Incorporation and Validation of Flexibility Modelling Tools for *De Novo* Enzyme Design

Ilya Sergeevich Dementyev

Thesis submitted to the University of Ottawa in partial fulfilment of the requirements for the Master of Science (MSc.)

Department of Chemistry and Biomolecular Sciences
Faculty of Science
University of Ottawa

Table of Contents

List of Abbreviations

AI – Artificial Intelligence

BR – Backrub

CED – Computational Enzyme Design

CPD – Computational Protein Design

EGM – Ensemble-Generating Method

ER – Ensemble Refinement

IGP – Indole-3-glycerophosphate

MD – Molecular Dynamics

MG – Motif Generation

MM – Molecular Mechanics

MSA – Multiple Sequence Alignment

PM/PertMin – Perturbation-Minimization

PDB – Protein Data Bank

QM – Quantum Mechanics

RMSD – Root-Mean-Squared Deviation

RP – Repacking

TS – Transition State

TSA – Transition State Analogue

TSR – Transition State, (*R*)-enantiomer

TSS – Transition State, (*S*)-enantiomer

WT – Wild-type

List of Tables

List of Figures

List of Equations

Acknowledgements

First, thank you to Dr. Nathan Luedtke, without whom I would have never discovered my interest and curiosity for computational biochemistry. Thank you to Dr. Roberto A. Chica, who was kind enough to take me under his wing as a graduate student and allowed me to show him what I am capable of. A big thank you to my fellow lab members, whose guidance and kindness never went unnoticed. A special thanks to my lab member, mentor, and colleague, Rojo V. Rakotoharisoa, whose extensive help during grad school was extremely kind and led me to the position I am in today. Lastly, a very warm thank you to my friends and family, in particular my mother and father, Viktoriya Dementyeva and Sergey Dementyev, whose guidance and words of encouragement helped me tremendously.

"*Don't sink, just swim towards the storm, and once again you'll be reborn*"

- Daughter

Abstract

Computational *de novo* enzyme design is a rapidly evolving field, involving the bottom-up design of an active enzyme for an important chemical reaction, starting from an inactive protein scaffold. Many methods and pipelines have been in development for decades, and it has been shown that designing enzymes from a multitude of input templates (ensemble design) rather than a single one (static design) usually leads to better results. Methods for generating ensembles are not perfect and are prone to error as well as biases depending on the specific method used. In this work, backrub's (BR) efficacy as an alternative ensemble generation method for the development of enzymes is explored and utilized to recapitulate pre-existing enzymes with known catalytic efficiencies approaching that of wild-type enzymes, as well as backrub serialized with other methods. These methods are compared to pre-existing ones that do not utilize backrub, and advantages as well as disadvantages are discussed and explained. Further on, backrub ensembles are used to design a new retro-aldolase, TyRA95.0, whose design is then computationally characterized. To conclude, recommendations are given to the next generation of enzyme engineers given what was learned from Backrub's recapitulations and TyRA95.0 design, and new metrics for improved ensemble analysis are discussed.

Chapter 1 – Introduction

1.1 Enzyme Design

To start, enzymes are proteins capable of catalyzing chemical reactions. Such proteins can be thought of as Nature's nano machines designed by evolution for biocatalysts, where some enzymes can increase reaction rates by impressive factors of up to 10^{21}.[1] Eventually, scientific research unearthed the valuable nature of such nano machines, and the first industrial implementations of an enzyme to catalyze a relevant chemical reaction began in the seventies.[2]

This was revolutionary, due to catalysis generally being performed using toxic and environmentally harmful inorganic compounds at concentrations relevant for industrial process chemistry.[3,4,5] However, an industrial enzyme capable of catalyzing the same reaction at comparable speeds is usually not as toxic to handle and is much more environmentally friendly. Generally, these enzymes are redesigned or evolved directly from pre-existing enzymes that perform a similar function,[6,7] using rational design or directed evolution methodologies respectively. Rational design involves the modification or insertion of amino acids or entire domains into a pre-existing protein to achieve the desired result that the modification or insertion normally provides.[8] Of course, this depends on the quality of data available,[9] and sometimes mutations may not perform a desired function. To circumvent this, one could use directed evolution, a process by which an enzyme's sequence is randomly mutated in an iterative fashion, creating a library of variants which can then be expressed and selected for desired characteristics, discarding other sequences (similar to natural evolution).[10,11] The improved genes are then subjugated to the same mutagenesis, and the process can be repeated until one has the desired level

of activity. While these methods are powerful, they rely on the enzyme initially having catalytic function – an inactive enzyme's sequence cannot be used as a starting point for the iterative mutagenesis, or rational design.[12] Hence, sometimes an enzyme must be designed from scratch, especially when there are cases of industrial reactions that don't have a natural biochemical analog catalyzed by a known enzyme.[13] For that case, one must create a *de novo* ("of new") enzyme, one that doesn't already exist in nature. In general, the *de novo* enzyme creation pipeline involves input structure selection, theozyme design, theozyme placement, and repacking.[14,15,16]

1.2 Computational Enzyme Design

Computational Enzyme Design (CED) is the process of engineering a catalytic protein from an inactive one for a reaction that has no natural bio-catalyzed analogue. Since this process creates a new enzyme, rather than re-designing an enzyme that already exists, it is also referred to as *de novo* CED, as mentioned earlier. Generally, this is performed via a bottom-up approach,[13,16] by creating a stabilized configuration of the ligand and catalytic sidechains involved in the catalysis, using quantum mechanical (QM) methods that can calculate the optimum geometry of such a system.[17] This set of geometrically optimized atoms is then grafted onto a potential active site within a protein backbone scaffold using a matching algorithm.[18] The residues around the catalytically relevant species are then optimized to ensure stability.[19] For the reader's ease, each step in the general pipeline for *de novo* enzyme creation is laid out below in further detail.

1. **Input Preparation:** To design an enzyme using computational methods, one must first decide on the input protein template.[20] Generally, this is a PDB structure of a protein,

or any coordinate file that contains the non-hydrogen atoms of the protein structure as well as any bound ligand. Then, before running anything, the file is standardized.[14] This involves removing all water molecules, ligands and counter-ions. For some standardizations, this will also include removing ligand molecules. Then, the program used by the researcher will add hydrogens the system, include any residues whose sidechains must be protonated.[21] Generally, this protein structure is chosen from a collection of protein templates that are able to accommodate the theozyme while ensuring that the theozyme's geometry is not significantly modified, nor should the theozyme be engaged in any steric clashes.[15] Sometimes, this protocol is followed-up by a quick minimization of the system. Sometimes, one desires a collection of similar backbone structures of the same protein, in order to represent the inherent flexibility of the protein in aqueous media. This is known as ensemble design[22] and is discussed further in Chapter 1.3.

2. **Theozyme Design:** Directly after the input preparation follows theozyme design,[17] in which the reactant ligand and the catalytic residue sidechains (Figure 1.1 – bottom pink box) are isolated from a protein environment and geometrically optimized using QM methods leading to a collection of functional groups with a transition-state-stabilizing geometry, fixed in space by arbitrary constraints.[17] This array is known as the theozyme.
3. **Theozyme Placement:** After the theozyme has been generated, the next step involves its incorporation into the backbone input structure(s) generated in the first step,[23] in absentia of sidechains within or close to the site designed by the enzyme engineer. The software attempts this by running a matching algorithm (e.g. PhoenixMatch[24]) that can

identify certain locations on the backbone scaffold that would allow the theozyme's pre-defined geometries to be satisfied. Afterwards, the best lowest-energy structures are chosen as the starting inputs for the next step.

4. **Repacking:** Finally, after the theozyme has been successfully incorporated into the backbone, the residues located on the design positions (residue positions that are important for enzyme activity) are optimized to stabilize the sequence,[25] by deriving the proper rotamer from a backbone-dependent, backbone-independent, or a continuous library of allowed rotamers. Such positions are usually chosen by relative proximity to the ligand,[15] by analyzing which residues' sidechain atoms lie within a radius of a sphere centered on the ligand. Residues whose sidechains point away from the active site are often removed from consideration.

A visual summary of this pipeline can be found in Figure 1.1. Generally, visual inspection at all stages is also recommended to ensure that bond lengths and atomic positions are chemically appropriate.[15]

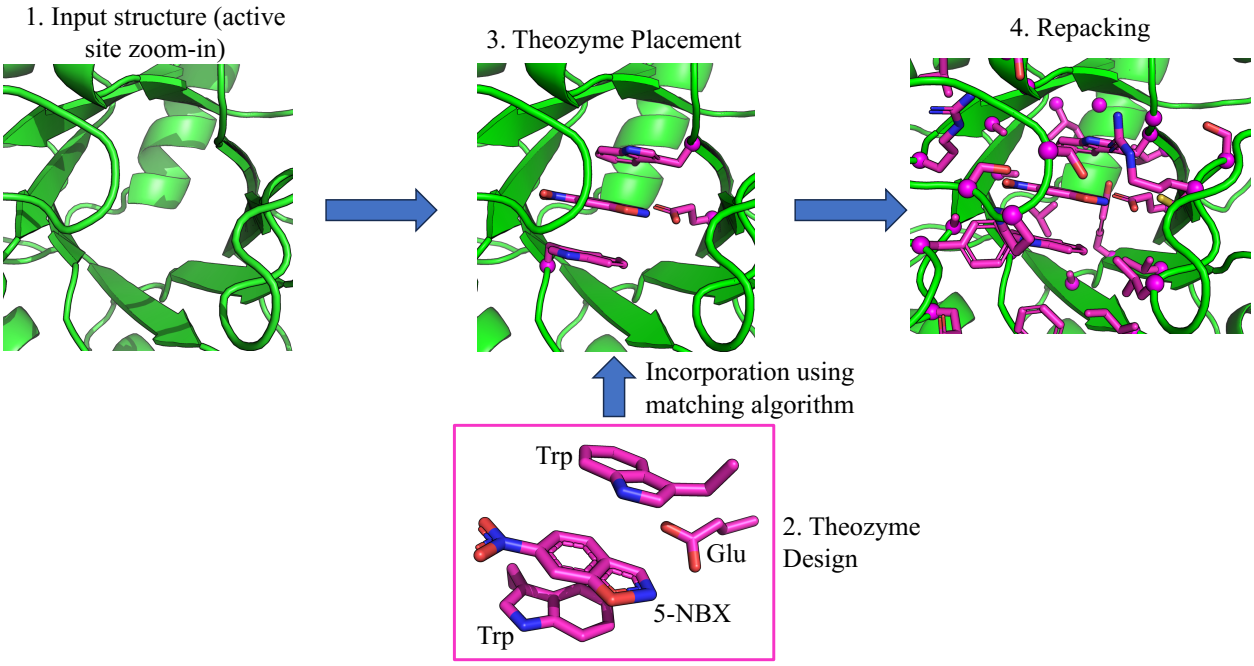


Figure 1.1: Visualized pipeline for de novo enzyme design. In this particular example, the theozyme is a 5-nitrobenzisoxazole ligand (5-NBX), with the reactive glutamate, sandwiched between 2 tryptophans to stabilize the transition state (TS) via π-stacking. The α-carbons are emphasized as spheres in step 4. The input structure chosen is from 1A53-core, but any input structure could have been chosen as an example.

Using the CED pipeline (Figure 1.1), many enzymes have been created in a *de novo* fashion.[17,26,27,28] One of the most notable examples is the retro-aldolase RA95 by Althoff et. al.[29]. This was an impressive feat, as at the time, it was the most mechanistically complex enzyme created from scratch and catalyzed an industrially relevant carbon-carbon bond breaking reaction. Unfortunately, the enzyme's efficiency was drastically lower than the average natural enzyme, and needed a follow-up directed evolution experiment to improve its activity, which luckily the enzymes are quite amenable to.[30] Hence, although such a design pipeline can confer activity to a previously inactive protein backbone, the activity itself may not be as high as preferable, especially in the context of industrial processes, where enzymes need high efficiencies for reactions to proceed at an acceptable timescale.[31,32] Although the directed evolution (DE) does eventually lead

to significant efficiency improvements, it takes up time and costly resources to do so. For instance, the most difficult step in DE is developing the screening method,[33] which may be impractical for many enzyme activities of interest. Hence, although directed evolution can close the catalytic efficiency gap between de novo designed enzymes and natural enzymes, it is important to improve the *de novo* pipeline such that it can design high efficiency enzymes in a one-shot manner.

One such way is to transition from a single backbone template for the input structure, to using an ensemble of backbone templates for the same step. Althoff et. al. used a single backbone template to design the enzyme. Although this is a valid part of the design process (Step 1: Input Structure Design) and helps to improve computational efficiency, the use of a single backbone *foregoes any potential modelling of the enzyme's inherent backbone flexibility*, an undeniably important aspect of an enzyme's general flexibility (even in crystalline conditions[34]), as only one structure is allowed to represent the protein. This causes sidechain rotamers that may normally be allowed if the backbone was slightly shifted to be removed from consideration, resulting in a false negative, and that sequence being discarded. Also, a single-state backbone can cause the predicted transition state to not be as tightly bound,[35] resulting in lower catalysis. Furthermore, conformational entropic effects are not modelled, as well as second-shell interactions (interactions between the β-barrel and the surrounding α-helices), and flexibility important for efficient turnover.[28] All of these factors may lead to a worse prediction of the final active site structure, as seen in previous studies.[14] Sometimes, using a single-state design might lead to an enzyme with absolutely no measurable activity, as was the case with HG-1 designed by Privett et. al.[26] It was only after considering the flexibility of HG-1 via MD that the causes for the lack of catalytic activity was identified: (1) large active site entrance led to an increase in solvent molecules which

diminished activity, and (2) active site residues were too flexible to ensure proper rotamer positioning beneficial for catalysis. Once these issues were addressed in a new variant (HG-2), activity was detected.[26]

To improve this approach, one method used is an ensemble approach, whereby an array of backbone structures are used as templates, rather than a single input structure. This approach can lead to improved predictions in ligand placement and rotamer recapitulation. To test this, one can perform a recapitulation, where a target enzyme of known structure and sequence is designed using a given list of input backbone structures, as well as the sequence to be predicted. The ligand and rotamer structure are then predicted using protein design software such as Triad. To compare the recapitulation qualities, one of the ways to do so is by comparing the accuracy of recapitulations is by checking the number of predicted rotamers that fit within the same bin as the target crystal structure. Since rotamers lie within relatively well-defined probability sets (bins),[36] if the predicted rotamer fits within the same defined bin as the crystal's rotamer, then we may consider that a successful recapitulation. Furthermore, whichever predicted ligand's heavy-atom RMSD (relative to the crystal structure's ligand) is closer to 0, the more accurate the prediction. The recapitulations of a Kemp Eliminase is shown below, using both single and ensemble template approaches (Figure 1.2).

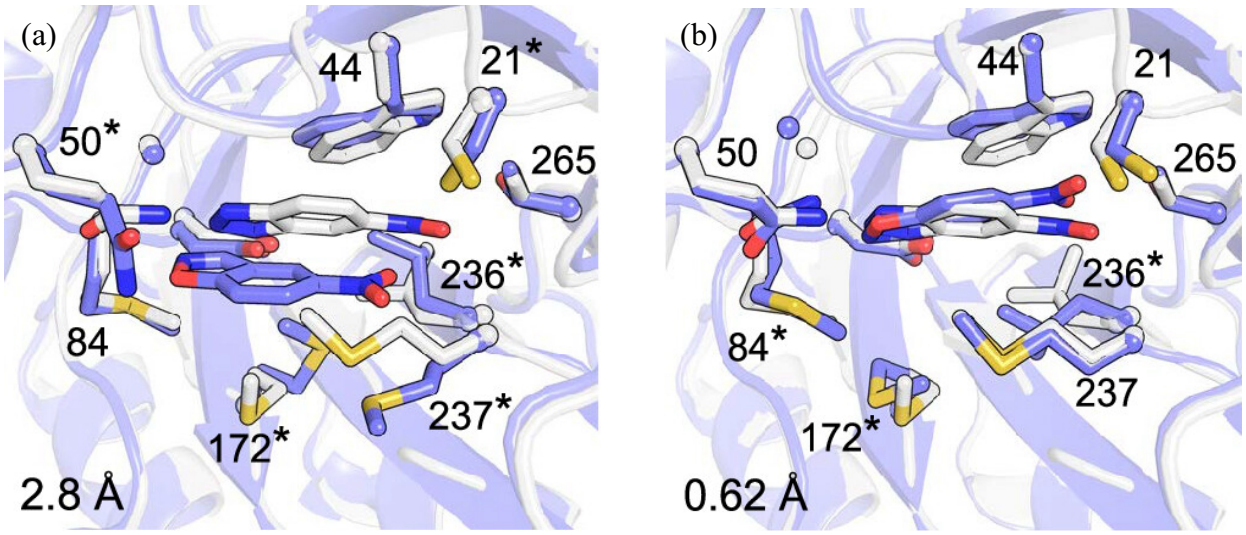


Figure 1.2: Single-backbone design has been shown to lead to prediction issues. Comparison between (a) single-state (1 backbone) and (b) ensemble-state (multiple backbone) active site recapitulations of HG185,[14] with original crystal structure shown in gray carbons and the design in purple, with heavy-atom ligand RMSDs in the bottom left corner. Note the high RMSD disagreement for the final structure generated from single-state design. Asterisks represent the residues who's predicted rotamers deviated from the bins of the crystal rotamers. The number of asterisks is noticeably higher in the single-state design (a), indicating less correct rotamers predicted. Figure used with permission from Rakotoharisoa et. al.[14]

As one can see from Figure 1.2, the prediction inaccuracy is made visible by the high ligand RMSD, and the larger number of incorrect rotamers (Figure 1.2a). Although this was a recapitulation, and the issues are more shown in a more obvious fashion, doing a design with a single input structure when the final target structure is not known can lead to an exacerbation of the issues previously mentioned, especially for a new structure whose backbone is not as well-studied. Hence, it may be pertinent to use an ensemble of input template protein structures for engineering a new enzyme, and to know how to design such an ensemble would be crucial to increase active site predictions between an *in silico* generated enzyme and the target.

1.3 Sampling and Modeling Protein Movement for Enzyme Design

There are many ways to sample the physically feasible geometric transformations that occur in a protein, the most obvious involve taking collections of 3D structure files directly from experimental data. One way to do this is by collecting NMR data of proteins in solution, the first of which was done by Wüthrich's lab in 1985.[37] Briefly, a protein structure is isotopically labelled (using ^{15}N and ^{13}C), then analyzed inside an NMR spectrometer, after which structural calculations and refinement is performed, in order to assess the structural quality.[38] The protein sample can be simply suspended in an aqueous buffer, or placed into more complex systems, such as a reverse micelle. NMR for protein structure determination can also be done in solid-state, which is particularly useful for studying protein-membrane interactions and conformational changes.[39] These factors make NMR a versatile method for analyzing protein structures, especially when x-ray crystallographic data is unavailable, and can even lead to similar results for protein designs under various simulation conditions.[40] While NMR is effective at analyzing *in vitro* or *in vivo* protein structural information, ensembles generated from it have been shown to perform poorly for sequence designs by Davey et. al.,[41] resulting in sequences with lower stabilities than WT folds. This was hypothesized to occur due to NMR ensembles being "off-target" (i.e. their templates had low structural similarity to the crystal structure).[41] Also, NMR-derived ensemble members generally have lower rotamericity (high deviations of X_1 and X_2 angles) relative to X-ray crystallographic structures, which decreases the likelihood of fitting even the native amino acid into its corresponding location in the structure,[40] and other studies showed inaccuracy of prediction of WT sequences leading to energies similar to that of unfolded proteins.[42] Furthermore, new sequences derived from Gβ1 NMR data by Allen et. al.[42] resulted in structures adopting non-native

folds (or not folding at all). Hence, NMR is a versatile ensemble design tool, capable of representing an *in vitro* fold accurately but can lead to improper designs for new sequences, a worrying issue for the goal of *de novo* enzyme design. This was not seen with molecular dynamics (MD).

MD is a computer simulation method for modelling atom movements in a system according to Newton's laws of motion, usually involving several molecules whose interactions are modelled using a forcefield.[43] Essentially, the atomic trajectories are found by solving Newton's classical equations of motion, which is easily done by computing the gradient of the potential energy, derived directly from a chosen forcefield used to model the energy of the system. This will then run a simulation causing the movement of the atoms at each timestep. Snapshots during set intervals are taken, in order to generate templates for the final ensemble. Choosing the proper system, as well as the proper parameters such as forcefield specifications, can lead to accurate predictions of important protein flexibility,[44] hence leading to ensembles that may lead to more accurate designs than previously mentioned ensemble generation methodologies. Unfortunately, conformational changes that occur on millisecond timescales are difficult to model given that most MD simulations run at 1-2 fs timesteps.[45] Furthermore, utilizing Newton's laws ignores quantum effects, which can be mitigated with joint quantum-mechanics/molecular mechanics methods (QM/MM)[45,46] but runs into similar issues with large processing requirements as with the previous case.

Another way to model conformational changes and acquire a structural ensemble is through ensemble refinement (ER). This is a protein ensemble-generating method (EGM) that utilizes a

protein crystal's electron density data, along with time-averaged molecular dynamics (MD) to generate an ensemble of proteins that is restrained by electron density parameters taken from the MTZ file in the protein's PDB entry.[47] This data is superior to the static single-structure PDB, as ER can model local atomic fluctuations without oversampling the global disorder of the protein, and whose MD is restricted to what is experimentally observed. Generally, the R_{free} (measure of agreement between crystallographic model and a subset of the data[48]) and the R_{work} (measure of agreement between crystallographic model and data[49]) values for the protein tends further to zero after having undergone refinement than before, indicating an increased quality of the generated structures to the experimental data.[50] It is also known that using ensembles created from constrained MD as in ER can lead to improved sequences for immunoglobulins as suggested by Davey et. al.[41]. This method's main limitation is requiring crystallographic data to implement it during protein design, not just the coordinate file deposited in the Protein Data Bank. Furthermore, it is difficult to locally control the magnitude of perturbations that occur in different regions of the protein, which can diminish a researcher's ability to control the output.

Another method that can be used to generate ensembles from a starting input structure is the Perturbation-Minimization Protocol (PertMin). PertMin is a protein design module designed by Davey et. al.[41] that randomly perturbs each non-hydrogen atom in a protein input structure by around ± 0.001 Å along each Cartesian axis. As one might expect, the geometry of these structures is virtually unchanged, having an RMSD diversity of 0.0017 Å. Afterwards, the perturbed structures are minimized via a truncated Newton algorithm, which causes the initially identical structures to diverge along the geometric and energy landscape, creating the necessary diversity in the ensemble while exploring the space locally, never straying far from the original structure.

PertMin is highly useful for sequence design, as its minimization capabilities lead to stabilizations in the protein backbone, allowing for better predictions of low-energy sequences that lead to proper folding, as was seen in Davey et. al.'s experiments with WT folding of Gβ1 using novel sequences.[41,51,52] However, as a logical consequence of PertMin's minimization methodology, it cannot explore other potential energy surface (PES) wells beyond the one it is already located in, resulting in poor modelling of high-energy conformational changes. Furthermore, it is a holistic structural manipulation tool, and cannot be used for minimizing a specific region of an input protein.

Recently, the development of AI tools for protein science such as AlphaFold[53] and RF*diffusion*[54] has led to a significantly easier approach to modelling and designing proteins from a bottom-up perspective, whereupon the one-dimensional information of a protein (i.e. its sequence) is enough to generate a near-accurate three-dimensional structure in the former software, and the latter can design a protein programmatically, allowing the user to build a structure containing a certain sequence with a specified symmetry group, or graft a secondary structure onto a disordered domain, all with a single command line execution. Furthermore, in this work, it was decided to focus on physics-based methods that can manipulate the backbone structure of a PDB input, as these methods rely on first principles and do not require prior training, making them adaptable to diverse systems. All methods described have been included as a visual summary in Figure 1.3.

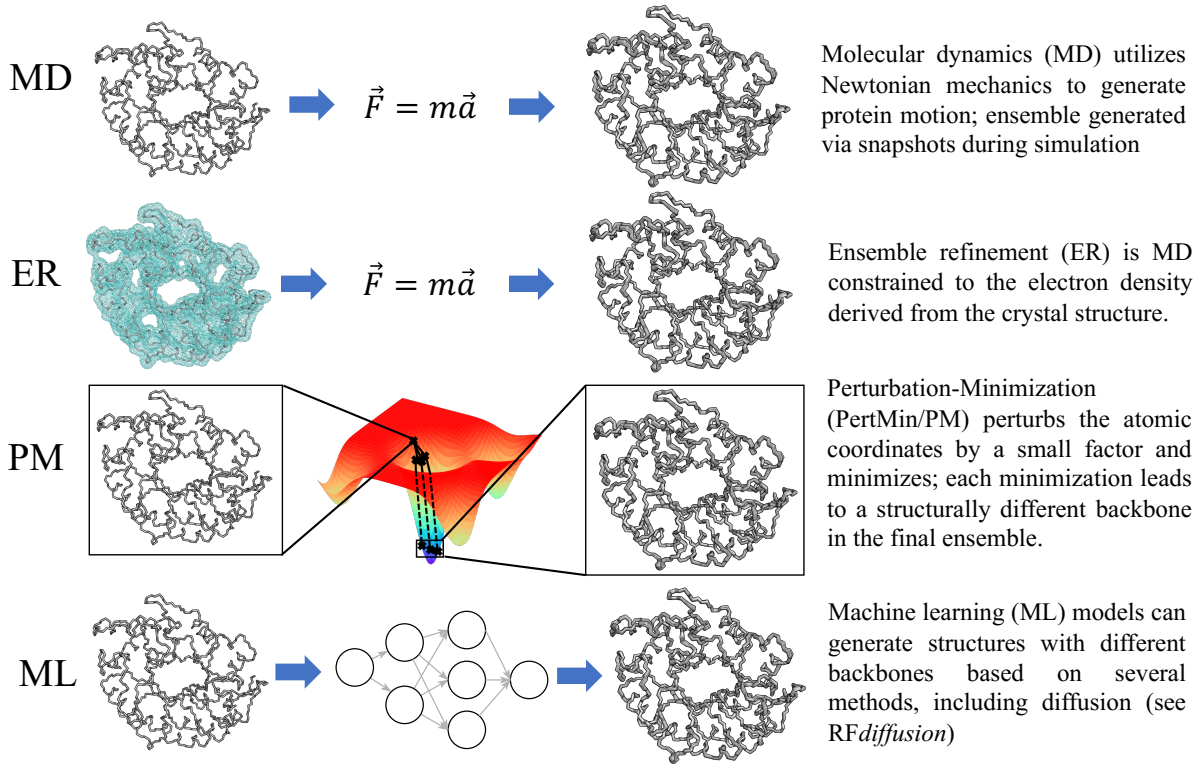


Figure 1.3: A summary of the methods described for ensemble generation.

Each method models flexibility in different manners. NMR-modelled backbone structures are generally diverse, but exhibits low structural similarity to crystallized structures of the same protein, leading to undesirable downstream effects.[41] Ensembles derived from ER reflect the flexibility of the protein within the crystal, meaning segments with undefined secondary structure (e.g. loops) will have a high diversity relative to other secondary structure elements like α-helices. For PM, since the perturbation is not significant, most of the structural diversity from a PM ensemble arises from the minimization, which affects secondary structural elements slightly less than unstructured elements, although like ER, it is a holistic manipulation tool and cannot be used to minimize individual segments of a protein.

Finally, a method that is used frequently to model conformational transitions, yet to the author's knowledge, has rarely been used within *de novo* enzyme design,[30] is the Backrub (BR), the method which this work is mainly focused on and deserves its own section. Although many methods were mentioned earlier, the list is not exhaustive, and reviews for each can be found in the literature.[55,56,57,58,22]

1.4 Backrub

In 2006, the paper "The Backrub Motion: How Protein Backbone [sic.] Shrugs When a Sidechain Dances" was published,[34] outlining the BR motion as a form of backbone plasticity in which a local consecutive sequence of amino acid residues (usually more than 3) is rotated slightly around the axis made by the first and last Cα in the sequence (Figure 1.4), sometimes accompanied by a rotamer change within the sequence. The residues containing the first and last Cα have relatively small rotations and are thus treated as immobile during the rotation, which usually occurs at a magnitude of around 5°. This was discovered in the *in vitro* protein crystal structures, and was later adapted into its algorithmic analogue by the same team which developed the BACKRUB tool for KiNG.[59]

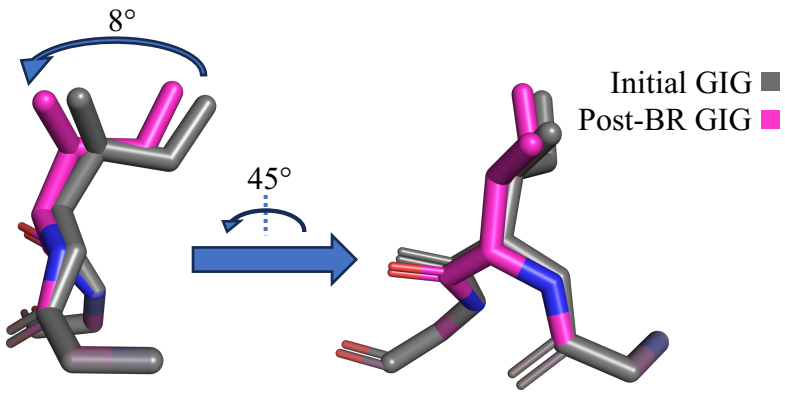


Figure 1.4: An 8° BR performed in Triad on a consecutive GIG sequence (PDB: 1D9J, Model 9). The N-C termini vector points into the screen. The rotation direction can be found via the right-hand rule, with the index finger pointing in the direction of the termini vector, and the thumb into the direction of the sidechain branch. Original peptide in gray, backrubbed colored. Hydrogens removed for clarity.

Later, Smith and Kortemme published their version of BR, integrated into Rosetta,[60] a well-known protein design software suite, for more broader applications to protein engineering. Since then, BR has been effectively incorporated into designs and analysis of various systems, including protein-protein interface design,[61] protein-ligand specificity,[62] protein-protein docking,[63] *de novo* design of high-affinity antibody variable regions,[64] and other simulations where accurate conformational sampling is required, as it can still provide reliable structures that in some cases fully recapitulate conformational flexibility of proteins, which other methods like NMR could not. For instance, BR was successfully used to recapitulate the open/closed states of a triosephosphate isomerase loop by Smith and Kortemme in 2008.[60] Also, BR can be used to perform accurate sequence design while simultaneously modelling the backbone via flexible backbone design, something that NMR ensembles were not useful for.[65] Furthermore, BR can be easily redesigned to only target certain locations during a simulation for constructing an ensemble, something that is noticeably more difficult to perform in the case of MD, ER, and PertMin. To the author's knowledge, this approach has not been done before. The BR is applied to a specific region in order to remodel backbone structures for improved sidechain-ligand interactions. To elaborate, most methods like PertMin are holistic (as mentioned in Section 1.3 Sampling and Modeling Protein Movement for Enzyme Design), and don't offer control over being able to design an ensemble of only a local region of a protein. It is also easier to explore an energy landscape given the Metropolis Monte Carlo (MC) algorithm (Figure 1.5) is used to push the simulation forward without

significantly increasing the energy, without any minimization performed like in PertMin or MD and avoiding being too entrenched in an energy well. It is also not restricted by the protein's electron density, allowing for more free motion than ER. Hence, given the advantages of BR relative to other EGMs, and its relatively few disadvantages, the incorporation of BR into the design of protein ensembles for de novo enzyme design could result in improved predictions compared to previous methods seen in Section 1.3 Sampling and Modeling Protein Movement for Enzyme Design. Within this work, it will be explored whether BR on its own can generate good predictions for designs (and help conserve simplicity in the pipeline), or if joining it with another EGM is necessary to generate acceptable final predictions. To do this, one must first choose a template PDB to use as a model.

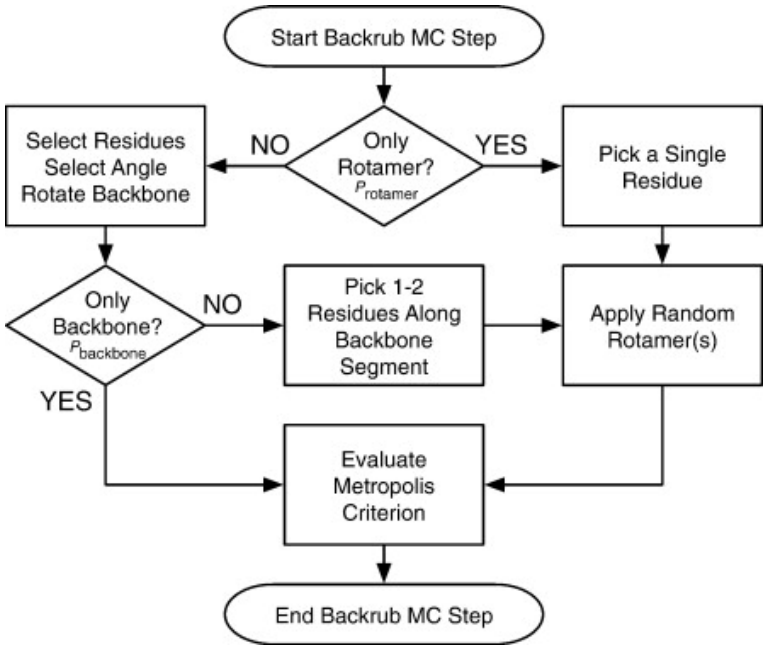


Figure 1.5: The original Backrub Metropolis MC Workflow by Smith and Kortemme, representing a single MC step. The number of MC steps, probability of choosing only a rotamer path, the residue number and the maximum rotation angle are all user-defined. For consistency, this was the same workflow used to design the backrub for Triad. Figure reproduction permissions obtained from Elsevier (via Copyright Clearance Centre).

1.5 Model Reactions for Studying De Novo Enzyme Design

To study the effects of new ensemble generation methods on de novo enzyme design, one can perform a recapitulation of a previously-designed de novo enzyme. In this context, this involves the design of a target enzyme with a known backbone and sequence, using an ensemble with different backbones and a sequence that matches the target's. In this work, the Kemp Eliminases 1A53-core and HG4 were recapitulated to evaluate the efficacy of different ensemble generation techniques, assuming that the exact same design protocol is used (Figure 1.1). After successfully recapitulating an enzyme using an ensemble method, one can proceed to attempt engineering of a new de novo enzyme based off of a previous design, which is the retro-aldolase RA95 in this work.

1.5.1 Kemp Elimination

The Kemp elimination is a base-mediated deprotonation reaction of a nitrobenzisoxazole (Figure 1.6), undergoing one transition state that can be stabilized by a hydrogen-bonding donor, without any intermediates, forming a 2-cyanophenolate product.[66] The reaction works best in a hydrophobic and basic environment, making an enzyme's active site a theoretically amenable location for reaction progression.[67] This model reaction is well-studied in *de novo* enzyme engineering, primarily due to its simplicity as a model for proton abstraction by carbon. Also, no known WT enzyme catalyzes this reaction, reducing the likelihood of false positives. Furthermore, the isoxazole ring opening can be quickly analyzed using UV/Vis spectroscopy, leading to rapid

characterization of the enzyme's kinetics.[68] One enzyme that catalyzes such an elimination is 1A53-core (PDB: 8FOQ), designed by Zarifi et. al. (to be published).

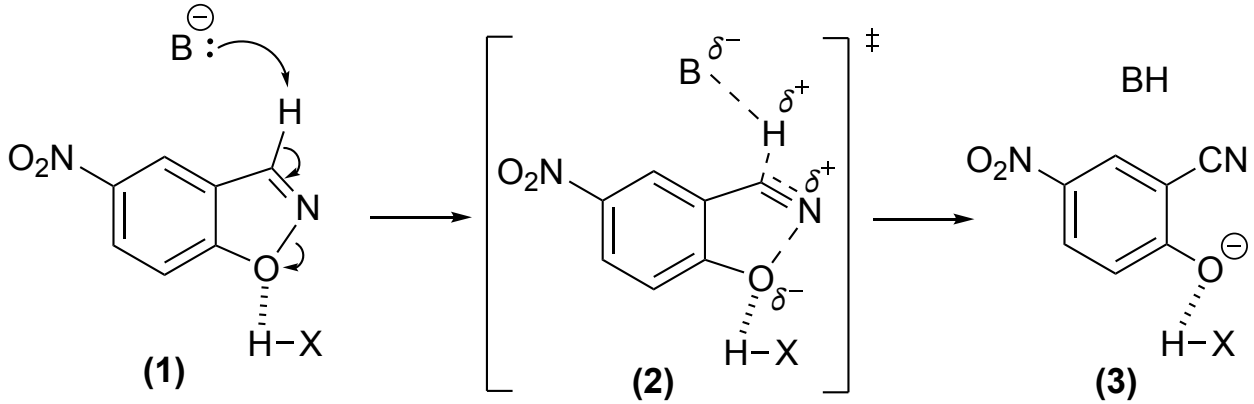


Figure 1.6: A general Kemp elimination of 5-nitrobenzisoxazole (1) to form 2-cyano-5-nitrophenolate (3) via proton abstraction using a generic base B. Throughout the reaction, the reactant and transition state (2) is stabilized via hydrogen bonding using a hydrogen bond donor. For 1A53-core, the base is Glu178's carboxylate.

