EnzyHTP: A High-Throughput Computational Platform for Enzyme Modeling

Qianzhen Shao[1], Yaoyukun Jiang[1] and Zhongyue J. Yang[1-4,*]

[1]*Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235, United States*

[2]*Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37235, United States*

[3]*Vanderbilt Institute of Chemical Biology, Vanderbilt University, Nashville, Tennessee 37235, United States* [4]*Data Science Institute, Vanderbilt University, Nashville, Tennessee 37235, United States*

ABSTRACT: Molecular simulations, including quantum mechanics (QM), molecular mechanics (MM), and multiscale QM/MM modeling, have been extensively applied to understand the mechanism of enzyme catalysis and to design new enzymes. However, molecular simulations typically require specialized, manual operation ranging from model construction to post-analysis to complete the entire life-cycle of enzyme modeling. The dependence on manual operation makes it challenging to simulate enzymes and enzyme variants in a high-throughput fashion. In this work, we developed a Python software, EnzyHTP, to automate molecular model construction, QM, MM, and QM/MM computation, and analyses of modeling data for enzyme simulations. To test the EnzyHTP, we used fluoroacetate dehalogenase (FAcD) as a model system and simulated the enzyme interior electrostatics for 100 FAcD mutants with a random single amino acid substitution. For each enzyme mutant, the workflow involves structural model construction, 1 ns molecular dynamics simulations, and quantum mechnical calculations in 100 MD-sampled snapshots. The entire simulation workflow for 100 mutants was completed in 7 hours with 10 GPUs and 160 CPUs. EnzyHTP is expected to improve the efficiency and reproducibility of computational enzyme, facilitate the fundamental understanding of catalytic origins across enzyme families, and accelerate the optimization of biocatalysts for non-native substrate transformation.

Keywords: high throughput, enzyme modeling, automation

1. Introduction

Natural enzymes promote the transformation of native substrates with superior efficiency and selectivity compared to aqueous solution.[1] In contrast, lower catalytic competence is frequently observed when applying natural enzymes to transform non-native substrates.[2-4] As such, identifying new enzyme variants for non-native chemical transformations presents a "holy-grail" in academia and industry because it will allow late-stage functionalization of drug-like molecules,[5] polymer upcycling,[6, 7] degradation of environmental pollutants,[8, 9] and treatment of food allergies.[10] Experimental high-throughput screening techniques[11] has popularized directed evolution as a tool to optimize enzyme variants for function improvement by iterative rounds of random mutations.[12-14] However, because the relationship between enzyme sequence, structure, and function is unknown, the number of iterative screening rounds and the improvement of functional performance are highly sequence- and substrate-dependent.[15] Molecular simulation methods, including quantum mechanics (QM), molecular mechanics (MM), and multiscale QM/MM modeling, have been extensively applied towards directed evolution to guide the selection of function-enhancing mutations[16]. These simulations inform the mechanistic detail underlying the variation of rate and selectivity upon mutagenesis in an enzymatic reaction, inspiring the development of new design principles that pinpoint the beneficial mutations.[17-19]

To maximize the potential of molecular simulations in biocatalyst discovery, it is essential to perform enzyme modeling in an automatic and high-throughput fashion. A computational high-throughput platform parallelizes the computation of a large number of enzyme models, which allows understanding of enzyme catalytic mechanisms across a large number enzymes in a protein family, enables the virtual screening of enzyme mutants, and collects electronic structure and dynamics data for building data-driven predictive models.[20] Computational workflows have been established to automate general-purpose protein simulations. For example, Doerr et. al. developed HTMD[21] for high-throughput MD simulations with the ACEMD[22] engine. Parton et. al. developed Ensembler[23] to enable high-throughput simulations of members in a protein superfamily combining MD and template-based structural prediction.

Unlike general-purpose protein simulations that typically emphasize protein scaffold alone (i.e., HTMD[21] and Ensembler[23]), enzyme modeling requires the treatment of enzyme-substrate pre-reaction complexes, transition state, and other reacting species. To complete a life-cycle of enzyme modeling, specialized manual input is needed to build the pre-reaction complex; conduct the workflow of electronic structure or classical simulations; and analyse the simulation data to evaluate the impact of mutation on the enzyme electric field, substrate positioning dynamics, and the activation free energy of an elementary step. Amrein et al. pioneered in developing CADEE[24] to conduct free energy calculations to assess the activation free energy barriers in different enzyme variants to facilitate optimization of enzyme variants. Similarly, other protocols to automate free energy simulations have also been developed, including BFEE2,[25] PyAutoFEP,[26] and BRIDGE[27]. These tools have been employed to accelerate the discovery of functional enzyme variants. However, a holistic platform that focuses on the entire lifecycle of enzyme modeling still remains undeveloped.

In this work, we developed a new computational platform, EnzyHTP, that aims to automate the entire life-cycle of enzyme modeling. EnzyHTP allows the users to conduct molecular mechanics (MM), quantum mechanics (QM), and multiscale QM/MM simulations in a high-throughput manner. EnzyHTP consists of four modules that facilitate preparation of enzyme structure models, generation of enzyme mutants, sampling of conformations, and calculation of energies. As a proof of concept, we tested EnzyHTP by evaluating the mutation effects on enzyme electrostatics in 100 fluoroacetate dehalogenase (FAcD) mutants with a random single mutation. Each simulation was conducted by using a single, facile python script to assemble the modules aforemetioned. The software is expected to facilitate the fundamental understanding of catalytic origins across enzyme families and to accelerate the optimization of biocatalysts for non-native substrates.

2. Implementation

Design Architecture of EnzyHTP The framework of EnzyHTP involves four levels of operation that follows a top-down hierarchy that consists of protein preparation, mutant generation, geometry variation, and energy engine (**Figure 1**). The outcomes of each level of operation serves as an input for the next level. First, protein preparation emphasizes building computational models for known enzyme structures that are derived from either X-ray crystallography experiments or computational predictions (i.e., AlphaFold2[28] or RoseTTAFold[29]). These enzyme structures involve a diverse range of binding substrates, sequences, cofactors, coenzymes, protonation states, stoichiometry numbers, and structure quality (e.g., missing loop, presence of hydrogen atoms). Second, mutant generation emphasizes generating new enzyme variants based on a common enzyme sequence and scaffold. The mutation of an existing amino acid changes the sidechain type and conformation. These variations may also perturb the protonation of nearby enzyme residues. Third, geometry variation emphasizes the change of enzyme conformation and substrate reaction states. The catalytic proficiency of enzymes critically depends on protein dynamics and the interplay between protein dynamics and substrate reacting states. Fourth, energy engine emphasizes conducting the MM, QM, or QM/MM calculations. To balance accuracy and efficiency, different levels of theory should be integrated to simulate enzymes' catalytic functions.

