

# **Computational Design and Modeling of Synthetic Enzymes: Biochemistry *in silico***

Harrison W. Kuhn

In partial fulfillment of CHEM 305, Fall 2020  
Jared R. Mays, Ph.D.

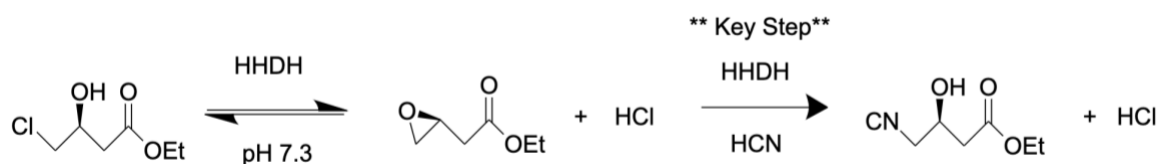
Proteins are an integral class of functional molecules essential to sustain life. They contribute to building physical structure, transporting molecules, and catalyzing reactions. Decades of research performed by biochemists around the world has contributed to understanding of the functional possibilities within a linear sequence of amino acids. Utilizing the wealth of accumulated knowledge and the rise of computational resources in recent years, enzyme structures are able to be proposed *in silico* and tested for functionality without even touching a micropipette. The exciting new field of computational biochemistry presents several advantages such as reduced physical laboratory requirements, reagents, and personnel needs, all while accelerating the development timeline and production of catalytic enzyme candidates. Three models of computational design, with differing requirements and challenges are presented in this work.

## **I. Beginnings and Evolutionary Methods**

A 1983 paper published in Nature by Pabo suggested that use of available X-ray diffraction data of backbone protein structures can seed the first instances of enzyme design and protein modification.<sup>1</sup> *de Novo* prediction of folding and resultant tertiary structures was considered impossible during the time period, however, an “inverted” approach was proposed to circumvent these limitations. Using a known structure and sequence as baseline compensated for the lack of long-range interaction information and thus allowed replacement of single amino acids with similar charge and structure properties to generate a desired function or structure.

Early protein engineering techniques serve as a springboard towards computational design of enzymes but do not directly answer the question posed by and able to be answered by *de Novo* design; “What specific sequence of amino acids gives exceptional catalytic activity?”

A process drawing inspiration from natural evolution on a broad time scale, laboratory directed evolution (LDE) enables the accelerated creation of new proteins and enzymes that evolve from standard model (Figure 3) generated precursors indicating slight affinity towards side reactions using an atypical substrate.<sup>2</sup> In an iterative process, mutations are introduced into protein sequences, developing new functionality, which are then screened with antibody assays for viable candidates, which are subsequently returned for additional rounds of mutagenesis.<sup>3</sup> At the end of mutagenesis, a refined enzyme shows activity for its new function. This technique in conjunction with ProSAR (**protein sequence activity relationships**) data has been proven to produce exceptional biocatalysts, most notably in the case of replacing traditional nucleophilic cyanation with (R)-4-cyano-3-hydroxybutyrate and a modified halohydrin dehydrogenase (HHDN) to produce precursors for synthesis of Lipitor at industrial efficiency levels.<sup>4</sup>



**Figure 1:** Modified HHDH catalyzed Cyanation scheme of (R)-4-cyano-3-hydroxybutyrate. This scheme is outlined by Fox et al. in their work to improve catalytic function of HHDN. Modified HHDN was taken from *Agrobacterium radiobacter* and recombinantly expressed in *Escherichia coli* where 18 subsequent rounds of LDE were performed to result in industrially viable HHDN catalyzing the reaction above.<sup>4</sup>

LDE is known to be most effective when viable candidates are initially given. Likened to refining crude oil, LDE directs (refines) computationally designed enzyme candidates with low activity to mutate into enzymes with higher activity, i.e. the resultant high-octane gasoline used

in internal combustion to do work.<sup>2</sup> These initial candidates are generated with data provided by modeling and simulation methods, outlined in the following section.




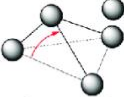


## II. Modeling Atomic and Electronic Interactions in Active Sites

The work of Thomas, Fermi, and Dirac provided fundamental modeling equations of kinetic energy resultant of systems containing many electrons.<sup>5</sup> These modeling functions have consolidated over time to what is known as density functional theory (DFT). In simplistic terms, DFT maps electron density over structures using an approximation of the renown Schrödinger equation (eqn. 1), a wavefunction that increases exponentially in complexity with additional electron count, quickly hoarding computational resources, and thus, limiting its effective use to small atoms.