1.5.1.1 1A53-core

The enzyme 1A53-core was developed by computational design, directed evolution (DE) and rational design (RD) process from its evolutionary ancestor indole-3-glycerolphosphate (IGP) synthase (PDB: 1A53)[69] (Figure 1.7). and is the first enzyme that is the subject of the recapitulation experiments. First, the IGP synthase was computationally designed into 1A53-2 (PDB: 3NZ1) by Privett et. al.,[26] and later Bunzel et. al.[70] subjected the latter to 5 rounds of saturation mutagenesis and DNA shuffling, resulting in 1A53-2.5 (PDB: 6NW4), which was co-crystallized with the transition state analog (TSA) (Figure 1.8) . Bunzel et. al. also subjugated 1A53-2.5 to 4 more rounds of error-prone PCR and DNA shuffling, leading to the 1A53-2.9 enzyme. While this enzyme was unable to be crystallized, its core mutations (changes in residues between 1A53-2 and 1A53-2.9 that are located within a 4 Å sphere of the ligand's geometric center) were incorporated

directly into 1A53-2 to form 1A53-core, during an investigation by Zarifi et. al. to confirm if only core mutations are necessary for high-efficiency catalysis.

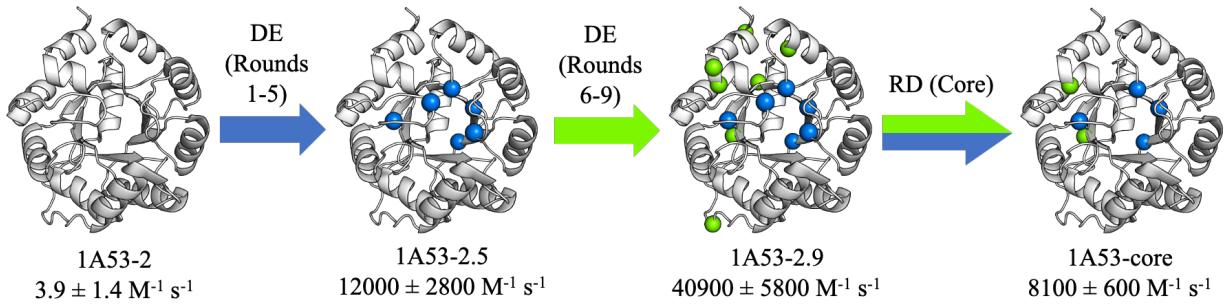


Figure 1.7: Visualized evolution of the 1A53-series enzymes up to 1A53-core. Mutations created in the evolution from 1A53-2 to 1A53-2.5 are shown with the corresponding residue α-carbons in blue, and mutations acquired during the evolution from 1A53-2.5 to 1A53-2.9 shown in green α-carbons. Each enzyme has their name and corresponding catalytic efficiency (in M^{-1} s^{-1}) shown below each structure.

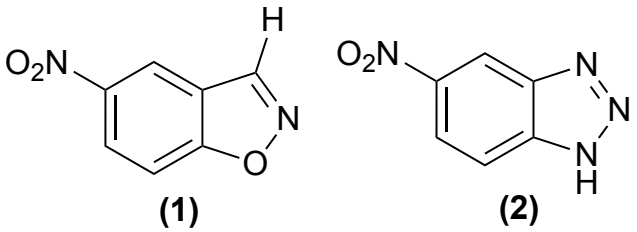


Figure 1.8: The reactant 5-nitrobenzisoxazole (1) and the Transition State Analogue (TSA) (2). The benzisoxazole (1) was used as the ligand in all relevant design simulations.

There are a few reasons for why this Kemp Eliminase was chosen, as opposed to other suitable candidates.[71,72,73] Firstly, this enzyme was chosen for recapitulation studies as it was rationally designed *in silico*, it is a TIM-barrel protein (amenable to a high sequence diversity while retaining WT folding), as well as having been well-studied by Zarifi et. al. Furthermore, 1A53-core contains only near-active site residue mutations, as it was shown previously by the same authors that one can focus solely on active site design without changing distal mutations can

lead to enhanced activity. Also, during past designs of 1A53-core, the predicted model's tyrosine 159 residue did not match the rotamer that appeared in the crystal structure (Figure 1.9b). This is an issue, as the rotamer must be correct in order to facilitate the necessary preorganization of the catalytic E178 via hydrogen-bonding stabilization to improve 1A53-core's elimination efficiency.[74] Upon further investigation, it was hypothesized that an ensemble generated using BR could potentially recapitulate the correct rotamer for Y157 which known to have important hydrogen-bonding interactions.[70] This hypothesis arises from the structure of residue 157 in 1A53 to 1A53-core (Figure 1.9a), and how it appears to be a BR, showing similarity to the model BR demonstrated in Figure 1.3. Hence, it is likely that implementing BR during ensemble design could aid in approximately modeling such motion, leading to an improved prediction accuracy for invaluable intramolecular interactions. To generalize, if it is possible to recapitulate the active site structure of this protein with BR, or the BR integrated with other computational protein design methods, it would suggest that designing an enzyme with a rationally engineered active site via BR is possible.

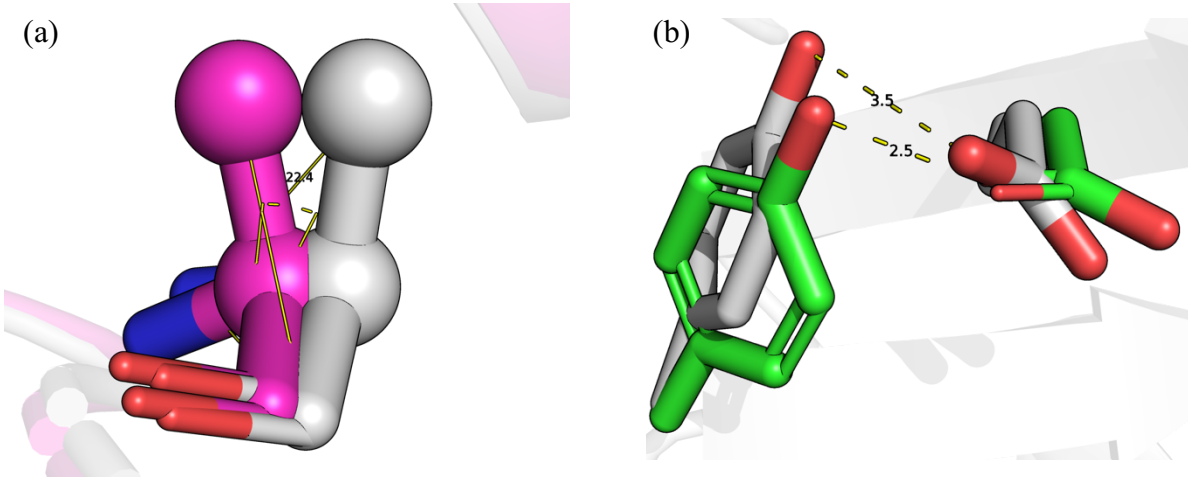


Figure 1.9: BR can potentially perform the movement necessary in the protein backbone to recapitulate the desired Y157 rotamer. (a) A ball-and-stick model of residue 157 α and β carbons from 1A53-core and 1A53, showing a potential ~22° BR motion (the actual value is likely smaller due to the deviations between backbone nitrogen and carbon positions). 1A53-core in magenta, 1A53 in gray. This appears to be a BR motion and is similar to the model shown in Chapter 1.4 (Fig. 1.3) (b) The designed Y157 rotamer (green) is not in the same bin as the crystal structure (gray). This is what is seen in previous ensemble generation attempts, and what will be hopefully circumvented via BR.

Another enzyme that was chosen for recapitulation studies within this work was HG4.

1.5.1.2 HG4

In 2012, Privett et. al.[26] engineered a *de novo* Kemp Eliminase named HG-1 from a thermostable xylanase discovered in *T. aurantiacus* (PDB: 1GOR) using *in silico* methods previously developed.[24] Unfortunately, HG-1 displayed no observable catalytic activity during *in vitro* characterization, leading to the discovery of two issues that contributed to the lack of activity, namely the presence of water molecules in the crystallized active site and the high flexibility of catalytically-relevant residues. Upon their resolution, the next generation HG-2 enzyme was catalytically active, albeit orders of magnitude less than a WT enzyme. Further MD analysis showed that a single point mutation from HG2 (S265T) that led to the HG-3 design (PDB: 5RG4) would improve catalysis through higher preorganization, which was confirmed *in vitro*. Broom et. al.[75] continued the design in 2020 using a combination of rational engineering and directed evolution to design HG4 (PDB: 5RGF) (Figure 1.10), which had improved catalysis from its evolutionary predecessor HG3, after widening the entrance to the active site and further increasing active site residue pre-organization. HG4's catalytic efficiency was recorded at 103,000 M^{-1} s^{-1}, one of the highest-efficiency Kemp Eliminases designed.

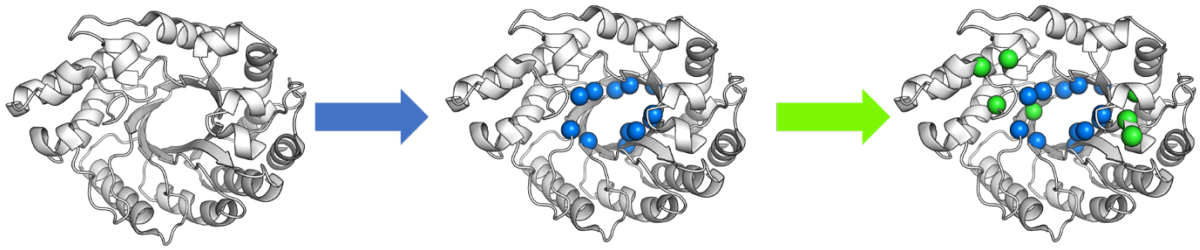


Figure 1.10: Visualized evolutionary trajectory of the HG-series enzymes up to HG4. Mutations introduced during the design of *T. aurantiacus* xylanase (TAX) to HG3 are shown with their α-carbons in blue, and mutations acquired during the evolution from HG3 to HG4 shown in green α-carbons. HG2 was not shown given that it differed from HG3 by a single point mutation. Each protein's catalytic efficiency is shown below their name.

In this work, HG4 (Figure 1.10) was chosen for recapitulation as it and its ancestors have already been well-studied (see Broom et. al.[75] and Privett et. al.[26]). Furthermore, this enzyme is highly optimized, approaching catalytic efficiencies present in some WT enzyme counterparts. Hence, being able to recapitulate the structure directly from a non-active thermostable WT protein (PDB: 1GOR in this work) using either BR or BR joined with other methods, would demonstrate the theoretical ability to design a high-activity enzyme from inactive templates solely using non-ML computational methods, leading the field of protein engineering closer to efficiently designing high-activity custom enzymes relevant for industrial applications. As such, one of the main goals addressed in this work (Section 1.6 and 2) is the demonstration of BR's recapitulation ability, as well as its potential for diversifying novel enzyme sequences (Section 3) to create enhanced activity from pre-existing *in silico* enzymes.

1.5.2 Retro-aldol Reaction

The retro-aldol reaction is the reverse reaction of the well-studied aldol condensation in organic chemistry, which can be catalyzed by a variety of species, including acids,[76] bases,[77] and ionic liquids.[78] The forward reaction involves the addition of an enolate to an aldehyde or ketone, forming a β-hydroxy aldehyde or ketone with a newly-formed carbon-carbon bond (Figure 1.11). Hence, the reverse (retro) reaction involves the breaking of a carbon-carbon bond to reform the enol and aldehyde or ketone reactants. One of the enzymes designed to catalyze the retro-aldol reaction is RA95.0, belonging to a class of enzymes known as retro-aldolases.

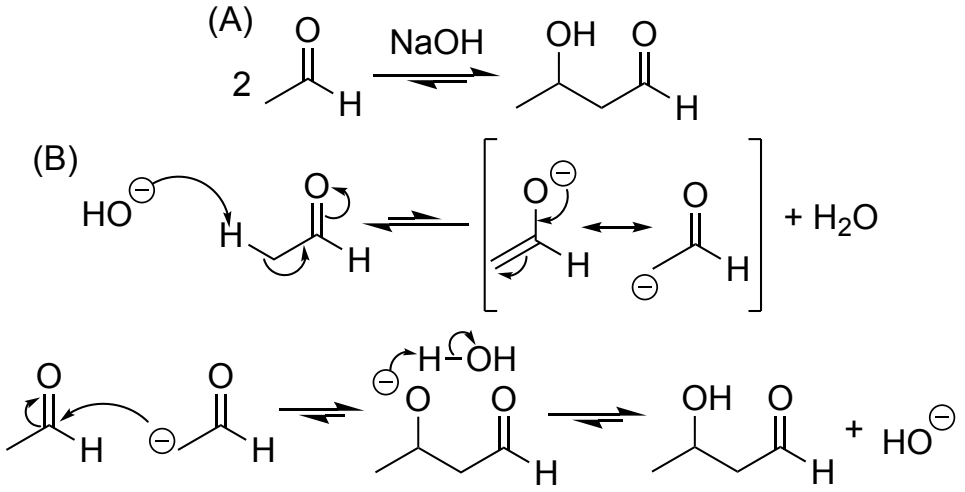


Figure 1.11: General reaction and mechanism of an aldol reaction. (A) General base-catalyzed aldol reaction involving the reaction of 2 aldehydes to form a β-hydroxy aldehyde. Here, sodium hydroxide is used as a catalyst, but any Lewis base can be used. (B) The aldol reaction mechanism involving 2 aldehydes.

1.5.2.1 RA95.0

The retro-aldolase RA95.0 was designed *in silico* from the non-catalytic IGP synthase from *S. solfataricus* (PDB: 1LBL)[69] protein template using the Rosetta3 software.[79] It catalyzes the reversible retro-aldol reaction of (R)/(S)-methodol (4-hydroxy-4-(6-methoxy-2-naphthyl)-2-

butanone), via a multi-step pathway involving Schiff-base intermediate formation (Figure 1.12). This variant catalyzes with efficiency of 0.17 M^{-1} s^{-1},[80] which was subsequently improved via several mutagenesis rounds to give RA95.5-8,[81] catalyzing the retro-aldol reaction with an efficiency of 1600 M^{-1} s^{-1}, though was unable to be crystallized, although recent efforts by Yu et. al.[82] have successfully done so (PDB: 8XYN). Importantly, this version gave rise to a completely different catalytic site, moving the catalytic lysine from position 210 to 83 and increasing its complementarity, likely a major factor in increasing efficiency (Figure 1.13), which is retained in further rounds of evolution. Finally, using fluorescence-activated droplet sorting (FADS) as the screening method, further evolution led to RA95.5-8F, an enzyme whose efficiency climbs even further to a WT enzyme with k_{cat}/K_M 34000 M^{-1} s^{-1}.[83]

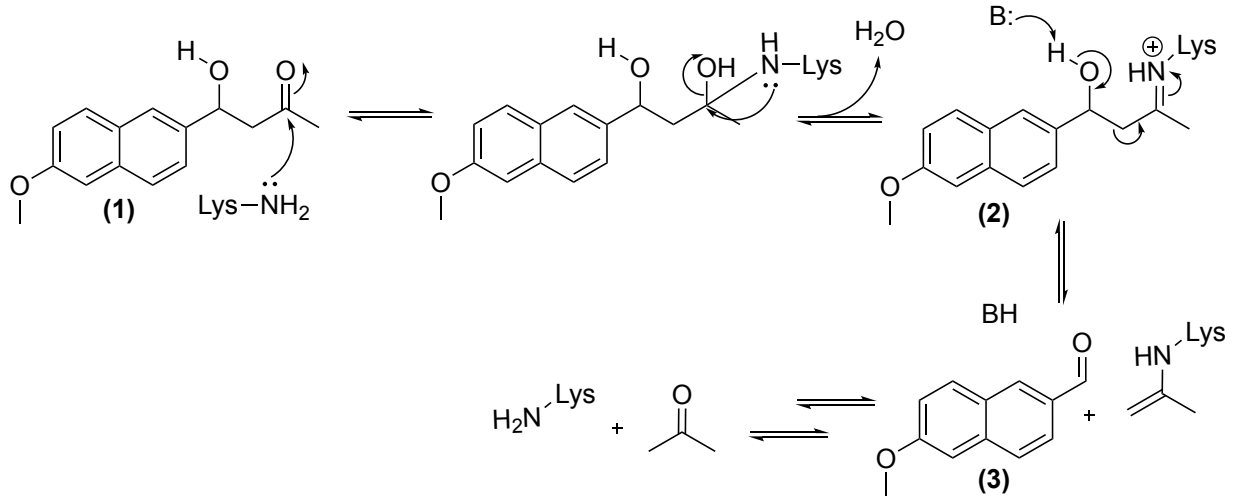


Figure 1.12: Retro-aldolase-catalyzed mechanism for the cleavage of (±)-methodol via the lysine-methodol Schiff-base intermediate. The methodol reactant (1) undergoes nucleophilic attack from a lysine to form a Schiff-base intermediate (2) to form the final aldehyde product (3). Modified from Giger and colleagues, 2013.[81]

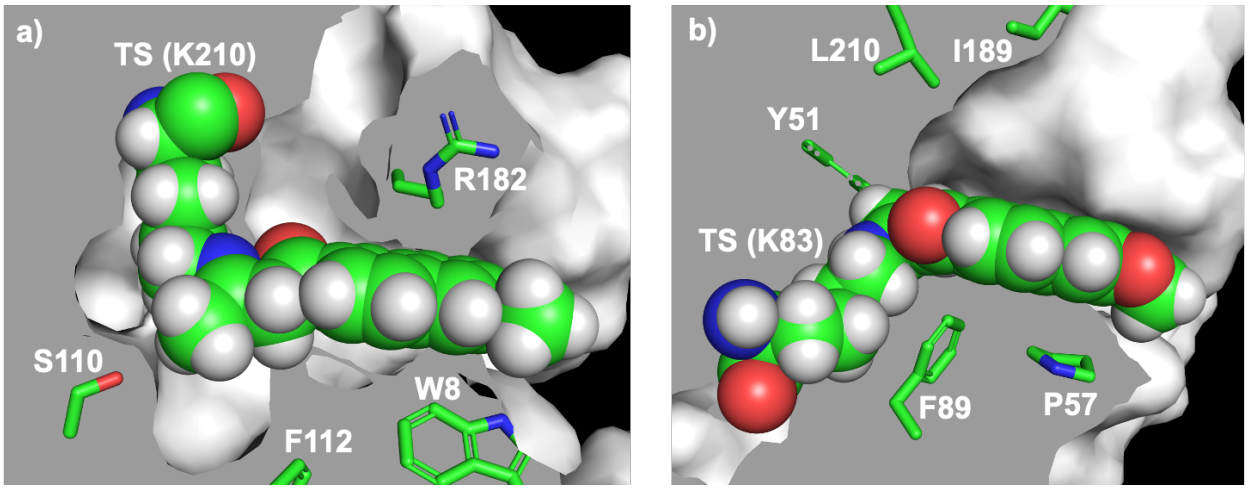


Figure 1.13: During DE, the complementarity for the intermediate was increased. Cut-away representation of the ligand burial of the intermediate in (a) RA95.0 and (b) RA95.5-8F. Schiff base intermediate represented as spheres, residue sidechains that fill the surface volume around the intermediate shown as sticks, rest of the protein as a white surface. As evolution progresses, the location of the lysine sidechain on which the Schiff base intermediate (TS) is located changes from 210 to 83, rendering the entire intermediate buried further into the active site, with increased complementarity for the binding pocket of the intermediate. The original K210 is mutated to a leucine in the final design.

This present work uses RA95.0 (Figure 1.14, second enzyme) as a starting point for designing a new tyrosine-mediated hydrogen bond via backrub-generated ensembles for the eventual goal of stabilizing the transition state and increase catalysis, inspired by the tyrosine hydrogen bond network located in the catalytic tetrad in RA95.5-8F, mainly due to it being generated solely *in silico*. More specifically, it is hypothesized that using backrub ensembles as the main flexibility model is enough to predict more diverse sequences for an enzyme with a hydrogen-bonding ligand interaction relative to a single-template approach. RA95.0 is chosen as it is a TIM-barrel protein, affording it all of the design benefits as the listed predecessors. As the main goal was to design a new enzyme, recapitulations of RA95.0 were not performed, apart from positive control experiments. Subsequent *in silico* improvements to the reactivity, if any, can be

considered as computationally designed, proving a more efficient engineering pipeline than *in vitro* or *in vivo* methods.

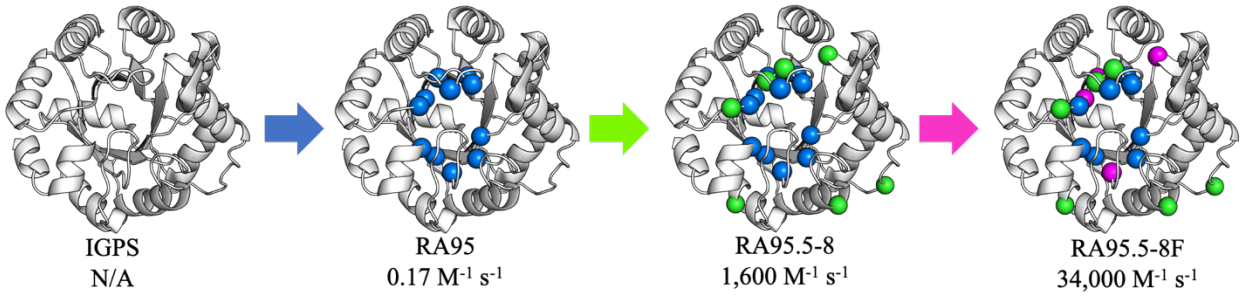


Figure 1.14: Visualized evolutionary trajectory of the RA95 series up to RA95.5-8F. Mutations created during the evolution from enzymes IGPS to RA95 are shown with their α-carbons in blue, mutations acquired during the evolution from RA95 to RA95.5-8 shown in green α-carbons, and mutations acquired during the evolution of RA95.5-8F from RA95.5-8 shown in magenta α-carbons.

1.6 Thesis Objectives

To reiterate, the main hypothesis of the work produced herein is that BR can generate ensembles that improve the design of high-activity enzymes, which is explored through the main goal of developing an improved pipeline for designing better enzymes in a *de novo* fashion. This goal is broken down into 2 steps – determining if backrub ensembles can be utilized for accurate active site structure predictions of high-activity enzymes, and investigating if BR (and BR-serialized methods) can be used to diversify sequence designs for a new enzyme entirely. To achieve the former, Kemp eliminases 1A53-core and HG4 (both high-activity) will be attempted to be recapitulated. For the latter, a new retro-aldolase dubbed TyRA95.0 will be designed, where a tyrosine is hydrogen-bonded to the ligand (Section 3), in order to stabilize the transition state, as was seen in later evolutions of RA95.0. To finish, results will be stated, discussed and summarized, with new metrics being proposed at the end to help improve ensemble engineering. Each chapter's purpose is discussed in detail below.

In Section 2, the main goal is to demonstrate BR's ensemble's recapitulation ability, given the hypothesis that BR used in the pipeline of ensemble generation increase the accuracy of active site structural prediction. Recapitulations of enzymes and their active sites will be shown and discussed. The enzymes to recapitulate are 1A53-core and HG4, designed by Zarifi et. al. (to be published) and Broom et. al.[75] The protein templates 1A53, 3NYZ and 3NZ1 are chosen for the recapitulation of 1A53-core, and 1GOR is used for re-designing HG4. From the 1A53-core starting templates, ensembles will be generated using a variety of methods except backrub, from which a random number of ensembles are selected to generate backrub ensembles. These final ensembles are then minimized and run through the Triad software for theozyme placement and repacking. For HG4, ensembles are generated through similar methods including backrub, and then these structures are minimized and the same procedure as 1A53-core is performed. Afterwards, the final structures are compared with the crystal structures of their respective enzyme targets to determine the best predictions.

In Chapter 3, the main goal is to use BR for re-designing a new retro-aldolase, given the hypothesis that BR allows a more diverse set of sequences to be generated relative to the single-state design analog. A new enzyme is designed *in silico* from the crystal structures of the pre-existing RA95.0, termed TyRA95.0 for its tyrosine-TSA hydrogen bond interaction, with the results shown and discussed. Due to time constraints, *in vitro* verification of the enzyme's kinetics and activity was unable to be performed.

The last chapter will go over the successes and failures of the previous 2 chapters and will discuss future directions for enzyme engineers wishing to add backrub into their design pipeline, as well as key lessons learned from both chapters to ensure protein scientists are equipped with a better understanding of *in silico* practices for improved designs. All aforementioned steps are summarized visually in Figure 1.15.

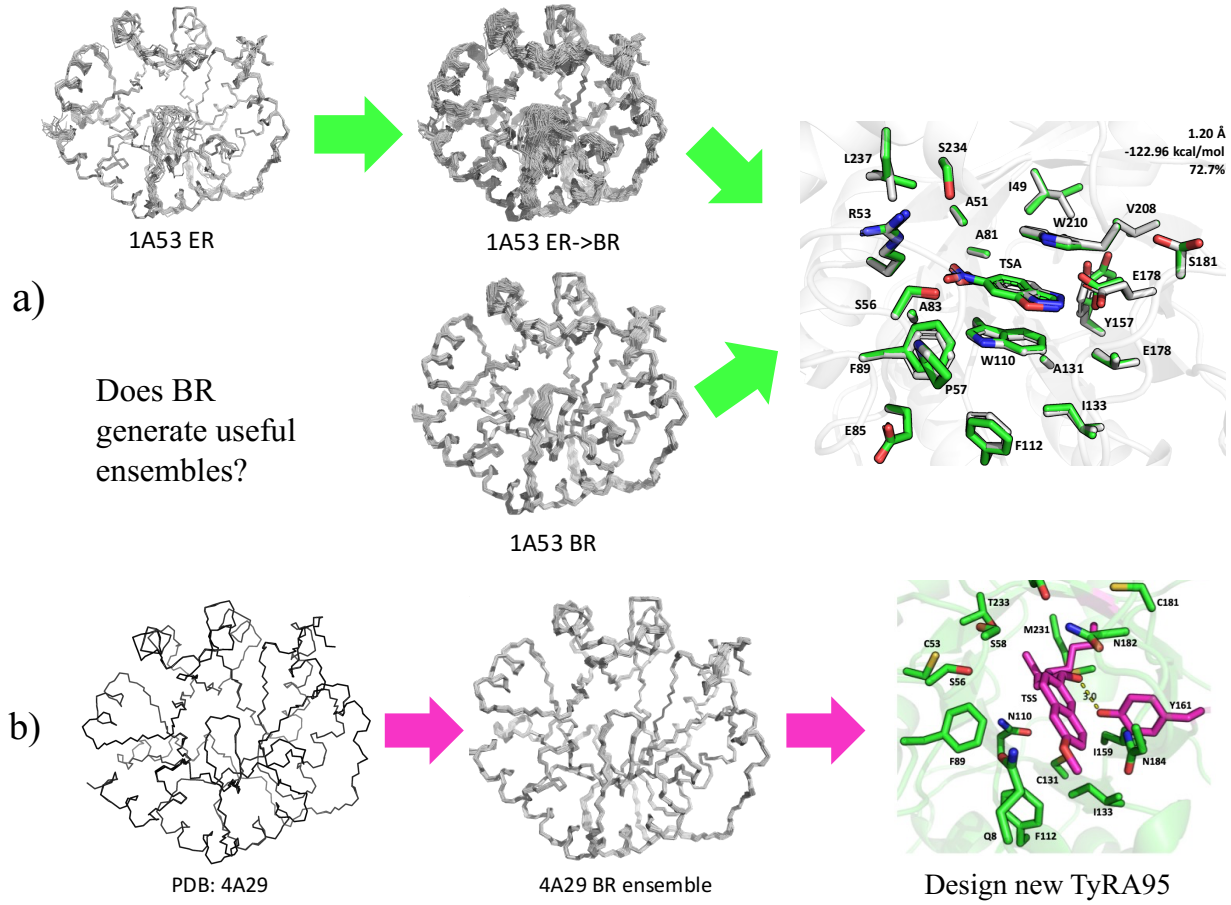


Figure 1.15: Visual summary of thesis objectives. a) The second chapter will describe BR's potential to increase recapitulation accuracy by hopefully designing better ensembles, either through serializing BR ensemble generation with another method like ER (shown at top) or performing BR on its own (a, bottom). The 1A53-core recapitulation is shown as an example. b) The third chapter concerns itself with using BR to redesign a new retro-aldolase dubbed TyRA95, with a TS-stabilizing hydrogen bond with a tyrosine, inspired by the catalytic tetrad in RA95.5-

8F. It will be determined if BR ensembles help to generate a more diverse list of sequences than a single-state design approach.

Chapter 2 – BR Algorithm Implementation and Recapitulation of 1A53-core and HG4

2.1 Statement of Contribution

The Triad Backrub module was designed and coded mainly by Ilya S. Dementyev, with help from Paul M. Chang from Protabit (Pasadena, CA, USA). ER and MD ensembles were generated and used with permission from R. V. Rakotoharisoa and N. T. Hang Pham. HG4 simulations performed by (and used with permission from) R. V. Rakotoharisoa and N. T. Hang. Pham. HG4 simulations were redone by Ilya S. Dementyev to confirm. All other work performed by Ilya S. Dementyev.

2.2 Introduction

In this work, BR's potential as a valuable ensemble-generating method (EGM) is explored, both on its own and serialized with other EGMs like ER (Section 2.3.4). One reason is BR's likely involvement in the Kemp Eliminase 1A53-core's design (Figure 1.7), where a crucial residue undergoes a backbone rotation similar to others seen in the literature[34] during the evolution of the enzyme from 1A53 to 1A53-2.5 (to be published). It is also capable of performing localized rotations, something that most EGMs cannot do. Furthermore, to necessitate catalysis, residues generally move within a range of 0.15 – 1 Å during the transition state formation in order to perform the catalysis.[84] Furthermore, the serialization of EGMs with BR (e.g. performing BR on randomly chosen ER templates) is explored for the potential ability of the serialization to explore

structural space in a unique way. This work is not the first to introduce serialization in ensemble design workflows,[41] but its use here for the increased traversal of Cartesian space is a novel approach that has the potential for more accurate recapitulations, and thus improved *de novo* design of entirely new enzymes. Allowing more than one type of movement on an input would allow for more complicated flexibilities to be modelled, potentially leading to reduced challenges in modelling difficult conformational changes, leading to increased recapitulation accuracy, while also maintaining relative energetic stability due to BR's Metropolis criterion ensuring that the backbone does not have significant structural deformities. Hence, if one could use BR (either on its own, or serialized) to recapitulate this deviation while designing ensemble templates for the redesign of 1A53-core, and even apply it to recapitulate other high-efficiency enzymes like HG4,[75] then it would confirm BR ensemble generation's ability to improve the prediction accuracy of the 1A53-core or HG4 Kemp Eliminase active site structure.

2.3 Methods

For all structures, solvent and counter-ion atoms were removed, while keeping any bound ligands intact. Any identified missing residues in crystal structures were rebuilt using Coot (Medical Research Council Laboratory of Molecular Biology, Cambridge, UK).[85] Structural standardization was done using the *addH.py* module in Triad (Protabit, Pasadena, CA, USA). The module also assigned HIS protonation states (HIE/HID), flipped ASN/HIS/GLN sidechains, and respectively rotated hydroxyl, sulfhydryl, methyl, and amino groups in SER, THR, CYS, MET, and LYS, if required. All skeletal structures of backbone ensembles in Sections 2.3.3-2.3.6 were

generated using PyMOL, whose displayed backbones are the atoms N, C or CA as named in the PDB. Before beginning recapitulations using different templates, one must first determine if Triad is able to recapitulate the target sequence (with proper rotamers) given the target backbone structure, especially in areas with the defined active sites. All steps in the ensemble generation and de novo pipeline (including theozyme placement and repacking) were performed using the phoenix forcefield.

2.3.1 Control Recapitulations

To test Triad's efficacy at recapitulating enzyme structure, positive and negative control experiments were set up, where the backbone chosen as the template is that of 1A53-core (PDB: 8FOQ) and 1A53 WT (PDB: 1A53) respectively for 1A53-core. For HG4, the positive and negative control backbones were HG4 and *Thermoascus aurantiacus* xylanase WT 10Å (PDB: 1GOR). The negative controls were chosen to investigate the ability of a protein having no catalytic activity for our reaction to recapitulate the structure of our active targets given its backbone, relative to the positive control which determines how accurate recapitulation is when the input is the target structure's backbone. Hence, results from both controls indicate the expected range of recapitulation accuracy (as indicated by chosen metrics of energy, ligand RMSD and percentage of correct rotamers) that can be performed by the software. Both targets' residue positions to recapitulate were the design positions, which were chosen according to their proximity to the ligand. For 1A53-core, residues were chosen if any of their atoms lied within a 4 Å sphere of the ligand's center-of-mass, as well as residues within a 4 Å sphere of the catalytic residues (W110, E178, W210). A similar process was used to determine design positions for HG4, where the catalytic residues for HG4 were D127 and Q50, and design positions were chosen to be within a

12 Å sphere from the ligand. A275 was included as it is known to play a role in facilitating substrate channel opening.[75] In all cases, residues whose sidechains pointed outward from the active site were excluded.

Before running theozyme placement, repacking, and hollowing, the structures were standardized (Section 2.3 introduction). Theozyme placement was performed using Triad's *phoenixMatch.py* module, generating a maximum of 4 structures per input, using the *phoenix* forcefield (non-covalent), with optimum theozyme geometry defined in Table 2.1, taken from 8FOQ's active site structure. The hollowing for the protein takes all design positions (with ligand removed from the active site) and mutates them to glycine, then calculates the final energy of this mutated structure. After theozyme placement had finished, repacking occurs by programming Triad (using the same *phoenixMatch.py* module) to mutate the residues at the design locations to the respective amino acids found in 1A53-core (Table 2.2) or HG4 (Table 2.3). Geometries for the theozyme are maintained but are not defined for the rest of the amino acids, letting the software design the rotamers it considers the most probable at each position. If Triad successfully generates a structure close to the crystal in the positive control, and a poor (or no) structure for the negative control, one can then use this as evidence that Triad can indeed recapitulate enzyme active site structure (Section 2.4.1). Structures were determined to be successful recapitulations if they were at or below a ligand RMSD of ± 1.45 Å for 1A53-core, and 0.69 Å for HG4. Hits also must include a correct rotamer percentage above 66% and 50% for 1A53-core and HG4 recapitulations respectively. Such values were chosen based on the extracted metrics from the positive controls of each recapitulation, with a buffer value added. For ligand RMSDs, a value of 0.25 Å was added to each original extracted metric (1.20 Å for 1A53-core and 0.44 Å for HG4), and for correct rotamer

percentages, a buffer value of 6-8% was subtracted from the original values (73% for 1A53-core and 58% for HG4). These values were based off of previous experiments (to be published) performed in the lab to add a leeway to allow for more hits to be considered, rather than those which achieve target values with similar precision to the positive control, which is difficult to achieve in practice. It is acknowledged that more strict or lenient buffer values could be chosen, which would lead to a smaller or larger hit number respectively. Future investigations could benefit from a thorough benchmarking approach to achieve better buffers by analyzing hit frequencies and determining an optimal trade-off between buffer leniency and hit number.

Table 2.1: 1A53-core theozyme geometry bias values. Interaction names are string variables set by the user to keep track of the different interactions. The first bias, base, defines important geometries between the ligand (Substrate) and the catalytic base E178 (Template). The second bias (base2) is identical, except accounting for carboxylic acid rotation (OE1 interacting with ligand rather than OE2). Atoms interacting correspond to their naming in the PDB, and either belong to the substrate or the template (S/T respectively). Interaction type defines the kind of geometric measurement. Each range's lower and upper values were defined by the corresponding value in the crystal structure, with added values of ± 10° for angles and dihedrals, as well as ± 0.2 Å for distances. Due to poor agreement with the crystal structure in initial positive control designs, a larger range was chosen for the base and base2 distances.

Interaction Name	Species interacting	Atoms interacting	Interaction type	Ranges*
base	6BX (S), E178 (T)	S/H3, T/OE1	Distance	1.5, 2.2
		S/H3, T/OE1, T/CD	Angle	65.8, 85.5
		S/C3, S/H3, T/OE1	Angle	119.2, 139.2
		S/H3, T/OE1, T/CD,T/OE2	Dihedral	11.6, 31.6
		S/C3, S/H3, T/OE1, T/CD	Dihedral	167.6, 187.6
		S/N2, S/C3, S/H3, T/OE1	Dihedral	-175.0, -155.0
base2	6BX (S), E178 (T)	S/H3, T/OE2	Distance	1.5, 2.2

		S/H3, T/OE2, T/CD	Angle	65.8, 85.5
		S/C3, S/H3, T/OE2	Angle	119.2, 139.2
		S/H3, T/OE2, T/CD,T/OE2	Dihedral	11.6, 31.6
		S/C3, S/H3, T/OE2, T/CD	Dihedral	167.6, 187.6
		S/N2, S/C3, S/H3, T/OE2	Dihedral	-175.0, -155.0
pi1	6BX (S), W110 (T)	(T/CE2, T/CD2, T/CE3, T/CZ2, T/CZ3, T/CH2), (S/C4','S/C5','S/C6','S/C7','S/C8','S/C9)	Distance (Centroid)	3.3, 3.7
		(T/CE2, T/CD2, T/CG, T/CD1, T/NE1), (S/C4, S/C5, S/C6, S/C7, S/C8, S/C9)	Distance (Centroid)	3.4, 3.8
		(T/CE2, T/CD2, T/CE3, T/CZ2, T/CZ3, T/CH2), (S/C4, S/C3, S/N2, S/O1, S/C9)	Distance (Centroid)	3.4, 3.8
		(T/CE2, T/CE3, T/CG), (S/C4, S/C6, S/C8)	Angle (planar)	(-6.9, 13.1), (166.9, 186.9)
pi2	6BX (S), W210 (T)	(T/CE2, T/CD2, T/CE3, T/CZ2, T/CZ3, T/CH2), (S/C4, S/C5, S/C6, S/C7, S/C8, S/C9)	Distance (Centroid)	3.7, 4.1
		(T/CE2, T/CD2, T/CG, T/CD1, T/NE1), (S/C4, S/C5, S/C6, S/C7, S/C8, S/C9)	Distance (Centroid)	3.8, 4.2
		(T/CE2, T/CD2, T/CG, T/CD1, T/NE1), (S/C4, S/C3, S/N2, S/O1, S/C9)	Distance (Centroid)	3.5, 3.9
		(T/CE2, T/CE3, T/CG), (S/C4, S/C6, S/C8)	Angle (planar)	(-10.0, 10.0), (170.0, 190.0)

*Distances in Å, angles and dihedrals in degrees (°).