In particular, QM characterization of enzyme's functional sites and reacting species is essential to understanding and predicting the catalytic actions of enzyme catalysis.

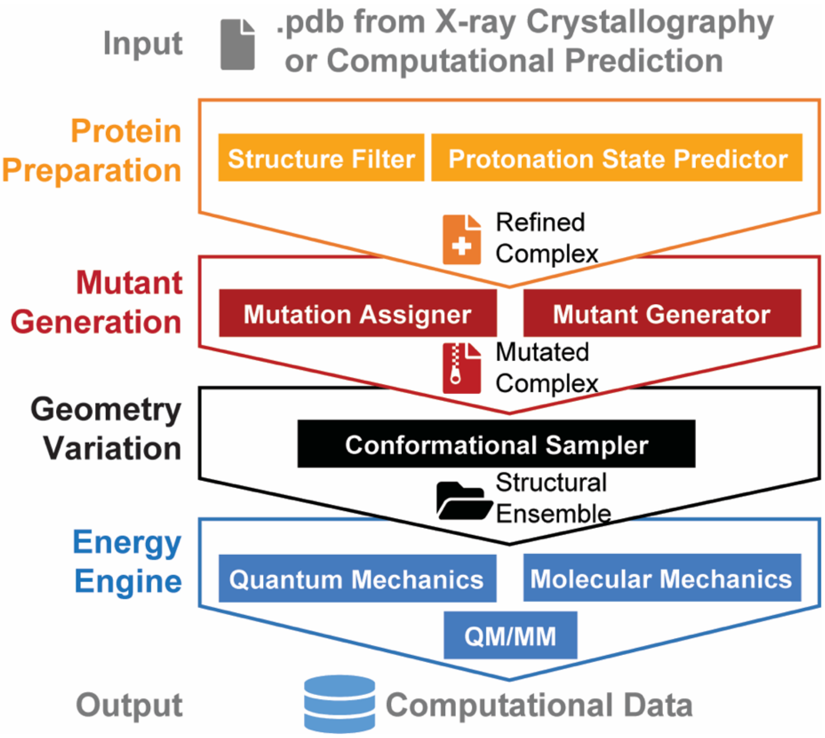


Figure 1. The hierarchical structure of high-throughput enzyme modeling. The framework involves four levels of operation, including protein preparation, mutant generation, geometry variation, and energy calculation. The framework takes in the enzyme structure as an input and delivers computational modeling data as an output.

Notably, existing HTP software typically automates the operations of one or two specific levels within the HTP framework laid out in Figure 1. For example, BFEE2,[11] PyAutoFEP,[12] and BRIDGE[13] automate the free energy calculation and geometry sampling but depend on manually-curated enzyme structures as an input. HTMD[21] and CADEE[24] automate mutant generation, but the free energy calculation is based on molecular mechanics or empirical valenace bond theory. Ensembler[23] automates protein structure preparation (i.e., protonation state assignment and miss loop remedy) but does not support mutant generation. Additionaly, the free energy calculation is also conducted at the MM level.

Modules of EnzyHTP EnzyHTP consists of four modules (**Figure 2**). Each module handles a particular level of operation laid out in the design framework, including protein preparation, mutant generation, geometry variation, and energy engine (**Figure 1**). First, we developed an enzyme preparation module to automatically convert an initial enzyme structure to computational models with a standardized input file for the subsequent operations. The protein preparation module contains two primary functions: 1) structure filter and 2) protonation state predictor. The structure filter takes an enzyme complex structure as an input and then performs structural curation. The structure filter removes the structures containing missing loops, removes common co-crystallization reagents and solvent, and keeps user-defined, chemically-relevant small-molecule ligands in the complex, including inhibitors, reactants, intermediates, products, and structural analogs. In cases where no optimal enzyme complex can be found in the protein structure databases, a user-curated enzyme complex is required. An on-going work for the protein preparation module involves developing a new enzyme structure fixer function that can 1) remedy missing loops or residues,[30] 2) modify the analog to a corresponding substrate, and 3) construct enzyme-reacting species' pre-reaction complexes via docking.

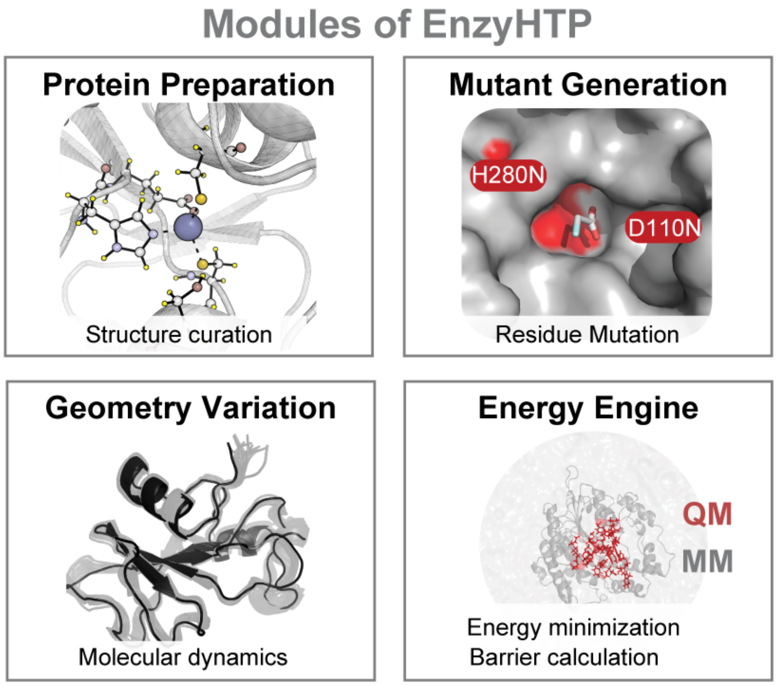
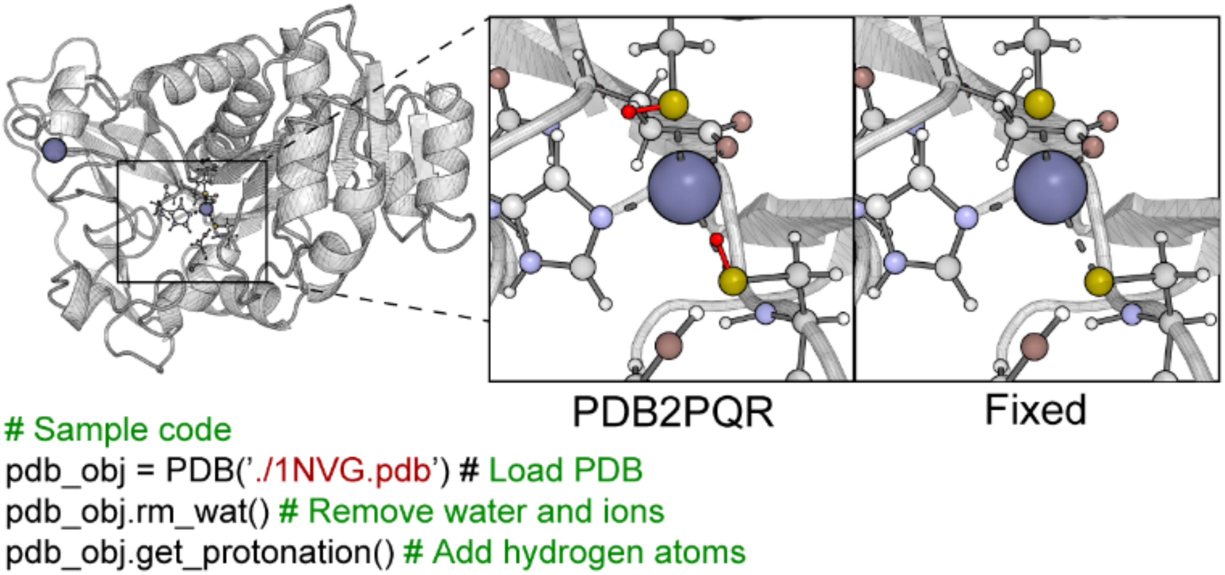


