comparison of the calculated energy surfaces of the first two χ angles of cysteine

Now we have something similar to our original variational theorem except now we are calculating the first order perturbative correction. From here we can now focus on minimizing the right side of the inequality. A major difference in H_1 compared to H_0 is the replacing of the scalar potential with the electric field. Thus, our molecular integration will be different from the developed technology. Our problem comes down to the calculation of the following integral,

$$\int_{-\infty}^{\infty} d^3r \, \langle \mathbf{r}|\psi\rangle^2 \, \frac{\delta\mathbf{R}\cdot(\mathbf{r}-\mathbf{R})}{|\mathbf{r}-\mathbf{R}|^3} \tag{9}$$

2.3 Minimization Problem

In this section, we will formalize the statement of the minimization problem involved in reference to equation (8). We first let the basis containing the full spanning set of molecular orbitals be labeled as, $\mathcal{P} = \{|\psi^A_{nlm}\rangle\}$. Where we have used A to highlight the atom for which the orbital is centered and subscripted the orbital's quantum numbers. In practice, we will use contracted Gaussian-type orbitals (GTOs) where the nlm will be replaced by ijk, a. Now, we will let $\mathcal{A}$ be the set of nuclei that are perturbed. We can then define,

$\tilde{\mathcal{P}} = \{|\psi^A_{nlm}\rangle : A \in \mathcal{A}\}$. Our minimization problem is then given by,

$$\min \sum_{|\psi^A_{ijk,a}\rangle \in \tilde{\mathcal{P}}} \frac{\int_{-\infty}^{\infty} d^3r \left\langle \mathbf{r} \middle| \psi^A_{ijk,a} \right\rangle^2 \frac{\delta \mathbf{R}_a \cdot (\mathbf{r}-\mathbf{R}_A)}{|\mathbf{r}-\mathbf{R}_A|^3}}{\int_{-\infty}^{\infty} d^3r \left\langle \mathbf{r} \middle| \psi^A_{ijk,a} \right\rangle^2} \tag{10}$$

Our task now is in solving the integrals in (10). We will use as our basis functions contracted GTOs. They are defined as,

$$\langle \mathbf{r} | \psi^A_{ijk,a} \rangle = x_A^i y_A^j z_A^k \exp\left(-a\mathbf{r}_A^2\right) \tag{11}$$

Thus, the denominator of (10) evaluates to,

$$\int_{-\infty}^{\infty} d^3r \langle \mathbf{r} | \psi^A_{ijk,a} \rangle^2 = (\frac{\pi}{2a})^{3/2} \frac{(2i-1)!!(2j-1)!!(2k-1)!!}{(4a)^{i+j+k}} \tag{12}$$

We next need to solve the integral in the numerator, and write

$$\begin{aligned} &\int_{-\infty}^{\infty} d^3r \langle \mathbf{r} | \psi^A_{ijk,a} \rangle^2 \frac{\delta \mathbf{R}_a \cdot (\mathbf{r}-\mathbf{R}_A)}{|\mathbf{r}-\mathbf{R}_A|^3} = \\ &\int_{-\infty}^{\infty} dz z^{2k} \exp\left(-2az^2\right) \left(\int_{-\infty}^{\infty} dy y^{2j} \exp\left(-2ay^2\right) \left(\int_{-\infty}^{\infty} dx x^{2j} \exp\left(-2ax^2\right) \frac{\delta \mathbf{R}_a \cdot (\mathbf{r}-\mathbf{R}_A)}{|\mathbf{r}-\mathbf{R}_A|^3}\right)\right) \end{aligned} \tag{13}$$

From which we can solve by solving the following integral family,

$$F_n(x) = \int_{-\infty}^{\infty} \frac{x^{2n} \exp\left(-2ax^2\right)}{((x-C_1)^2 + C_2^2)^{3/2}} dx \tag{14}$$

The $\frac{1}{r^3}$ from the electric field operator compared to the $\frac{1}{r}$ from the coulomb potential operator necessitates a different solution than what has previously been solved and used in current *ab initio* methods.

3 Discussion and Future Work

This work is significant in several ways. For one, it provides a way to bound energy calculations in a way that is useful for Dead-end elimination algorithms. The Hartree-Fock method provides a less expensive (compared to other methods) energy upper-bound of the conformation. This upper-bound can act as a lower-bound for the pruning criteria of the combinatorial protein design algorithms. Furthermore, freezing non-perturbed orbitals significantly cuts down on the computational cost of exploring the energy surface of a residue's conformational space. Our perturbation method is also residue separable. That is, the sum of orbitals can be grouped based on the residue it is a member of. This could be useful in getting the energy into a residue-pairwise form. More work is needed, however, in carrying these methods out and putting them all together. As it stands, this is more of a scattering of tools and stray ideas that need to be focused and organized. Further work into calculating the free energy of the whole protein system in an efficient way is warranted as well. The use of capping segments of the protein and calculating contributions has been used in interaction energy calculations and could be useful here.

References

[1] B. Donald. "Algorithms in Structural Molecular Biology". In: 2011.