Table 2.2: Amino acid positions and identities for recapitulated residues of 1A53-core. Theozyme residues are bolded. The ligand (ID: 500) is not shown.

Residue ID	Amino acid
49	I
51	A
53	R
56	S
57	P
58	S
81	A
83	A
85	E
89	F
108	L
110	**W**
112	F
131	A
133	I
157	Y
159	V
178	**E**
180	C
181	S
182	R
184	L
208	V
210	**W**
211	G
231	G
233	G
234	S
237	L

Table 2.3: Amino acid positions and identities for recapitulated residues of HG4. Theozyme residues are bolded. The ligand (ID: 500) is not shown.

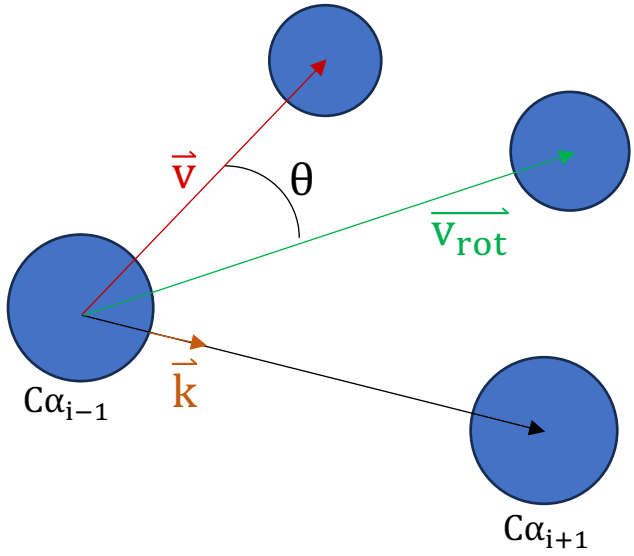
Residue ID	Amino acid
16	V
17	Y
21	A
42	M
44	W
46	E
47	N
50	**Q**
79	L
81	G
83	G
84	C
87	W
90	F
125	T
127	**D**
130	G
170	Y
172	M
207	Q
209	H
234	S
236	L
237	M
239	D
265	T
267	M
275	A
276	F

After this was done, the ensembles could be designed. For each XX-BR design pipeline, 20 structures from each PDB generated using the XX method (where XX is MD, PM or ER) were randomly sampled using Python's random module to be then used for the backrub process.

2.3.2 General Description of Backrub Parameters

To generate ensembles consisting of backrubbed protein templates, the backrub algorithm needed to be implemented into the protein-design software Triad. The written backrub application creates an ensemble of back-rubbed protein structures, by taking a high-resolution crystal structure of a protein as input and running it through the backrub application, creating a collection of structures which are each formed by backrubbing the input several times. Triad's backrub axes are defined by carbon α atoms, and a minimum of 3 residues is required for a backrub to occur. The number of backrub-generated ensemble members equalled the number of Ensemble Refinement structures generated. For one backrub "step", moves are performed during a "Monte-Carlo (MC) step", in which a Metropolis probability criterion ($k_B T = 0.6 \frac{kcal}{mol}$, $T = 302$ K) decides if a backrub move should be accepted or rejected, based on the energy of the previous pre-backrub conformation.[60] All backrubs were performed with 2000 MC steps each simulation. Probability values used were 0.0001 for the rotamer-only pathway ($P_{rotamer}$, Figure 1.4), ensuring a backbone backrub rotation ($P_{backbone}$, Figure 1.4) was guaranteed to occur at each step. For the Metropolis criterion, the energy of the system before and after each MC step was calculated automatically via the Phoenix forcefield,[26] with the van-der-Waals radii of atoms scaled to 0.9 allowing for more space for backbones to rotate.[26] For each step, a backbone solution is generated and is used as the input for the next step, until the user-specified number of MC steps had been reached. Structural validation (ensuring no steric clashes, backbone breaks or improper rotamer placements) was conducted with the PHENIX macromolecular structure determination software (v1.18.2-3874) (Berkeley, CA, USA).[86] While this application shares similarities with Smith and Kortemme's

backrub written separately for Rosetta, it has a few differences. First, Triad's version uses Rodrigues rotation (Figure 2.1) in Cartesian coordinates to perform the rotation, whereas Rosetta's version makes use of internal coordinate methods. Furthermore, Triad's user-level app is written in only Python, whereas Rosetta's can be used through Python or other methods, similar to writing an XML file. The C++ implementation of the Rodrigues Rotation and main rotator are added in the Appendix.



$$\overrightarrow{v_{rot}} = \vec{v}\cos\theta + (\vec{k}\times\vec{v})\sin\theta + \vec{k}(\vec{k}\cdot\vec{v})(1-\cos\theta)$$

Figure 2.1: Rodrigues rotation of a vector $\vec{v}$, around an arbitrary rotation axis defined by $\vec{k}$, to form $\overrightarrow{v_{rot}}$. This is the simplest Rodrigues rotation of a backbone, involving a sequence of 3 residues $(i-1, i, i+1)$ where the rotation vector is defined by the difference in positional vectors of $C\alpha_{i+1}$ and $C\alpha_{i-1}$. The angle of rotation is defined by θ.

In this work, after the BR was performed with the necessary pre-defined settings, each structure was refined using the *proteinProcess* module in Triad, which modestly minimized the backbone using the *cphoenix* (covalent *phoenix*) forcefield to account for N and C-termini artifacts left behind by the *backrub.py* module, but only after all other crucial design steps have occurred, as the minimization can force the protein further down into its local minimum, especially with

algorithms like steepest descent.[87] Due to time constraints, for the HG4 recapitulation, only the BR and control results are shown and discussed in detail.

2.3.3 PertMin-BR Ensemble Generation

To start, PDB crystal structures 1A53, 3NYZ and 3NZ1 all went through the *addH.py* preparation process (Section 2.3 intro). Then, the *generateEnsemble.py* module in Triad was used to perform ±0.001 Å deviations in all cartesian atoms and coordinates, to create 100 independent structures, each of which were energy-minimized over 3000 iterations. The ±0.001 Å value was chosen based on previous experiments performed using PertMin ensembles[22] and determining that such a value led to an optimal balance between increasing the diversity and deviation while keeping the structural integrity intact relative to the original input. The minimization used the conjugate gradient approach and incorporated the Phoenix energy forcefield with covalent terms (*cphoenix*).

After the structures were generated, 20 were chosen at random using Python's *random* module for the structures to be backrubbed. Twenty input structures were a good starting point to generate backrubs as the backrub designs will result in a significant increase in the total number of structures generated. There have been successful recapitulations done in the past for ensembles with template sizes greater than 100, so 20 was chosen to account for that value, as the initial starting point should be modest to account for the combinatorial growth that would occur as a result of doing 20 x 10 x 4 x 1 = 800 structures each (after repacking).

For each randomly sampled structure as the input, 10 backrubs were performed using the backrub.py module in Triad, with 2500 MC steps in total, a maximum rotational window of 12 consecutive residues, and a maximum rotational angle of 5°. The best MC step number lies within 2000-2500. This was chosen after extensive experimentation the settings to determine which values for their respective parameters would function best. Going higher than 2500 leads to abnormally long distances between backbone atoms, but less than 2000 leads to low sampling of geometric space. The 5° was chosen for the same justification. After all 200 structures were generated for each PDB input, they were modestly minimized using Triad's *proteinProcess.py* module. As their deviations relative to their original input structures for each PDB code had increased, these 200 structures were chosen to proceed with theozyme placement and repacking, discussed in the Repacking section below.

2.3.4 ER-BR Ensemble Generation

Briefly, ER is a form of restricted MD, in which the simulation is confined within the electron density as described by the x-ray crystallographic data deposited with the protein, usually available as an MTZ file in the PDB. ER was performed using the PHENIX macromolecular structure determination software (v1.18.2-3874)[47], in which the parameters specified were p_{TLS}, t_x, and $w_{x\text{-}ray}$, which are the atom fraction included in TLS fitting, relaxation time, and the weight of the electron density data respectively. Each crystal structure was edited to remove low-occupancy conformers, after which the remaining conformer was assigned a value of 1. Hydrogens were added to the protein using *phenix.ready_set*, after which parallel simulations were run using values for p_{TLS} (0.6, 0.8, 0.9, 1.0), τ_x (0.5, 1.5, 2.0), and $w_{x\text{-}ray}$ (2.5, 5.0, 10.0), upon which the best ensemble was chosen. As the optimal value of some variables cannot be determined *a priori* (e.g.

p_{TLS}), it was necessary to run a few collaborative simulations with variable parameters, after which the ensemble with the lowest R_{free} value was selected, while ensuring that the $R_{work} - R_{free}$ differences remained within less than 5%.

Once again, 10 BRs were performed with the same settings as the PertMin BRs, on 20 randomly chosen ensembles for each PDB. After all 600 structures were generated, they were modestly minimized using Triad's *proteinProcess.py* module to remove N- and C-termini artifacts. As their deviations relative to their original input structures for each PDB code were increased, these 200 structures were chosen to proceed with theozyme placement and repacking.

2.3.5 MD-BR Ensemble Generation

To generate ensembles, 20 simulations were performed, each 1 ns (2 fs timestep), were performed using the AMBER99sb forcefield.[88] The Particle Mesh Ewald method was used for the long-range electrostatics (>1 nm).[89] The Antechamber package was used for calculating ligand parameters.[88,90] All crystal structures (ligand-bound or unbound) were protonated via the H++ server at pH = 7.0 (http://newbiophysics.cs.vt.edu/H++/index.php).[21] All protein structures were inserted into a dodecahedral symmetry box with periodic boundary conditions, with a 1 nm distance between the protein surface and the dodecahedron's edge. The system was neutralized with 0.15 M of Na^+ and Cl^- counterions, and all water molecules were modelled explicitly using the TIP3P model.[91] Water molecules found in the crystal structure were kept in the inputs. Minimization of the prepared structures was performed with steepest descent to a 1000 kJ mol^{-1} nm^{-1} maximum force. Before equilibrating, the system was heated to 300 K from 0 K iteratively using 50 ps steps. Afterwards, the system was equilibrated using the canonical ensemble using the

Nose-Hoover thermostat with heavy-atom position restraints.[92] Next, an NPT equilibration was performed with identical restraints, at a temperature and pressure of 300 K and 1 bar respectively. Pressure was held constant with the Berendsen barostat.[93] After removing position restraints, the 1 ns production runs began, using Parrinello-Rahman as the barostat.[94] Ensembles were generated by extracting models of variable template numbers by concatenating the twenty 1 ns simulations into one 20-ns run. Template numbers depend on the number generated by ER, as the latter method's numbers cannot be controlled or pre-determined.

After the ensembles were generated, 20 were chosen at random (Fig. 2.7) for the structures to be backrubbed. For each structure as the input, 10 backrubs were performed using the backrub.py module in Triad, with 2500 MC steps in total, a maximum rotational window of 12 consecutive residues, and a maximum rotational angle of 5 degrees. After all 200 structures were generated, they were modestly minimized using Triad's *proteinProcess.py* module. For this method, the deviations relative to their original input structures for each PDB code had decreased (or stayed constant), thus these 200 structures were not chosen for motif generation nor repacking for any of the 1A53, 3NYZ or 3NZ1 PDB structures.

2.3.6 BR-only Ensemble Generation

For recapitulation of 1A53-core using a backrub ensemble, unlike previous methods which integrated a BR ensemble generating step after the initial ensemble generation (such as ER-BR), here the BR was performed all at one step, with N=200 final templates for each PDB. Hence, only one set of diversity and deviation calculations were obtained. As before, BR was performed using 2000 MC steps, with a temperature of 0.6 kcal mol^{-1}, a maximum window of 12 residues was

chosen for rotation, and a 5° maximum allowed rotation angle. After all 600 templates were created, all were subjected to minimization with Triad's *proteinProcess.py* module, with the covalent Phoenix forcefield (*cphoenix*). Final backbone diversities and deviations were analyzed for all 3 sets of 200 templates. Since there was no previous step enforcing backbone deviation improvements, all templates were chosen to proceed with theozyme placement, hollowing, repacking and final analysis.

To recapitulate HG4, ensembles were generated with 80 templates from 1GOR, 50 from 5RG4, and 84 from 5RGA. These numbers were selected to match the number of conformers produced by ER, as the exact number of conformers generated by this method depends on the refinement procedure and cannot be precisely controlled. Consequently, all non-serialized ensemble generation methods require matching template numbers in each ensemble to align with ER.

2.3.7 BR Benchmarking – Comparison with Rosetta

To check if Triad's BR works similarly to other software versions, a BR ensemble generation experiment was set up with a well-established protein design software, Rosetta (Seattle, WA, USA). The input PDBs used were 1A53, 3NYZ, 3NZ1, 1GOR, 5RG4 and 5RGA. Before running, all waters, solvents and counterions were removed from the coordinate file. Then, standardization (including protonation) was performed with Rosetta. No minimization was performed before running Rosetta's BR. Afterwards, the Rosetta BR execution file was set up to ensure the same parameters were used as in Section 2.2.6. Any protonated histidine residues (HID/HIE) were renamed to HIS, as Rosetta deals with protonation without the need for user's

specification, and because HID/HIE were not recognized by the software as valid residues. Then, the ensemble generations were run in 6 different folders on the Graham cluster (Digital Research Alliance of Canada), each corresponding to the PDBs used. Each PDB was used to generate 200 BR templates. Afterwards, to smooth out termini artifacts, the outputted coordinate files were modestly minimized using Triad's *proteinProcess.py* module. Backbone diversities and deviations were then analyzed using a custom Python script. Only geometric metrics of diversity and deviation were considered for the benchmark analysis.

2.3.8 Energy Calculations

For 1A53-core, residue positions 49, 51, 53, 56, 57, 58, 81, 83, 85, 89, 108, 110, 112, 131, 133, 157, 159, 178, 180, 181, 182, 184, 208, 210, 211, 231, 233, 234, 237 were mutated to glycine, and all ligands from the system were removed. These positions were chosen due to their close proximity to the TSA (all residues whose atoms overlap a 4 Å radius from TSA's centroid), after pruning for residues whose sidechains pointed away from the active site. The energy calculations were again performed with the *phoenixMatch.py* module with the same simulated annealing approach, although this step was much faster as glycine has the simplest sidechain modelling.

For HG4, residue positions 16, 17, 21, 42, 44, 46, 47, 50, 79, 81, 83, 84, 87, 90, 125, 127, 130, 170, 172, 207, 209, 234, 236, 237, 239, 265, 267, 275, 276 were mutated to glycine, and all ligands from the system were removed. These positions were chosen due to their close proximity to the TSA (these correspond to all residues whose atoms overlap a 12 Å radius from TSA's

centroid), after pruning for residues whose sidechains pointed away from the active site. All following procedures for hollow energy calculations were performed identically to 1A53-core.

These mutated proteins are used to calculate the strength of the energy stabilization by the interactions between the designed residues and ligand that is conferred onto the enzyme during design. The energy calculated also serves as the energy of a final design, and is used as one of the metrics to determine recapitulation quality, discussed in Section 2.3.1.

2.3.9 Theozyme Placement (Motif Generation)

For 1A53-core, residue positions 49, 51, 53, 56, 57, 58, 81, 83, 85, 89, 108, 110, 112, 131, 133, 157, 159, 178, 180, 181, 182, 184, 208, 210, 211, 231, 233, 234, 237 were chosen due to their close proximity to the TSA. Positions 110, 178, 210 were mutated to TRP, GLU, TRP respectively. The ligand incorporated was 5-nitrobenzisoxazole in all cases, and the design geometries (relevant distances, angles, torsions for properly recapitulating the theozyme) are in Table 2.4. All other positions not mentioned were not changed during the design. The design simulations were performed using Triad's *phoenixMatch.py* module, which performed 4 parallel runs using a standard fixed backbone sidechain repacking optimization in a Monte Carlo simulated annealing approach. Structures whose theozyme placements were above 1000 kcal/mol were not considered for repacking.

For the HG4 recapitulation, residue positions 16, 17, 21, 42, 44, 46, 47, 50, 79, 81, 83, 84, 87, 90, 125, 127, 130, 170, 172, 207, 209, 234, 236, 237, 239, 267, 275, 276, 265 were chosen. The difference in sphere radius is for maintaining consistency with previous designs performed by

colleagues R. V. Rakotoharisoa and N. T. Hang Pham. All further aspects to theozyme placement for HG4 were identical to 1A53-core's procedure, although the HG4-specific geometries can be found in Table 2.4.

Table 2.4: HG4 theozyme geometry bias values. The first bias, base, defines important geometries between the NBX ligand (Substrate) and the catalytic base D127 (Template). The second bias (base2) is identical, except accounting for carboxylic acid rotation (OD1 interacting with ligand rather than OD2). Atoms interacting correspond to their naming in the PDB, and either belong to the substrate or the template (S/T respectively). Interaction type defines the kind of geometric measurement. Each range's lower and upper values were defined by the corresponding value in the crystal structure, with added values of ± 10° for angles and dihedrals, as well as ± 0.3 Å for distances between the base and ligand. The range of ± 0.55 Å was chosen for the acid based on previous in-house designs performed (not published).

Interaction Name	Species Interacting	Atoms Interacting	Interaction Type	Ranges*
base	NBX (S), D127 (T)	T/OD2, S/H3	Distance	1.0, 1.6
		T/CG, T/OD2, S/H3	Angle	109, 131
		T/OD2, S/H3, S/C3	Angle	159, 180
		T/CB, T/CG, T/OD2, S/H3	Dihedral	-21, 21
base2	NBX (S), D127 (T)	T/OD1, S/H3	Distance	1.0, 1.6
		T/CG, T/OD1, S/H3	Angle	109, 131
		T/OD1, S/H3, S/C3	Angle	159, 180
		T/CB, T/CG, T/OD1, S/H3	Dihedral	-21, 21
acid	NBX (S), Q50 (T)	T/1HE2, S/O1	Distance	1.2, 2.3
		T/NE2, T/1HE2, S/O1	Angle	145, 157
		T/1HE2, S/O1, S/N2	Angle	120, 140
		T/1HE2, S/O1, S/N2, S/C3	Dihedral	160, 200
acid2	NBX (S), Q50 (T)	T/2HE2, S/O1	Distance	1.2, 2.3
		T/NE2, T/1HE2, S/O1	Angle	145, 157
		T/2HE2, S/O1, S/N2	Angle	120, 140
		T/2HE2, S/O1, S/N2, S/C3	Dihedral	160, 200

*Distances in Å, angles and dihedrals in degrees (°).

2.3.10 Repacking

After the motif generation is performed, all other design positions were mutated to the design amino acids for the respective recapitulation targets, 1A53-core and HG4. The same respective geometries were used for this step, this time for constraining rotamers in the theozyme positions, as well as the ligand to prevent movement deteriorative to theozyme structure during the simulation. For all recapitulations, designs were performed using Triad's *phoenixMatch.py* module. Energies were calculated by finding the energy of the hollow enzyme (Section 2.3.7), then the designed one created during the repacking, and subtracting the hollow energy from the final design's energy. During post-repacking analysis of the structural results, ligand RMSD was calculated using PyMOL's *rms_cur* utility. The percentage of correct rotamers was determined through analysis of individual rotamers of all designed positions using *phenix.rotalyze*, a function from the PHENIX software suite (Berkeley, CA, USA) and a custom Python script for percentage calculations.

2.4 Results

To recap, control recapitulations were performed to determine the values of the criteria used for evaluating further recapitulation quality. Then, all ensembles are generated, followed by their use in the de novo pipeline to predict active site structures for 1A53-core and HG4. As an aside, the BR ensembles generated from Triad were compared to Rosetta's BR on the same inputs using RMSD metrics including diversity and deviation, to analyze the difference between the 2 implementations. Potential reasons for 1A53-core's recapitulation failures were also analyzed.

2.4.1 Control Experiments

In Figure 2.2, one can see the level of recapitulation that Triad can achieve for both (a) 1A53-core and (b) HG4. Evidently, Triad can recapitulate structures to the original designed template in question, indicating its competency to proceed with engineering 1A53-core using non-1A53-core backbone starting ensembles. These results are similar to what was seen in previous recapitulations (to be published). Most of the amino acids not shown in Figure 2.2a and b contribute to the correct rotamer percentage attenuation. The ligand RMSD is a measure of how far the predicted ligand is from the crystal's, with a larger number indicating a larger "distance", showing that the prediction for 1A53-core was slightly worse than HG4's ligand recapitulation. Such values were chosen for filtering of "hits" – recapitulations that can count as successful redesigns, where 1.45 Å was used as a cutoff for ligand RMSD and 66% for the correct rotamers for 1A53-core, whereas 0.69 Å and a correct rotamer percentage of 50% was chosen for the HG4 (adding a 0.25 Å RMSD leeway and a 6-7% rotamer correctness leeway for both) for the reasons mentioned in Section 2.3.1.

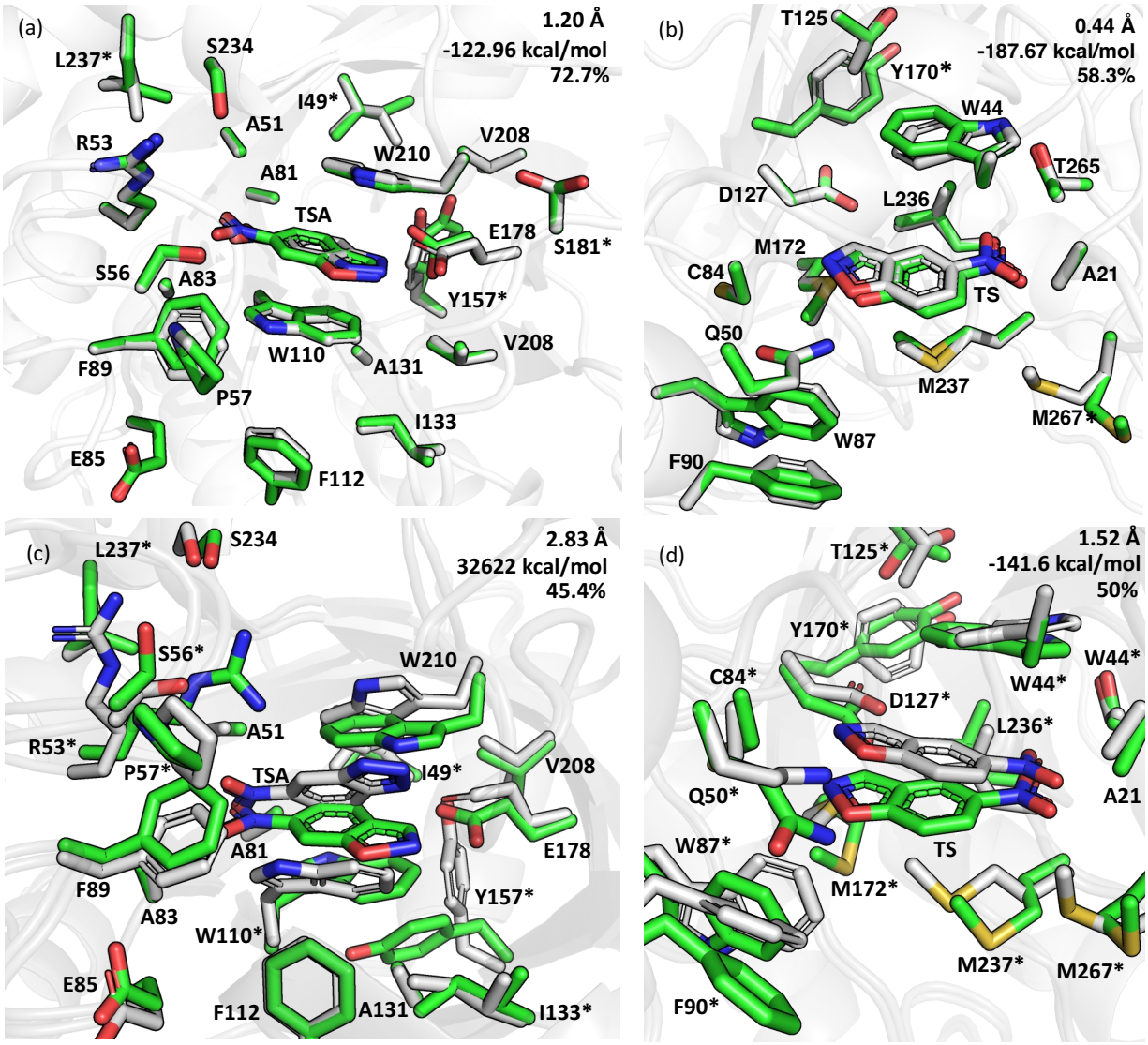


Figure 2.2: Control recapitulations of 1A53-core and HG4 using Triad. 1A53-core (a) positive and (c) negative control recapitulations, as well as HG4 (b) positive and (d) negative control recapitulations with the ligand RMSD, energy and percentage of correct rotamers shown in upper right corner (each entry written on the figure sequentially as described). For clarity, design residues 16, 17, 42, 46, 47, 79, 130, 207, 209, 234, 239, 276 are not shown in (b). Designed enzyme with the transition state colored with green carbons, crystal structure of enzyme with TSA colored with gray carbons. The modest percentage of correct rotamers can be attributed to the hydrophobic residues, whose sidechains provide ample degrees of freedom, especially for the χ_2 angle. Incorrectly predicted rotamers were denoted with an asterisk at the end of the label. Theozyme rotamers are fully aligned (in the same bin) with the crystal structure. Hydrogens removed for clarity. The high energy in (c) is due to steric clashes between A131 and Y157.

2.4.2 Ensemble Generation

To start, PertMin ensembles were generated, and 20 random templates taken from each ensemble (Figure 2.3) of 1A53, 3NYZ, 3NZ1 inputs was used for BR, generating the PM-BR serialized backbones (Figure 2.4), in order to see if this serialization can recapitulate 1A53-core from the 3 PDB inputs mentioned. Here, the ensembles generated from both PM and PM-BR display a relatively low deviation relative to all other ensembles (Table 2.6), likely due to PM's nature of only slightly perturbing the backbone atoms (on the milli-Ångstrom scale) and minimizing. Qualitatively, this is seen in the backbone structures in Figure 2.3-Figure 2.4, where the backbones are modestly more diverse in Figure 2.4 relative to Figure 2.3, as the backbones occupy a larger volume of the space in Figure 2.4. Fortunately, BR is able to overcome the low deviation as evidenced by the increase in all 3 deviations in Figure 2.3-Figure 2.4. Hence, this ensemble was used for motif generation and repacking for the purposes of 1A53-core recapitulation.

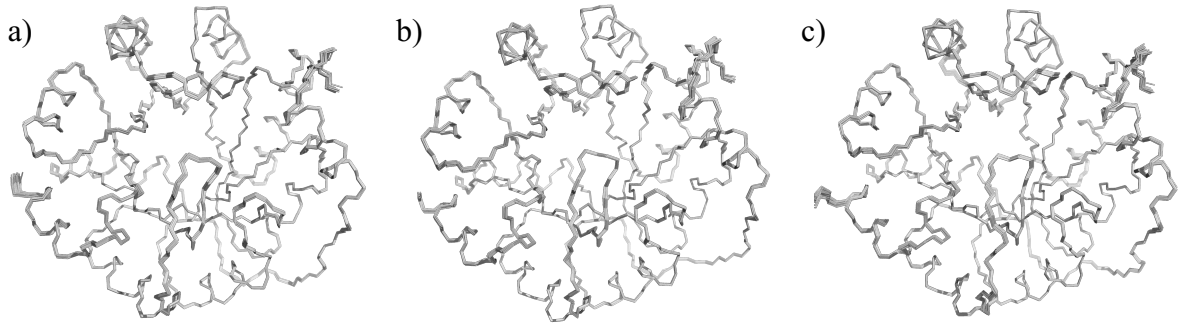


Figure 2.3: Backbone structures of the N=20 randomly chosen PM structures from each PDB. (a) 1A53, (b) 3NYZ and (c) 3NZ1 used for generating the PM-BR ensemble. The diversity and deviation for each PDB are as follows: 1A53 (0.18 ± 0.06 Å) and (0.23 ± 0.01 Å), 3NYZ (0.16 ± 0.05 Å) and (0.24 ± 0.01 Å), 3NZ1 (0.18 ± 0.06 Å) and (0.24 ± 0.02 Å).

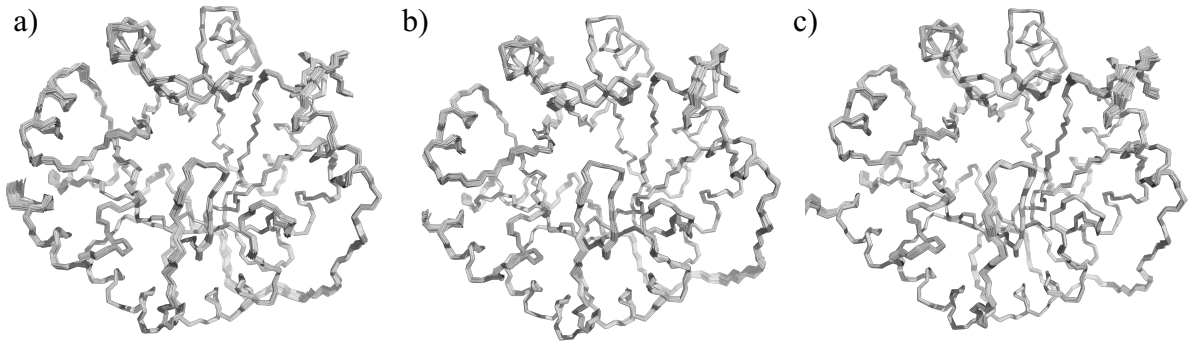


Figure 2.4: Backbone structures of the final PM-BR backbones (N=200 each), post-processed from each PDB: (a) 1A53, (b) 3NYZ, and (c) 3NZ1. The diversity and deviation for each PDB respectively are as follows: (0.23 ± 0.04 Å) and (0.36 ± 0.03 Å), (0.22 ± 0.04 Å) and (0.43 ± 0.01 Å), (0.21 ± 0.04 Å) and (0.39 ± 0.01 Å).

Contrary to the PM-BR results in the previous section, the ER and ER-BR ensembles display higher deviations, in large part due to the difference in ER relative to PM (electron density-constrained MD and small perturbation-minimization respectively). The highest deviation is found for 3NYZ, which is likely due to the structure being the apo analog of 3NZ1, hence leading to a much higher deviation relative to the original input. As evidenced by the values for Figure 2.5 and Figure 2.6, BR is able to generate a larger deviation for the ensemble after ER, hence the final 600 templates were used for recapitulation of 1A53-core as well.

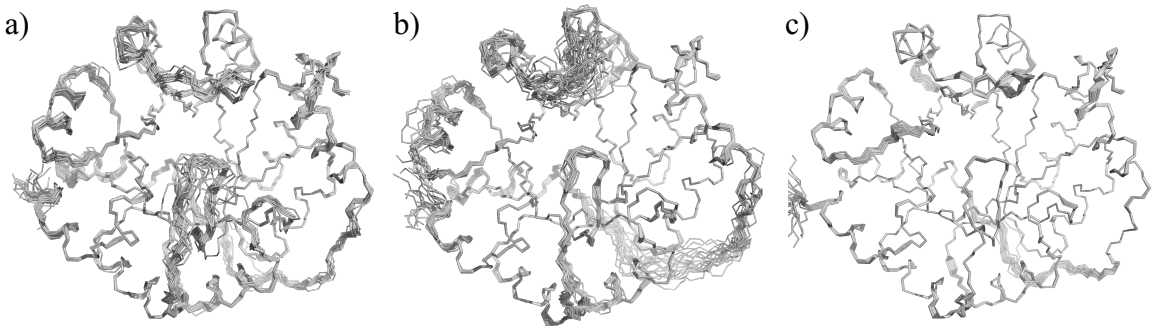


Figure 2.5: Backbone structures of the N=20 randomly chosen ER structures from each PDB, (a) 1A53, (b) 3NYZ and (c) 3NZ1 used for generating the ER-BR ensemble. The diversity and deviation for each PDB are as follows: 1A53 (0.5 ± 0.2 Å) and (0.61 ± 0.04 Å), 3NYZ (0.9 ± 0.3 Å) and (1.1 ± 0.1 Å), 3NZ1 (0.23 ± 0.07 Å) and (0.33 ± 0.02 Å).

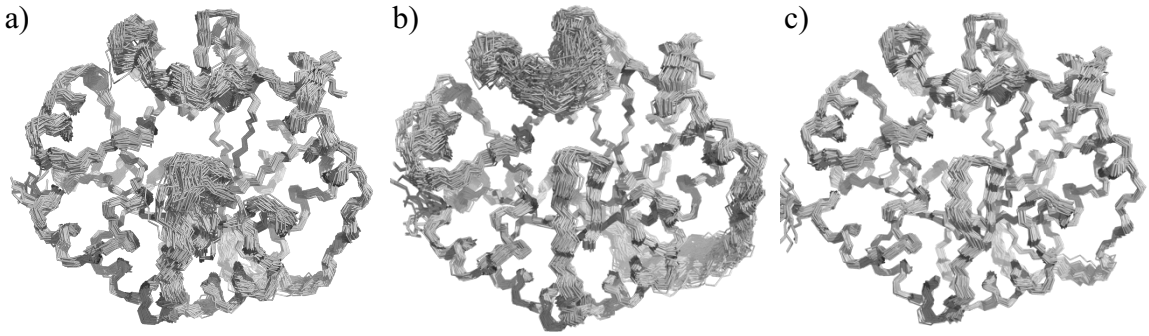


Figure 2.6: Backbone structures of the final ER-BR backbones (N=200 each), post-processed from each PDB using *proteinProcess.py*. (a) 1A53, (b) 3NYZ, and (c) 3NZ1. The diversity and deviation for each PDB respectively are as follows: (0.7 ± 0.1 Å) and (0.51 ± 0.07 Å), (1.0 ± 0.2 Å) and (1.0 ± 0.1 Å), (0.47 ± 0.06 Å) and (0.35 ± 0.04 Å).

As can be seen from Figure 2.7-Figure 2.8, as well as their respective deviations, the MD pipeline has the largest deviations for all 3 PDB inputs. Although it may appear as though the deviations are larger for the MD-BR than the BR (especially between Figure 2.7b and Figure 2.8b), the deviations seen in MD-BR (Figure 2.8) are lower or equal to MD (Figure 2.7). The likeliest cause is the energy minimization step in MD, which will guide the protein's scaffold towards a local minimum, thereby reducing the potential for movement to be generated during BR, leading to an ensemble deviation similar to MD's. Energy minimization was chosen as it is part of every default pipeline involving MD. In hindsight, using minimization (not just for MD but all methods) also faced similar issues, but was not identified until after the experiments were performed. Future work should only utilize minimization protocols for final ensemble cleanup. Hence, these

templates were not used for follow-up design, including hollow energy calculations, theozyme placement or repacking.

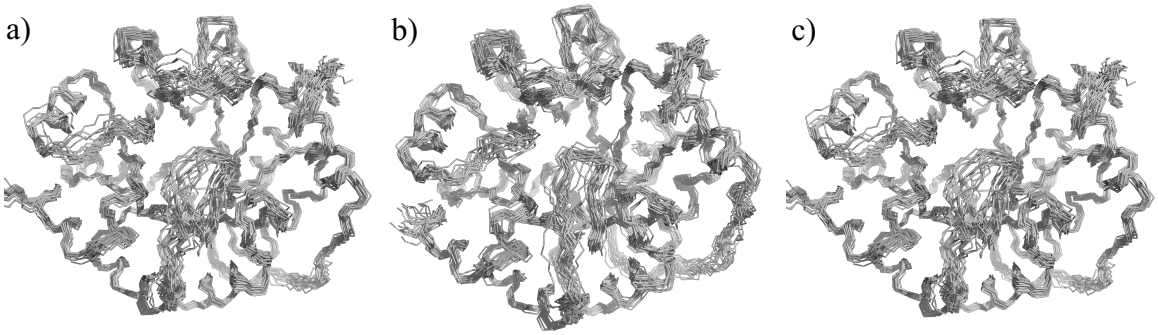


Figure 2.7: Backbone structures of the N=20 randomly chosen MD structures from each PDB. (a) 1A53, (b) 3NYZ and (c) 3NZ1 used for generating the PM-BR ensemble. The diversity and deviation for each PDB are as follows: 1A53 (0.8 ± 0.3 Å) and (0.8 ± 0.1 Å), 3NYZ (0.9 ± 0.3 Å) and (0.89 ± 0.09 Å), 3NZ1 (0.9 ± 0.3 Å) and (0.9 ± 0.1 Å).