Figure 2. Four modules of EnzyHTP.

The protonation state predictor integrates the functions of PDB2PQR[31] (i.e., version 3.2.2) and OpenBabel[32] to determine the protonation state of protein residues and ligands, respectively. PDB2PQR employs PROPKA3[33] to determine the protonation state of each titratable residue by computing the empirical pK_a and the hydrogen-bonding network in a user-defined pH and dielectric constant. PDB2PQR then optimizes the hydrogen bonding network to get the final structure. However, the influence of metal ion cofactors on protein protonation states is not considered in either program. This may lead to unphysical protonation states for protein residues that coordinate to the metal ions. For example, in the alcohol dehydrogenase (PDB ID: 1NVG), the Zn^{2+}-coordinating cystines are determined to be protonated (i.e., –SH) by PDB2PQR, which is thermodynamically unstable (**Figure 3**). To fix the protonation state for metal ion-coordinating residues, we designed a checker function for enzymes that contain metal ions as a cofactor. The checker function first detects the metal ion-coordinating residues based on the atomic radius of a metal ion or a user-defined distance cut-off (Table S1, Supporting Information), and then attempts to deprotonate the metal ion-coordinating residues if PDB2PQR has not done so yet. The checker function can be easily adapted by the user who intends to customize the protonation rules for a set of enzymes of interest. An on-going work to further develop the protonation state predictor involves the implementation of residue pK_a prediction software H++[34], which accounts for the influence of metal ion in the pK_a prediction.

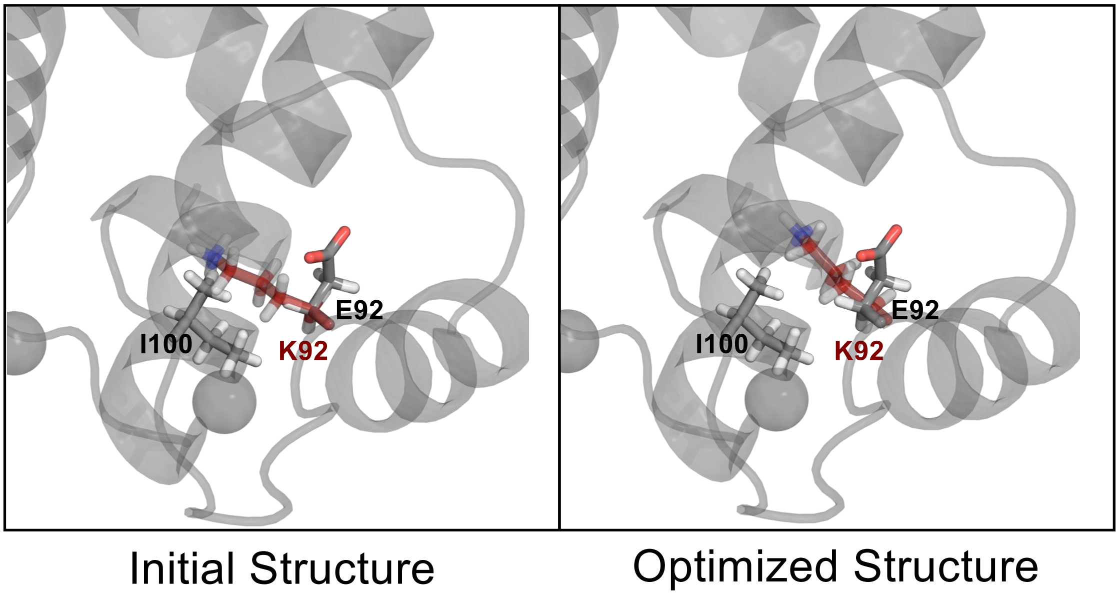


```
# Sample code
pdb_obj = PDB('./1NVG.pdb') # Load PDB
pdb_obj.rm_wat() # Remove water and ions
pdb_obj.get_protonation() # Add hydrogen atoms
```

Figure 3. Protonation of metal ion-coordinating residues in EnzyHTP. PDB2PQR does not account for the influence of metal ion in the protonation state prediction, which causes the metal ion-coordinating residues to populate in a thermodynamically-unstable protonation state (shown in red). This problem can be fixed by a built-in checker function in enzyHTP. The sample code is provided.