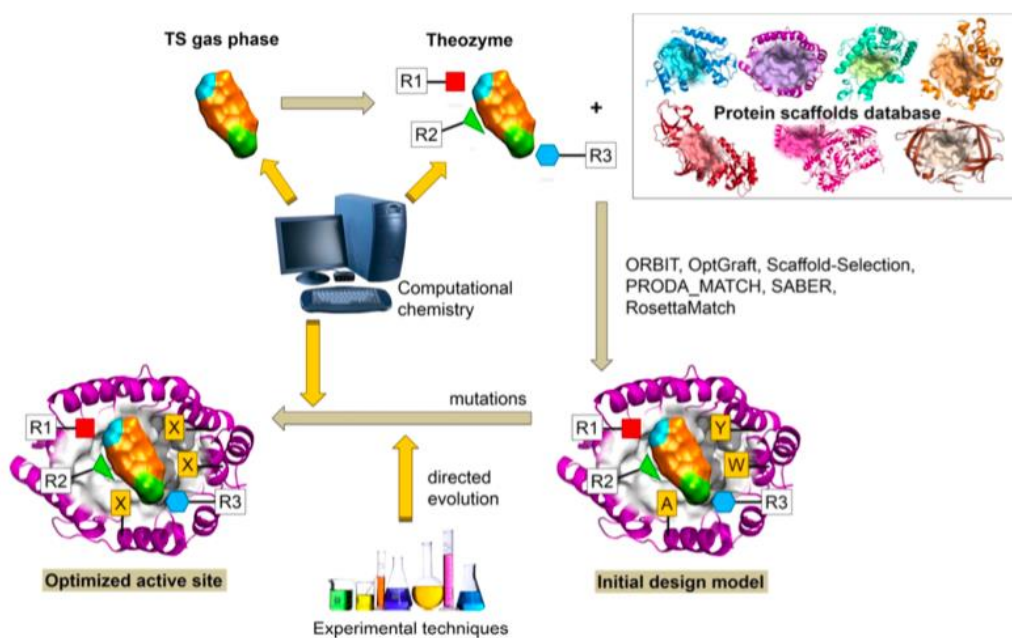
$$\hat{H}\Psi = E\Psi \quad (1)$$

To best approximate (1), computation of spatial density of electrons as an approximation to the psi term of (1) significantly reduces resource requirements and allows accountancy for as many known interactions contributing to the overall picture of electrons in space. Results of these calculations provide localized atomic charge data for the creation of a force field for molecular dynamics (MD) simulations. MD simulations rely on classical Newtonian mechanics, utilizing an iterative calculation cycle to predict to the physical movement of atoms

**Figure 2:** Interaction energies calculated by molecular dynamics simulations. In a typical MD cycle, energies are computed to generate initial motion of each atom, then forces are computed to determine acceleration which is then used to translate atoms for each iteration in time. Note that the energies in some of the equations are largely dependent on the  $r$  term, representing 3D distance between atoms. *Figure from Chang et al.*

$U(R) = \sum_{bonds} k_r (r - r_{eq})^2$	<i>bond</i>	
$+ \sum_{angles} k_\theta (\theta - \theta_{eq})^2$	<i>angle</i>	
$+ \sum_{dihedrals} k_\phi (1 + \cos[n\phi - \gamma])$	<i>dihedral</i>	
$+ \sum_{impropers} k_\omega (\omega - \omega_{eq})^2$	<i>improper</i>	
$+ \sum_{i < j}^{atoms} \epsilon_{ij} \left[ \left( \frac{r_m}{r_{ij}} \right)^{12} - 2 \left( \frac{r_m}{r_{ij}} \right)^6 \right]$	<i>van der Waals</i>	
$+ \sum_{i < j}^{atoms} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$	<i>electrostatic</i>	

by simultaneously computing the interaction energies illustrated in (Figure 2)<sup>6</sup> between atoms in 3D space.<sup>5</sup> Limitations of traditional Born-Oppenheimer molecular dynamics (BOMD) simulations arise in computational overhead associated with accountancy of all electronic interactions such as those contained in bonds and in electron clouds of adjacent non-reactive structures. Thus, Ab-initio molecular dynamics (AIMD), widely used in the standard model, was proposed by Car and Parrinello to account for only valence electrons used in chemical bond formation. These methods and equations have been widely refined over many decades and give accurate approximations of bond energies (error of 0.2-0.3 eV) using the generalized gradient approximation (GGA).<sup>2</sup>