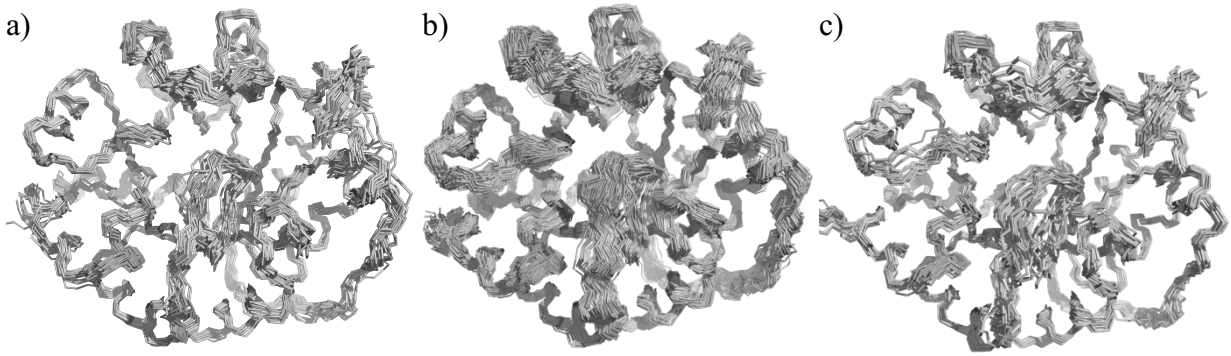


Figure 2.8: Backbone structures of the final MD-BR backbones (N=200 each) post-processed from each PDB: (a) 1A53, (b) 3NYZ, and (c) 3NZ1. The diversity and deviation for each PDB respectively are as follows: (0.8 ± 0.2 Å) and (0.8 ± 0.1 Å), (0.8 ± 0.2 Å) and (0.89 ± 0.09 Å), (0.8 ± 0.2 Å) and (0.9 ± 0.1 Å).

In Figure 2.9, one can see the results for the final minimized BR-only backbones. As usual, all disordered loops have higher apparent deviation than the regions with defined secondary structure, due to the lack of stabilizing hydrogen bonds to prevent BR occurring as frequently. As

there was no follow-up BR step after the initial BR (Figure 2.9), no deviation comparisons were done, and the minimized BR templates were used as-is for the theozyme placement and repacking design steps.

In Figure 2.10, the same minimized BR-only backbone outputs are shown, this time with a focus on the HG4-family proteins 1GOR, 5RG4 and 5RGA. The differing structural data output relative to Fig. 2.8 shown makes sense, as the structure of the TIM barrels for the HG4 family are different than the 1A53-core backbones. Finally, the diversities and deviations of both 1A53-core and HG4 protein ensembles seem relatively similar, as was an expected byproduct due to running the same BR process with the same default values on relatively similar TIM barrel structures. To check the competency of Triad's BR, it was decided to run BR on Rosetta and compare those ensembles with Triad's by analyzing diversities and deviations of both.

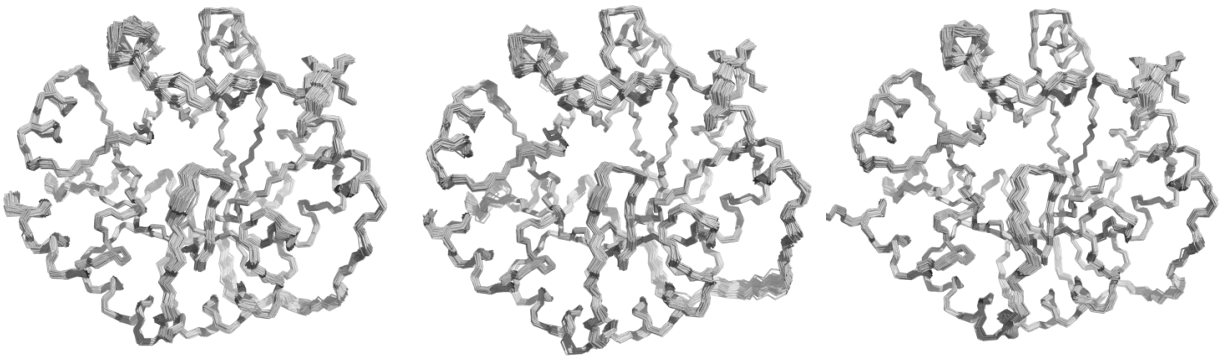


Figure 2.9: Backbone structures of the final BR-only backbones (N=200 each) post-processed from each PDB, (a) 1A53, (b) 3NYZ, and (c) 3NZ1. The diversity and deviation for each PDB respectively are as follows: (0.36 ± 0.05 Å) and (0.38 ± 0.02 Å), (0.36 ± 0.05 Å) and (0.41 ± 0.02 Å), (0.30 ± 0.04 Å) and (0.37 ± 0.02 Å).

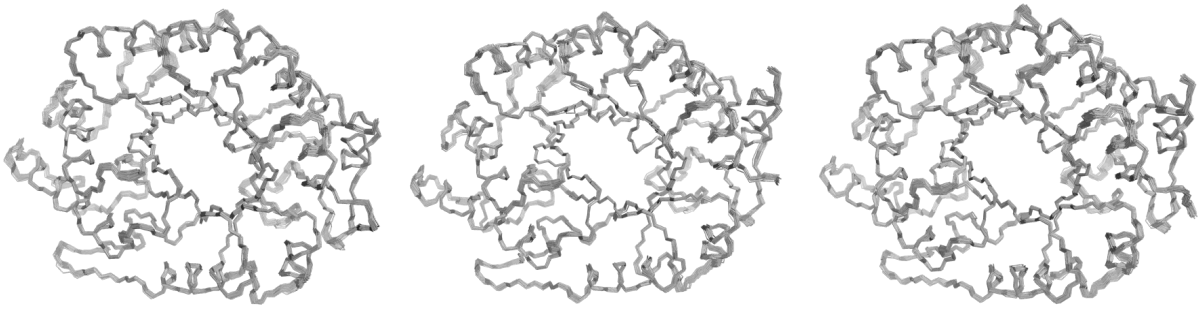


Figure 2.10: Backbone structures of the final BR-only backbones post-processed from each PDB, (a) 1GOR (N=80), (b) 5RG4 (N=50), and (c) 5RGA (N=84). The diversity and deviation for each PDB respectively are as follows: (0.30 ± 0.05 Å) and (0.40 ± 0.02 Å), (0.30 ± 0.06 Å) and (0.33 ± 0.02 Å), (0.30 ± 0.05 Å) and (0.32 ± 0.02 Å). A potentially problematic artifact is circled in red and discussed below.

Table 2.5: Diversities of backbone ensembles generated using various methods.

PDB	PM (Å)	ER (Å)	MD (Å)	BR (Å)	PM-BR (Å)	ER-BR (Å)	MD-BR (Å)
1A53	0.18 ± 0.06	0.5 ± 0.2	0.8 ± 0.3	0.36 ± 0.05	0.23 ± 0.04	0.7 ± 0.1	0.8 ± 0.2
3NYZ	0.16 ± 0.05	0.9 ± 0.3	0.9 ± 0.3	0.36 ± 0.05	0.22 ± 0.04	1.0 ± 0.2	0.8 ± 0.2
3NZ1	0.18 ± 0.06	0.23 ± 0.07	0.9 ± 0.3	0.30 ± 0.04	0.21 ± 0.04	0.47 ± 0.06	0.8 ± 0.2
1GOR	-	-	-	0.30 ± 0.05	-	-	-
5RG4	-	-	-	0.30 ± 0.06	-	-	-
5RGA	-	-	-	0.30 ± 0.05	-	-	-

Table 2.6: Deviations of backbone ensembles generated using various methods.

PDB	PM (Å)	ER (Å)	MD (Å)	BR (Å)	PM-BR (Å)	ER-BR (Å)	MD-BR (Å)
1A53	0.23 ± 0.01	0.61 ± 0.04	0.8 ± 0.1	0.38 ± 0.02	0.36 ± 0.03	0.51 ± 0.07	0.8 ± 0.1
3NYZ	0.24 ± 0.01	1.1 ± 0.1	0.89 ± 0.09	0.41 ± 0.02	0.43 ± 0.01	1.0 ± 0.1	0.89 ± 0.09
3NZ1	0.24 ± 0.02	0.33 ± 0.02	0.9 ± 0.1	0.37 ± 0.02	0.39 ± 0.01	0.35 ± 0.04	0.9 ± 0.1
1GOR	-	-	-	0.40 ± 0.02	-	-	-

5RG4	-	-	-	0.33 ± 0.02	-	-	-
5RGA	-	-	-	0.32 ± 0.02	-	-	-

2.4.3 Rosetta BR Benchmark Ensembles

As shown in Figure 2.11, Triad's BR ensembles are quite similar to Rosetta's, if the deviation of the ensembles relative to each respective original input file is considered. For instance, 3NYZ and 3NZ1 (Figure 2.11c, d) both have deviations of 0.39 ± 0.01 Å and 0.36 ± 0.01 Å (**Error! Reference source not found.**), which are extremely close to Triad's BR-only 3NYZ and 3NZ1 ensemble deviation values of 0.41 ± 0.02 Å and 0.37 ± 0.02 Å. The values for 3NZ1 lie within each other's standard deviations. However, the original 1A53's deviation for Triad is 0.38 ± 0.02 Å, which is larger than Rosetta's 0.33 ± 0.02 Å deviation, where both values do not intersect within one standard deviation. It was unable to be determined why 1GOR and 4A29 had an abnormally large deviation for Rosetta's and Triad's BR respectively, considering it is at least 8-fold larger than the 2nd-largest deviation belonging to 3NYZ. Rosetta's diversities usually were below Triad's values, including 1 standard deviation.

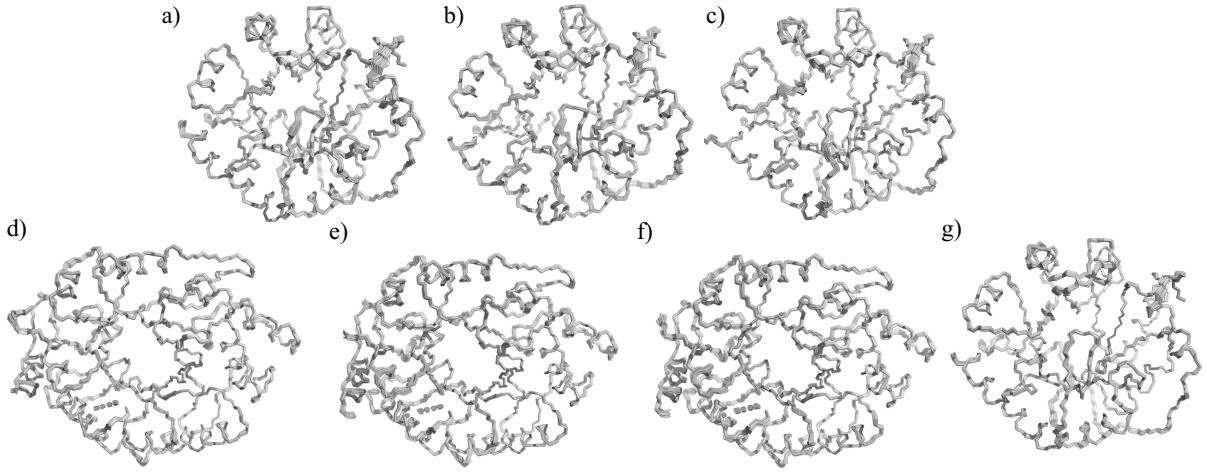


Figure 2.11: Triad's BR ensembles are similar to Rosetta's BR ensemble variant. Rosetta BR ensembles generated from (a) 1A53, (b) 3NYZ, (c) 3NZ1, (d) 1GOR, (e) 5RG4, (f) 5RGA, (g) 4A29 inputs. The number of templates in each ensemble, as well as the respective diversities and deviations are listed in Table 2.3.

Table 2.7: Diversities and deviations of Rosetta and Triad BR. Triad values are shown for comparison. Only backbone atoms (N, Cα, C) were considered in the calculation, for residues 2-248 each. Errors correspond to standard deviations. To be consistent with template numbers in Chapter 2.3.5, the number of templates generated using Rosetta were identical to that of Triad's.

PDB	Diversity (Å), Rosetta	Diversity (Å), Triad	Deviation (Å), Rosetta	Deviation (Å), Triad	N
1A53	0.24 ± 0.03	0.36 ± 0.05	0.33 ± 0.02	0.38 ± 0.02	200
3NYZ	0.25 ± 0.03	0.36 ± 0.05	0.39 ± 0.01	0.41 ± 0.02	200
3NZ1	0.22 ± 0.03	0.30 ± 0.04	0.36 ± 0.01	0.37 ± 0.02	200
1GOR	0.20 ± 0.03	0.30 ± 0.05	0.33 ± 0.01	0.40 ± 0.02	80
5RG4	0.21 ± 0.04	0.30 ± 0.06	0.30 ± 0.01	0.33 ± 0.02	50
5RGA	0.22 ± 0.03	0.30 ± 0.05	0.29 ± 0.01	0.32 ± 0.02	84
4A29	0.22 ± 0.02	0.32 ± 0.06	0.34 ± 0.01	1.39 ± 0.01	100

2.4.4 Final Recapitulated Designs

Overall, the 1A53-core recapitulations were not successful, and the HG4 recapitulations showed considerable success for the BR templates. To start, for the 1A53-core, all 3 serialized input templates (MD-BR, ER-BR, PM-BR) were unsuccessful for various reasons. MD-BR's templates could not increase their deviation from MD-only, leading to their templates being removed from consideration as inputs. PM-BR, although successful in increasing deviation relative to PM-only, did not generate any final structures post-repacking. It was difficult to compare the templates with PM-only templates as they had the same issue. ER-BR templates were the only structures that passed the deviation test and had outputs generated by the software after the recapitulation was complete. That being said, it is clear from the best structures outputted that the recapitulation with ER-BR templates (Figure 2.12) had failed. Comparing the ER-BR results (Figure 2.12a) to ER-only (Figure 2.12b) showed that while there are no significant structurally predictive improvements between the two ensembles, the ER-BR has improved energetics, likely due to reduced steric issues.

Further analyzing the energies for all templates (Figure 2.15-Figure 2.17), showed that very few templates (apart from the positive control) had a recapitulation energy below zero. This was a further reason why none of the templates resulted in recapitulations that lied within the criteria. To investigate why the recapitulations failed for all input cases, analysis was performed for the α-carbon deviations. From Figure 2.14-Figure 2.15, as well as Table 2.8-Table 2.9, it is clear that the level of accuracy between the serialized Cα-Cα theozyme predictions to the target 1A53-core enzyme active site do not correlate to the inability of Triad to predict 1A53-core's (PDB: 8FOQ) crystal structure, especially for ER-BR structure and theozyme deviation results (Figure 2.12, Figure 2.14). Firstly, one would expect to see an improvement of prediction for ER-

BR generated ensembles from 1A53 to 3NYZ, given the sharp decrease in deviations (Table 2.8), as the decrease in deviations with 3NYZ indicates a closer α-carbon alignment for relevant theozyme residues to the target crystal structure, meaning that the backbone structure proximal to the theozyme residues is be relatively closer to the target structure's, indicating improvement in theozyme-associated backbone structural prediction, but the only viable prediction generated by Triad was from a 1A53 input, with all 1A53-2 inputs leading to null output structures.

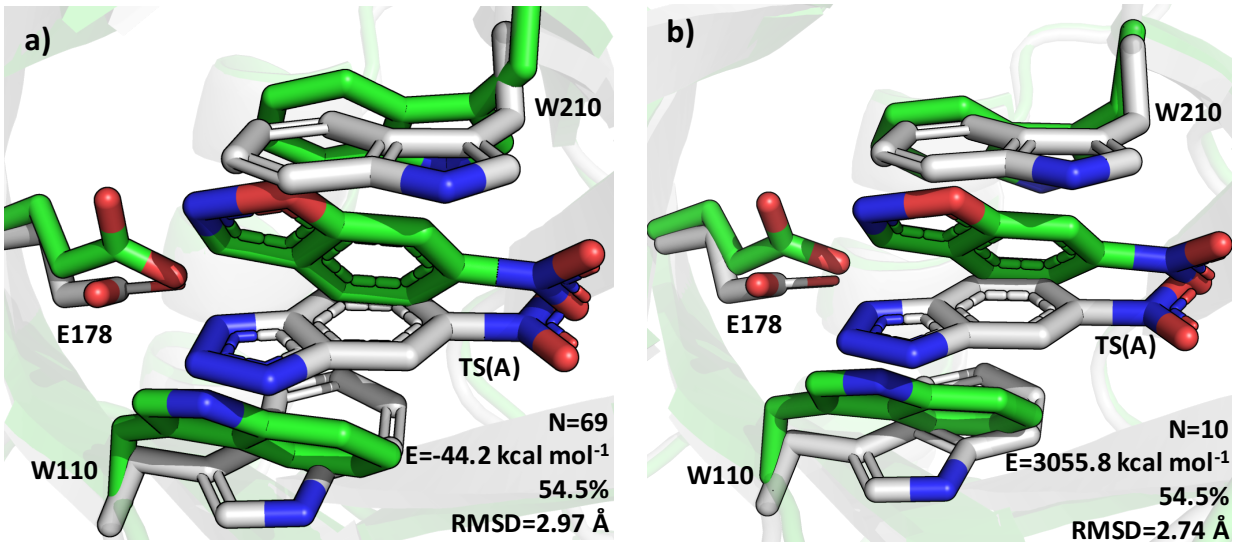


Figure 2.12: The combined ER-BR pipeline did not increase recapitulation accuracy. Final recapitulated 1A53-core theozyme of the (a) ER-BR pipeline and (b) the ER-only pipeline (shown as a control) with original PDB: 1A53 (3NYZ and 3NZ1 outputted no structures). Designed structure shown in green carbons, and crystal structure shown in gray carbons. Residues and TS, as well as crystal ligand TS analogue are labelled as TS(A). Number of viable structures (N, out of 800 for (a) and 80 for (b)), calculated energy, percent correct designed rotamers and ligand RMSD all shown in the bottom right of each figure. Hydrogens removed for clarity.

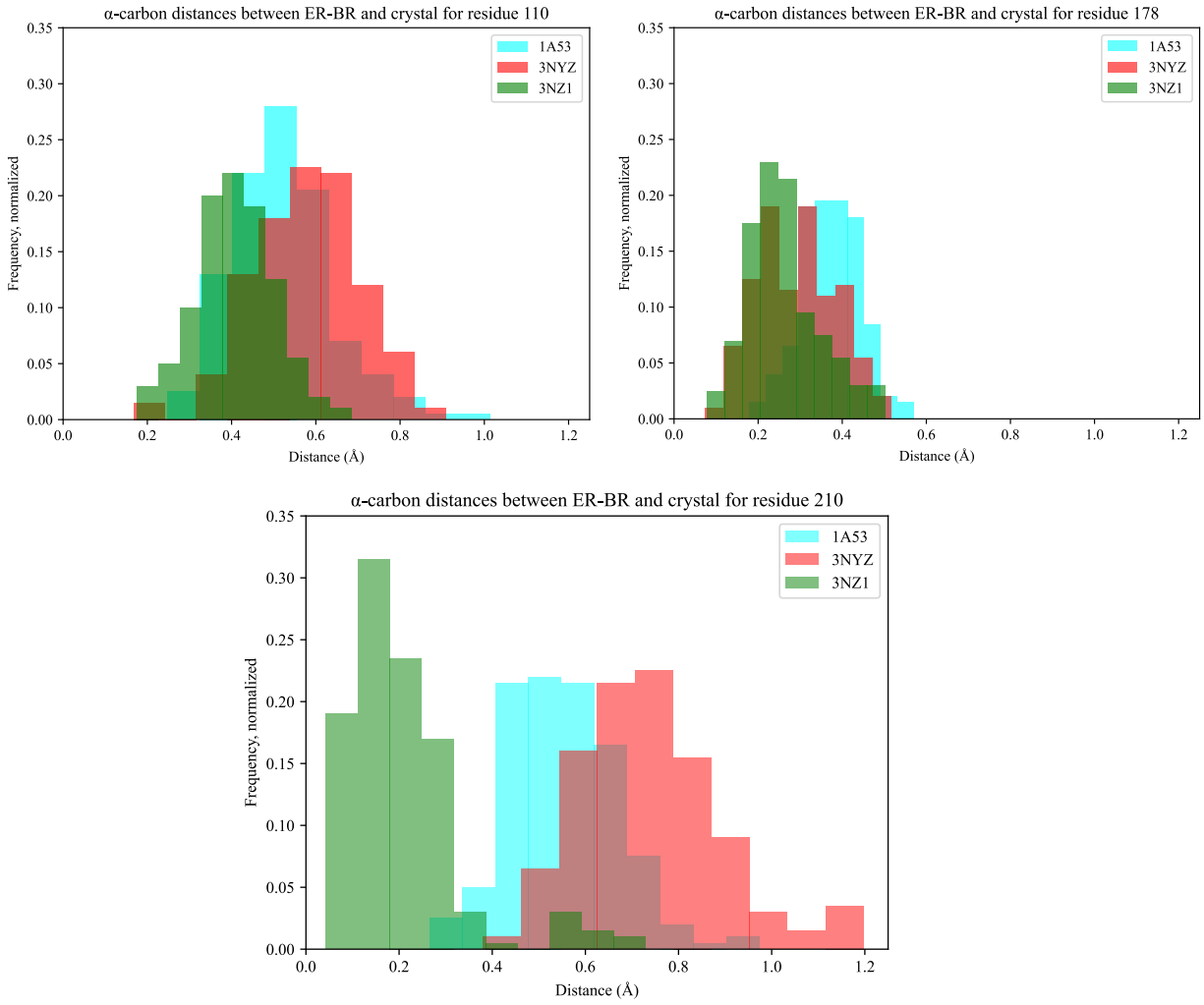


Figure 2.13: The distribution of carbon-to-carbon α-carbon distances for the ER-BR designs and the crystal structure do not explain the success of 1A53 and the failure of 1A53-2 structures to successfully design outputs. Histograms of 1A53, 3NYZ and 3NZ1 distributions for α-carbon-α -carbon distances for residues (a) 110, (b) 178 and (c) 210. In all cases, 3NZ1 has a lower arithmetic mean, whereas 1A53 is either situated in the middle, or has the largest mean.

Table 2.8: ER-BR Cα-Cα theozyme mean distances. These values approximately correspond to the centers of the histogram distributions in Fig. 2.13. Errors equal the standard deviation.

PDB \ Res	110	178	210
1A53	0.5 ± 0.1	0.37 ± 0.07	0.6 ± 0.1
3NYZ	0.6 ± 0.1	0.29 ± 0.09	0.7 ± 0.2
3NZ1	0.41 ± 0.09	0.26 ± 0.08	0.2 ± 0.1

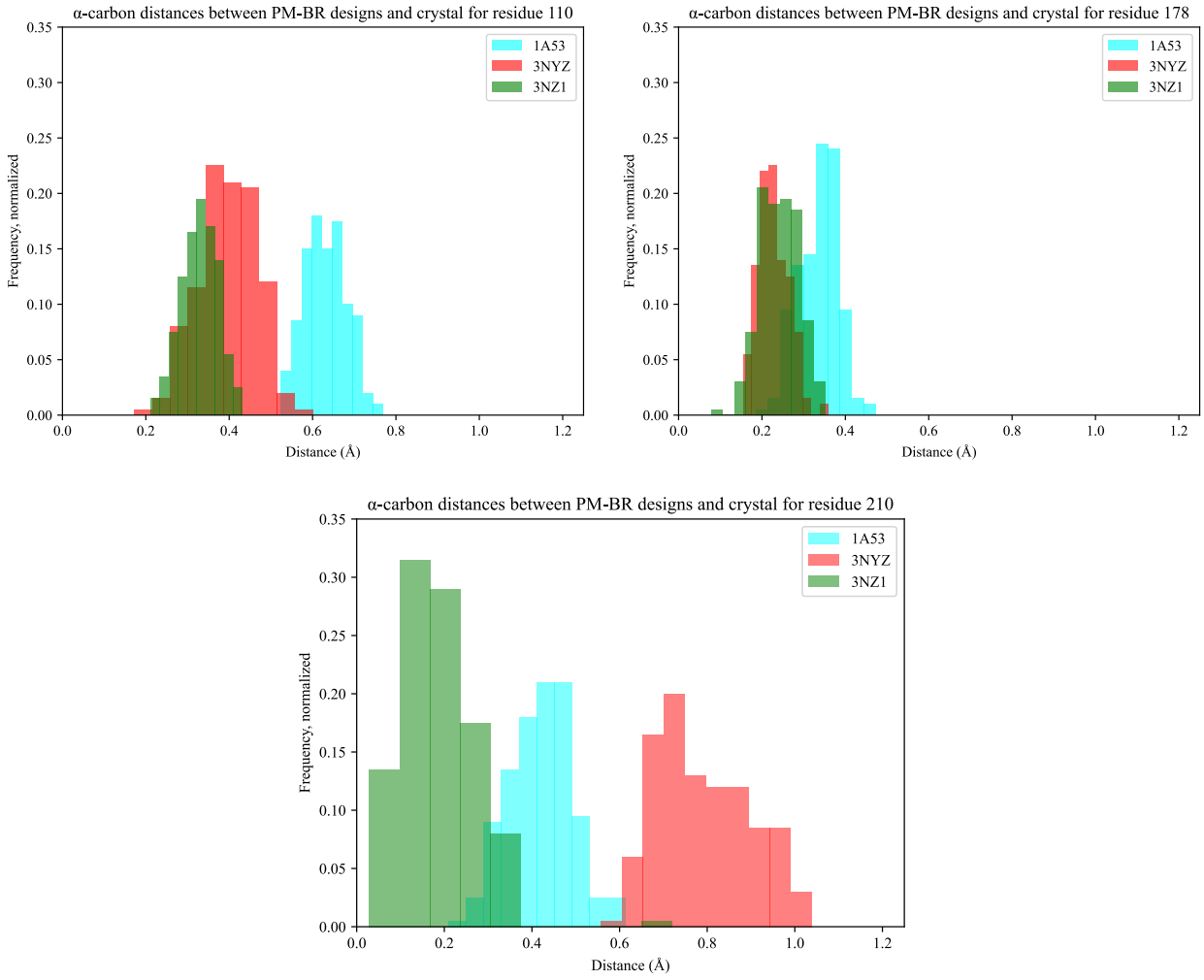


Figure 2.14: The distribution of carbon-to-carbon α-carbon distances for the PM-BR designs and the crystal structure do not explain the success of 1A53 and the failure of 1A53-2 structures to successfully design outputs. Histograms of 1A53, 3NYZ and 3NZ1 PM-BR distributions for α-carbon-α-carbon distances for residues (a) 110, (b) 178 and (c) 210.

Table 2.9: PM-BR Cα-Cα theozyme mean distances. Values approximately correspond to the centers of the histogram distributions in Fig. 2.14. Errors equal the standard deviation.

PDB \ Res	110 (Å)	178 (Å)	210 (Å)
1A53	0.63 ± 0.05	0.34 ± 0.05	0.42 ± 0.07
3NYZ	0.40 ± 0.07	0.23 ± 0.04	0.8 ± 0.1
3NZ1	0.33 ± 0.04	0.24 ± 0.05	0.19 ± 0.08

As seen from Figure 2.15-Figure 2.17, there were no hits (green bars) generated. In the case of 1A53-core, a hit would have reflected a predicted structure whose energy is below 0 kcal/mol, a ligand RMSD below 1.45 Å, and a percentage of correct rotamers higher than 66%. Unfortunately, this means that there were no viable predictions generated as defined by the filtering.

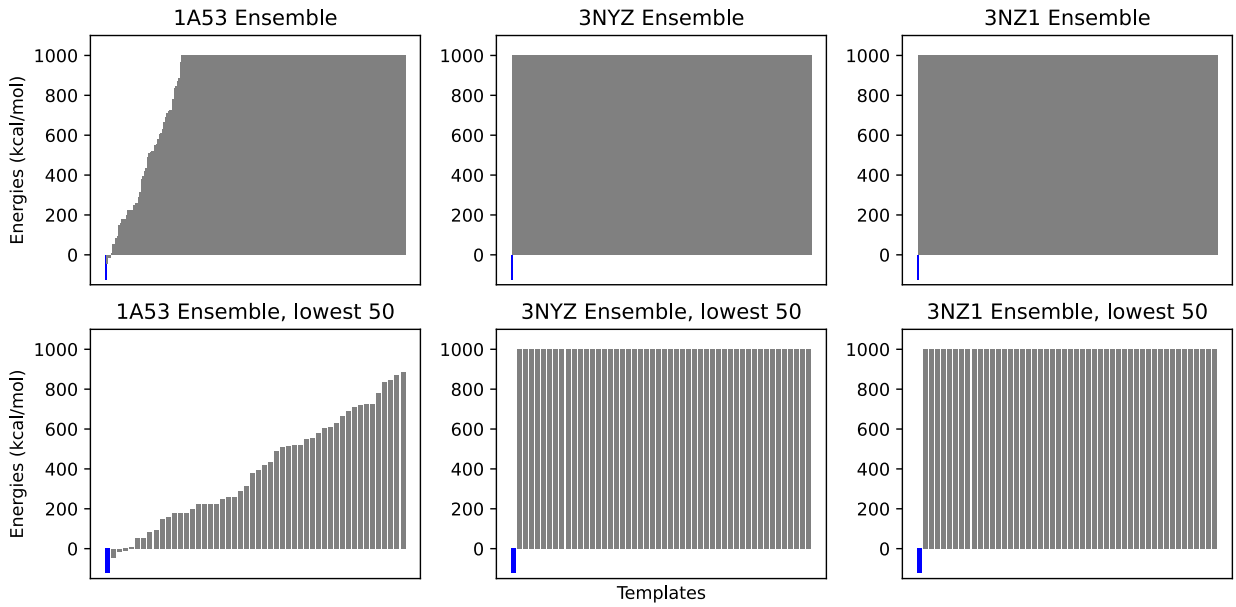


Figure 2.15: Energies of final 1A53-core ER-to-BR recapitulation corresponding to each template in the ensemble for all input PDBs. All N=200 templates are shown for the top graphs, and the 50 lowest-energy structures from each ensemble are shown as the bottom graphs. Any templates that were unable to be generated during recapitulation (e.g. all 3NYZ and 3NZ1 repacking structures) had their energy capped at 1000 kcal/mol for completion's sake. Blue bars indicate the positive control. Gray bars indicate templates that were not hits. Since no structure was generated for the negative control, the respective bar is not shown.

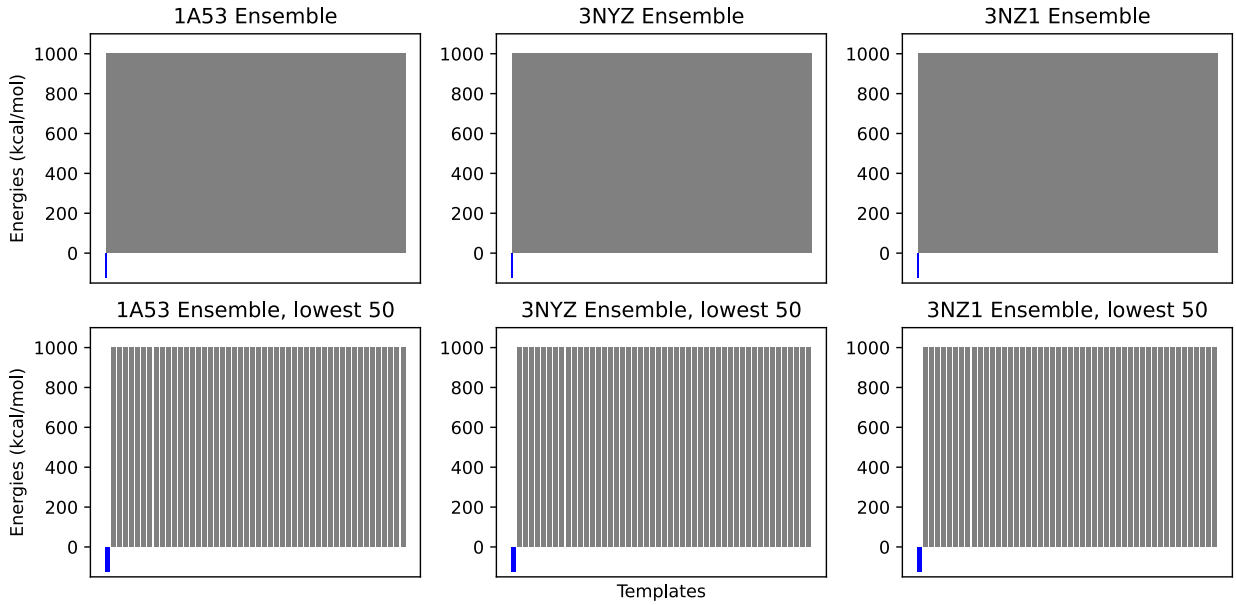


Figure 2.16: Energies of final 1A53-core PM-to-BR recapitulation corresponding to each template in the ensemble for all input PDBs. All N=200 templates are shown for the top graphs, and the 50 lowest-energy structures from each ensemble are shown as the bottom graphs. Any templates that were unable to be generated during recapitulation (e.g. all structures) had their energy set at 1000 kcal/mol for completion's sake. Blue bars indicate the positive control and its energy. As no structure was generated for the negative control, the respective bar is not shown.

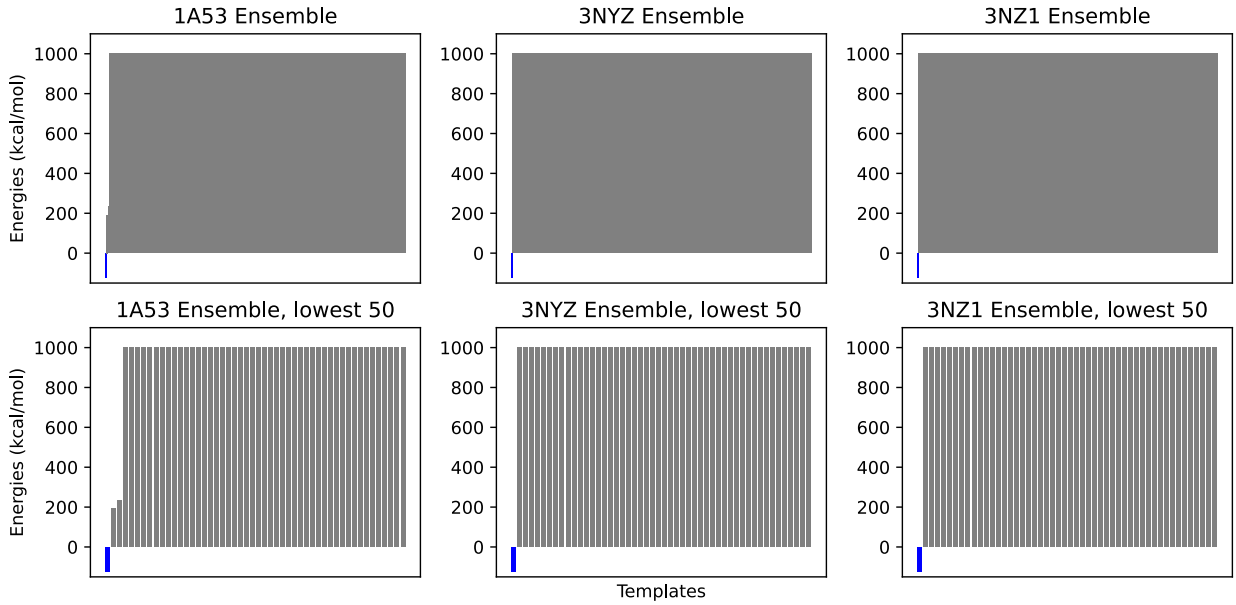


Figure 2.17: Energies of final 1A53-core BR-only recapitulation corresponding to each template in the ensemble for all input PDBs. All N=200 templates are shown for the top graphs,

and the 50 lowest-energy structures from each ensemble are shown as the bottom graphs. Any templates that were unable to be generated during recapitulation (e.g. all 3NYZ and 3NZ1 templates) had their energy set at 1000 kcal/mol for completion's sake. Blue bars indicate the positive control and its energy. As no structure was generated for the negative control, the respective bar is not shown.

Continuing on to BR-only inputs, it is clear from Figure 2.18, that 1A53-core was unable to be recapitulated given the inputs of 1A53, 3NYZ, and 3NZ1 used to generate BR templates. Not only are the rotamer percentages rather low, but the ligand RMSD is also much higher than the 1.20 A value determined from the positive control (Figure 2.2a). Again, it is strange to see the 1A53 protein leading to outputs, whereas the 3NYZ and 3NZ1 structures, which are evolutionarily closer to 1A53-core, not generating any output structures. Admittedly, there seems to be a pattern with 1A53 ensembles leading to outputs (even if they are far from preferred recapitulation accuracy) and the evolved 1A53-2 inputs not leading to any, suggesting that there might be an inherent structural issue in the backbones from 1A53-2 structures and the rotamers that are attempted to be designed (Table 2.2). This is explored further in 2.5.2.2.