Second, we developed a mutant generation module to generate computational models for enzyme variants in a random or user-defined fashion. The mutant generation module also contains two primary functions: 1) mutation assigner and 2) mutant generator. The mutation assigner function creates a list of mutations either randomly or in a user-defined fashion. The random generator function randomly mutates a residue in the reference enzyme complex to another type of amino acid. Additionally, the mutant generator function also allows the user to define rules to limit the scope of mutation to the desired subgroup of amino acids (Text S1 and Table S2-S3, Supporting Information). For instance, the user can specify mutating residues with a hydrophobic side chain, mutating bulkier residues to smaller residues, mutating residues within or beyond a certain range of spatial proximity to a reference residue, protecting specific residues from mutation, and so on. Alternatively, the user can also provide a list of "X##Y" flags to specify the desired mutations, where X refers to the residue prior to mutation, ## refers to the residue index, and Y refers to the mutation residue.

The mutant generator function replaces a selected residue by one of the 19 remaining types of canonical amino acids using the tLEaP submodule of Amber[35]. The newly generated mutant structure is minimized at the MM level to remove local steric friction caused by the change of amino acid volume during the computational mutation. The structural relaxation is particularly important in the circumstance where a smaller-sized amino acid mutates to a bulkier-sized amino acid, helping the new sidechain to fit into the local pocket (**Figure 4**). Notably, in some rare cases, unphysical geometries emerge from the mutation (e.g., Lys penetrating the phenyl ring of Phe, Figure S1, Supporting Information), which structural optimization may fail to fix. To create a more thermodynamically stable variant structure, we are developing an alternative mutation function that directly generates structure for a certain enzyme variant using SWISSMODEL[36] or AlphaFold2[28]. This will ensure smooth operation of high-throughput enzyme modeling.



Initial Structure

Optimized Structure

```
# Sample Code
pdb_obj = PDB('2kz2.pdb') # Load PDB
pdb_obj.Add_MutaFlag('r') # Rondom assign Mutation
pdb_obj.PDB2PDBwLeap() # Generate Mutant
# Perturb protonation state
pdb_obj.rm_allH()
pdb_obj.get_protonation()
# Relax structure by minimization
pdb_obj.PDB2FF()
pdb_obj.PDBMin(engine='Amber_pmemd_gpu')
```

Figure 4. An example of mutation from smaller to a larger residue. The sidechain of original residue (Glu), mutated residue (Lys), and the residue (Ile) that has local friction with the mutated residue are shown in sticks. The carbon atoms of the Glu and Ile are shown in grey with full opacity, and those of the Lys are shown in red with 50% transparency. The nitrogen and oxygen are shown in blue and red with full opacity, respectively. The sample code is provided.

Third, we developed the geometry variation module to sample enzyme conformations by molecular dynamics and to map reaction energy or free energy landscape. The input structures of this module are direct results from either one of the first two modules. The geometry variation module contains one primary function, which is a conformational sampler. The conformation sampler prepares the simulation parameter files and conduct molecular dynamics (i.e., AMBER[35]) simulations to construct a conformational ensemble. Users can set the default values for the parameters (zip, Supporting Information). With the conformational ensemble, the user can further apply the built-in functions of EnzyHTP to call CPPTRAJ[37] to evaluate the flexibility (i.e., RMSF) of an enzyme scaffold or catalytically important loops or residues. Notably, the geometry variation module can employ enhanced sampling methods implemented in AMBER to accelerate the sampling of enzyme conformers or to sample along a reaction coordinates that involve significant energy variation (e.g., chemical bond breaking). An ongoing development in this module is to automatically construct the complexes between an enzyme scaffold and different reacting species, which is expected to ease the job of mapping energy pathways for catalytic reactions.

Fourth, we developed the energy engine module to prepare the input file and set up the parameters for a given type of simulation task. The energy engine module contains three primary functions: 1) MM engine, 2) QM engine, and 3) QM/MM engine. MM calculations allow efficient assessment of the interaction energies between substrate and enzyme using force field parameters; QM calculations allow a more accurate description of substrate-active site residue interactions that involve charge transfers, polarization, and long-range electrostatic effects; while multiscale QM/MM calculations can describe chemical reactions in a realistic enzyme and solvent environment with a balanced efficiency and accuracy. EnzyHTP integrates MM (e.g. AMBER[35]), QM, and QM/MM engines (e.g. Gaussian16[38]) and wavefunction analysis tools (e.g. Multiwfn[39]) along with built-in analysis functions to calculate the energy and perform electronic structure analysis (zip, Supporting Information). This enables the calculation of many energy/free energy-related features like MM-based binding free energy, QM or QM/MM-based local bond dipole, electron density, transition state barrier, and so on. This complements previous software like

CADEE[24] and HTMD[21]. We are also actively developing new EnzyHTP interfaces with other modeling software, which will enable broader choices of energy calculation engines.

3. Results and Discussion

To demonstrate the high-throughput capability of EnzyHTP, we employed EnzyHTP to study the impact of single mutations on the electrostatic environment of the enzyme scaffold in 100 fluoroacetate dehalogenase (FAcD) variants. *Rhodopseudomonas palustris* FAcD hydrolyzes the C–F bond of fluoroacetate (FAc) with a turnover rate of tens of seconds.[40-44] The enzyme active site involves a catalytic triad (i.e., Asp^{110}–Asp^{134}–His^{280}) that is common to hydrolases (Figure S2, Supporting Information). In the first step of the catalyzed reaction, upon the binding of the substrate, the Asp^{110} attacks the C–F bond in S_N2 manner and forms the covalent intermediate. This cleavage of the strong C–F bond contributes to the rate-determining step.[45] The process of breaking the C–F bond involves charge separation, which can be stabilized by the interior electrostatic environment of FAcD. Upon mutation, the electrostatic environment is likely perturbed, which can influence the catalytic efficiency. To quantify the strength of the electrostatic environment in the enzyme, we employed a physical descriptor, electrostatic stabilization energy (i.e., ΔG_{elec}), which is computed by the dot product between the electric field and the C–F bond dipole (**Figure 5**). The descriptor was introduced by Fried et al.[46, 47] and has been shown to correlate with activation free energy in ketosteroid isomerase,[48] Kemp eliminase,[49, 50] methyltransferase,[51] and P450 enzymes.[52] Accordingly, we presume that the descriptor ΔG_{elec} informs the variation of catalytic competence upon the mutation of FAcD.