**Figure 3:** Computational Standard Model of Enzyme Design. Figure from Swiderek et al.<sup>7</sup> Using quantum-mechanics (QM) methods, the reaction's transition state is predicted with stabilizing amino acid side chains in a unit known as a theozyme. At this stage, precursor data in the form of template active site scaffolds is an essential component of the design process and are sourced from external databases such as the protein databank (PDB). The initial design model is then initially evaluated in an AIMD simulation and subsequently refined for increased catalytic activity with additional simulations. LDE as the finishing step, further optimizes the enzyme's catalytic properties.

### **III. Assessing Functionality and Limitations of Standard Model Designed Enzymes**

As the scheme in Figure 3 was pioneered, initial limitations were quickly identified. Three areas of poor design performance are outlined in a review by Welborn and Head-Gordon.<sup>2</sup>

The first area of limitation was discovered in plateaued LDE results from poorly designed enzymes. Further research determined that LDE did not result in enzymes that mimic the traditional stabilization of a transition state, rather, the synthetic enzymes showed functionality in stabilizing the reactant state.<sup>2</sup> Attaining the proper transition state is critical in thermodynamic performance and overall viability of the enzyme, which gave means to reconsider designing the active site in the presence of the interactions provided by a supportive scaffold. For algorithms utilizing feedback of their performance to improve future designs, faulty data from this unintended function hinders subsequent performance of the algorithms to produce ideal transition state stabilization. This issue of tainted data is also common in the machine learning and AI areas of computer science, where substantial efforts to provide accurate and meaningful training data are of high priority. This concern relates back to a requirement of successful enzyme design: accurate starting data. Machine learning integration in enzyme design is further discussed in section V of this work.

A second area of limitation was found in the maintenance of hydrogen bonding interactions throughout the catalytic process. This limitation has since been accounted for in work performed by Rajopalan et al. by modifying software packages such as RosettaMatch and RosettaDesign to include necessary steric packing and pi-pi stacking interactions occurring in a catalytic triad active site through additional rounds of QM runs.<sup>2</sup>

A third area of limitation, discovered from analyzing MD simulation output, determined that designed enzymes lacked backbone flexibility found in their evolutionary counterparts.

From introductory biochemistry courses, it is made apparent that conformational change and flexibility is paramount to the functions performed by enzymes, as observed in glycogen phosphorylase and in citrate synthase.<sup>8</sup> To incorporate additional flexibility, advancements in QM models consider backbone requirements based on the conformational changes from substrate(s) to transition state to product. This is achieved by introducing non-traditional amino acids into the backbone sequence.<sup>2</sup> Limitations with flexibility (in the form of entropic effects) of the designed protein were also encountered when attempting *de Novo* design.

#### **IV. Complete *de Novo* Design**

Work into complete *de Novo* design requires incorporation and accurate modeling of the wholistic environment, contrasting the standard model (Figure 3). Initial approaches to tackling this daunting task began with modeling tertiary structures, such as alpha-helical barrels contained within metalloenzymes.<sup>2</sup> Performed by DeGrado et al., design of A<sub>2</sub>B<sub>2</sub> heterotetrametric diiron bundles required an algorithm capable of producing the desired alpha-helix structure while minimizing the possibility of folding an alternative structure. They developed such an algorithm that included stabilizing effects of a secondary coordination sphere. Laboratory characterization of the designed protein using thermal denaturation and spectrophotometric-monitored ferroxidase activity showed thermal stability and enzymatic activity slightly-less than evolutionarily-designed counterpart bacterioferritin.<sup>9</sup> This was one of the first examples of a successful *de Novo* design.