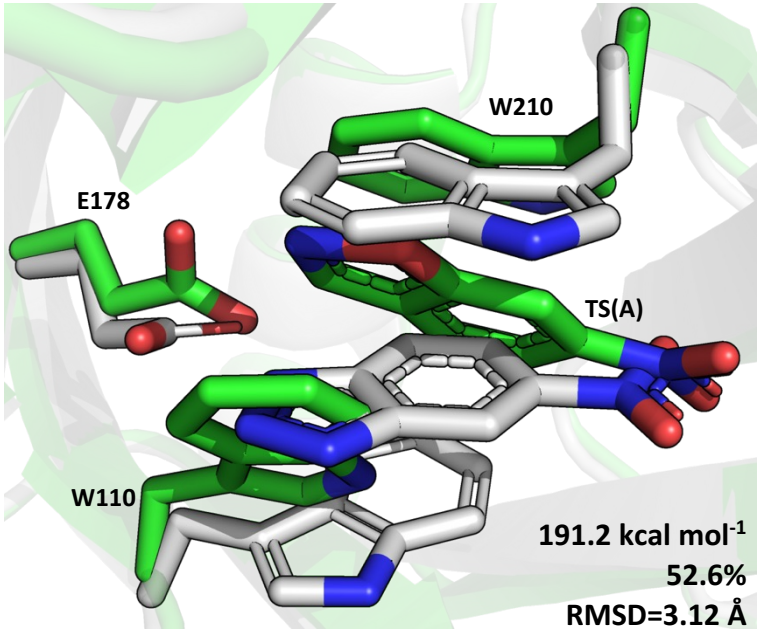


Figure 2.18: Lowest-Energy final repacking theozyme structure of 1A53-core derived from BR-only templates of 1A53, 3NYZ, 3NZ1. A 1A53 input template is shown here. Only 1A53 inputs were successful in generating results. Number of total templates, energy, percentage of correct rotamers and ligand RMSD shown in bottom right-hand corner.

To contrast, the HG4 recapitulations were more successful. As one can see from Figure 2.19, the BR was successful in recapitulating the structure of HG4 using BR templates generated from the WT 1GOR structure. This is notable, as not only was there a successful recapitulation of a high efficiency (>100,000 M^{-1} s^{-1}) enzyme, but this was also done by using just BR on a WT protein with no previous Kemp Eliminase activity (Figure 2.19a). This is a significant improvement in recapitulation over the issues seen with 1A53-core, where neither WT, nor evolved enzyme, nor serialized XX-BR methods could redesign the necessary rotamers and ligand positioning. Furthermore, this suggests that it is possible (at least in Kemp Eliminase's case) to acquire a high-efficiency enzyme through a WT non-catalytic variant by generation of ensembles

from such variant using a simple mechanical manipulation tool, without having to resort to more complex tools such as AlphaFold.

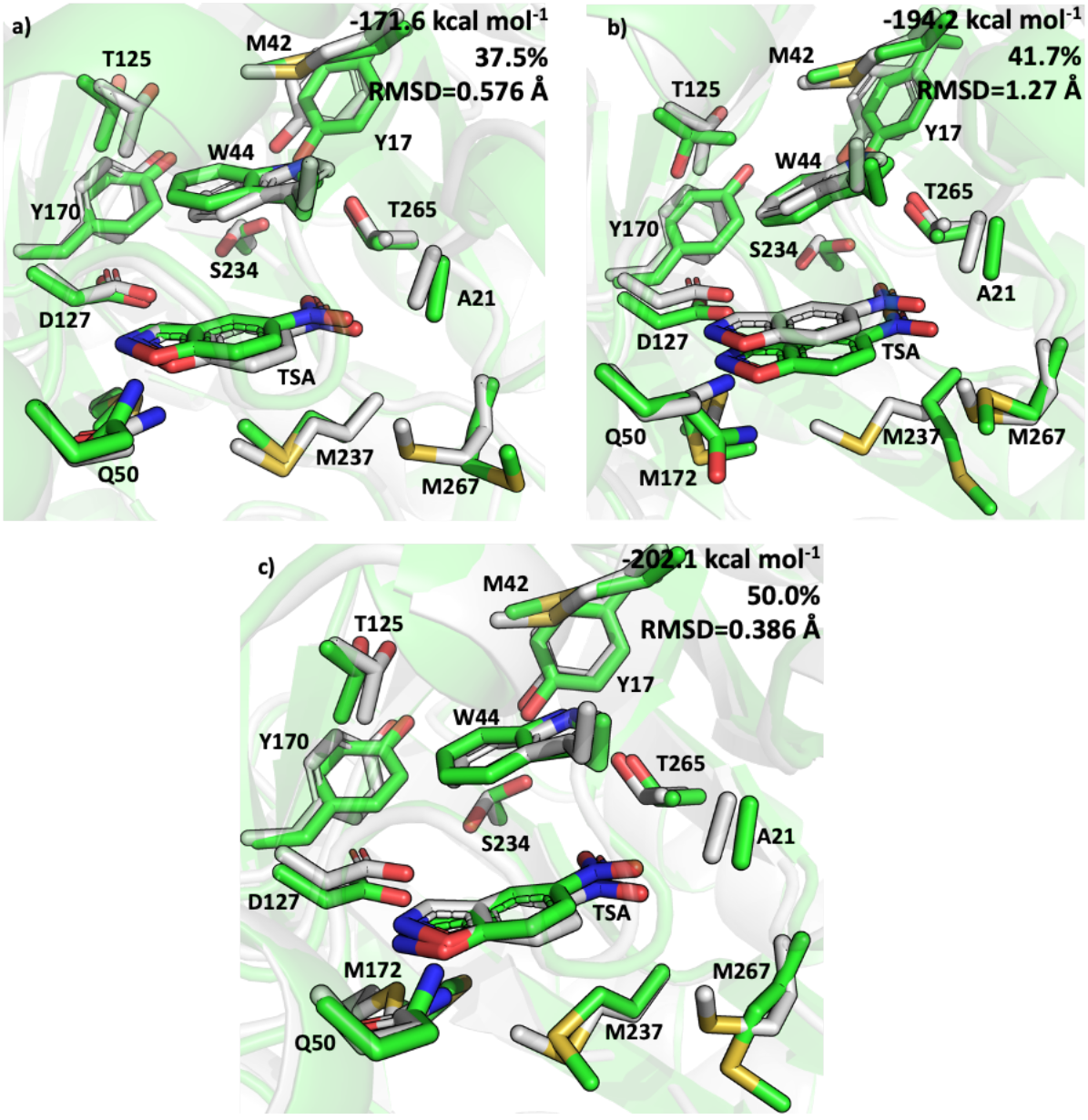


Figure 2.19: Lowest-Energy final repacking structure of HG4 derived from BR-only templates of (a) 1GOR, (b) 5RG4, (c) 5RGA. Final active site energy (design residues and ligand), percentage of correctly predicted rotamers and ligand RMSD relative to crystal structure written in top right-hand corner. In all cases, residues 16, 46, 47, 79, 84, 87, 90, 207, 209, 236, 239, 275 and 276 were removed for clarity. All of the 3 protein PDB inputs were able to generate

results, and because of this, all 3 figures to represent the lowest energy design for each PDB is shown. For 5RGA (c), the ligand RMSD is lower relative to the positive control, indicating excellent ligand prediction.

For Figure 2.20, the total number of hits for the 1GOR, 5RG4 and 5RGA results are as follows (respectively): 19, 0, 29. The 48 total hits are much higher than the zero hits resulting from the 1A53-core pipeline. A few reasons for why this may be the case is discussed further in Section 2.5.2. The 0.69 Å quantity for the filter was chosen based on the positive control's ligand RMSD (0.44 Å) as well as an extra 0.25 Å added to not under-sample potential hits. As seen from Figure 2.21 and Table 2.6, 1GOR has the highest deviation for residue 50, but the lowest for aspartate 127. As for 5RG4 and 5RGA, both display significant overlap in both residues, with residue 50 featuring a lower deviation on average relative to 1GOR, compared to residue 127 where both PDB final repacks displaying a higher deviation. The significance of this is explored further in the discussion section (Section 2.5.2.2).

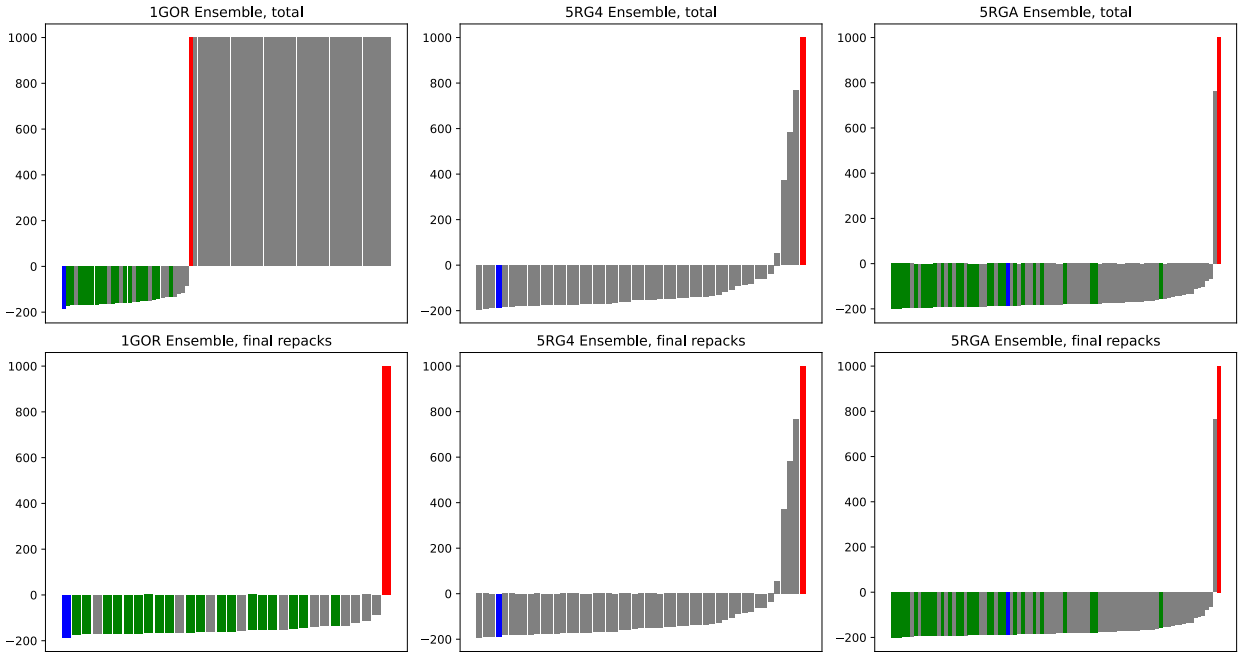


Figure 2.20: Bar graphs of final repacking energies by template, organized by ascending energies. Blue bars represent the positive control (HG4 recapitulation using HG4's backbone – see Fig. 2.2c), red bars represent the negative control (since Fig. 2.2d has no output, this was set to 1000 kcal/mol by default), and green bars represent those designed repacking outputs that fell within crucial criteria (energy below zero, and ligand RMSD below 0.69). Final repacks (second row) were all of the final repacks that were outputted by Triad at the end of the simulation. Some inputs did not generate an output, hence those templates are set to 1000 kcal/mol and are featured in the total graphs (first row, after negative control) for a total of N=80, 50, 84 templates each in the first-row graphs (1GOR, 5RG4, 5RGA respectively).

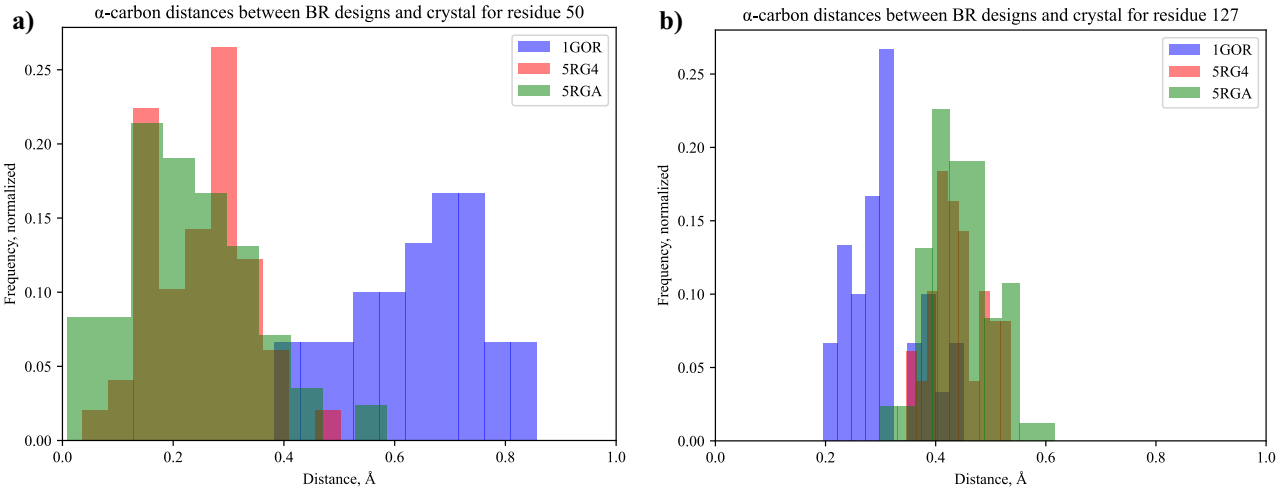


Figure 2.21: The distribution of carbon-to-carbon Cα distances for the BR designs and the crystal structure do not correlate to the number of hits. Histograms of 1GOR, 5RG4 and 5RGA BR distributions for α-carbon-α-carbon distances for residues (a) 50 and (b) 127.

Table 2.10: BR-only Cα-Cα theozyme mean distances. Values approximately correspond to the centers of the histogram distributions in Fig. 2.20. Errors equal the standard deviation.

PDB \ PID	50 (Å)	127 (Å)
1GOR	0.6 ± 0.1	0.30 ± 0.07
5RG4	0.25 ± 0.09	0.44 ± 0.05
5RGA	0.2 ± 0.1	0.45 ± 0.06

2.4.5 1A53-core Theozyme Backbone Deviation

In Sections 2.4.2-2.4.4, it was explained that for serialized methods to BR, only the ensembles whose deviations increased after BR would be accepted. However, the RMSD values were derived from the entire protein backbone, rather than the theozyme backbone atoms. To test if there would be a difference in ensemble acceptance if only the theozyme atoms were considered, diversity and deviation calculations were performed for all 1A53, 3NYZ and 3NZ1 ensembles

(excluding BR-only as that pipeline had no method before BR) on nitrogen and carbon backbone atoms (N, CA, C as named in PDB files) belonging to residues 110, 178 and 210.

From Table 2.11-Table 2.12, the first noticeable difference between the full-protein and theozyme-only calculations is that the latter's magnitudes are smaller (relative to the values seen in Section 2.4.2-2.4.4 figure legends). This is anticipated, since the theozyme is located within the protein's beta-barrel, which tends to deviate less than other parts of the protein (e.g. loops) due to the beta-strand hydrogen bonding strength and thus will have lower diversities relative to other templates in the ensemble (Table 2.11), as well as lower deviations relative to the input (Table 2.12). From Table 2.12, one can see that if only theozyme deviations were considered for ensemble acceptance, MD-BR would not be accepted as the deviations are similar to or lower than their MD equivalents for each input PDB, exactly what is seen in Section 2.4.4. Furthermore, ER-BR would be accepted, as the deviations are higher than their ER counterparts, similar to Section 2.4.3. This is due to the criteria for acceptance of an XX-BR ensemble for it to have a higher deviation than the XX ensemble relative to the input structure for it to be utilized as input in the de novo pipeline. However, the PM-BR deviations here have the same problem as the MD-BR, which is not what was seen in Section 2.3.3. Hence, the major change would be that PM-BR ensembles would not be accepted, like MD-BR. Since the ensembles created by PM-BR did not lead to any valuable outputs as seen in 2.4.4, it would have conserved time to not focus on PM-BR ensembles as inputs for 1A53-core recapitulation, and a more thorough investigation for why ER-BR failed to recapitulate 1A53-core could have been explored.

Table 2.11: Diversities of theozyme backbone atoms derived from various single and serialized method-generated ensembles.

PDB	MD (Å)	MD-BR (Å)	PM (Å)	PM-BR (Å)	ER (Å)	ER-BR (Å)
1A53	0.3 ± 0.1	0.3 ± 0.1	0.07 ± 0.03	0.08 ± 0.04	0.05 ± 0.02	0.16 ± 0.07
3NYZ	0.4 ± 0.2	0.3 ± 0.1	0.07 ± 0.03	0.10 ± 0.04	0.05 ± 0.02	0.16 ± 0.06
3NZ1	0.4 ± 0.2	0.3 ± 0.2	0.07 ± 0.04	0.07 ± 0.04	0.08 ± 0.03	0.15 ± 0.06

Table 2.12: Deviations of theozyme backbone atoms derived from various single and serialized method-generated ensembles.

PDB	MD (Å)	MD-BR (Å)	PM (Å)	PM-BR (Å)	ER (Å)	ER-BR (Å)
1A53	0.32 ± 0.08	0.33 ± 0.09	0.20 ± 0.02	0.19 ± 0.04	0.08 ± 0.02	0.13 ± 0.04
3NYZ	0.3 ± 0.1	0.3 ± 0.1	0.14 ± 0.02	0.15 ± 0.02	0.05 ± 0.01	0.12 ± 0.04
3NZ1	0.32 ± 0.09	0.31 ± 0.09	0.12 ± 0.02	0.14 ± 0.02	0.06 ± 0.02	0.12 ± 0.04

2.4.6 Analysis of 1A53-core Theozyme Cα-Cβ RMSDs

In Section 2.4.4, an analysis of theozyme Cα deviations relative to the crystal structure target (1A53-core) was discussed. However, using just the α-carbon could ignore valuable information about the rotamer's conformation, which could be solved by considering the Cα and Cβ deviations from the crystal structure (i.e. RMSD). It was also shown previously (article to be published) that larger agreements between input and target rotamer geometry for catalytic residues as measured by the Cα-Cβ vector agreement leads to improved prediction.

For Figure 2.22, the histogram of PDB 1A53 (the best performing input for ER-BR recapitulation) lies between 3NYZ and 3NZ1 in all cases, except for residue 210 where one can see a bimodal distribution. The ER-BR histograms seen in Section 2.4.4 are similar to Figure 2.22, except for residue 178, for which 1A53 has the highest deviation as seen in Figure 2.13.

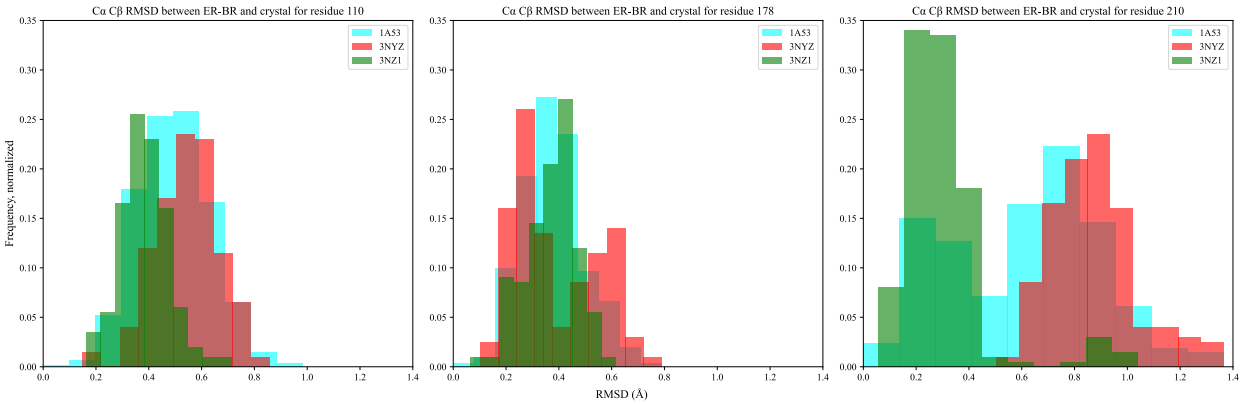


Figure 2.22: Combined ER-BR Cα-Cβ RMSDs for all 3 theozyme residues. RMSDs were calculated for the input templates relative to the target 1A53-core enzyme.

Table 2.13: The RMSD averages for Cα and Cβ atoms of ER-BR templates relative to 1A53-core. This table is a quantitative summary of Fig. 2.21.

PDB \ Res	110 (Å)	178 (Å)	210 (Å)
1A53	0.5 ± 0.1	0.4 ± 0.1	0.6 ± 0.3
3NYZ	0.5 ± 0.1	0.4 ± 0.2	0.9 ± 0.2
3NZ1	0.38 ± 0.09	0.4 ± 0.1	0.3 ± 0.2

For Figure 2.23, one can see similar distributions that are found from previous analyses of Cα deviations (Figure 2.14), except in the former, residue 178's 1A53 is approximately found lying in between 3NYZ and 3NZ1, whereas the distribution for 1A53 in Figure 2.14 was centered the furthest away from zero. Again, the PM-BR distances do not explain why all 3 inputs could not generate successful recapitulations from the PM-BR ensemble input.

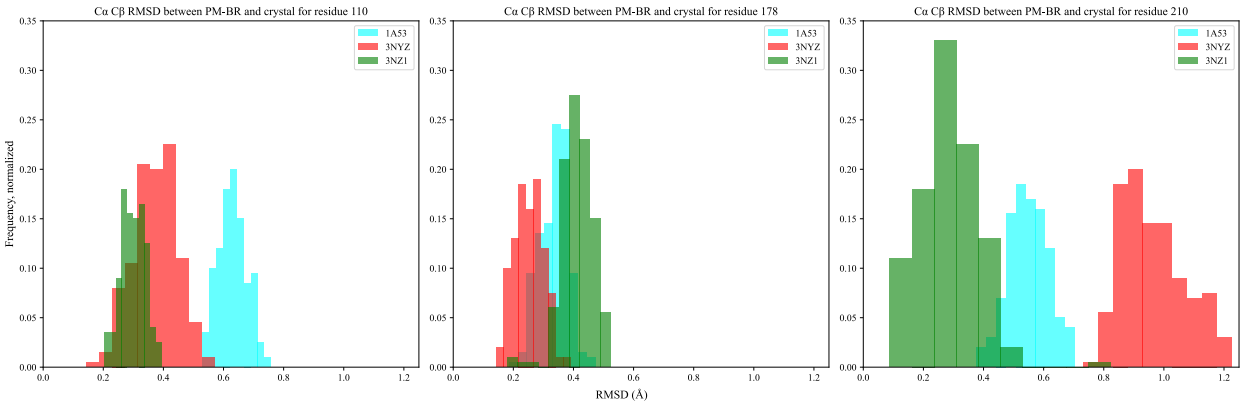


Figure 2.23: Combined PM-BR Cα-Cβ RMSDs for all 3 theozyme residues. RMSDs were calculated for the input templates relative to the target 1A53-core enzyme.

Table 2.14: The RMSD averages for Cα and Cβ atoms of PM-BR templates relative to 1A53-core. This table is a quantitative summary of Fig. 2.22.

PDB \ Res	110 (Å)	178 (Å)	210 (Å)
1A53	0.63 ± 0.05	0.34 ± 0.05	0.55 ± 0.07
3NYZ	0.37 ± 0.07	0.25 ± 0.05	1.0 ± 0.1
3NZ1	0.30 ± 0.04	0.41 ± 0.05	0.3 ± 0.1

2.5 Discussion

2.5.1 Generated Ensembles

All structures generated (Figure 2.3-Figure 2.10) had deviations that lied within 0.15-1 Å, which are favorable for approximating the sub-angstrom distances required for an enzyme to properly catalyze a reaction.[84,95] The structures with the largest deviations were the last N=600 MD-BR structures. Ironically, these coordinate files when comparing deviations did not have a noticeable increase in deviation relative to the initial MD ensembles (Figure 2.7), leading these inputs to be discarded during the selection for templates to be used in theozyme placement and subsequent repacking. It is unclear why this was the case, especially when qualitatively (Figure 2.8) they appeared to have larger deviations than the N=20 ensemble (Figure 2.7). One possible

explanation is that due to the minimization protocol performed just before the production MD step, the proteins were subjected to a low energy well, that made it difficult for the BR to effectively sample any backbone changes, due to the Metropolis MC criterion (Section 2.3.2) failing more frequently as a result of the higher energetic penalty incurred via the protein's location in a local minimum. However, this seems unlikely given that the minimization protocol performed this role for all replicates and frames, and led to similar results for all protein inputs. Also, the diversities and deviations calculated for all ensembles used a full-protein approach, where all backbone atoms from residues 2-248 were included for the calculations for all proteins. There could be an increase in accepted ensembles chosen for subsequent recapitulation, if the deviations and diversities of the active site backbone atoms were the sole coordinates used for RMSD calculations. In future designs, this would be a more appropriate approach, as these backbone atoms would influence the reaction's efficiency the most, compared to the entirety of the protein. To be thorough, one might also consider the residues within 1 residue of each active site residue position, as their interactions with the active site would likely be relevant.

With regards to the Rosetta BR benchmarks, the ensembles generated are similar to those of Triad's, considering the deviations seen in Table 2.7. However, the same table shows the lack of agreement between Rosetta and Triad's ensemble diversities, with Rosetta's being lower than that of Triad's. Seeing as all BR variables (e.g. MC step number) were controlled for, this is likely a result of the different energy functions between Rosetta and Triad, *REF15*[96] and *phoenix*[26] respectively. *REF15* has covalent terms, while *phoenix* is non-covalent by default. In principle, this would allow for less restricted movement between MC steps, which would explain the higher diversity for Triad BR, but this effect should also extend to the deviation as well. While the

deviation is higher in all cases (except for the 1GOR outlier), the difference's magnitude is lesser than what is seen for all PDB diversities, indicating that another effect could be present in Rosetta's BR that is absent from Triad's. The precise nature of this discrepancy is unclear, but likely arises from a combined interplay between Rosetta's energy function and algorithm, which both differ from Triad's. The algorithm itself likely plays a more significant role here, but the exact cause is difficult to deduce, as Triad's backrub implementation was designed to be as faithful as possible to Kortemme's, which was later implemented in Rosetta.

2.5.2 Final Repacking

2.5.2.1 1A53-core

After a rigorous examination of the ensemble designs' quality, one must consider how effective they were at recapitulating their intended target structures of 1A53-core and HG4. From Figure 2.12, one can see the best recapitulations for the ER-BR pipeline. Unfortunately, no final structures were generated for 3NYZ nor 3NZ1 initial protein input templates. The likeliest reason for this is the Y157 rotamer necessary for hydrogen bonding with the catalytic glutamate would not be able to fit given the geometry of the W210 rotamer (Figure 2.24), producing a steric clash if the two were forced to adopt such a configuration, affecting all pipelines (not just ER-BR). It is likely that Triad's method for analyzing sets of rotamers discovered this, and as such could not provide the correct Y157 rotamer, leading to no output in some instances. Ironically, this also suggests that the 1A53-core enzyme is a difficult one to fully recapitulate from evolutionarily similar enzymes like 1A53-2, as even if ligand and non-catalytic rotamer conformations may be well-predicted, the tyrosine (which forms an important hydrogen bond with a catalytic glutamate)

would never be fully redesigned. As the original 1A53 does not have W210, the tryptophan rotamer clash analysis for 1A53 WT was not performed.

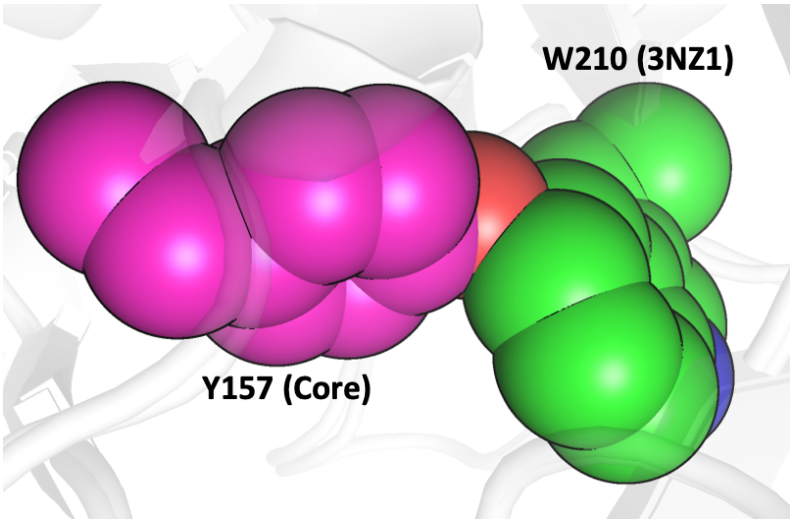


Figure 2.24: The Y157 core rotamer clashes with the required W210 rotamer necessary for proper π-bonding of the TS. An atom sphere model showing the tyrosine residue at position 157 from 1A53-core (magenta carbons) clashing with the tryptophan rotamer (green carbons) from 1A53-2 (holo). The backbone of 3NZ1 was superimposed onto 1A53-core's backbone using PyMOL's *align* feature, using residues 2-246 for both proteins. Cartoon backbone was made transparent for clarity. Hydrogens removed for clarity.

Furthermore, one can observe a similar issue with the sidechain steric issue from the predictions of the BR-only approach using PDB: 1A53 as input, where the W110 of the predicted design clashes with the Y157 of the core, likely due to the higher positioning of the backbone from the perspective of Figure 2.25, as well as the clash of W210 with the TSA where the reactive ligand would be located during the elimination. If the backbones were better aligned with the crystal structure's, it's likely that the rotamers would adopt the necessary conformations and not result in any clashes that would compromise the prediction. This is also seen in Figure 2.26 where the 1A53-2 structures were used as input for the BR-only ensemble, and a similar rotameric clash is

seen between Y157 and W210. Hence, BR was unable to model the flexibility needed to achieve the backbone of the target enzyme.

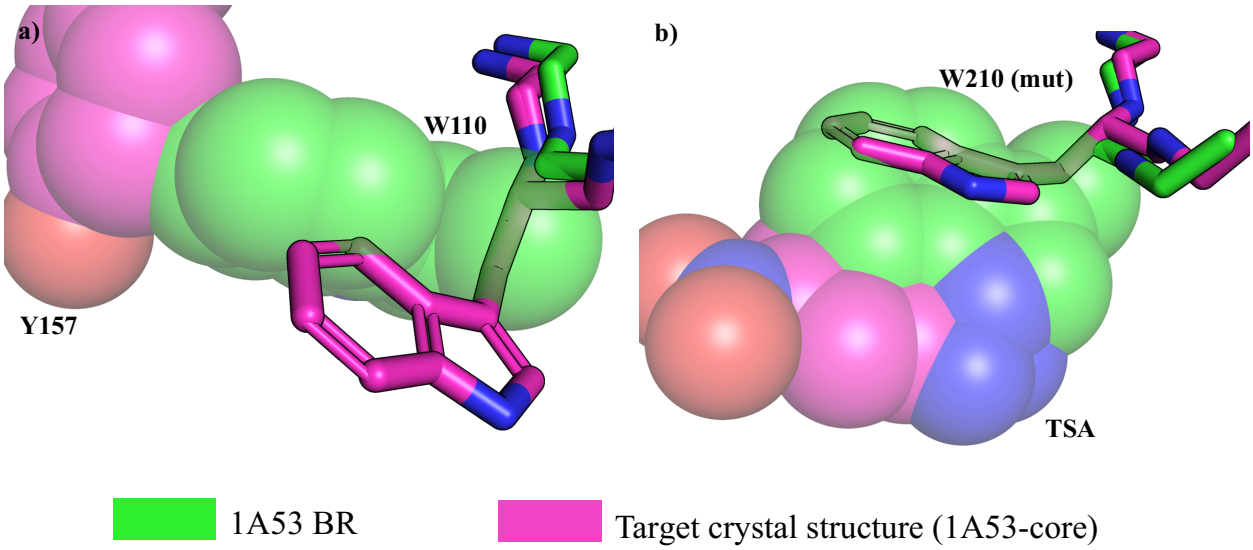


Figure 2.25: Sidechain rotamer clashes between a) W110 and Y157, as well as b) W210 and the TSA between predicted 1A53 BR structures and the original 1A53-core target. The (mut) indicates that the tryptophan was added using PyMOL's 2010 Dunbrack backbone-dependent library to demonstrate the impact of the backbone's deviation from the target.

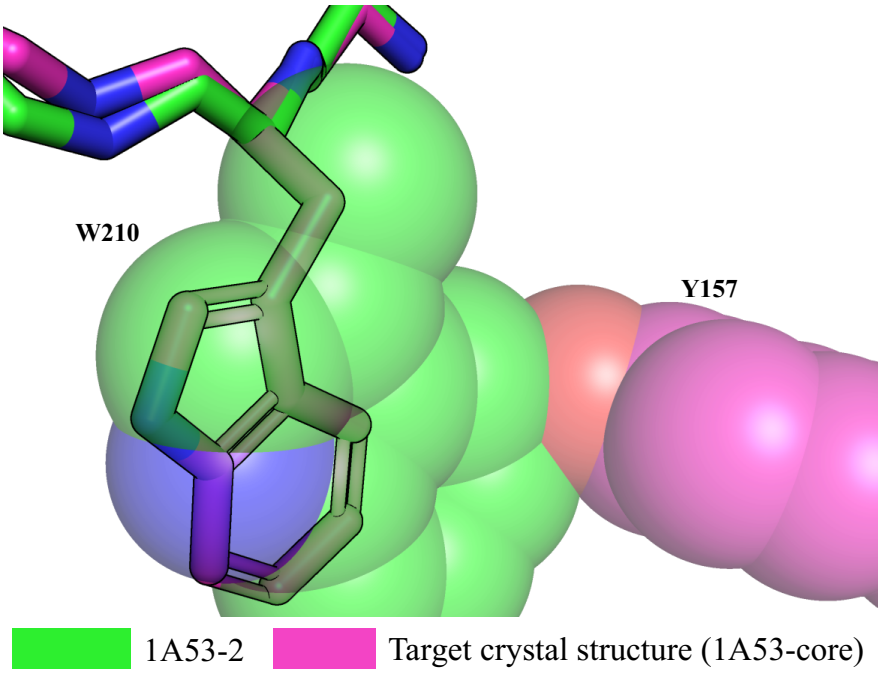


Figure 2.26: Sidechain rotamer clash between W210 from the 1A53-2 BR prediction and Y157 from the target crystal structure.

Nonetheless, an advantage of using backrubs for ensemble design coupled with existing methods is clearly visible – the final predicted energy of the active site is below zero and smaller by a few orders of magnitude relative to the ER-only pipeline, indicating structural and intramolecular energetic improvements in ER-BR's final design.

Finally, to further investigate the failure of ER-BR in recapitulating the active site crystal structure, the α-carbon distances from each of the theozyme residues 110, 178 and 210 were analyzed from 1A53, 3NYZ, 3NZ1, relative to the 1A53-core crystal structure. None of the 3 residues have expected active site residue Cα deviation distributions (Figure 2.13a to c), where one would predict 1A53 to have the lowest mean given the outputs, thus being able to perform theozyme placement much easier. All histograms (Figure 2.13) show 3NZ1 with the lowest mean distance, indicating that it should be the most optimal template group of the 3 PDBs for theozyme

placement, yet the software fails to output valid active site structures for this protein. This is also supported by Figure 2.22 which shows a similar trend for 3NYZ and 3NZ1, which might suggest they are more optimally suited to recapitulate the structure, but again, that is not seen.

Next, for the PM-BR pipeline, no final RP structures were generated after running the necessary design simulations. There can be a few reasons as to why this is the case. First, PM is capable of exploring the protein's energy landscape and travelling further into it, but it is impossible for the protein to exit out of an energy well if it already inhabits one. Hence, it is likely that the input's starting structure already existed within a well and "slid" further into it via PM, resulting in poor backbone orientation for the theozyme and subsequently all other residues that exist on 1A53-core. Nevertheless, the same theozyme α-carbon analysis was performed on the PM-BR backbones, to determine if the distributions could provide insight for the backbones' failure to generate adequate active site structures. Again, those histograms (Figure 2.14), including the histograms that analyzed the Cα-Cβ RMSDs (Figure 2.23) do not demonstrate why design outputs were not generated. Between ER-BR and PM-BR, one can see that the Cα deviation distributions are similarly centered when comparing the same residues. Furthermore, the shapes of the distribution are also alike, indicating that theozyme positions of the backbone are not likely to be the cause of the lack of data output, due to the similarity between distributions. Similarities between ER-BR and PM-BR Cα-Cβ RMSDs were seen for residues 110 and 178 as well, indicating similar results for both pipelines, especially since the distributions for the important catalytic glutamate (residue 178) were similar. Hence, it is unclear why PM-BR backbones were not amenable for structure generation, especially 1A53 inputs, which were successful during the ER-BR pipeline, and had no success for PM-BR. Therefore, designing with PM is valid for exploration

of a protein in the local minimum, but should not be used for the main ensemble generation in a protein design process, especially if the goal is to explore the energetic landscape outside of the local energy well. While the ER design had valid outputs, its energy was highly destabilizing, showing improvements with respect to energies (Figure 2.12) when coupled to another ensemble generating method like BR, indicating that the destabilization was likely caused by backbone steric issues, as the rotamer and ligand placements stayed relatively similar – there was substantial improvement seen in the energy from the ER-BR relative to the ER ensemble. Finally, one should consider that each ensemble (pre-BR) had a different number of templates. Hence, there could be fewer low-energy backbones explored by an ensemble containing fewer templates than ensembles with higher numbers of protein backbones. Hypothetically, if there are 2 ensembles, one with N templates, and another with N+1, assuming the N templates are identical in both ensembles, any sequences that will be tested within an N+1 ensemble will generally result in a lower weighted probability for a shared sequence structure pair relative to the N ensemble using a weighted Boltzmann approach. Hence, this could alter the types of sequence-structure designs chosen for final testing, yielding a potential false negative.