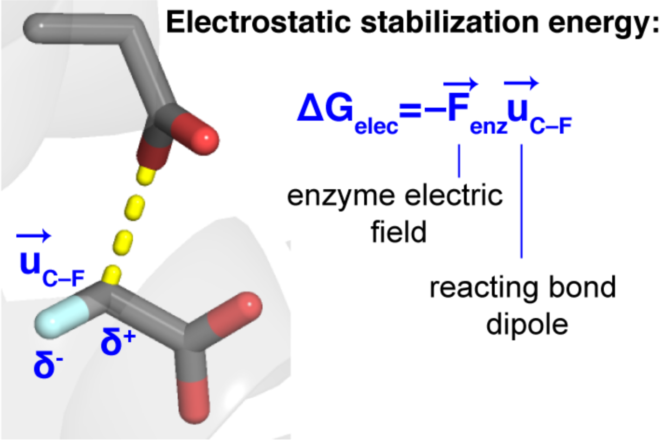

Electrostatic stabilization energy:
$\Delta G_{elec} = -\vec{F}_{enz} \vec{u}_{C-F}$
enzyme electric field
reacting bond dipole
$\vec{u}_{C-F}$
δ^-
δ^+

Figure 5. The definition of the electrostatic stabilization energy ΔG_{elec} of the C–F bond of the pre-reaction FAcD-FAc complex. The substrate FAc and nucleophile Asp110 are shown in sticks.

Enabled by EnzyHTP, we designed a Python workflow to compute electrostatic stabilization energy values for 100 FAcD variants with random single amino acid substitution (**Figure 6** and Text S2 Supporting Information). For each variant, we computed the ensemble average of ΔG_{elec} values (denoted by $<\Delta G_{elec}>$) using 100 conformational snapshots extracted from a 1 ns MD trajectory. We purposefully employed a short propagation time for the MD simulations to ensure that the sampled enzyme geometries bear high resemblance to the crystal structure. Under this circumstance, the computed $<\Delta G_{elec}>$ value should reflect the impact of mutation on the active-site electrostatic environment with little perturbation by the protein conformational changes.

The workflow first creates 100 variants using the mutant generation module based on a curated FAcD crystal structure (PDB ID: 6QHQ), in which the co-crystal reagents were removed, and the protein residue side chains were protonated (Table S4, Supporting Information). The structure involves a pre-reaction complex in which the residue Asp^{110} is aligned with the substrate C–F bond for a potential S_N2 attack. During the mutation operation, the catalytic residue Asp^{110} is not mutated because of its direct participation in the reaction. Second, the workflow conducts MD simulation for each variant and samples 100 conformers from a 1 ns MD production run (zip, Supporting Information). These conformers constitute a structural ensemble. Third, the workflow computes the electrostatic stabilization energy (ΔG_{elec}) in each sampled conformation by multiplying the bond dipole of the substrate C–F bond and the electric field strength of the entire enzyme projected onto the C–F bond (excluding substrate and Asp^{110}). The bond dipole is computed using a single-point QM cluster calculation (HF/6-31G(d)) that consists of the substrate and the catalytic residue Asp^{110}, followed by the wavefunction-based localized molecular orbital (LMO) analysis using Multiwfn. The electronic field strength of the enzyme is computed based on the point charges of enzyme amino acids using Coulomb's law. Notably, differences in the final value in different conformations reflect the fluctuation of the wavefunction and thus the charge transfer. Fourth, the workflow averages over the ΔG_{elec} values of all 100 conformations for each variant and stores the averaged $<\Delta G_{elec}>$ values in a Python-compatible data format. Through the workflow, we evaluated the $<\Delta G_{elec}>$ values for 100 FAcD variants in 7 hours with 10 GPUs (NVIDIA V100 SMX2) and 160 CPUs (Xeon Gold 6248). In contrast, the manual operation time to model 100 enzyme variants would approximately take weeks to complete due to tedious processes of mutant structure curation and file preparation plus computational runtime.

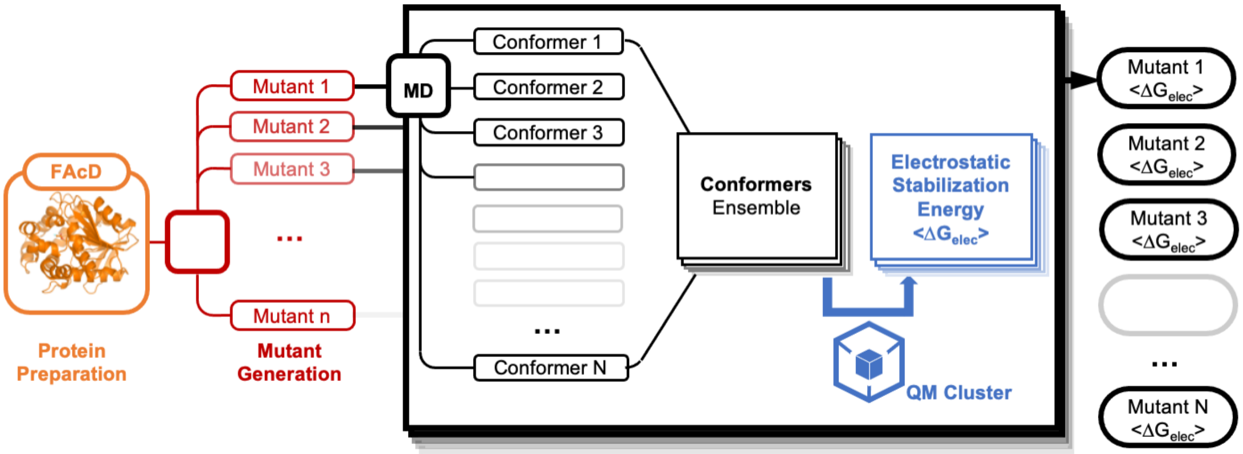


Figure 6. A workflow to compute electrostatic stabilization energy values (i.e., <ΔG_{elec}>) for 100 FAcD variants with random single amino acid substitution. The workflow is constructed by EnzyHTP modules in Python.