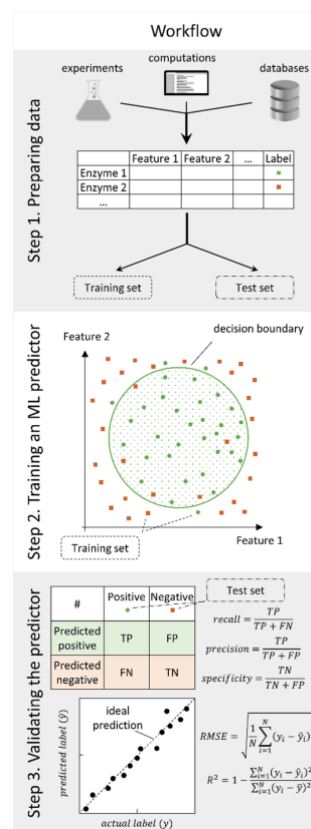
Compensating for entropic effects on free energy of the transition state may greatly improve catalytic performance beyond enzymes designed using the standard model.<sup>2</sup> The “Circe Effect”, proposed by Jencks in 1975, hypothesizes that free energy gained during favorable binding of the substrate is used to compensate for destabilization of the substrate’s ground

state.<sup>10</sup> While not verified in practice, this hypothesis inspired studies using room-temperature X-ray crystallography and NMR relaxation methods to determine that large entropic reserves, in the form of conformational states, contributes greatly to the overall thermodynamic stability of an enzyme's activation energy but showed reduced catalytic activity.<sup>2</sup> This inverse effect is an important consideration for *de Novo* enzymes used in industrial processes where high activity and stability are both required simultaneously.

From these current limitations, further developments and enhancements into algorithmic approaches are required to refine and validate the complicated *de Novo* design process.

## V. Machine Learning Implementations

As machine learning (ML) algorithms and artificial intelligence (AI) are becoming increasingly commonplace in consumer services such Siri, Google Home, and Amazon Alexa, scientists are also beginning to utilize these algorithms to better analyze MRI imaging, human genome datasets, and drug candidates. A third approach to enzyme design, the use of ML coupled with large protein structure-function datasets may offer viable solutions that differ from *de Novo* or standard model results. A scheme for this process is shown in figure 4. In essence, ML is directed to detect existing patterns in “training sets” and identify new ones in provided “test sets”. A distinct advantage of ML resides in overall



**Figure 4 (at right):** Workflow of machine learning in enzyme design. Partial figure from Mazurenko et al. Utilizing data from protein databases containing annotated sequences, algorithms are “trained” to create a correlation boundary. The trained algorithm is then tested against a known outcome, and further refined with slight adjustments to model sensitivity. From a given structure the ML algorithm predicts an enzyme structure and function.

generalizability of trained algorithms to quickly make predictions without construction and refinement of a new model, as is commonplace in the *de Novo* and standard models. The success of this technique also relies upon stringent requirements related to the quality and standardized reporting of data.

Data for machine learning is sourced from available databases such as the PDB (mentioned earlier), Brenda, ExCatDb, M-CSA and many others outlined by Mazurenko and colleagues.<sup>11</sup> Issues arise from inconsistent reporting terminology and differing methods used to obtain data, partly because this data was not originally intended for use in the ML capacity. For future data reporting, the author cites next-generation sequencing, high throughput screening, and deep mutational screening as reliable and robust methods for providing comprehensive ML datasets.<sup>11</sup> Projects such as STRENDA currently exist to standardize enzyme data, much like IUPAC unified naming conventions for molecules, and mitigate laborious manual culling of datasets used for analysis. To further complicate matters, inherent biases also arise in the datasets, as reported data generally are only of successful experiments and best performers, leaving out unviable candidates for the ML algorithms to account for.

Currently, the field of ML predicted enzyme catalysis and structure property relationships is still largely in its infancy. Community-wide experimental competitions such as CASP, COMBREX, and CAFA serve to accelerate model development efforts by vetting community sourced models that produce inaccurate predictions.<sup>11</sup> At the CASP13 competition, the AlphaFold algorithm predicted two-thirds of the target structures with topologically correct results. A less-than-ideal result for complete design, continual improvement over previous competitions shows potential for this class of ML structure-predictors. Mazurenko et al. suggest



potential in coupling MD simulations to ML algorithms, creating a hybrid-like prediction method akin to the standard model mentioned earlier.<sup>11</sup>

While machine learning has much potential, robust datasets, proven methods and algorithms, and correct results will be required for this technique to rival that of the standard model and *de Novo* design.