For the BR-only pipeline, only 2 of the 600 templates (0.33% hit rate) in the 3 ensembles were able to create a final structure, both of which were derived from 1A53 WT input (Figure 2.18). Final energy of the lowest-energy structure was positive, indicating poor prediction. Again, 3NYZ and 3NZ1 offered no final repacking outputs, indicating an issue with 1A53-2, likely the same one mentioned at the beginning of this section.

Overall, from what was discussed in this section, BR does not perform well for recapitulating the 1A53-core enzyme given the WT and 1A53-2 templates. This is unfortunate, given BR's success in designing new enzymes with modest activities.[35] Given the discovery found from Figure 2.24, this is likely due to the 1A53-core structure being difficult to work with, given its bulky tryptophan side chains within the active site. Hence, BR on its own, or even combining it with other EGMs, does not seem to improve enzyme predictions **in general**, and should be carefully considered when designing new enzymes, especially ones whose active sites are sterically hindered by large side chain rotamers like tryptophan. However, while BR may perform poorly for generating ensembles that will lead to good predictions of some enzymes (e.g. 1A53-core), it does quite well for others, like HG4.

2.5.2.2 HG4

From the results of HG4 recapitulation (Figure 2.19a-c), it is clear that the BR pipeline worked more successfully for this family of TIM-barrels compared to that of the 1A53-core group. To investigate, the α-carbons of the catalytically relevant residues D127 and Q50 were examined, and their deviations relative to the crystal structure of HG4 (Figure 2.3c). The distributions of the carbons (Figure 2.21) do not correlate to the number of hits found for each of the PDBs. Relatively speaking, one would expect 1GOR's distribution to be located in the middle of 5RG4 and 5RGA, with 5RGA located to the left (having the lowest mean deviation) and 5RG4 has the highest deviation due to it having the lowest number of hits. However, this is not what was seen for either residue 50 (Figure 2.21a) or 127 (Figure 2.21b). The 1GOR repacks had the lowest deviation for the aspartate, suggesting a relatively higher agreement between the crystal and designed backbones at around that location, but it had the highest deviation for the glutamate at position 50, suggesting

poorer backbone positioning. Just like in 1A53-core's pipeline, this could not be used as an accurate prediction of the number of hits (or whether a structure was generated in 1A53-core's case). In both situations, 5RG4 and 5RGA have relatively similar distributions. This was not expected, because although 5RGA is simply the holo variant of 5RG4, the difference in the backbone's adjustments to the ligand in both scenarios should have been enough of an RMSD difference to force the distributions apart.

It is also interesting to note that rather than the hit number steadily increasing from 1GOR to 5RGA, it starts relatively high at 19 for 1GOR, then goes to zero for 5RG4, and then to 29 for 5RGA (Figure 2.21). Since 5RG4 is closer to the HG4 protein than 1GOR, one would expect to see an increase going from 1GOR to 5RG4 (due to 5RG4 being evolutionarily closer to the HG4 target), but that is not what is seen. However, via the same analysis, the highest number of hits occurring for 5RGA relative to 1GOR and 5RG4 does indeed fit what is expected. An independent recapitulation was performed by I. S. Dementyev, the number of hits was similar (14, 1 and 27 for 1GOR, 5RG4, and 5RGA respectively), hence the original data for the hits (Figure 2.20) is provided as the representative results. The bar graphs showing the hits for the recollected data is shown in Figure 2.27.

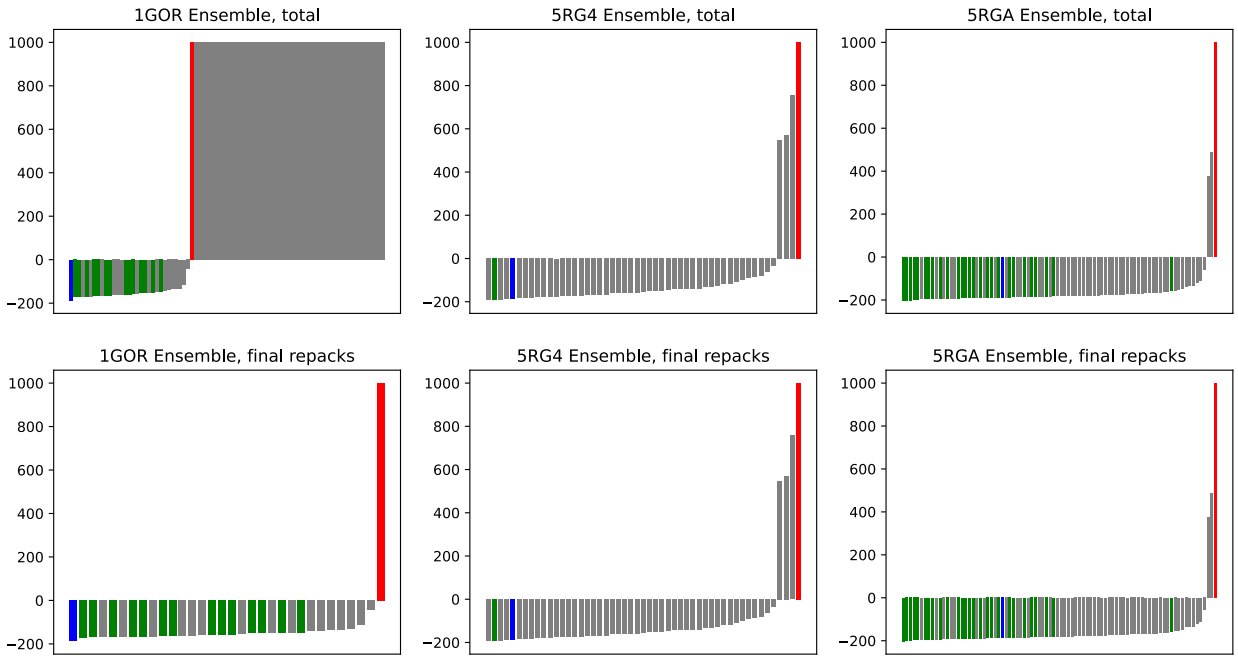


Figure 2.27: Bar graphs of final recollected repacking energies by template, organized by ascending energies. All colour schemes and hit definitions are the same as in Fig. 2.19. To recap, the total number of hits in this recollection are 14, 1 and 27 for 1GOR, 5RG4, 5RGA respectively.

Another important detail that should be discussed is the smaller percentage of correct rotamers in all 3 final repacked PDBs (Figure 2.3) relative to the results from 1A53-core (Figure 2.12, Figure 2.18). This is also clearly visible in the positive controls (Figure 2.3a, c), where the HG4 positive control has a 14.4% difference relative to 1A53-core and is also smaller than the 1A53-core enzyme. There can be a few reasons as to why this is. Firstly, the HG4 structure deals with many more residues whose sidechains have a larger number of degrees of freedom (e.g. more methionines as opposed to residues like valine). This renders simulated predictions difficult, as the number of viable rotamers increases, resulting in the software with the challenging task of predicting the exact same rotamer bin for the prediction as for the crystal structure. The increase in such sidechains would inevitably decrease the percentage accuracy, leading to a lower correct rotamer value. It is unlikely that the lower percent is caused by the absolute number of residues

designed, as in both 1A53-core and HG4, the numbers are equal to each other, at N=29 residue positions tested. Furthermore, the percentage calculation is inherently a normalization of sorts, so the percentage values should be close to each other, especially for similar or equal numbers of absolute values tested.

Lastly, it is important to mention that the number of hits could vary given the filtering criteria used. Since the 0.25 Å ligand RMSD value for the leeway (added onto the 0.44 Å positive control) as well as the 66% correct rotamer percentage were chosen rather arbitrarily, future studies would benefit from iteratively testing different ranges of such values to ensure that higher accuracy predictions are generated, without necessarily decreasing the number of predictions accepted for final design, and vice versa. Using a higher value in this work would obviously lead to an increased number of hits, but the number of correct rotamers would have potentially decreased, and the predicted energy could increase, leading to a reduced prediction quality due to overfocus on 1 variable of the design quality. A lower value would have led to improved accuracy from the total results, at the expense of decreasing the number of final predictions to analyze. A similar pattern would be seen from manipulating the correct rotamer percentage cutoff.

Throughout the HG4 recapitulations, it was clear that although BR could not recapitulate other Kemp Eliminases like 1A53-core (Section 2.5.2.1), it was capable of re-designing HG4 to a high degree, given the defined filtering criteria for prediction hits as mentioned previously. Hence, the success of BR for HG4 recapitulation was largely used as motivation for pursuing the diversification of sequence design for the creation of a new retro-aldolase, TyRA95, derived from the RA95 enzyme.[29]

Chapter 3 – Design of TyRA95 Sequences

3.1 Statement of Contribution

Boilerplate simulation patch file for theozyme placement and repacking were provided by Serena Hunt. All other work performed by Ilya S. Dementyev.

3.2 Introduction

In Chapter 2, the strength of backrub (BR) as an ensemble-generating method to improve structural prediction of de novo enzymes was explored. It was found that for some enzymes like 1A53-core, it showed no measurable improvements in the accuracy of design, yet for other enzymes like HG4, the structure of the high-efficiency enzyme was predicted from a BR ensemble of a wild-type (WT) protein (PDB: 1GOR). In this chapter, the use of BR for designing a transition state-stabilizing hydrogen bond is explored, along with using the same ensemble for increasing sequence diversity in the final design.

To recap, the design of high-efficiency enzymes with multiple transition states has improved considerably over the last two decades. One of the biggest successes is the retro-aldolase RA95, designed via computational enzyme design (CED) by Althoff et. al.[29] RA95 (PDB: 4A29) is a TIM-barrel-fold retro-aldolase capable of catalyzing the cleavage of (R)/(S)-methodol (4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone), via a multi-step pathway involving Schiff-base intermediate formation (Figure 3.1).[81] Crucially, the enzyme catalyzes a reversible carbon-carbon

bond at a rate increase of x15000 over background,[81] highlighting computational design's important role in enzyme design, and its potential to design meaningful biocatalysts. Several rounds of evolution led to RA95.5-8F, which featured the TS moving from position 210 to 83 (further into the barrel – Figure 1.12) and the TS is engaged in a hydrogen bonding network with several tyrosines (including Y51 and Y180). Apart from facilitating the interactions necessary for important secondary structure formations (e.g. α-helices), hydrogen bonds are known to stabilize TS formation[97] and improve preorganization,[98] leading to improved catalytic efficiencies in enzymes. Hence, creating a hydrogen bond to stabilize a TS which previously had no such interactions is important to consider for *de novo* design. Previously, hydrogen bonding was rationally designed using methods like HBNet,[99] and machine learning models like 'family-wide hallucination' to generate de novo luciferases by Yeh et. al.[100] In all cases, the hydrogen bonds were generated from single-input structures as opposed to an ensemble. In this work, the backrub is used to generate ensembles of RA95 for the explicit purpose of designing a new tyrosine-TS hydrogen bond interaction. The same ensemble is also then used to verify if the final designed sequence set is more diverse than a single-state design approach. It is important to increase diversity of a sequence set during design, as new sequences may offer enhanced activity that may not have been initially obvious.[101] However, since deleterious mutations always have the potential to be introduced during design, the sequence can be filtered or designed using other methods like RosettaRemodel.[102] As a result, the method described here focuses on merely improving the initial set of sequences that offer structural and intramolecular stability, which can then be iterated using a different method to improve function. All goals are summarized in the objectives.

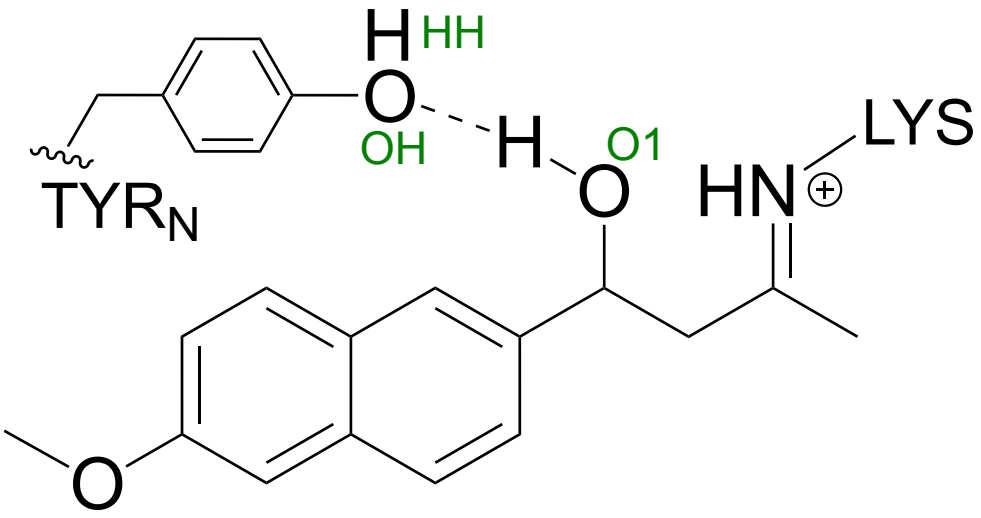


Figure 3.1: A hypothetical tyrosine mutation at position N allowing for a hydrogen bond-mediated stabilization of the Schiff base intermediate. Through the hydrogen-bond between the tyrosine's hydroxyl to the methodol's, the stability of the lysine-methodol complex should increase and improve rate of Schiff base formation. Names of important atoms involved in geometry definitions (Table 3.2) denoted in green, written beside respective atoms. For the tyrosine mutation, N=51 in RA95.5-8F and N=210 in RA95.0. Irrelevant hydrogens removed for clarity.

3.2.1 Objectives

In the work of this chapter, the goals are two-fold: to see if BR ensembles can generate a hydrogen bond-stabilized TS with a tyrosine, and to see if ensemble design via BR can generate more diverse sequences for the new TyRA95 than a single-state design approach of the same enzyme. Although further-optimized enzymes exist, such as RA95.5 (PDB: 4A2S) and RA95.5-8F (PDB: 5AN7), they were not chosen for design as they are derived from RA95.0 through *in vitro* means, including directed evolution in the latter's case. The motivation was to start from the computationally designed enzyme RA95.0 (PDB: 4A29) and rationally engineer potentially beneficial mutations *in silico*, using *in vitro* methods for characterization and verification. TSS and TSR refer to the (S)-TS and (R)-TS enantiomers of the transition state respectively.

Due to time constraints, the full design of a new TyRA95 (Tyrosine hydrogen-bonded Retro-Aldolase 95) that catalyzes the reaction for the TSR was unable to be fulfilled, with this work highlighting the efforts done for TSS instead. Furthermore, *in vitro* characterization of TyRA95 were not achieved given similar constraints. To start, the one must rationally decide which input to use, and where to place a tyrosine to stabilize the methodol transition state.

3.3 Methods

3.3.1 Rational Design of Tyrosine's Mutation Position

To determine viable positions of this mutation while mitigating possible deleterious effects, all residues containing at least one atom lying within an 9.5 Å sphere centered at the acceptor oxygen O1 (Figure 3.1) were determined using PyMOL (Figure 3.2a), using the structure of RA95. This value was chosen to account for the average hydrogen bond distance of 3.0 Å (2.7-3.3 Å, as described in literature),[103] in addition to the average distance between a tyrosine's side-chain oxygen to its β-carbon (6.5 Å), which considers the side-chain's length. Using this filter, 18 positions were identified (Figure 3.2b, Table 3.1). Some of these positions were removed from consideration due to their sidechains unproductively pointing away from the active site (e.g. G59). Prolines were also omitted from the list, as they are considered integral to backbone structure and tampering with the sequence at such positions could lead to unintended backbone disturbances. From preliminary visual analysis using PyMOL's mutagenesis wizard, in which Dunbrack's 2010 backbone-dependent rotamer library was utilized,[104] G212 was determined to have a high likelihood of an advantageous hydrogen bond formation (Figure 3.3a) as seen by visual inspection, along with L159 (Figure 3.3b), although L159Y was hypothesized to be less prioritized due to the

higher likelihood for clashes for similar sequences. The final list of design positions is listed in Table 3.1.

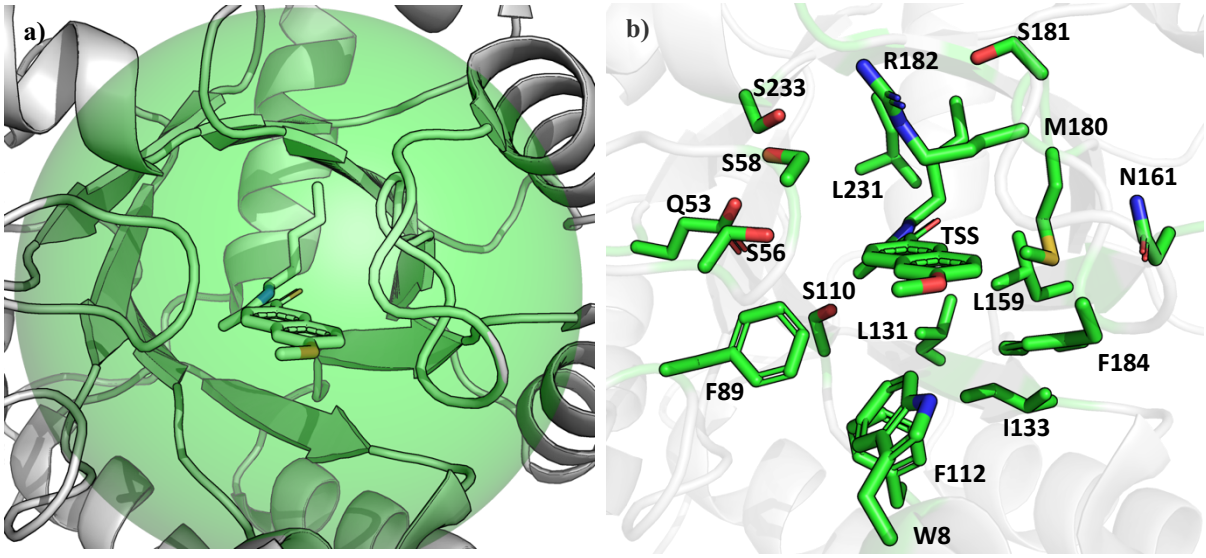


Figure 3.2: Rational design of mutation positions led to 18 design positions total. (a) Sphere centered at the methodol's oxygen O1 (Fig. 3.1) with a radius of 9.5 Å. All residues whose atoms lied within this sphere were considered, excluding unproductive side chains. (b) Final 18 residues chosen after filtering out unproductive residues, the ligand is bound to K210, creating TSS. Hydrogens removed for clarity. All structures shown were from PDB: 4A29.

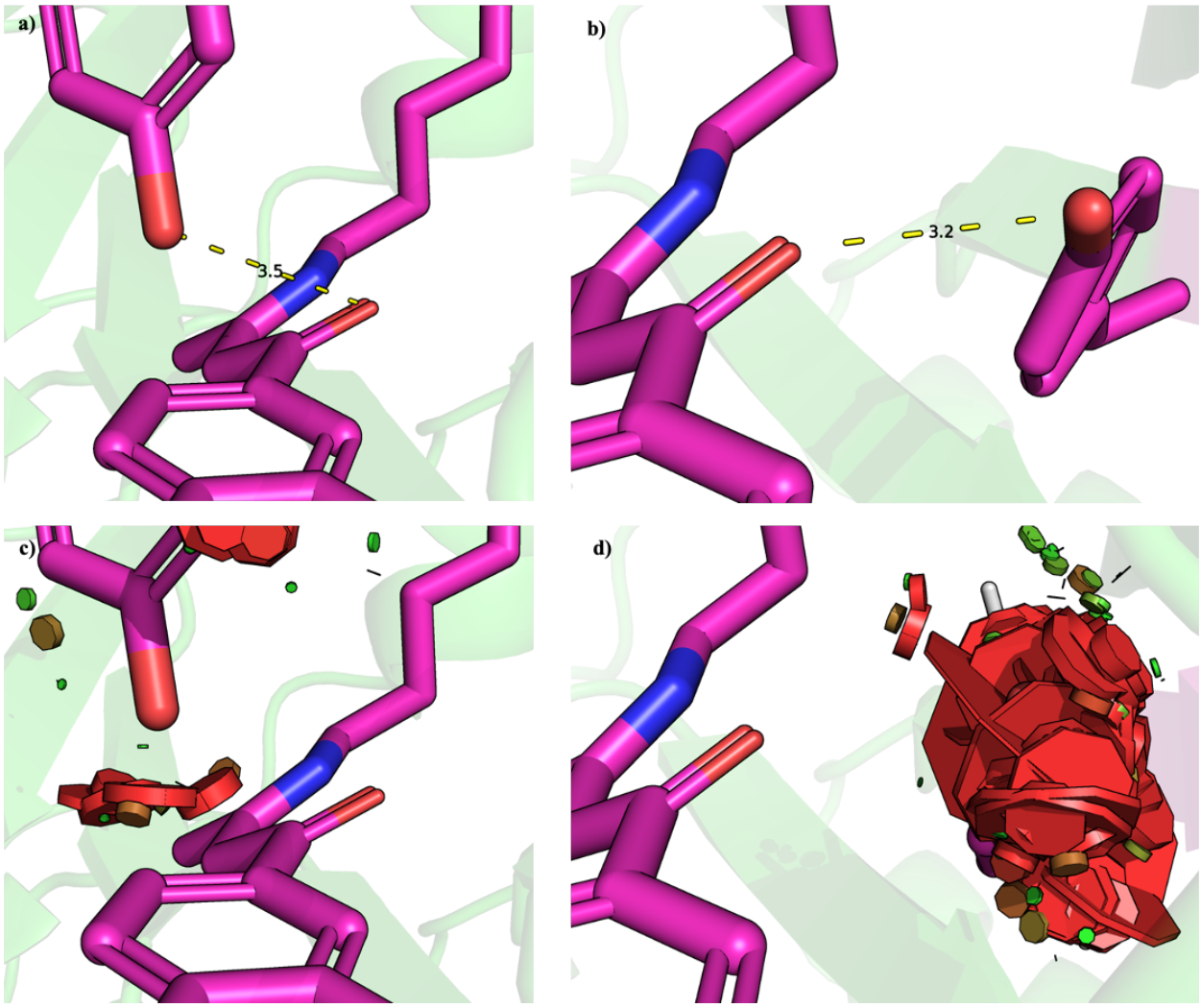


Figure 3.3: The G212Y mutation would be superior to L159Y for hydrogen bond formation to the ligand. Hydrogen bond formation of the (a) G212Y mutation and the (b) L159Y mutation after application of PyMOL's mutagenesis wizard, with corresponding strains shown for (c) G212Y and (d) L159Y. The G212Y hydrogen bond is slightly longer than what is usually seen in the literature (2.7-3.3 Å), whereas the L159Y variant lies neatly within the range. PyMOL's default backbone-dependent library was chosen for the mutagenesis.

Table 3.1: Tyrosine design mutations for the new enzyme, TyRA95. For any 1 tyrosine mutation, the theozyme is the TSS/TSR TS analogue (in bold) as well as the corresponding tyrosine position. The amino acid tabulated is the wild-type residue located at the respective position from the original RA95.0 enzyme (PDB: 4A29)

Residue ID	Amino Acid
8	W
53	E
56	S
58	S
89	F
110	S
112	F
131	L
133	I
159	L
161	N
180	M
181	S
182	R
184	F
210	**TSS/TSR**
212	G
231	L
233	S

3.3.2 Control Recapitulations of RA95

For positive and negative controls, the PDB: 4A29 and 1A53 backbones were chosen as starting inputs, respectively. The negative control was chosen due to its lack of similarity in structure, sequence and function to RA95-family enzymes. Recapitulation of the starting input was performed as a positive control, because it was reasoned that if the input could not be recapitulated, it would be futile to attempt designs using such a protein, especially if the ligand was poorly placed.

Before running any simulations, both PDBs were stripped of solvent molecules, ions, and any cofactors used during crystallization or WT ligands. No minimizations were performed. Afterwards, the lysine-methodol (S) and (R) transition states (TSS and TSR respectively) were grafted into position 210 using PyMOL scripts provided by S. Hunt. The 4 structures were then subjected to theozyme placement, hollowing and repacking, with design residues affected (Table 3.1) and all others remained unchanged. All design steps were performed with Triad, using the *phoenix* forcefield and rotamer packing options enabled for theozyme placement and repacking. For theozyme placement, 4 structures were generated from the starting input, and repacking used all 4 structures and generated 16 final structures. At the end of repacking, the final energy of the lowest-energy structure was subtracted from the hollow protein energy, where all design residues (Table 3.1) were mutated to glycine, including the TSS and TSR at position 210.

3.3.3 Backrub Ensemble Generation

The PDB: 4A29 was chosen as input. Before running Triad's backrub application, any solvents and counterions were removed. No pre-backrub minimization was performed. Then, 100 backrub structures were generated from the input file, using 2000 Monte-Carlo steps, a maximum angle of 5°, a maximum residue window length of 12, and a temperature of $k_B T = 0.6 \frac{\text{kcal}}{\text{mol}}$ was used for energetics calculations with the *phoenix* forcefield. Afterwards, all N=100 templates generated were subjected to minimizations using the *proteinProcess.py* module, using the covalent *phoenix* forcefield model for energy calculations. Following minimization, the lysine-methodol

(S) and (R) transition states (TSS and TSR respectively) were grafted into position 210 using PyMOL scripts provided by S. Hunt. Since there is only one ensemble generation step, none of the templates were discarded due to a low deviation, as was the case in the 2nd chapter.

3.3.4 Theozyme Placement (Motif Generation)

For the sake of completion, all 18 residues (Table 3.1) were tested *in silico* using Triad's PhoenixMatch algorithm[24] to determine the most probable location for the tyrosine, for a theoretical total of 18 × 100 × 4 = 7200 structures tested. In practice, only around 1 out of 4 valid structures were generated for each input, leading to around 1800 final structures for generating the theozyme. Briefly, a theozyme is a set of functional groups aligned in a certain geometry that are predicted to stabilize the transition state as per Pauling theory.[17] In this context, the theozyme is the tyrosine mutated at one of the 18 positions, as well as the (S)/(R)-TS ligand covalently linked to K210. The optimized geometry for the hydrogen bond was defined as shown in Table 3.2. After the placement was performed, the total number of output structures would be subject to a filter, which were the same criteria as the design requirements (Table 3.2). At this stage, neither the energy nor the SASA was checked as a criterion for filtering. All structures which passed the filter were selected for the follow-up sequence design.

Table 3.2: TyRA95 design bias values. The only bias present is "hbond", representing the hydrogen bond between a tyrosine in design position N (Y_N), and the TSS transition state with the covalently bound ligand. The donor-acceptor atom (T/OH and S/O1) distance is defined by value ranges commonly seen in the literature. A similar definition is applied to the angle between the oxygen on TSS and the hydrogen and oxygen on the tyrosine's hydroxyl groups.

Interaction name	Species interacting	Atoms interacting	Interaction Type	Ranges*
hbond	Y_N (T), TSS210 (S)	S/O1, T/OH	Distance	2.7, 3.3
		S/O1, T/HH, T/OH	Angle	100, 180

*Distances in Å, angles and dihedrals in degrees (°).

3.3.5 Sequence Design

Sequence design for TyRA95 are performed in a similar fashion to theozyme placement. As usual, the *phoenix* non-covalent forcefield is used for energy calculations, with a 0.90 van-der-Waals atom radius scaling factor. All 20 amino acids, including variants of histidine with different protonation locations (δ-nitrogen and ε-nitrogen, HID and HIE respectively), are explicitly set as design residues for all positions as listed in Table 3.1, excluding the tyrosine mutation and the TSS (position 210).

To compare the sequence diversity qualitatively, a sequence logo was generated using WebLogo (Berkeley, CA, USA)[105] from a list of final sequence designs for only the designed residue positions for each pipeline, with the height of each residue corresponding to its respective frequency. To compare the diversities of sequences generated by the design procedure, a quantitative measure describing the sequence variability is used: the Shannon entropy. Briefly, this entropy is a measure of amino acid diversity at a single position,[106] given by the following:

$$D_i = -\sum_{j=1}^{20} P(x_{i,j}) \log_2 P(x_{i,j}), i = 1, 2, \ldots, K \qquad \textbf{(Equation 1)}$$

Where $x_{i,j}$ represents a sequence with the i^{th} residue mutated to j (one of the 20 standard amino acids), and $P(x_{i,j})$ represents their respective frequency in the list.[107] The base-2 logarithm is chosen to maintain consistency with standard methods.[108] The equation suggests that the Shannon entropy of an entire sequence list (the average value for all D_i) is correlated to the diversity of the sequence list – the more diverse the sequences, the higher the total Shannon entropy of the list. Hence, one expects to see a higher total entropy for the ensemble design's final sequence list, as we expect to see a higher sequence diversity for ensemble-derived design sequences than for the single-state approach. A custom Python script was used for calculating all diversities.

3.4 Results

3.4.1 Control Recapitulations of RA95

For both TSR and TSS positive controls (Figure 3.4-Figure 3.5), the ligand RMSD is more than 2 angstroms higher than the usual ~0.5-1 Å seen for recapitulations in the previous chapter. There are likely a few reasons for this. Firstly, the TSS and TSR ligands do not have the exact same chemical structure as the ligand seen in the original 4A29 crystal structure. For instance, the carbon bonded to the O1 oxygen (Figure 3.1) in the 4A29 crystal is sp^2 hybridized, but it is sp^3 in the grafted ligand. Hence, the RMSD will increase as a result of the geometry difference. Furthermore, the degrees of freedom in this ligand are much higher than for the 5-nitrobenzisoxazole, as the latter does not contain a modest hydrocarbon chain like lysine's in the TS. Previous works have seen lower RMSD values, which was originally thought to result from considering all active site sidechains, but performing a similar analysis using the positive control show a higher RMSD at 4.18 Å (calculated using PyMOL's rms_cur), indicating a different source

of the lower value. Given this result, RMSD calculations were performed in a different manner relative to that work. Hence, it was decided to proceed with TyRA95 designs nonetheless. For the negative controls, the likeliest reason for the high rotamer disagreement is due to the backbone structural differences, especially for loops, such as the one containing residue S181, as well as the loop with F89. These backbone differences, along with rotamer-backbone and rotamer-rotamer clashes avoided by Triad, likely led to the low rotamer agreement. After the control structures were generated, it was crucial to perform a backrub on the input to generate the BR ensemble that would be necessary for further design for the TyRA95.

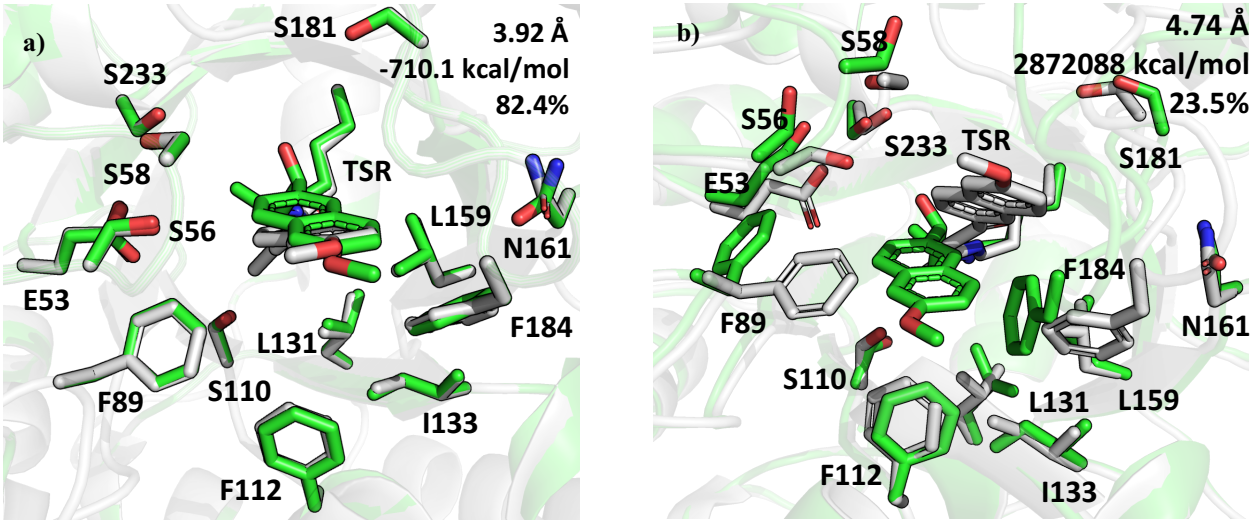


Figure 3.4: Positive and negative RA95 TSR controls performed using Triad. Final (a) positive and (b) negative control post-repacking structures of 4A29 with the TSR ligand incorporated. Positive control uses the RA95 backbone as input, whereas the negative control uses the inactive 1A53 WT protein backbone as input. Ligand RMSD (relative to crystal's ligand), final energy (relative to hollow structure) and percentage of design rotamers correctly predicted relative to the crystal structure shown in upper right-hand corner. Some residues omitted from the figure for sake of visual clarity (8, 180, 182, 212, 231), but all were kept for rotamer calculations (glycines and alanines excluded). Hydrogens removed for clarity.

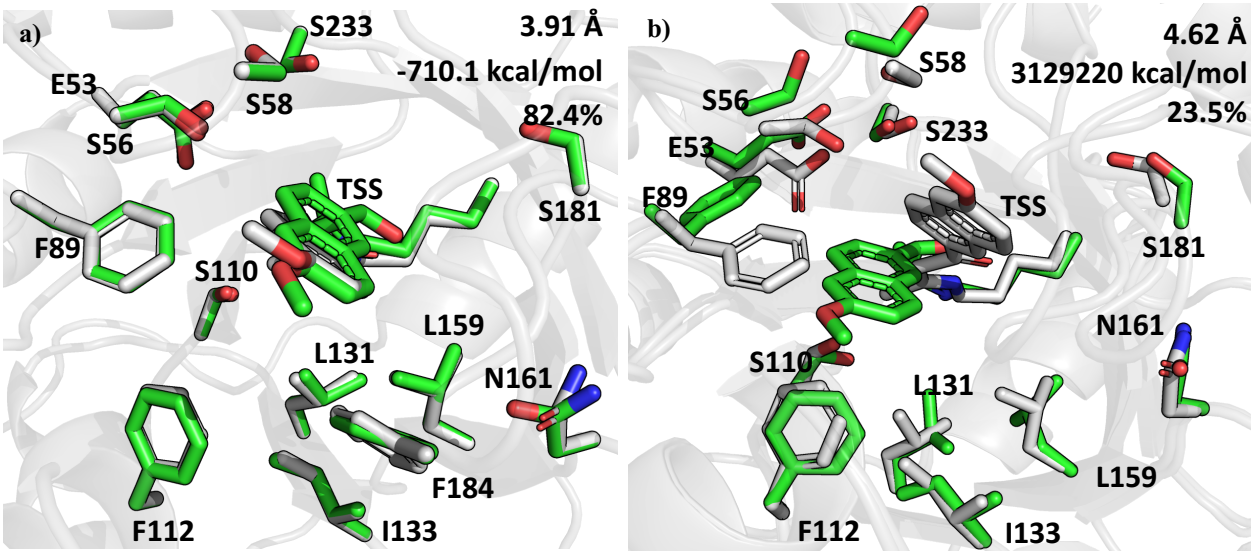


Figure 3.5: Positive and negative RA95 TSS controls performed using Triad. Final (a) positive and (b) negative control post-repacking structures of 4A29 with the TSS ligand incorporated. Positive control uses the RA95 backbone as input, whereas the negative control uses the inactive 1A53 WT protein backbone as input. Ligand RMSD (relative to crystal's ligand), final energy (relative to hollow structure) and percentage of design rotamers correctly predicted relative to the crystal structure shown in upper right-hand corner. Some residues omitted from the figure for sake of visual clarity (8, 180, 182, 212, 231, 184 in (b)), but all were kept for rotamer calculations (glycine and alanine excluded). Hydrogens removed for clarity.

3.4.2 RA95 BR Ensemble

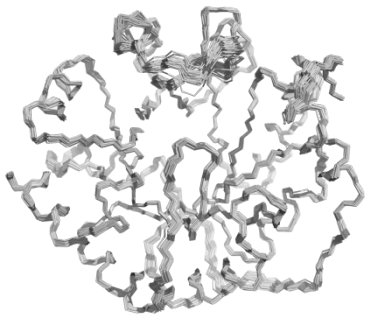


Figure 3.6: Triad-generated BR ensembles of PDB: 4A29 (N=100). The diversity and deviation are 0.32 ± 0.06 Å and 1.39 ± 0.01 Å respectively.