Figure 7 shows the distribution of mutation positions and <ΔG_{elec}> values for 100 FAcD variants. The mutation spots spatially distribute over the entire FAcD enzyme scaffold (**Figure 7a** and Table S4, Supporting Information). The center of mass (COM) distances between the mutation spot and the substrate range from 7 Å (i.e., first coordination shell) to 32 Å (remote mutations). These mutations implicate different types of polarity change (Table S5, Supporting Information), including neutral-neutral (49 mutants), neutral-charged (19 mutants), charged-neutral (24 mutants), and charged-charged (8 mutants) mutations. Notably, the impact of a mutation on the <ΔG_{elec}> value does not appear to depend on its COM distance to the substrate (**Figure 7b**). The <ΔG_{elec}> values range from -1.1 kcal/mol to 8.2 kcal/mol (**Figure 7c**). Using the <ΔG_{elec}> of the wild-type (WT) FAcD as reference (i.e., 0.5 kcal/mol), we observed that a smaller amount of mutations (~10%) induce the <ΔG_{elec}> value to descrease, and most mutations (~90%) have an opposite effect. Decreasing the <ΔG_{elec}> value reflects a more favorable electrostatic environment to stabilize the developing C–F dipole (including the TS) in the FAcD mutant than that in the wild-type FAcD. As such, the mutation that results in a decreased <ΔG_{elec}> value may be more prone to enhance catalytic competence, while the <ΔG_{elec}>-increasing mutations are likely to be rate-deleterious. Noticeably, FAcD is a natural enzyme that has been well-evolved in nature – this explains the observation that most mutations leads to a less favorable electrostatic environment for the reaction.

The mutant with the largest negative <ΔG_{elec}> value is K83D (i.e. descreases by –1.7 kcal/mol relative to the WT). This residue resides along the axis of the reacting C–F bond and is

12 Å away (i.e., second coordination shell) from the substrate. The K83D mutation involves a change from a positively-charged residue (i.e. lysine) to a negatively-charged residue (i.e.asparate), which significantly perturbs the electric field projected on the C–F bond, changing the sign of the <ΔG_{elec}> value to negative. Additionally, we observed mutations that lead to a significant descrease in the <ΔG_{elec}> values but do not involve a change of the residue charge state, including: Y200V (i.e., by –1.5 kcal/mol), A186W (i.e., by –1.4 kcal/mol) and I14C (i.e., by –1.2 kcal/mol). These mutations may change the electrostatic environment through perturbing local protein dynamics in the sub-ns time scale. Notably, we do not intend to overstate the implication of the current results because of the simplied model (i.e., pre-reaction complex) used for describing the substrate reaction state, the moderate QM region size and level of theory used in evalutating the <ΔG_{elec}> value, and the lack of consideration of mutant expressibility and solubility. Nonetheless, this case study demonstrates the potential of EnzyHTP to facilitate the identification of beneficial mutations for biocatalyst discovery by leveraging enzyme modeling of different theoretical levels.

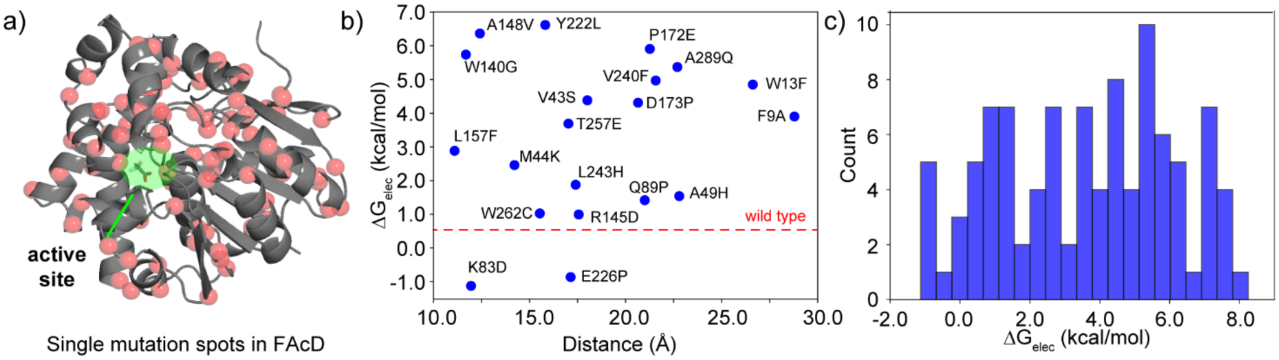


Figure 7. The distribution of mutation positions and <ΔG_{elec}> values for 100 FAcD variants. a) Spatial distribution of singe mutation spots of FAcD variants. The positions of C_α of the mutated residues are shown in red sphere. b) The relationship between <ΔG_{elec}> values and mutation

positions in 20 randomly selected FAcD mutants. The mutation position is quantified using the center of mass distance between a mutated residue and the substrate FAc. The ΔG_{elec} value for the wild-type FAcD is shown in the red dashed line. c) The distribution of ΔG_{elec} values for 100 FAcD mutants.

Conclusion

We developed a high-throughput computational platform, EnzyHTP, that automates the life-cycle of enzymatic modeling via incorporating state-of-the-art software. EnzyHTP consists of four main modules: protein preparation, mutant generation, geometry variation, and energy engine. We tested the performance of EnzyHTP using fluoroacetate dehalogenase (FAcD) as a model enzyme. We built a Python workflow in EnzyHTP to simulate the enzyme interior electrostatics for 100 FAcD mutants with a random single amino acid substitution. From a single PDB input file, the workflow refines the structure, determines protonation states, generates mutant structures, performs molecular dynamics and quantum mechanics simulations, and calculates the electrostatic stabilization energies $<\Delta G_{elec}>$. The entire simulation workflow was completed in 7 hours of wall clock time with 10 GPUs and 160 CPUs. This work enables high-throughput modeling of enzyme catalysis with a combination of QM, MM, and QM/MM simulations. EnzyHTP sets the basis for *in silico* high-throughput screening that identifies beneficial enzyme variants, which can accelerate the development cycle of new biocatalysts that catalyzes non-native substrates. EnzyHTP also helps generate computational data for our database IntEnzyDB that guides future statistical understanding and machine learning.

ASSOCIATED CONTENT

Supporting Information. Parameters of metal ion radius; customized mutation rules set in

EnzyHTP; residue category of polarity; residue rank of volume; a typical example of bad contact involved in the mutation; active site structure of wild-type FAcD; computational details for the simulations of FAcD variants; electrostatic stabilization eneriges for 100 FAcD variants; different categories of residue polarity change in the FAcD variants (PDF)

Default input files for the QM, MM, and electronic structure analysis calculations; parameter and input coordinate files generated by EnzyHTP (ZIP).

Data and Software Availability. The code and sample input for EnzyHTP framework is publically available at https://github.com/ZJYgrp/EnzyHTP. The input files and structures are provided as part of the SI files. AMBER 18 is available from http://ambermd.org/. Gaussian 09 is available from https://gaussian.com/.