[2] Joël Janin et al. "Conformation of amino acid side-chains in proteins". In: *Journal of Molecular Biology* 125.3 (1978), pp. 357–386. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(78)90408-4. URL: https://www.sciencedirect.com/science/article/pii/0022283678904084.

[3] J. Desmet et al. "The dead-end elimination theorem and its use in protein side-chain positioning". In: *Nature* 356 (1992), pp. 539–542. DOI: https://doi.org/10.1038/356539a0.

[4] P. E. Hart, N. J. Nilsson, and B. Raphael. "A Formal Basis for the Heuristic Determination of Minimum Cost Paths". In: *IEEE Transactions on Systems Science and Cybernetics* 4.2 (1968), pp. 100–107. DOI: 10.1109/TSSC.1968.300136.

[5] Ryan H. Lilien et al. "A Novel Ensemble-Based Scoring and Search Algorithm for Protein Redesign and Its Application to Modify the Substrate Specificity of the Gramicidin Synthetase A Phenylalanine Adenylation Enzyme". In: *Journal of Computational Biology* 12.6 (2005). PMID: 16108714, pp. 740–761. DOI: 10.1089/cmb.2005.12.740. eprint: https://doi.org/10.1089/cmb.2005.12.740. URL: https://doi.org/10.1089/cmb.2005.12.740.

[6] Ivelin Georgiev, Ryan H. Lilien, and Bruce R. Donald. "The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles". In: *Journal of Computational Chemistry* 29.10 (2008), pp. 1527–1542. DOI: https://doi.org/10.1002/jcc.20909. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20909. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20909.

[7] Mark A. Hallen and Bruce R. Donald. "comets (Constrained Optimization of Multistate Energies by Tree Search): A Provable and Efficient Protein Design Algorithm to Optimize Binding Affinity and Specificity with Respect to Sequence". In: *Journal of Computational Biology* 23.5 (2016). PMID: 26761641, pp. 311–321. DOI: 10.1089/cmb.2015.0188. eprint: https://doi.org/10.1089/cmb.2015.0188. URL: https://doi.org/10.1089/cmb.2015.0188.

[8] Pablo Gainza, Kyle E. Roberts, and Bruce R. Donald. "Protein Design Using Continuous Rotamers". In: *PLoS Computational Biology* 8.1 (2012), e1002335. DOI: 10.1371/journal.pcbi.1002335. URL: https://app.dimensions.ai/details/publication/pub.1052975160%20and%20https://journals.plos.org/ploscompbiol/article/file?id=10.1371/journal.pcbi.1002335&type=printable.

[9] Mark A. Hallen, Daniel A. Keedy, and Bruce R. Donald. "Dead-end elimination with perturbations (DEEPer): A provable protein design algorithm with continuous sidechain and backbone flexibility". In: *Proteins: Structure, Function, and Bioinformatics* 81.1 (2013), pp. 18–39. DOI: https://doi.org/10.1002/prot.24150. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.24150. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.24150.

[10] Pablo Gainza, Kyle E. Roberts, and et al. "OSPREY: protein design with ensembles, flexibility, and provable algorithms". In: *Methods Enzymol* 523.1 (2013), pp. 87–107. DOI: doi:10.1016/B978-0-12-394292-0.00005-9.

[11] Mark A. Hallen, Pablo Gainza, and Bruce R. Donald. "Compact representation of continous energy surfaces for more efficient protein design". In: *J Chem Theory Comput* (2015), pp. 2292–2306. DOI: doi:10.1021/ct501031m.

[12] Mark A. Hallen, Jonathan D. Jou, and Bruce R. Donald. "LUTE (Local Unpruned Tuple Expansion): Accurate Continuously Flexible Protein Design with General Energy Funcions and Rigid Rotamer-Like Efficiency". In: *Journal of Computational Biology* (2017), pp. 536–546. DOI: doi:10.1089/cmb.2016.0136.

[13] Donald BR Gainza P Roberts KE. "Protein Design Using Continouse Rotatmers". In: *PLoS Comput Biol* (2012). DOI: 10.1371/journal.pcbi.1002335.

[14] K Kunzmann P. Hamacher. "Biotite: a unifying open source computational biology framework in Python". In: *BMC Bioinformatics* (2018). DOI: https://doi.org/10.1186/s12859-018-2367-z.

[15] Justin M. Turney et al. "Psi4: an open-source ab initio electronic structure program". In: *WIREs Computational Molecular Science* 2.4 (2012), pp. 556–565. DOI: https://doi.org/10.1002/wcms.93. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.93. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.93.

[16] V. Fock. "Naherungsmethode zur Losung des quantenmechanischen Mehrkorperproblems". In: *Zeitschrift für Physik* 61 (1930), pp. 126–148. DOI: http://dx.doi.org/10.1007/BF01340294.

[17] Milton S. Møller Christian; Plesset. "Note on an Approximation Treatment for Many-Electron Systems". In: *Phys. Rev.* 46 (1934).

[18] P. Hohenberg and W. Kohn. "Inhomogeneous Electron Gas". In: *Phys. Rev.* 136 (3B Nov. 1964), B864–B871. DOI: 10.1103/PhysRev.136.B864. URL: https://link.aps.org/doi/10.1103/PhysRev.136.B864.