3.4.3 Final Sequence Design Using Ensembles

For all possible 1800 initial designs (18 Tyr design positions x 100 BR each) generated from backrub ensembles, only 31 final designs were able to be generated, with the best energy displayed in Figure 3.7 (respective sequence in Table 3.3 and Table 3.4). As one can see from Figure 3.7, BR ensembles were indeed successful in helping to design a tyrosine hydrogen bond to the TS ligand. The transition-state rotamer is pitched slightly downwards relative to the figure as opposed to the rotamer in the original crystal structure. The hydrogen bond is successfully created, evidenced by the donor-acceptor distance (3.0 Å) accentuated in Figure 3.7, as well as the donor-acceptor angle of 162.8°, both calculated via PyMOL. Energies are orders of magnitude lower than what was seen in 1A53-core's recapitulations. Due to the nature of sequence design, correct rotamer percentages are not shown, as this is a novel enzyme design, hence pre-existing structures for rotamer bin match analysis do not exist.

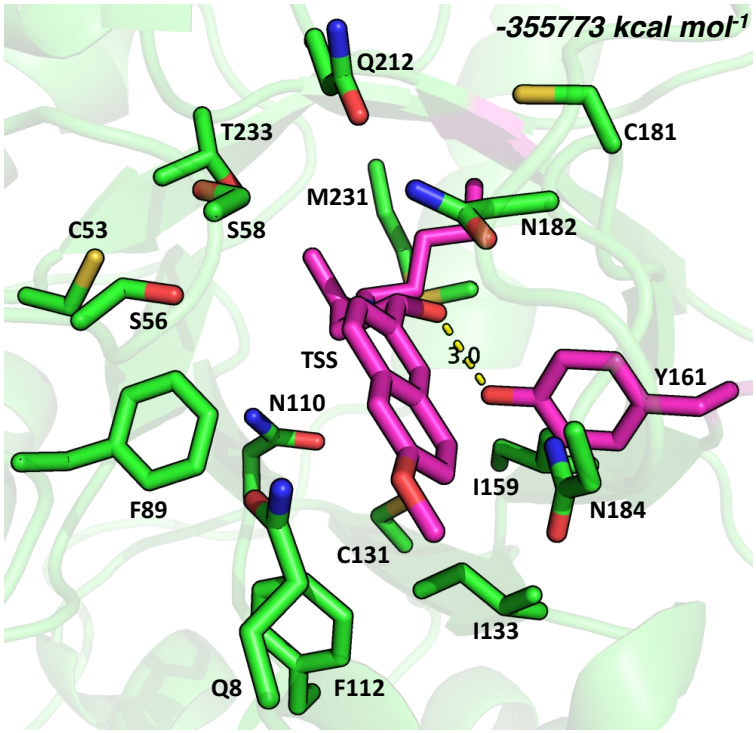


Figure 3.7: The best energetically preferred location for a tyrosine mutation stabilization is at residue 161 (formerly asparagine). The structure of the lowest-energy sequence design for a potentially higher-activity RA95, displaying a hydrogen bond between the designed tyrosine at residue 161, and the oxygen of interest on the ligand TSS. The tyrosine and methodol TSS are highlighted in magenta carbons for clarity, all other residues shown with green carbons. Hydrogens removed for clarity.

Table 3.3: Top 10 lowest-energy TyRA95 sequences using ensemble-based design. For brevity, only the amino acids engineered in the design positions are shown, then concatenated to form the sequence entry on each row. An equivalent table which allocates each amino acid to its respective position is shown below. The last sequence belongs to the original RA95.

Sequence	Energy, E (kcal/mol)
WCSSHNFCIWNMSCQYCS	-936.791
RCSSFNFLILYASTAQMT	-900.394
QCSSHNFCIINMSCWYQS	-850.874
RCSSHHFLINYGTNPQQS	-725.261
QCSSHNFCIITSSCWYMS	-722.922
QCSSHNFCIIYAQIAQMT	-722.634
ACSSFHFLIYNGSNHQQD	-717.007
QCSSHNFCIWYGNTPQCS	-716.932
QCSSHNFCIITSHCAYLS	-716.109
QCSSHNFCIISMSCWYMS	-710.385
WESSFSFLILNMSRFGLS	-710.1

Table 3.4: Top 10 sequence designs of TyRA95 using an ensemble-based approach. Sequences sorted top-down by increasing energy. For brevity, only design positions are shown.

Residue ID	8	53	56	58	89	110	112	131	133	159	161	180	181	182	184	212	231	233
Amino acid	W	C	S	S	H	N	F	C	I	W	N	M	S	C	Q	Y	C	S
	R	C	S	S	F	N	F	L	I	L	Y	A	S	T	A	Q	M	T
	Q	C	S	S	H	N	F	C	I	I	N	M	S	C	W	Y	Q	S
	R	C	S	S	H	H	F	L	I	N	Y	G	T	N	P	Q	Q	S
	Q	C	S	S	H	N	F	C	I	I	T	S	S	C	W	Y	M	S
	Q	C	S	S	H	N	F	C	I	I	Y	A	Q	I	A	Q	M	T
	A	C	S	S	F	H	F	L	I	Y	N	G	S	N	H	Q	Q	D
	Q	C	S	S	H	N	F	C	I	W	Y	G	N	T	P	Q	C	S
	Q	C	S	S	H	N	F	C	I	I	T	S	H	C	A	Y	L	S
	Q	C	S	S	H	N	F	C	I	I	S	M	S	C	W	Y	M	S

The diversity of the final sequence design is represented qualitatively in Figure 3.8, where three positions have 100% agreement with all designs – C53, F112 and I133. It is unclear if this

relates to the importance of the residue in the enzyme's function, or if this reflects the design's quality, with the software unable to appropriately design alternative residues in those positions. Not all of these residues are present in the original input, with C53 originally being E53. It is interesting to note that all residues that are conserved are of a hydrophobic nature and are found closer to the active site.

Figure 3.8: Total sequence design logo for all 31 TSS TyRA95 sequences. Sequence design logo featuring the design positions and original residue identity from the input crystal structure (written on the bottom) as well as the distribution of residues designed at each position, represented via its 1-letter amino acid code. Green, blue, and black colors correspond to neutral, hydrophilic and hydrophobic sidechains respectively. Height determined by frequency of appearance within the respective position in the residue list.

For Table 3.3, due to the work-around for calculating energies discussed in Section 3.3, the energies are abnormally low. However, the relative order is still correct, even if the energy values are not, since the error could be fixed by a simple addition knowing the actual energy of the hollow protein. Energy values notwithstanding, given the success of designing a new tyrosine

hydrogen-bonding interaction with the ligand, this was then compared with the sequence design of the single-state backbone design route.

3.4.4 Final Sequence Design Using Single-State Backbone

As one can see from Figure 3.9, using a single PDB structure of RA95.0 (PDB: 4A29) as input was insufficient in generating a hydrogen bond for the structure with the lowest-energy sequence. The tyrosine at position 89 does not have a valid hydrogen bond generated (Figure 3.9b), as the distance (5.8 Å) is too large relative to the literature upper limit value of 3.3 Å. Furthermore, the tyrosine is located at an unproductive position, one that could not possibly lead to a productive intramolecular interaction between the tyrosine's and ligand's oxygens. Finally, all 180 structures had a positive energy, and a lack of an appropriate hydrogen bond, contrary to what was seen with ensemble design of TyRA95. This indicates that the designed sequences are relatively unstable compared to the hollow energies calculated from the template, with all designs apart from TSS mutated to glycine. Potential reasons for this are explored further in the discussion (Section 3.5.3).

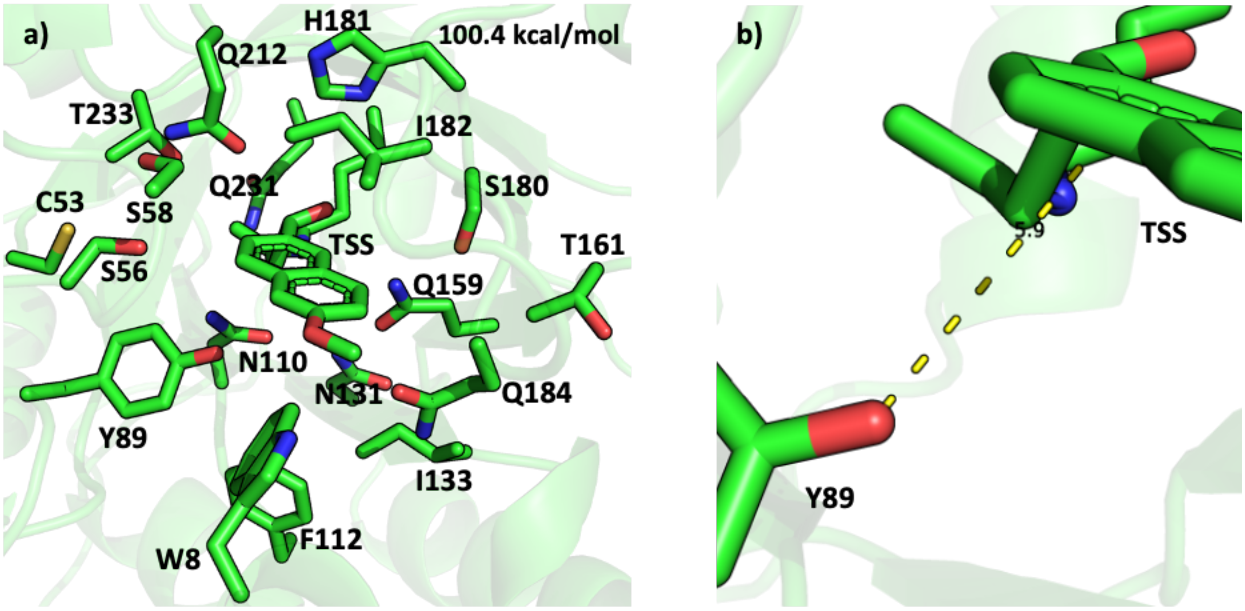


Figure 3.9: Single state design was unsuccessful in producing the proper hydrogen-bonding interaction between the tyrosine and the ligand. (a) Lowest-energy sequence structure of designed TyRA95.0 using single state design (with the original RA95.0 as input). For this sequence, tyrosine is mutated at position 89. As this is a novel sequence, correct rotamer and RMSD calculations were not performed as in Chapter 2. (b) The hydrogen bond designed for the lowest energy structure shown in (a).

To show the final energies of all designed sequence predictions, Table 3.5 was created that displayed the best 10 lowest-energy structures derived from the single-state design pipeline. The energies seen are more reasonable compared to the values seen in Table 3.3, even if the values are all above zero, indicating a poor stabilization. To compare the diversity between single-state and ensemble-design sequences, a Shannon entropy graph (akin to a quantitative equivalent of Figure 3.8 and Figure 3.10) was plotted for both.

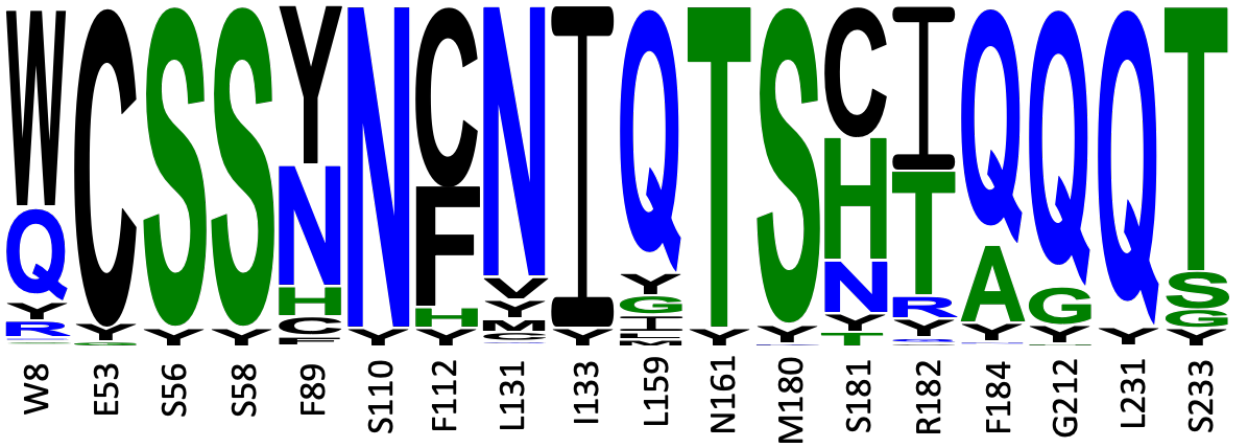


Figure 3.10: Total sequence design logo for all 180 TSS TyRA95 sequences from single-state design. Sequence design logo featuring the design positions and original residue identity from the input crystal structure (written on the bottom) as well as the distribution of residues designed at each position, represented via its 1-letter amino acid code. Green, blue, and black colors correspond to neutral, hydrophilic and hydrophobic sidechains respectively. Height of individual 1-letter codes was determined by frequency of appearance within the respective position in the residue list.

Table 3.5: Top 10 lowest-energy TyRA95 sequences using single-state-based design. For brevity, only the amino acids engineered in the design positions are shown, then concatenated to form the sequence entry in each row. An equivalent table which allocates each amino acid to its respective position is shown in the Appendix (Table 3.6). The last sequence belongs to the original RA95 (not shown in Table 3.6).

Sequence	Energy (kcal mol^{-1})
WCSSYNFNIQTSHIQQQT	100.432
WCSSYNFNIQTSCTQQQT	104.89
WCSSYNFNIQTSNTQQQT	105.348
WCSSYNFNIQTSCRQQQT	105.643
WCSSNNYNIQTSCTQQQT	105.785
WCSSYNCNIQTSCYQGQT	106.128
WCSSNNYNIQTSNTQQQT	106.366
WCSSYYCQIITSCRQQQT	106.994
QCSSYNFNIQTSCIAQQT	107.256
WCSSYNFNIQTSCTQQQS	107.633
WESSFSFLILNMSRFGLS	-710.1

Table 3.6: Top 10 sequence designs of TyRA95 using a single-backbone-based approach. Sequences sorted top-down by increasing energy. For brevity, only design positions are shown.

Residue ID	8	53	56	58	89	110	112	131	133	159	161	180	181	182	184	212	231	233
Amino acid	W	C	S	S	Y	N	F	N	I	Q	T	S	H	I	Q	Q	Q	T
	W	C	S	S	Y	N	F	N	I	Q	T	S	C	T	Q	Q	Q	T
	W	C	S	S	Y	N	F	N	I	Q	T	S	N	T	Q	Q	Q	T
	W	C	S	S	Y	N	F	N	I	Q	T	S	C	R	Q	Q	Q	T
	W	C	S	S	N	N	Y	N	I	Q	T	S	C	T	Q	Q	Q	T
	W	C	S	S	Y	N	C	N	I	Q	T	S	C	Y	Q	G	Q	T
	W	C	S	S	N	N	Y	N	I	Q	T	S	N	T	Q	Q	Q	T
	W	C	S	S	Y	Y	C	Q	I	I	T	S	C	R	Q	Q	Q	T
	Q	C	S	S	Y	N	F	N	I	Q	T	S	C	I	A	Q	Q	T
	W	C	S	S	Y	N	F	N	I	Q	T	S	C	T	Q	Q	Q	S

3.4.5 Sequence Diversity Comparisons

As one can see from Figure 3.11, the overall diversity of sequences designed from an ensemble of templates is greater than the diversity of sequences designed with a single template, verifying the hypothesis made earlier in Section 1.6, although both sets can be considered relatively conserved as their $D_i \leq 2.0$ in both cases.[109] Here, Figure 3.11 (b) and (a) can be thought of as the quantitative equivalent to Figure 3.8 and Figure 3.10 respectively, describing the diversity in a quantitative manner. One can see that although there are individual positions that have a larger diversity for the single-state design (e.g. F112, Fig. 3.11), the overall diversity of the sequence set is lower than the design which used ensembles. Furthermore, the maximum residue entropy of an individual amino acid is higher for the ensemble state design than single state, as seen for F184. Implications for such results are discussed further in Section 3.5.2.

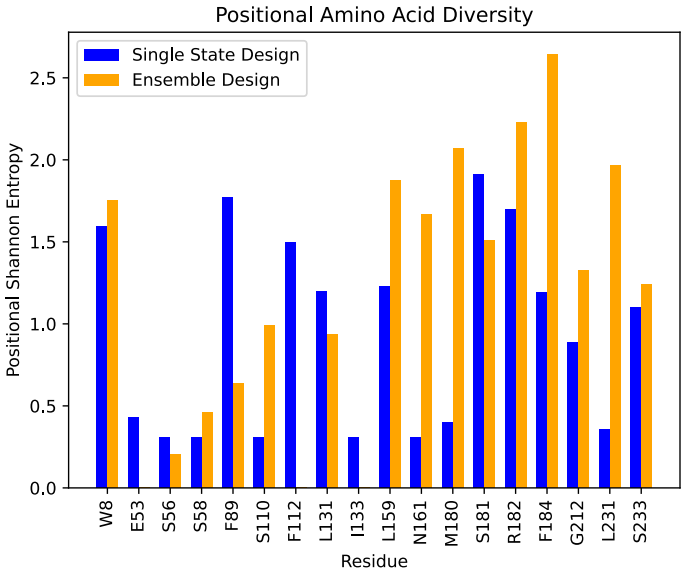


Figure 3.11: The absolute Shannon entropy for the ensemble state design is larger than the single-state backbone design approach. Positional amino acid diversities (entropies) for sequences designed from single state and ensemble backbone design approaches. For the single state, the total entropy was 0.93, whereas the entropy for ensemble state design was 1.20. The original 4A29 WT sequence was not included in the calculations.

3.5 Discussion

3.5.1 Positive and Negative Control Recapitulations

Although the positive control's ligand RMSD (Figure 3.4a and Figure 3.5a) is higher than one would hope, Triad managed to predict the general placement for both the TSS and TSR ligand enantiomers using the 4A29 as a backbone in the positive controls (Figure 3.4a and Figure 3.5a), especially when compared to the negative control ligand predictions (Figure 3.4b and Figure 3.5b) where the structural disagreement is obvious, especially for rotamers like residue F89. This suggests Triad can closely recapitulate rotamers (both residues and TS) given a backbone, indicating that it be used for the development of new RA95's with more complex ligand-residue

interactions, such as a hydrogen bond, dubbed TyRA95, which both had an ensemble-based and single-state-based design approach discussed in detail below.

3.5.2 Final Sequence Design Using Ensembles

As predicted, the BR ensemble pipeline is able to not only generate a proper hydrogen bond between a tyrosine and the TS ligand, but also generate a higher diversity in the final sequence set compared to single-state design. Also, the G212Y mutation has one of the best energies for the top 10 TyRA95 sequence designs, as predicted (Figure 3.3). Yet, the N161Y mutation having the lowest energy out of all possible sequences was not predicted to be the case and warrants further discussion. The likeliest reason for this outcome is the stereochemistry of the TSS ligand. Note that the (S)-methodol's oxygen is pointed away from position 212, and towards 161 (Figure 3.5a and Figure 3.6). Such an analysis could not have been performed directly from the 4A29 crystal input, as it did not contain the proper covalently bound TS that was later grafted onto position 210. In fact, it is likely that G212Y would be the more optimal position for a tyrosine-ligand hydrogen bond interaction to occur for the (R)-methodol isomer, which suggests the possibility of rationally designing a stereoselective version of TyRA95.

Delving deeper into the sequences designed, one can see the rather large number of polar sidechains in the lowest-energy sequence, relative to the total number of design positions. The location of these polar sidechains is closer to the opening of the TIM-barrel "pore", and further from the active site center than the non-polar sidechains, which is expected as polar sidechains usually exist on the solvent-exposed area of the enzyme, and more non-polar sidechains lie within the active site to improve stability and catalytic activity. Furthermore, some of the designed

residues appear to contain hydrogen bonding bridges between their sidechains, like the sidechain oxygen of Q212 and the sidechain nitrogen of N182.

As suggested by the sequence design logo (Figure 3.7), there are some residues that should be conserved when designing sequences for TSS catalysis. For instance, C53, F112 and I133 should likely be used when engineering new retro-aldolases with a tyrosine mutation to stabilize the ligand. Comparing this with the original sequence of RA95 (Table 3.1), we see that E53, F112 and I133 are the original residues in those respective positions. This suggests that E53, F112 and I133 might play an important role in the retro-aldol reaction, perhaps not directly as catalytic residues but by introducing stabilizing effects. In fact, past studies have shown that E53 was originally designed to facilitate water-mediated proton abstraction, yet mutation to alanine showed no measurable difference in activity.[81] Furthermore, the software could have designed the cysteine due to the glutamate's longer sidechain affecting the sterics of the loop on which it is located – the cysteine mutation reduces the potential hindrance that might otherwise be caused by the relatively longer glutamate sidechain, causing the energy to decrease and the sequence to be preferred (Figure 3.9a). The E53C mutation could suggest that a more hydrophobic residue is preferred at the location as it is nearer to the active site than other sidechains such as N184. These residues are also likely to be involved in increasing substrate specificity and active site preorganization, which has been shown before to increase enzyme reactivity when such factors are optimized.[75,110]

In retrospect, one aspect that should have been more considered was the Solvent-Accessible Surface Area (SASA).[111] Briefly, the SASA is a metric for determining the packing of a ligand by the active site sidechains – the larger the packing, the lower access that water would

theoretically have to the ligand, thereby increasing the catalytic efficiency by preventing unproductive hydrogen bonding (or reactions) with water and decreasing the SASA value. This is also a useful metric that can substitute certain recapitulation metrics during an actual design, as information like correct rotamer percentage and ligand RMSD may not always be present. Although this project could have used RA95.5-8F as a design target, the catalytic tetrad in that enzyme significantly differs from the active site structure of TyRA95 (Section 3.4.3-3.4.4), including the position of the transition state ligand, and thus it would have been inappropriate to consider as a design target. Unfortunately, the exclusion of SASA was an oversight that was not considered during the initial design pipeline (though ultimately, the goal of the hydrogen bond and increased sequence diversity were reached), but can easily be done so for future work by running a similar positive and negative control calculation as in Section 3.3.2, then choosing a range that allows for slight deviation from the positive control to ensure that false negatives are minimized.

Overall, in the case of designing the TSS TyRA95, it seems that BR ensembles are indeed capable of being used to design a tyrosine-TS ligand hydrogen-bond interaction, and to increase the diversity of sequences designed, at least *in silico*. Of course, it would be imperative to not only consider the lowest-energy sequences, but also the sequences that appear the most frequent as per the sequence logo, assuming that the total number of designed sequences is large and difficult to study in even high-throughput scenarios, and to test such sequences within an *in vitro* setting. However, in the current context, it would be best to study all 31 sequences as such a task would not be difficult to undertake, nor would it require a significant investment of time and resources. Regardless, the BR-generated ensemble design approach offered more sequence diversity in the final design predictions, as compared to the single-state pipeline discussed below. Future studies

would benefit from comparing BR-generated ensembles to ensembles generated using other methods like ER or MD, or perhaps serializing 2 or more methods together to create the ensemble, though this must be done with caution due to the recapitulation issues mentioned in Section 2.4.2.1.

3.5.3 Consequences for Using a Single-State Approach

As seen in Section 3.4.4, single-state design failed to produce any negative energy designs, with the best structure reaching 100.4 kcal/mol. Although it was expected that single-state design would have a smaller number of predicted viable structures, due to a lack of geometric space exploration by the backbone, it was not foreseen that this number would be nil. Initially, this was assumed to be an issue in the Triad software itself. However, it is usually the case that if a defined geometry is unable to be satisfied, the software would not output a final structure, as seen with the ER-BR structures in Section 2.4.3, or the negative control attempts in Section 2.4.1. Even when outputted structures do not meet defined criteria, a portion of the structures still maintain such defined geometries, as seen in Section 3.4.3. Hence, further considering that all but one variable (the number of input templates) was controlled when comparing the ensemble- and single-state design, it is unlikely that this is a fault within the coding of Triad.

The next potential source was determined to be the input file itself, the 4A29 structure. Due to the single-state nature of the experiment, only one backbone structure was used for theozyme placement and repacking. Since flexibility was unable to be modelled (apart from rotamer motion), this had multiple downstream effects: for instance, a potential rotamer clash was identified by the software during design (either with another residue or the backbone), which could

have been potentially avoided by introducing flexibility into the backbone through ensemble design, as was done with BR.

Finally, one could argue that the issue could have resulted from the low number of structures generated. As 180 structures were generated from the single state design (as opposed to the 1800 from ensemble design), it could have been the case that the sheer number of templates from the ensembles in Section 3.4.3 is what led to viable final designs compared to the single-state design which had 10-fold less structures, and the 10-fold reduction in structures used for single-state design led to zero final repacking structures with proper hydrogen bonding, through elimination of structures that could accommodate the necessary hydrogen bonding. However, this hypothesis is flawed, as no matter what sequences are designed (and which rotamers are therefore engineered), this would not change the designed hydrogen bond length from the theozyme to the repacking, and the only way to do so would be to change the structure of the backbone, which was indeed done in Section 3.4.3, and results with excellent hydrogen bonding distances that fit within the original range were seen.

For future designs of a TyRA95-like enzyme, one would need to focus on several aspects of the design pipeline not seen here. Firstly, using an ensemble-design approach will likely lead to better results as seen in Section 3.5.2. Also, rather than utilizing the average hydrogen-bond values seen in the literature, one may simply choose to use the hydrogen bond ranges seen in the designed RA95.5-8F, as those interactions are known to stabilize the transition state and increase pre-organization, or one may simply investigate different donor-acceptor distances for all designed retro-aldolases, and extract a mean with ranges from those values. Furthermore, the

rotamer library may not be large enough to recapitulate or design the necessary transition state rotamer, as well as the tyrosines necessary for hydrogen bonding to the specific oxygen. As seen in Figure 3.4a and Figure 3.5a, the rotamer of the transition state is decently predicted, but there is always room for more improvement. Furthermore, the results in Figure 3.9b show that the number of design positions could be decreased, as there are positions (e.g. 56) that simply would not be able to generate a good hydrogen bond to the ligand's oxygen, even if a continuous rotamer library was used. A different rotamer library (perhaps even one that is continuous[112]) used could improve design quality, one that is both recent and contains more options than the Dunbrack set used in this chapter.

Overall, the generation of a more diverse set of sequences via BR-generated ensemble design as compared to single-state design was successful, as measured by entropies of 1.20 and 0.93 respectively (Figure 3.11). This indicates that while BR may not necessarily be the best ensemble design method developed in protein engineering (as evidenced by Section 2.4.2.1), it is important nonetheless (Section 2.4.2.2, 3.5.2) and could lead to improved designs of future enzymes. Implications on the future of enzyme engineering as a result of this work is discussed further in Section 4, as well as potentially new metrics to incorporate into the design pipeline to improve future predictions.

Chapter 4 – Conclusion

4.1 1A53-core and HG4 Recapitulations

In general, *de novo* enzyme design has become more popular with the boom in biotechnology research, more work remains for humanity to engineer a universal pipeline capable of generating an enzyme for a specific catalytic reaction, when only the reactive ligand and catalytic sidechains are provided as input. In this work, the topic of improving enzyme design using backrub, and BR-conjoined methods, was discussed. Through rigorous recapitulation studies of the evolved Kemp Eliminase 1A53-core using both XX-BR (serialized BR) and BR methods, it was shown that neither pipeline type was able to perform an accurate redesign of important active site residues and ligands. As mentioned previously, this could be an issue with BR itself, but it is more likely that the lack of successful recapitulations of 1A53-core could have also been caused by some inherent property of that enzyme, such as the orientation of the β-strand α-carbons near or at the active site, considering most ensembles generated even without BR did not result in a successful recapitulation either (Section 2.4.2.1). The potential causes of these effects were studied to determine a correlation between backbone deviation and repacking output quality, but as seen in the results and discussion section, none could be found, apart from the α-carbon deviation for Y157. In future works, more enzymes should be studied regarding the method conjoining, including other designed Kemp Eliminases such as KE07, KE59 and KE70. Of course, this is not an exhaustive list, and more could be added depending on resource availability.

Nonetheless, when the input template was switched to 1GOR, and the HG4 enzyme was chosen as the recapitulation target (Section 2.4.4), a larger number of viable hits were seen. This is significant, as it is (to the author's knowledge) one of the first instances of recreating a highly

efficient (>100,000 M^{-1} s^{-1}) enzyme using an inactive template, with only a few backbone transformations. This suggests that it is possible to generate a pipeline that can recreate such a scenario, but it would undoubtedly require a collection of EGMs (given that BR-generated templates successfully recapitulated HG4 but not 1A53-core), and perhaps an overarching ML method that can determine which ensemble generating method should be given priority given the desired outcome.

An explanation for HG4's successful recapitulation was attempted regarding the α-carbon deviations of the active site relative to the target, but such an exploration was not fruitful, since higher deviations of the ensemble template active site carbons relative to the crystal structure did not necessarily relate to a lower number of hits, as one might expect. In the future, it would be wise to consider β-carbons as well as α-carbons when attempting to explain the lack of correlation between predicted active site backbone deviations from the crystal structure and the correct number of hits predicted. This approach could shed more light on the relationship between prediction accuracy and backbone structure, by calculating the RMSD for the α and β carbons, without having to calculate more complex linear algebra metrics such as dot products or projections to determine deviation.

Given HG4's results, it is clear that ensembles have an improved approach to predicting high-efficiency de novo enzymes, and that future attempts regarding de novo enzyme design require an ensemble approach to improve their accuracy, although it is unclear what is the objectively best ensemble design method for this approach would be.

4.2 TyRA95 Designs

After the BR was performed, although the 1A53-core did not yield fruitful results, the success with HG4 recapitulation led us to believe that the design of a new, previously untested retro-aldolase was possible. Hence, we decided to investigate the *in silico* rational design of a hydrogen bond-stabilizing tyrosine mutation in the RA95.0 active site, leading to the enzyme dubbed "TyRA95.0". The selection of the amino acid for the mutation, as well as its position, was rationally developed and discussed. In the end, 31 final structures were determined for potential screening, from the ensemble-state design. Unfortunately, due to the time constraints presented by the long runtime of the sequence designs, the *in vitro* verification of the enzyme's folding and kinetics was unable to be achieved. Nonetheless, it was shown that the ensemble-state design was more beneficial than the single-state approach, shown in Sections 3.4, 3.5.2 and 3.5.3. Hence, the BR ensemble can successfully determine new positions beneficial for the tyrosine mutation to ensure a hydrogen bond to the TSS ligand.

In future works, *in silico* designs would benefit from starting with a structure whose ligand is buried further in the active site as seen with RA95.5-8F, in the hopes of increasing the likelihood of higher catalytic efficiency in the final TyRA95 design. Furthermore, follow up work would include the wet-lab culturing and enzyme extractions would be necessary to confirm if the design sees an improvement relative to the original RA95 that contained modest activity. Further design for the TSR analogue is also heavily encouraged, in order to determine potential residue matches between TSR and TSS analogs of TyRA95, allowing for a better prediction of stabilizing mutants, followed-up by similar *in vitro* verification.

4.3 Proposition of New Metrics for Improved Enzyme Design

Although methods and benchmarks like diversity and deviation have their place in enzyme design, with the development of more complex enzymes via multi-state design,[41] there calls for more powerful metrics capable of analyzing a larger superset of useful data from the same protein template ensemble. One proposed metric is the Wasserstein-based Statistical Tool to Compare Conformational Ensembles of Intrinsically Disordered Proteins (WASCO),[113] developed by CNRS. Briefly, the WASCO allows to define a conformational ensemble of proteins as a set of probabilities, allowing to compare 2 ensembles with each other to determine local and global differences between each other (Figure 4.1). Such a comparison is useful to determine which protein regions have more significant structural changes on a per-residue and all-residue basis, all in one workflow, allowing for deeper insight into which regions are more dynamic than others. This has already been applied to the study of intrinsically disordered proteins but can likely be just as applicable for the development of *de novo* enzymes with more coherent tertiary structures.

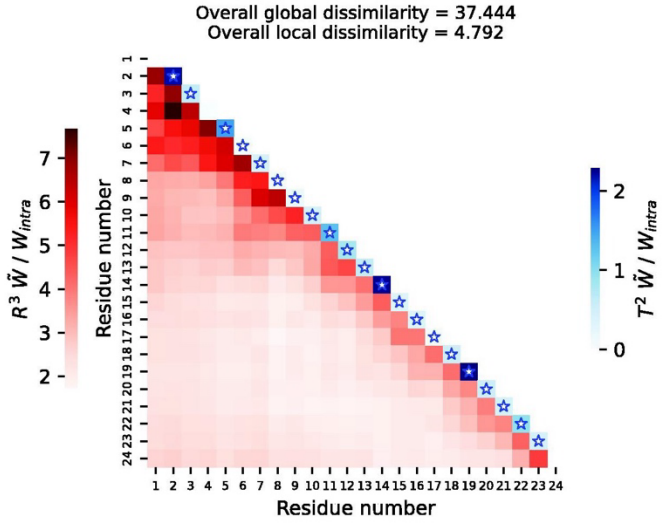


Figure 4.1: Wasserstein matrix heat map for inter-ensemble residue differences generated via MD simulations of Hst5. For non-diagonal values, the depth of the red color indicates how large the relative difference is of the inter-ensemble residues and its uncertainty. The significance of the values increasing closer to the diagonal indicate that globally (considering every ensemble), the most statistically relevant distances occur between residues close to each other in the primary structure. Lying on the matrix's diagonal, the values denoted in blue correspond to local differences (per ensemble), where stars also indicate that the difference is significantly ($p < 0.05$) greatly than zero. Figure from González-Delgado et. al.[113]

A similar metric for analyzing inter-residue distances was developed by Lazar et. al.,[114] who developed a composite heatmap-based approach to analyzing statistically significant distances between residues, to determine which regions have local or global similarity. One of the main benefits of such a method is that no superposition of protein structures is required to perform such RMSD calculations – this is especially useful in the study of tau proteins, which do not have a meaningful 3D structure where one may easily superimpose 2 different tau proteins together. While this method is similar to the previous, it provides some advantages, such as using medians rather than means to avoid skewing of averages by outlier values. It also utilizes the radius of gyration during its analysis, something that WASCO fails to accomplish. However, it is important

to note that WASCO provides a simpler data visualization approach than Lazar's, which is a caveat to consider for effective data communication in scientific papers. Unfortunately, the original PDB ID corresponding to the data used by Lazar and colleagues was unable to be retrieved or found within the original paper, hence a matrix is not provided here.

Overall, using the aforementioned bioinformatical tools will allow protein engineers to design improved ensembles by not only focusing on global diversity and deviation, but local versions of such metrics as well. Although certain ensemble generation tools like BR are not a one-size-fits-all solution as evidenced by this work, it is nevertheless beneficial and in most cases necessary to utilize an ensemble design approach for the *de novo* creation of improved enzymes for important biotechnological applications.

References

(1) Smejkal, G. B.; Kakumanu, S. Enzymes and Their Turnover Numbers. *Expert Rev. Proteomics* **2019**, *16* (7), 543–544. https://doi.org/10.1080/14789450.2019.1630275.

(2) Heckmann, C. M.; Paradisi, F. Looking Back: A Short History of the Discovery of Enzymes and How They Became Powerful Chemical Tools. *ChemCatChem* **2020**, *12* (24), 6082–6102. https://doi.org/10.1002/cctc.202001107.

(3) Wang, W.; Gao, J.; Li, Y.; Wang, S.; Shi, C.; Dai, X.; Jiang, T.; Sun, G. Thermal Hazard Assessment of Cyclohexane-Type Liquid Crystal Monomer from the Grignard Cross-Coupling Reaction by Different Calorimeters. *Thermochim. Acta* **2022**, *714*, 179225. https://doi.org/10.1016/j.tca.2022.179225.

(4) Zhao, C.; Xue, L.; Zhou, Y.; Zhang, Y.; Huang, K. A Microwave Atmospheric Plasma Strategy for Fast and Efficient Degradation of Aqueous *p*-Nitrophenol. *J. Hazard. Mater.* **2021**, *409*, 124473. https://doi.org/10.1016/j.jhazmat.2020.124473.

(5) Sinha, A.; Sahu, S. K.; Biswas, S.; Mandal, M.; Mandal, V.; Ghorai, T. K. Catalytic Use toward the Redox Reaction of Toxic Industrial Wastes in Innocuous Aqueous Medium and Antibacterial Activity of Novel CuxAgxZn1–2xO Nanocomposites. *ACS Omega* **2021**, *6* (44), 29629–29640. https://doi.org/10.1021/acsomega.1c03925.

(6) Avinash, V. S.; Pundle, A. V.; Ramasamy, S.; Suresh, C. G. Penicillin Acylases Revisited: Importance beyond Their Industrial Utility. *Crit. Rev. Biotechnol.* **2016**, *36* (2), 303–316. https://doi.org/10.3109/07388551.2014.960359.

(7) Stone, C. A.; Spiller, B. W.; Smith, S. A. Engineering Therapeutic Monoclonal Antibodies. *J. Allergy Clin. Immunol.* **2024**, *153* (3), 539–548. https://doi.org/10.1016/j.jaci.2023.11.018.

(8) Domingo-Espín, J.; Unzueta, U.; Saccardo, P.; Rodríguez-Carmona, E.; Corchero, J. L.; Vázquez, E.; Ferrer-Miralles, N. Engineered Biological Entities for Drug Delivery and Gene Therapy: Protein Nanoparticles. In *Progress in Molecular Biology and Translational Science*; Villaverde, A., Ed.; Nanoparticles in Translational Science and Medicine; Academic Press, 2011; Vol. 104, pp 247–298. https://doi.org/10.1016/B978-0-12-416020-0.00006-1.