AUTHOR INFORMATION

Corresponding Author

*Email: zhongyue.yang@vanderbilt.edu phone: 615-343-9849

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

ZJY thanks Ms. Reecan Juarez for proofreading the manuscript. This research was supported by the startup grant from Vanderbilt University. This work was carried out in part using computational resources from the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number TG-BIO200057.[53]

References

1. Knowles, J. R., Enzyme catalysis: not different, just better. *Nature* **1991,** *350* (6314), 121-124.

2. Bornscheuer, U. T.; Huisman, G. W.; Kazlauskas, R. J.; Lutz, S.; Moore, J. C.; Robins, K., Engineering the third wave of biocatalysis. *Nature* **2012,** *485* (7397), 185-194.

3. Burton, S. G.; Cowan, D. A.; Woodley, J. M., The search for the ideal biocatalyst. *Nature Biotechnology* **2002,** *20* (1), 37-45.

4. Madhavan, A.; Arun, K. B.; Binod, P.; Sirohi, R.; Tarafdar, A.; Reshmy, R.; Kumar Awasthi, M.; Sindhu, R., Design of novel enzyme biocatalysts for industrial bioprocess: Harnessing the power of protein engineering, high throughput screening and synthetic biology. *Bioresource Technology* **2021,** *325*, 124617.

5. Li, F.; Zhang, X.; Renata, H., Enzymatic CH functionalizations for natural product synthesis. *Current Opinion in Chemical Biology* **2019,** *49*, 25-32.

6. Knott, B. C.; Erickson, E.; Allen, M. D.; Gado, J. E.; Graham, R.; Kearns, F. L.; Pardo, I.; Topuzlu, E.; Anderson, J. J.; Austin, H. P.; Dominick, G.; Johnson, C. W.; Rorrer, N. A.; Szostkiewicz, C. J.; Copié, V.; Payne, C. M.; Woodcock, H. L.; Donohoe, B. S.; Beckham, G. T.; McGeehan, J. E., Characterization and engineering of a two-enzyme system for plastics depolymerization. *Proceedings of the National Academy of Sciences* **2020,** *117* (41), 25476-25485.

7. Rorrer, N. A.; Nicholson, S.; Carpenter, A.; Biddy, M. J.; Grundl, N. J.; Beckham, G. T., Combining Reclaimed PET with Bio-based Monomers Enables Plastics Upcycling. *Joule* **2019,** *3* (4), 1006-1027.

8. Wang, J.-B.; Ilie, A.; Yuan, S.; Reetz, M. T., Investigating Substrate Scope and Enantioselectivity of a Defluorinase by a Stereochemical Probe. *Journal of the American Chemical Society* **2017,** *139* (32), 11241-11247.

9. Goldman, P., The Carbon-Fluorine Bond in Compounds of Biological Interest. *Science* **1969,** *164* (3884), 1123-1130.

10. Gan, J.; Siegel, J. B.; German, J. B., Molecular annotation of food – Towards personalized diet and precision health. *Trends in Food Science & Technology* **2019,** *91*, 675-680.

11. Markin, C. J.; Mokhtari, D. A.; Sunden, F.; Appel, M. J.; Akiva, E.; Longwell, S. A.; Sabatti, C.; Herschlag, D.; Fordyce, P. M., Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. *Science* **2021,** *373* (6553), 411-+.

12. Arnold, F. H.; Volkov, A. A., Directed evolution of biocatalysts. *Current Opinion in Chemical Biology* **1999,** *3* (1), 54-59.

13. Wang, Y.; Xue, P.; Cao, M.; Yu, T.; Lane, S. T.; Zhao, H., Directed Evolution: Methodologies and Applications. *Chemical Reviews* **2021**.

14. Hammer, S. C.; Knight, A. M.; Arnold, F. H., Design and evolution of enzymes for non-natural chemistry. *Current Opinion in Green and Sustainable Chemistry* **2017,** *7*, 23-30.

15. Wrenbeck, E. E.; Azouz, L. R.; Whitehead, T. A., Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nature Communications* **2017,** *8* (1), 15695.

16. Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N., Computational Enzyme Design. *Angewandte Chemie International Edition* **2013,** *52* (22), 5700-5725.

17. Tishkov, V. I.; Pometun, A. A.; Stepashkina, A. V.; Fedorchuk, V. V.; Zarubina, S. A.; Kargov, I. S.; Atroshenko, D. L.; Parshin, P. D.; Shelomov, M. D.; Kovalevski, R. P.; Boiko, K. M.; Eldarov, M. A.; D'Oronzo, E.; Facheris, S.; Secundo, F.; Savin, S. S., Rational Design of Practically Important Enzymes. *Moscow University Chemistry Bulletin* **2018,** *73* (1), 1-6.

18. Bunzel, H. A.; Garrabou, X.; Pott, M.; Hilvert, D., Speeding up enzyme discovery and engineering with ultrahigh-throughput methods. *Current Opinion in Structural Biology* **2018,** *48*, 149-156.

19. Kries, H.; Blomberg, R.; Hilvert, D., De novo enzymes by computational design. *Current Opinion in Chemical Biology* **2013,** *17* (2), 221-228.

20. Yan, B.; Ran, X.; Jiang, Y.; Torrence, S. K.; Yuan, L.; Shao, Q.; Yang, Z. J., Rate-Perturbing Single Amino Acid Mutation for Hydrolases: A Statistical Profiling. *The Journal of Physical Chemistry B* **2021,** *125* (38), 10682-10691.

21. Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G., HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *Journal of Chemical Theory and Computation* **2016,** *12* (4), 1845-1852.

22. Godwin, R. C.; Melvin, R.; Salsbury, F. R., Molecular Dynamics Simulations and Computer-Aided Drug Discovery. Springer New York: 2015; pp 1-30.

23. Parton, D. L.; Grinaway, P. B.; Hanson, S. M.; Beauchamp, K. A.; Chodera, J. D., Ensembler: Enabling High-Throughput Molecular Simulations at the Superfamily Scale. *PLoS Comput Biol* **2016,** *12* (6), e1004728.

24. Amrein, B. A.; Steffen-Munsberg, F.; Szeler, I.; Purg, M.; Kulkarni, Y.; Kamerlin, S. C. L., CADEE: Computer-Aided Directed Evolution of Enzymes. *IUCrJ* **2017,** *4* (1), 50-64.

25. Fu, H.; Chen, H.; Cai, W.; Shao, X.; Chipot, C., BFEE2: Automated, Streamlined, and Accurate Absolute Binding Free-Energy Calculations. *J Chem Inf Model* **2021,** *61* (5), 2116-2123.