## **VI. Conclusions**

Over the previous decades, the rise of digital computers and advanced biochemistry techniques have created a blossoming new field of biochemistry. Computational enzyme design strives to answer questions that have long gone unanswered and generate functional molecules that enable greener reactions. This discipline is incredibly interdisciplinary, requiring knowledge of biochemistry, physics, statistics, mathematics, and computer science to understand and propose methods for generating these molecules. With diligent work in the field, these techniques may even replace traditional laboratory environments and substitute them for large data centers containing petaflops of compute horsepower. However, with the current state of modeling inaccuracies and limitations surrounding active site and wholistic enzyme interactions, there is still much work to be done. In the author's personal experience with simulating molecules and similarly with the opinion of Mazurenko et al., unification of algorithms and functional methods is a large factor into mainstream success of computational design, as each software package utilizes different file and naming formats, has different naming conventions, and has complex command-line interfaces, setting a barrier to entry for those not skilled in computer science. Nevertheless, great discoveries are poised to originate from computational design of enzymes.

## VII. References

- (1) Pabo, C. Molecular Technology: Designing Proteins and Peptides. *Nature* **1983**, *301* (5897), 200–200. <https://doi.org/10.1038/301200a0>.
- (2) Vaissier Welborn, V.; Head-Gordon, T. Computational Design of Synthetic Enzymes. *Chem. Rev.* **2019**, *119* (11), 6613–6630. <https://doi.org/10.1021/acs.chemrev.8b00399>.
- (3) Em, B.; Fh, A. Optimizing Non-Natural Protein Function with Directed Evolution. *Curr. Opin. Chem. Biol.* **2010**, *15* (2), 201–210. <https://doi.org/10.1016/j.cbpa.2010.11.020>.
- (4) Fox, R. J.; Davis, S. C.; Mundorff, E. C.; Newman, L. M.; Gavrilovic, V.; Ma, S. K.; Chung, L. M.; Ching, C.; Tam, S.; Muley, S.; Grate, J.; Gruber, J.; Whitman, J. C.; Sheldon, R. A.; Huisman, G. W. Improving Catalytic Function by ProSAR-Driven Enzyme Evolution. *Nat. Biotechnol.* **2007**, *25* (3), 338–344. <https://doi.org/10.1038/nbt1286>.
- (5) Car, R. Introduction to Density-Functional Theory and Ab-Initio Molecular Dynamics. *Quant. Struct.-Act. Relatsh.* **2002**, *21* (2), 97–104. [https://doi.org/10.1002/1521-3838\(200207\)21:2<97::AID-QSAR97>3.0.CO;2-6](https://doi.org/10.1002/1521-3838(200207)21:2<97::AID-QSAR97>3.0.CO;2-6).
- (6) Chang, C. A.; Huang, Y. M.; Mueller, L. J.; You, W. Investigation of Structural Dynamics of Enzymes and Protonation States of Substrates Using Computational Tools. *Catalysts* **2016**, *6* (6), 82. <https://doi.org/10.3390/catal6060082>.
- (7) Świderek, K.; Tuñón, I.; Moliner, V.; Bertran, J. COMPUTATIONAL STRATEGIES FOR THE DESIGN OF NEW ENZYMATIC FUNCTIONS. *Arch. Biochem. Biophys.* **2015**, *582*, 68–79. <https://doi.org/10.1016/j.abb.2015.03.013>.
- (8) Jared R. Mays. CHEM305 Lecture Notes.
- (9) Summa, C. M.; Rosenblatt, M. M.; Hong, J.-K.; Lear, J. D.; DeGrado, W. F. Computational de Novo Design, and Characterization of an A2B2 Diiron Protein. *J. Mol. Biol.* **2002**, *321* (5), 923–938. [https://doi.org/10.1016/S0022-2836\(02\)00589-2](https://doi.org/10.1016/S0022-2836(02)00589-2).
- (10) Jencks, W. P. Binding Energy, Specificity, and Enzymic Catalysis: The Circe Effect. In *Advances in Enzymology and Related Areas of Molecular Biology*; John Wiley & Sons, Ltd, 1975; pp 219–410. <https://doi.org/10.1002/9780470122884.ch4>.
- (11) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal.* **2020**, *10* (2), 1210–1223. <https://doi.org/10.1021/acscatal.9b04321>.