(9) Oroz-Guinea, I.; Zorn, K.; Brundiek, H. Chapter 2 - Protein Engineering of Enzymes Involved in Lipid Modification. In *Lipid Modification by Enzymes and Engineered Microbes*; Bornscheuer, U. T., Ed.; AOCS Press, 2018; pp 11–43. https://doi.org/10.1016/B978-0-12-813167-1.00002-5.

(10) Rehman, M. F. ur; Shaeer, A.; Batool, A. I.; Aslam, M. Chapter 2 - Structure-Function Relationship of Extremozymes. In *Microbial Extremozymes*; Kuddus, M., Ed.; Academic Press, 2022; pp 9–30. https://doi.org/10.1016/B978-0-12-822945-3.00023-3.

(11) Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed Engl.* **2018**, *57* (16), 4143–4148. https://doi.org/10.1002/anie.201708408.

(12) Arnold, F. H. The Nature of Chemical Innovation: New Enzymes by Evolution. *Q. Rev. Biophys.* **2015**, *48* (4), 404–410. https://doi.org/10.1017/S003358351500013X.

(13) Zanghellini, A. *De Novo* Computational Enzyme Design. *Curr. Opin. Biotechnol.* **2014**, *29*, 132–138. https://doi.org/10.1016/j.copbio.2014.03.002.

(14) Rakotoharisoa, R. V.; Seifinoferest, B.; Zarifi, N.; Miller, J. D. M.; Rodriguez, J. M.; Thompson, M. C.; Chica, R. A. Design of Efficient Artificial Enzymes Using Crystallographically Enhanced Conformational Sampling. *J. Am. Chem. Soc.* **2024**, *146* (14), 10001–10013. https://doi.org/10.1021/jacs.4c00677.

(15) Richter, F.; Leaver-Fay, A.; Khare, S. D.; Bjelic, S.; Baker, D. De Novo Enzyme Design Using Rosetta3. *PLOS ONE* **2011**, *6* (5), e19230. https://doi.org/10.1371/journal.pone.0019230.
(16) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N. Computational Enzyme Design. *Angew. Chem. Int. Ed.* **2013**, *52* (22), 5700–5725. https://doi.org/10.1002/anie.201204077.
(17) Tantillo, D. J.; Jiangang, C.; Houk, K. N. Theozymes and Compuzymes: Theoretical Models for Biological Catalysis. *Curr. Opin. Chem. Biol.* **1998**, *2* (6), 743–750. https://doi.org/10.1016/S1367-5931(98)80112-9.
(18) Zanghellini, A.; Jiang, L.; Wollacott, A. M.; Cheng, G.; Meiler, J.; Althoff, E. A.; Röthlisberger, D.; Baker, D. New Algorithms and an in Silico Benchmark for Computational Enzyme Design. *Protein Sci. Publ. Protein Soc.* **2006**, *15* (12), 2785–2794. https://doi.org/10.1110/ps.062353106.
(19) Gaines, J. C.; Virrueta, A.; Buch, D. A.; Fleishman, S. J.; O'Hern, C. S.; Regan, L. Collective Repacking Reveals That the Structures of Protein Cores Are Uniquely Specified by Steric Repulsive Interactions. *Protein Eng. Des. Sel.* **2017**, *30* (5), 387–394. https://doi.org/10.1093/protein/gzx011.
(20) Mak, W. S.; Siegel, J. B. Computational Enzyme Design: Transitioning from Catalytic Proteins to Enzymes. *Curr. Opin. Struct. Biol.* **2014**, *27*, 87–94. https://doi.org/10.1016/j.sbi.2014.05.010.
(21) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: A Server for Estimating pKas and Adding Missing Hydrogens to Macromolecules. *Nucleic Acids Res.* **2005**, *33* (Web Server issue), W368–W371. https://doi.org/10.1093/nar/gki464.
(22) Davey, J. A.; Chica, R. A. Multistate Computational Protein Design with Backbone Ensembles. In *Computational Protein Design*; Samish, I., Ed.; Springer: New York, NY, 2017; pp 161–179. https://doi.org/10.1007/978-1-4939-6637-0_7.
(23) Bertolani, S. J.; Carlin, D. A.; Siegel, J. B. Computational Introduction of Catalytic Activity into Proteins. In *Computational Design of Ligand Binding Proteins*; Stoddard, B. L., Ed.; Springer: New York, NY, 2016; pp 213–231. https://doi.org/10.1007/978-1-4939-3569-7_13.
(24) Lassila, J. K.; Privett, H. K.; Allen, B. D.; Mayo, S. L. Combinatorial Methods for Small-Molecule Placement in Computational Enzyme Design. *Proc. Natl. Acad. Sci.* **2006**, *103* (45), 16710–16715. https://doi.org/10.1073/pnas.0607691103.
(25) Korendovych, I. V.; DeGrado, W. F. De Novo Protein Design, a Retrospective. *Q. Rev. Biophys.* **2020**, *53*, e3. https://doi.org/10.1017/S0033583519000131.
(26) Privett, H. K.; Kiss, G.; Lee, T. M.; Blomberg, R.; Chica, R. A.; Thomas, L. M.; Hilvert, D.; Houk, K. N.; Mayo, S. L. Iterative Approach to Computational Enzyme Design. *Proc. Natl. Acad. Sci.* **2012**, *109* (10), 3790–3795. https://doi.org/10.1073/pnas.1118082108.
(27) Hanreich, S.; Bonandi, E.; Drienovská, I. Design of Artificial Enzymes: Insights into Protein Scaffolds. *ChemBioChem* **2023**, *24* (6), e202200566. https://doi.org/10.1002/cbic.202200566.
(28) Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D. Kemp Elimination Catalysts by Computational Enzyme Design. *Nature* **2008**, *453* (7192), 190–195. https://doi.org/10.1038/nature06879.

(29) Althoff, E. A.; Wang, L.; Jiang, L.; Giger, L.; Lassila, J. K.; Wang, Z.; Smith, M.; Hari, S.; Kast, P.; Herschlag, D.; Hilvert, D.; Baker, D. Robust Design and Optimization of Retroaldol Enzymes. *Protein Sci. Publ. Protein Soc.* **2012**, *21* (5), 717–726. https://doi.org/10.1002/pro.2059.
(30) Khersonsky, O.; Kiss, G.; Röthlisberger, D.; Dym, O.; Albeck, S.; Houk, K.; Baker, D.; Tawfik, D. Bridging the Gaps in Design Methodologies by Evolutionary Optimization of the Stability and Proficiency of Designed Kemp Eliminase KE59. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 10358–10363. https://doi.org/10.1073/pnas.1121063109.
(31) Savile, C. K.; Janey, J. M.; Mundorff, E. C.; Moore, J. C.; Tam, S.; Jarvis, W. R.; Colbeck, J. C.; Krebber, A.; Fleitz, F. J.; Brands, J.; Devine, P. N.; Huisman, G. W.; Hughes, G. J. Biocatalytic Asymmetric Synthesis of Chiral Amines from Ketones Applied to Sitagliptin Manufacture. *Science* **2010**, *329* (5989), 305–309. https://doi.org/10.1126/science.1188934.
(32) Li, T.; Liang, J.; Ambrogelly, A.; Brennan, T.; Gloor, G.; Huisman, G.; Lalonde, J.; Lekhal, A.; Mijts, B.; Muley, S.; Newman, L.; Tobin, M.; Wong, G.; Zaks, A.; Zhang, X. Efficient, Chemoenzymatic Process for Manufacture of the Boceprevir Bicyclic [3.1.0]Proline Intermediate Based on Amine Oxidase-Catalyzed Desymmetrization. *J. Am. Chem. Soc.* **2012**, *134* (14), 6467–6472. https://doi.org/10.1021/ja3010495.
(33) Cobb, R. E.; Chao, R.; Zhao, H. Directed Evolution: Past, Present and Future. *AIChE J. Am. Inst. Chem. Eng.* **2013**, *59* (5), 1432–1440. https://doi.org/10.1002/aic.13995.
(34) Davis, I. W.; Arendall, W. B.; Richardson, D. C.; Richardson, J. S. The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances. *Structure* **2006**, *14* (2), 265–274. https://doi.org/10.1016/j.str.2005.10.007.
(35) Khersonsky, O.; Röthlisberger, D.; Wollacott, A. M.; Murphy, P.; Dym, O.; Albeck, S.; Kiss, G.; Houk, K. N.; Baker, D.; Tawfik, D. S. Optimization of the in Silico Designed Kemp Eliminase KE70 by Computational Design and Directed Evolution. *J. Mol. Biol.* **2011**, *407* (3), 391–412. https://doi.org/10.1016/j.jmb.2011.01.041.
(36) Dunbrack, R. L. Rotamer Libraries in the 21st Century. *Curr. Opin. Struct. Biol.* **2002**, *12* (4), 431–440. https://doi.org/10.1016/S0959-440X(02)00344-5.
(37) Tabet, J.-C.; Rebuffat, S. [Nobel Prize 2002 for chemistry: mass spectrometry and nuclear magnetic resonance]. *Med. Sci. MS* **2003**, *19* (8–9), 865–872. https://doi.org/10.1051/medsci/20031989865.
(38) Hu, Y.; Cheng, K.; He, L.; Zhang, X.; Jiang, B.; Jiang, L.; Li, C.; Wang, G.; Yang, Y.; Liu, M. NMR-Based Methods for Protein Analysis. *Anal. Chem.* **2021**, *93* (4), 1866–1879. https://doi.org/10.1021/acs.analchem.0c03830.
(39) Retel, J. S.; Nieuwkoop, A. J.; Hiller, M.; Higman, V. A.; Barbet-Massin, E.; Stanek, J.; Andreas, L. B.; Franks, W. T.; van Rossum, B.-J.; Vinothkumar, K. R.; Handel, L.; de Palma, G. G.; Bardiaux, B.; Pintacuda, G.; Emsley, L.; Kühlbrandt, W.; Oschkinat, H. Structure of Outer Membrane Protein G in Lipid Bilayers. *Nat. Commun.* **2017**, *8* (1), 2073. https://doi.org/10.1038/s41467-017-02228-2.
(40) Schneider, M.; Fu, X.; Keating, A. E. X-Ray vs. NMR Structures as Templates for Computational Protein Design. *Proteins Struct. Funct. Bioinforma.* **2009**, *77* (1), 97–110. https://doi.org/10.1002/prot.22421.
(41) Davey, J. A.; Chica, R. A. Improving the Accuracy of Protein Stability Predictions with Multistate Design Using a Variety of Backbone Ensembles. *Proteins Struct. Funct. Bioinforma.* **2014**, *82* (5), 771–784. https://doi.org/10.1002/prot.24457.

(42) Allen, B. D.; Nisthal, A.; Mayo, S. L. Experimental Library Screening Demonstrates the Successful Application of Computational Protein Design to Large Structural Ensembles. *Proc. Natl. Acad. Sci.* **2010**, *107* (46), 19838–19843. https://doi.org/10.1073/pnas.1012985107.
(43) Polanski, J. Chemoinformatics☆. In *Comprehensive Chemometrics (Second Edition)*; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier: Oxford, 2020; pp 635–676. https://doi.org/10.1016/B978-0-12-409547-2.14327-6.
(44) Comba, P.; Remenyi, R. Inorganic and Bioinorganic Molecular Mechanics Modeling—the Problem of the Force Field Parameterization. *Coord. Chem. Rev.* **2003**, *238–239*, 9–20. https://doi.org/10.1016/S0010-8545(02)00286-2.
(45) Tzeliou, C. E.; Mermigki, M. A.; Tzeli, D. Review on the QM/MM Methodologies and Their Application to Metalloproteins. *Molecules* **2022**, *27* (9), 2660. https://doi.org/10.3390/molecules27092660.
(46) Groenhof, G. Introduction to QM/MM Simulations. In *Biomolecular Simulations*; Monticelli, L., Salonen, E., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2013; Vol. 924, pp 43–66. https://doi.org/10.1007/978-1-62703-017-5_3.
(47) Burnley, B. T.; Afonine, P. V.; Adams, P. D.; Gros, P. Modelling Dynamics in Protein Crystal Structures by Ensemble Refinement. *eLife* **2012**, *1*, e00311. https://doi.org/10.7554/eLife.00311.
(48) Rondeau, J.-M.; Schreuder, H. Chapter 22 - Protein Crystallography and Drug Discovery. In *The Practice of Medicinal Chemistry (Fourth Edition)*; Wermuth, C. G., Aldous, D., Raboisson, P., Rognan, D., Eds.; Academic Press: San Diego, 2015; pp 511–537. https://doi.org/10.1016/B978-0-12-417205-0.00022-5.
(49) Morris, A. L.; MacArthur, M. W.; Hutchinson, E. G.; Thornton, J. M. Stereochemical Quality of Protein Structure Coordinates. *Proteins Struct. Funct. Bioinforma.* **1992**, *12* (4), 345–364. https://doi.org/10.1002/prot.340120407.
(50) Wang, J. Estimation of the Quality of Refined Protein Crystal Structures. *Protein Sci. Publ. Protein Soc.* **2015**, *24* (5), 661–669. https://doi.org/10.1002/pro.2639.
(51) Davey, J. A.; Chica, R. A. Multistate Approaches in Computational Protein Design. *Protein Sci.* **2012**, *21* (9), 1241–1252. https://doi.org/10.1002/pro.2128.
(52) Davey, J. A.; Damry, A. M.; Goto, N. K.; Chica, R. A. Rational Design of Proteins That Exchange on Functional Timescales. *Nat. Chem. Biol.* **2017**, *13* (12), 1280–1285. https://doi.org/10.1038/nchembio.2503.
(53) 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. https://doi.org/10.1038/s41586-021-03819-2.
(54) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Hanikel, N.; Pellock, S. J.; Courbet, A.; Sheffler, W.; Wang, J.; Venkatesh, P.; Sappington, I.; Torres, S. V.; Lauko, A.; De Bortoli, V.; Mathieu, E.; Ovchinnikov, S.; Barzilay, R.; Jaakkola, T. S.; DiMaio, F.; Baek, M.; Baker, D. De Novo Design of Protein Structure and Function with RFdiffusion. *Nature* **2023**, *620* (7976), 1089–1100. https://doi.org/10.1038/s41586-023-06415-8.

(55) Friedland, G. D.; Kortemme, T. Designing Ensembles in Conformational and Sequence Space to Characterize and Engineer Proteins. *Curr. Opin. Struct. Biol.* **2010**, *20* (3), 377–384. https://doi.org/10.1016/j.sbi.2010.02.004.
(56) Janson, G.; Valdes-Garcia, G.; Heo, L.; Feig, M. Direct Generation of Protein Conformational Ensembles via Machine Learning. *Nat. Commun.* **2023**, *14* (1), 774. https://doi.org/10.1038/s41467-023-36443-x.
(57) Jacobs, D. J. Ensemble-Based Methods for Describing Protein Dynamics. *Curr. Opin. Pharmacol.* **2010**, *10* (6), 760. https://doi.org/10.1016/j.coph.2010.09.014.
(58) Levin, E. J.; Kondrashov, D. A.; Wesenberg, G. E.; George N Phillips, J. Ensemble Refinement of Protein Crystal Structures: Validation and Application. *Struct. Lond. Engl. 1993* **2007**, *15* (9), 1040. https://doi.org/10.1016/j.str.2007.06.019.
(59) Chen, V. B.; Davis, I. W.; Richardson, D. C. KING (Kinemage, Next Generation): A Versatile Interactive Molecular and Scientific Visualization Program. *Protein Sci.* **2009**, *18* (11), 2403–2409. https://doi.org/10.1002/pro.250.
(60) Smith, C. A.; Kortemme, T. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. *J. Mol. Biol.* **2008**, *380* (4), 742–756. https://doi.org/10.1016/j.jmb.2008.05.023.
(61) Smith, C.; Kortemme, T. Predicting the Tolerated Sequences for Proteins and Protein Interfaces Using RosettaBackrub Flexible Backbone Design. *PloS One* **2011**, *6*, e20451. https://doi.org/10.1371/journal.pone.0020451.
(62) Ollikainen, N.; Jong, R. M. de; Kortemme, T. Coupling Protein Side-Chain and Backbone Flexibility Improves the Re-Design of Protein-Ligand Specificity. *PLOS Comput. Biol.* **2015**, *11* (9), e1004335. https://doi.org/10.1371/journal.pcbi.1004335.
(63) Harmalkar, A.; Lyskov, S.; Gray, J. J. Reliable Protein-Protein Docking with AlphaFold, Rosetta, and Replica-Exchange. bioRxiv November 25, 2023, p 2023.07.28.551063. https://doi.org/10.1101/2023.07.28.551063.
(64) Boorla, V. S.; Chowdhury, R.; Ramasubramanian, R.; Ameglio, B.; Frick, R.; Gray, J. J.; Maranas, C. D. De Novo Design and Rosetta-Based Assessment of High-Affinity Antibody Variable Regions (Fv) against the SARS-CoV-2 Spike Receptor Binding Domain (RBD). *Proteins Struct. Funct. Bioinforma.* **2023**, *91* (2), 196–208. https://doi.org/10.1002/prot.26422.
(65) Ollikainen, N.; Smith, C. A.; Fraser, J. S.; Kortemme, T. Flexible Backbone Sampling Methods to Model and Design Protein Alternative Conformations. *Methods Enzymol.* **2013**, *523*, 61–85. https://doi.org/10.1016/B978-0-12-394292-0.00004-7.
(66) Casey, M. L.; Kemp, D. S.; Paul, K. G.; Cox, D. D. Physical Organic Chemistry of Benzisoxazoles. I. Mechanism of the Base-Catalyzed Decomposition of Benzisoxazoles. *J. Org. Chem.* **1973**, *38* (13), 2294–2301. https://doi.org/10.1021/jo00953a006.
(67) Acosta-Silva, C.; Bertran, J.; Branchadell, V.; Oliva, A. Catalytic Effect of Electric Fields on the Kemp Elimination Reactions with Neutral Bases. *ChemPhysChem* **2020**, *21* (22), 2594–2604. https://doi.org/10.1002/cphc.202000667.
(68) Risso, V. A.; Romero-Rivera, A.; Gutierrez-Rus, L. I.; Ortega-Muñoz, M.; Santoyo-Gonzalez, F.; Gavira, J. A.; Sanchez-Ruiz, J. M.; Kamerlin, S. C. L. Enhancing a de Novo Enzyme Activity by Computationally-Focused Ultra-Low-Throughput Screening †Electronic Supplementary Information (ESI) Available: Additional Simulation Details and Table of the Full List of Variants Predicted by FuncLib. See DOI: 10.1039/D0sc01935f. *Chem. Sci.* **2020**, *11* (24), 6134–6148. https://doi.org/10.1039/d0sc01935f.

(69) Hennig, M.; Darimont, B. D.; Jansonius, J. N.; Kirschner, K. The Catalytic Mechanism of Indole-3-Glycerol Phosphate Synthase: Crystal Structures of Complexes of the Enzyme from Sulfolobus Solfataricus with Substrate Analogue, Substrate, and Product. *J. Mol. Biol.* **2002**, *319* (3), 757–766. https://doi.org/10.1016/S0022-2836(02)00378-9.
(70) Bunzel, H. A.; Kries, H.; Marchetti, L.; Zeymer, C.; Mittl, P. R. E.; Mulholland, A. J.; Hilvert, D. Emergence of a Negative Activation Heat Capacity during Evolution of a Designed Enzyme. *J. Am. Chem. Soc.* **2019**, *141* (30), 11745–11748. https://doi.org/10.1021/jacs.9b02731.
(71) Lamba, V.; Sanchez, E.; Fanning, L. R.; Howe, K.; Alvarez, M. A.; Herschlag, D.; Forconi, M. Kemp Eliminase Activity of Ketosteroid Isomerase. *Biochemistry* **2017**, *56* (4), 582–591. https://doi.org/10.1021/acs.biochem.6b00762.
(72) Li, A.; Wang, B.; Ilie, A.; Dubey, K. D.; Bange, G.; Korendovych, I. V.; Shaik, S.; Reetz, M. T. A Redox-Mediated Kemp Eliminase. *Nat. Commun.* **2017**, *8* (1), 14876. https://doi.org/10.1038/ncomms14876.
(73) Genre-Grandpierre, A.; Tellier, C.; Loirat, M.-J.; Blanchard, D.; Hodgson, D. R. W.; Hollfelder, F.; Kirby, A. J. Catalysis of the Kemp Elimination by Antibodies Elicited against a Cationic Hapten. *Bioorg. Med. Chem. Lett.* **1997**, *7* (19), 2497–2502. https://doi.org/10.1016/S0960-894X(97)10003-8.
(74) Bunzel, H. A.; Anderson, J. L. R.; Hilvert, D.; Arcus, V. L.; van der Kamp, M. W.; Mulholland, A. J. Evolution of Dynamical Networks Enhances Catalysis in a Designer Enzyme. *Nat. Chem.* **2021**, *13* (10), 1017–1022. https://doi.org/10.1038/s41557-021-00763-6.
(75) Broom, A.; Rakotoharisoa, R. V.; Thompson, M. C.; Zarifi, N.; Nguyen, E.; Mukhametzhanov, N.; Liu, L.; Fraser, J. S.; Chica, R. A. Ensemble-Based Enzyme Design Can Recapitulate the Effects of Laboratory Directed Evolution in Silico. *Nat. Commun.* **2020**, *11* (1), 4808. https://doi.org/10.1038/s41467-020-18619-x.
(76) Casale, M. T.; Richman, A. R.; Elrod, M. J.; Garland, R. M.; Beaver, M. R.; Tolbert, M. A. Kinetics of Acid-Catalyzed Aldol Condensation Reactions of Aliphatic Aldehydes. *Atmos. Environ.* **2007**, *41* (29), 6212–6224. https://doi.org/10.1016/j.atmosenv.2007.04.002.
(77) Lippert, S.; Baumann, W.; Thomke, K. Secondary Reactions of the Base-Catalyzed Aldol Condensation of Acetone. *J. Mol. Catal.* **1991**, *69* (2), 199–214. https://doi.org/10.1016/0304-5102(91)80145-S.
(78) Wang, C.; Liu, J.; Leng, W.; Gao, Y. Rapid and Efficient Functionalized Ionic Liquid-Catalyzed Aldol Condensation Reactions Associated with Microwave Irradiation. *Int. J. Mol. Sci.* **2014**, *15* (1), 1284–1299. https://doi.org/10.3390/ijms15011284.
(79) Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K. W.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y.-E. A.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P. Rosetta3. In *Methods in Enzymology*; Johnson, M. L., Brand, L., Eds.; Computer Methods, Part C; Academic Press, 2011; Vol. 487, pp 545–574. https://doi.org/10.1016/B978-0-12-381270-4.00019-6.
(80) Zeymer, C.; Zschoche, R.; Hilvert, D. Optimization of Enzyme Mechanism along the Evolutionary Trajectory of a Computationally Designed (Retro-)Aldolase. *J. Am. Chem. Soc.* **2017**, *139* (36), 12541–12549. https://doi.org/10.1021/jacs.7b05796.

(81) Giger, L.; Caner, S.; Obexer, R.; Kast, P.; Baker, D.; Ban, N.; Hilvert, D. Evolution of a Designed Retro-Aldolase Leads to Complete Active Site Remodeling. *Nat. Chem. Biol.* **2013**, *9* (8), 494–498. https://doi.org/10.1038/nchembio.1276.
(82) Yu, M.-Z.; Yuan, Y.; Li, Z.-J.; Kunthic, T.; Wang, H.-X.; Xu, C.; Xiang, Z. An Artificial Enzyme for Asymmetric Nitrocyclopropanation of α,β-Unsaturated Aldehydes—Design and Evolution. *Angew. Chem. Int. Ed.* **2024**, *63* (25), e202401635. https://doi.org/10.1002/anie.202401635.
(83) Obexer, R.; Godina, A.; Garrabou, X.; Mittl, P. R. E.; Baker, D.; Griffiths, A. D.; Hilvert, D. Emergence of a Catalytic Tetrad during Evolution of a Highly Active Artificial Aldolase. *Nat. Chem.* **2017**, *9* (1), 50–56. https://doi.org/10.1038/nchem.2596.
(84) Nagel, Z. D.; Klinman, J. P. Update 1 of: Tunneling and Dynamics in Enzymatic Hydride Transfer. *Chem. Rev.* **2010**, *110* (12), PR41–PR67. https://doi.org/10.1021/cr1001035.
(85) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and Development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (Pt 4), 486–501. https://doi.org/10.1107/S0907444910007493.
(86) Liebschner, D.; Afonine, P. V.; Baker, M. L.; Bunkóczi, G.; Chen, V. B.; Croll, T. I.; Hintze, B.; Hung, L. W.; Jain, S.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R. D.; Poon, B. K.; Prisant, M. G.; Read, R. J.; Richardson, J. S.; Richardson, D. C.; Sammito, M. D.; Sobolev, O. V.; Stockwell, D. H.; Terwilliger, T. C.; Urzhumtsev, A. G.; Videau, L. L.; Williams, C. J.; Adams, P. D. Macromolecular Structure Determination Using X-Rays, Neutrons and Electrons: Recent Developments in Phenix. *Acta Crystallogr. Sect. Struct. Biol.* **2019**, *75* (Pt 10), 861–877. https://doi.org/10.1107/S2059798319011471.
(87) Steinmetz, H. L. USING THE METHOD OF STEEPEST DESCENT. *Ind. Eng. Chem.* **1966**, *58* (1), 33–39. https://doi.org/10.1021/ie50673a008.
(88) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.
(89) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092. https://doi.org/10.1063/1.464397.
(90) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25* (2), 247–260. https://doi.org/10.1016/j.jmgm.2005.12.005.
(91) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.
(92) Evans, D. J.; Holian, B. L. The Nose–Hoover Thermostat. *J. Chem. Phys.* **1985**, *83* (8), 4069–4074. https://doi.org/10.1063/1.449071.
(93) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690. https://doi.org/10.1063/1.448118.
(94) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52* (12), 7182–7190. https://doi.org/10.1063/1.328693.

(95) Knapp, M. J.; Rickert, K.; Klinman, J. P. Temperature-Dependent Isotope Effects in Soybean Lipoxygenase-1: Correlating Hydrogen Tunneling with Protein Dynamics. *J. Am. Chem. Soc.* **2002**, *124* (15), 3865–3874. https://doi.org/10.1021/ja012205t.
(96) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L. Jr.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13* (6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125.
(97) Vik, E. C.; Li, P.; Maier, J. M.; Madukwe, D. O.; Rassolov, V. A.; Pellechia, P. J.; Masson, E.; Shimizu, K. D. Large Transition State Stabilization from a Weak Hydrogen Bond. *Chem. Sci.* **2020**, *11* (28), 7487–7494. https://doi.org/10.1039/D0SC02806A.
(98) Warshel, A.; Papazyan, A. Energy Considerations Show That Low-Barrier Hydrogen Bonds Do Not Offer a Catalytic Advantage over Ordinary Hydrogen Bonds. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93* (24), 13665–13670.
(99) Boyken, S. E.; Chen, Z.; Groves, B.; Langan, R. A.; Oberdorfer, G.; Ford, A.; Gilmore, J.; Xu, C.; DiMaio, F.; Pereira, J. H.; Sankaran, B.; Seelig, G.; Zwart, P. H.; Baker, D. De Novo Design of Protein Homo-Oligomers with Modular Hydrogen Bond Network-Mediated Specificity. *Science* **2016**, *352* (6286), 680–687. https://doi.org/10.1126/science.aad8865.
(100) Yeh, A. H.-W.; Norn, C.; Kipnis, Y.; Tischer, D.; Pellock, S. J.; Evans, D.; Ma, P.; Lee, G. R.; Zhang, J. Z.; Anishchenko, I.; Coventry, B.; Cao, L.; Dauparas, J.; Halabiya, S.; DeWitt, M.; Carter, L.; Houk, K. N.; Baker, D. De Novo Design of Luciferases Using Deep Learning. *Nature* **2023**, *614* (7949), 774–780. https://doi.org/10.1038/s41586-023-05696-3.
(101) Wootton, J. C. Sequences with 'Unusual' Amino Acid Compositions. *Curr. Opin. Struct. Biol.* **1994**, *4* (3), 413–421. https://doi.org/10.1016/S0959-440X(94)90111-2.
(102) Huang, P.-S.; Ban, Y.-E. A.; Richter, F.; Andre, I.; Vernon, R.; Schief, W. R.; Baker, D. RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLOS ONE* **2011**, *6* (8), e24109. https://doi.org/10.1371/journal.pone.0024109.
(103) McREE, D. E. 3 - COMPUTATIONAL TECHNIQUES. In *Practical Protein Crystallography (Second Edition)*; McREE, D. E., Ed.; Academic Press: San Diego, 1999; pp 91-cp1. https://doi.org/10.1016/B978-012486052-0/50005-1.
(104) Mortensen, J. C.; Damjanovic, J.; Miao, J.; Hui, T.; Lin, Y. A Backbone-dependent Rotamer Library with High (ϕ, ψ) Coverage Using Metadynamics Simulations. *Protein Sci. Publ. Protein Soc.* **2022**, *31* (12), e4491. https://doi.org/10.1002/pro.4491.
(105) Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. WebLogo: A Sequence Logo Generator. *Genome Res.* **2004**, *14* (6), 1188–1190. https://doi.org/10.1101/gr.849004.
(106) Pan, K.; Deem, M. W. Quantifying Selection and Diversity in Viruses by Entropy Methods, with Application to the Haemagglutinin of H3N2 Influenza. *J. R. Soc. Interface* **2011**, *8* (64), 1644–1653. https://doi.org/10.1098/rsif.2011.0105.
(107) Sun, K.; Li, S.; Zheng, B.; Zhu, Y.; Wang, T.; Liang, M.; Yao, Y.; Zhang, K.; Zhang, J.; Li, H.; Han, D.; Zheng, J.; Coventry, B.; Cao, L.; Baker, D.; Liu, L.; Lu, P. Accurate de Novo Design of Heterochiral Protein–Protein Interactions. *Cell Res.* **2024**, 1–13. https://doi.org/10.1038/s41422-024-01014-2.
(108) Karaca, Y.; Moonis, M. Chapter 14 - Shannon Entropy-Based Complexity Quantification of Nonlinear Stochastic Process: Diagnostic and Predictive Spatiotemporal Uncertainty of Multiple Sclerosis Subgroups. In *Multi-Chaos, Fractal and Multi-Fractional Artificial*

Intelligence of Different Complex Systems; Karaca, Y., Baleanu, D., Zhang, Y.-D., Gervasi, O., Moonis, M., Eds.; Academic Press, 2022; pp 231–245. https://doi.org/10.1016/B978-0-323-90032-4.00018-3.
(109) *Theoretical and Experimental Insights into Immunology*; Perelson, A. S., Weisbuch, G., Eds.; Springer: Berlin, Heidelberg, 1992. https://doi.org/10.1007/978-3-642-76977-1.
(110) Alkema, W. B. L.; Dijkhuis, A.-J.; de Vries, E.; Janssen, D. B. The Role of Hydrophobic Active-Site Residues in Substrate Specificity and Acyl Transfer Activity of Penicillin Acylase. *Eur. J. Biochem.* **2002**, *269* (8), 2093–2100. https://doi.org/10.1046/j.1432-1033.2002.02857.x.
(111) Lee, B.; Richards, F. M. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55* (3), 379-IN4. https://doi.org/10.1016/0022-2836(71)90324-X.
(112) Gainza, P.; Roberts, K. E.; Donald, B. R. Protein Design Using Continuous Rotamers. *PLOS Comput. Biol.* **2012**, *8* (1), e1002335. https://doi.org/10.1371/journal.pcbi.1002335.
(113) González-Delgado, J.; Sagar, A.; Zanon, C.; Lindorff-Larsen, K.; Bernadó, P.; Neuvial, P.; Cortés, J. WASCO: A Wasserstein-Based Statistical Tool to Compare Conformational Ensembles of Intrinsically Disordered Proteins. *J. Mol. Biol.* **2023**, *435* (14), 168053. https://doi.org/10.1016/j.jmb.2023.168053.
(114) Lazar, T.; Guharoy, M.; Vranken, W.; Rauscher, S.; Wodak, S. J.; Tompa, P. Distance-Based Metrics for Comparing Conformational Ensembles of Intrinsically Disordered Proteins. *Biophys. J.* **2020**, *118* (12), 2952–2965. https://doi.org/10.1016/j.bpj.2020.05.015.

Appendix

The following is a code snippet from the BackrubMover.cc file:

```
bool
BackrubMover::
gatherRotations() {
  cleanData();
  bool changed = false;
  // Create error message to arise if proline is chosen
  std::string pro_err = "ERROR: Window residue cannot be proline.";
  // Call necessary parameters, including rotationa angle (thetaMain), starting and ending
residues (Wstart, Wend, respectively)
  std::size_t ResID = this->_ResID; // call ResID from constructor
  double thetaMain = this->_thetaMain;//* deg_to_rad; // call thetaMain from constructor
  std::size_t Wstart = this->_Wstart; // see above
  std::size_t Wend = this->_Wend; // see above
  tout << debug << "Number of sites: " << _conf->getNumSites() << std::endl << reset;
  // Initialize coordinates for left and right window Cas, and the axis vector between them.
  Coord leftCa, rightCa, axisVector;

  // extract CA variables using this loop -> we increase the scope of the variables,
  // make them more global
  std::size_t Start = Wstart;

  if (Wstart == 0) {Start = 1;}

  for (std::size_t i=0; i<_conf->getNumSites(); i++){
    if (i == Wend + 1) {break;}
    const std::size_t form = _conf->getForm(i);
    energy::pEnergyAtomSet eas = _model->getForm(i,form);

    for (std::size_t a = 0; a < eas->numAtoms(); a++) {
      molecular::pAtom atom = eas->getAtom(a);
      std::string name = eas->getName(a);

      std::string atomName = name.substr(name.rfind("/")+1);
      name = name.substr(0, name.rfind("/"));

      std::size_t residueNum = atom->getParent()->getResNum();
        name = name.substr(0, name.rfind("|"));

        std::string resNameFull = atom->getParent()->getName();
        std::string resName3 = resNameFull.substr(0, resNameFull.rfind("|"));
        //tout << debug << "Current residue (name): " << resName3 << std::endl << reset;
```

```
            if (residueNum != Wstart && residueNum != Wend) {
               continue;
            }
            else if (residueNum == Wstart && atomName == "CA") {
               if (resName3 == "PRO") {
                  throw std::runtime_error(pro_err);
               }
               leftCa = eas->getCoord(a);
               tout << debug << "leftCa acquired: " << leftCa << std::endl << reset;
            }
            else if (residueNum == Wend && atomName == "CA") {
               if (resName3 == "PRO") {
                  throw std::runtime_error(pro_err);
               }
               rightCa = eas->getCoord(a);
               tout << debug << "rightCa acquired: " << rightCa << std::endl << reset;
            //break;
            }
      }
   }

   axisVector = rightCa - leftCa;
   tout << debug << "axisVector: " << axisVector << std::endl << reset;

   // loop for performing a single backrub and “committing” the changes
   for (std::size_t i=0; i<_conf->getNumSites(); i++){
      //if (i == Wend + 1) {break;}
      const std::size_t form = _conf->getForm(i);
      energy::pEnergyAtomSet eas = _model->getForm(i,form);
      changed |= gatherRotations(eas,_model-
>getActiveSiteIndex(i),form,ResID,thetaMain,Wstart,Wend,leftCa,axisVector);

      if (_model->hasTemplateSite(i)) {
         eas = _model->getTemplateForm(i,form);
         changed |= gatherRotations(eas,_model->getTemplateSiteIndex(i),
                                    _model-
>getTemplateFormIndex(i,form),ResID,thetaMain,Wstart,Wend,leftCa,axisVector);
      }
   }

   return changed;
}

Coord
BackrubMover::
```

```
RodriguesRotation(const Coord& c, const Coord& axis, double thetaMain) {
  Coord unitAxis = axis * (1.0 / magnitude(axis)); // Find unit vector of Axis vector
  Coord cProduct = crossProductBR(c, unitAxis); // calculate cross product

  //tout << standard << "#######################################################" <<
std::endl << reset;
  //tout << standard << "RodRot beginning.." << std::endl << reset;
  //std::string sep = ", ";
  //tout << standard << "c = " << c << std::endl << reset;

  if (thetaMain <= 0.1) { // Taylor approximation for small angles
    return c + cProduct * thetaMain;
  }

  double cosTheta = std::cos(thetaMain);
  double sinTheta = std::sin(thetaMain);

  // debugging
  //tout << standard << "cos(thetaMain), sin(thetaMain) = " << cosTheta << ", " << sinTheta <<
std::endl << reset;

  //tout << standard << "axis = " << axis << std::endl << reset;
  //tout << standard << "unitAxis = " << unitAxis << std::endl << reset;

  double dProduct = innerProduct(unitAxis, c); // dot product

  // debugging
  //tout << standard << "Dot product: " << dProduct << std::endl << reset;
  //tout << standard << "Cross product: " << cProduct << std::endl << reset;

  return c * cosTheta + cProduct * sinTheta + unitAxis * dProduct * (1-cosTheta); // full Rod.
Rot. formula
}
```