26. Carvalho Martins, L.; Cino, E. A.; Ferreira, R. S., PyAutoFEP: An Automated Free Energy Perturbation Workflow for GROMACS Integrating Enhanced Sampling Methods. *J Chem Theory Comput* **2021,** *17* (7), 4262-4273.

27. Senapathi, T.; Suruzhon, M.; Barnett, C. B.; Essex, J.; Naidoo, K. J., BRIDGE: An Open Platform for Reproducible High-Throughput Free Energy Simulations. *J Chem Inf Model* **2020,** *60* (11), 5290-5295.

28. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021,** *596* (7873), 583-589.

29. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; Dijk, A. A. v.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D., Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021,** *373* (6557), 871-876.

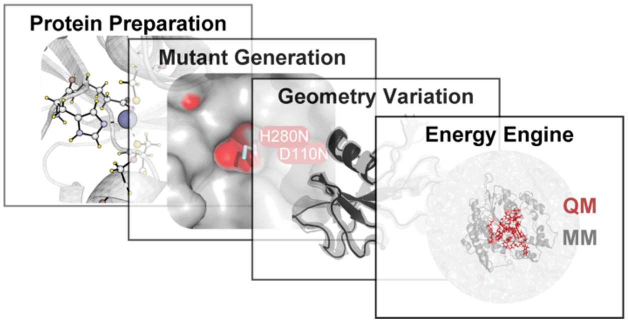
30. Jamroz, M.; Kolinski, A., Modeling of loops in proteins: a multi-method approach. *BMC Structural Biology* **2010,** *10* (1), 5.

31. Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A., PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research* **2007,** *35* (Web Server), W522-W525.

32. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011,** *3* (1), 33.
33. Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H., PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation* **2011,** *7* (2), 525-537.
34. Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V., H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Research* **2012,** *40* (W1), W537-W541.
35. D.A. Case, H. M. A., K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xue, D.M. York, S. Zhao, and P.A. Kollman Amber 2021.
36. Bienert, S.; Waterhouse, A.; Tjaart; Tauriello, G.; Studer, G.; Bordoli, L.; Schwede, T., The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Research* **2017,** *45* (D1), D313-D319.
37. Roe, D. R.; Cheatham, T. E., PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation* **2013,** *9* (7), 3084-3095.
38. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.
39. Lu, T.; Chen, F., Multiwfn: A multifunctional wavefunction analyzer. *Journal of Computational Chemistry* **2012,** *33* (5), 580-592.
40. Kim, T. H.; Mehrabi, P.; Ren, Z.; Sljoka, A.; Ing, C.; Bezginov, A.; Ye, L.; Pomès, R.; Prosser, R. S.; Pai, E. F., The role of dimer asymmetry and protomer dynamics in enzyme catalysis. *Science* **2017,** *355* (6322), eaag2355.
41. Makinen, M. W.; Fink, A. L., Reactivity and Cryoenzymology of Enzymes in the Crystalline State. *Annual Review of Biophysics and Bioengineering* **1977,** *6* (1), 301-343.
42. Mehrabi, P.; Di Pietrantonio, C.; Kim, T. H.; Sljoka, A.; Taverner, K.; Ing, C.; Kruglyak, N.; Pomès, R.; Pai, E. F.; Prosser, R. S., Substrate-Based Allosteric Regulation of a Homodimeric Enzyme. *Journal of the American Chemical Society* **2019,** *141* (29), 11540-11556.

43. Mehrabi, P.; Schulz, E. C.; Dsouza, R.; Müller-Werkmeister, H. M.; Tellkamp, F.; Miller, R. J. D.; Pai, E. F., Time-resolved crystallography reveals allosteric communication aligned with molecular breathing. *Science* **2019,** *365* (6458), 1167-1170.
44. Schulz, E. C.; Mehrabi, P.; Müller-Werkmeister, H. M.; Tellkamp, F.; Jha, A.; Stuart, W.; Persch, E.; De Gasparo, R.; Diederich, F.; Pai, E. F.; Miller, R. J. D., The hit-and-return system enables efficient time-resolved serial synchrotron crystallography. *Nature Methods* **2018,** *15* (11), 901-904.
45. Yue, Y.; Fan, J.; Xin, G.; Huang, Q.; Wang, J.-B.; Li, Y.; Zhang, Q.; Wang, W., Comprehensive Understanding of Fluoroacetate Dehalogenase-Catalyzed Degradation of Fluorocarboxylic Acids: A QM/MM Approach. *Environmental Science & Technology* **2021,** *55* (14), 9817-9825.
46. Fried, S. D.; Boxer, S. G., Measuring Electric Fields and Noncovalent Interactions Using the Vibrational Stark Effect. *Accounts Chem Res* **2015,** *48* (4), 998-1006.
47. Fried, S. D.; Boxer, S. G., Electric Fields and Enzyme Catalysis. *Annual Review of Biochemistry* **2017,** *86* (1), 387-415.
48. Welborn, V. V.; Head-Gordon, T., Fluctuations of Electric Fields in the Active Site of the Enzyme Ketosteroid Isomerase. *J. Am. Chem. Soc.* **2019,** *141* (32), 12487-12492.
49. Bhowmick, A.; Sharma, S. C.; Head-Gordon, T., The Importance of the Scaffold for de Novo Enzymes: A Case Study with Kemp Eliminase. *J. Am. Chem. Soc.* **2017,** *139* (16), 5793-5800.
50. Vaissier, V.; Sharma, S. C.; Schaettle, K.; Zhang, T.; Head-Gordon, T., Computational Optimization of Electric Fields for Improving Catalysis of a Designed Kemp Eliminase. *ACS Catal.* **2018,** *8* (1), 219-227.
51. Yang, Z.; Liu, F.; Steeves, A. H.; Kulik, H. J., Quantum Mechanical Description of Electrostatics Provides a Unified Picture of Catalytic Action Across Methyltransferases. *J. Phys. Chem. Lett.* **2019,** *10* (13), 3779-3787.
52. Bím, D.; Alexandrova, A. N., Local Electric Fields As a Natural Switch of Heme-Iron Protein Reactivity. *ACS Catal.* **2021,** *11* (11), 6534-6546.
53. Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D.; Roskies, R.; Scott, J. R.; Wilkins-Diehr, N., XSEDE: Accelerating Scientific Discovery. *Comput. Sci. Eng.* **2014**, *16*, 62-74.

Table of Contents Graphic


Protein Preparation
Mutant Generation
H280N
D110N
Geometry Variation
Energy Engine
QM
MM