

Engineering of multi-substrate enzyme specificity and conformational equilibrium using multistate computational protein design

Antony D. St-Jacques

Thesis submitted to the University of Ottawa

in partial fulfillment of the requirements for the

Doctorate in Philosophy degree in Chemistry

Department of Chemistry and Biomolecular Sciences

Faculty of Science

University of Ottawa

© Antony D. St-Jacques, Ottawa, Canada, 2018

Abstract

The creation of enzymes displaying desired substrate specificity is an important objective of enzyme engineering. To help achieve this goal, computational protein design (CPD) can be used to identify sequences that can fulfill interactions required to productively bind a desired substrate. Standard CPD protocols find optimal sequences in the context of a single state, for example an enzyme structure with a single substrate bound at its active site. However, many enzymes catalyze reactions requiring them to bind multiple substrates during successive steps of the catalytic cycle. The design of multi-substrate enzyme specificity requires the ability to evaluate sequences in the context of multiple substrate-bound states because mutations designed to enhance activity for one substrate may be detrimental to the binding of a second substrate. Additionally, many enzymes undergo conformational changes throughout their catalytic cycle and the equilibrium between these conformations can have an impact on their substrate specificity. In this thesis, I present the development and implementation of two multistate computational protein design methodologies for the redesign of multi-substrate enzyme specificity and the modulation of enzyme conformational equilibrium. Overall, our approaches open the door to the design of multi-substrate enzymes displaying tailored specificity for any biocatalytic application.

Acknowledgments

I would like to acknowledge my PhD supervisor, Dr. Roberto Chica. I want to say how grateful I am for providing me with the opportunity to pursue my academic and professional interests. Additionally, I'm very appreciative for your help regarding the preparation of presentations and manuscripts. The communication skills you have taught me will be very beneficial for the future.

Secondly, I would like to thank the Chica lab members, both past and present, for their support and feedback, especially in preparing presentations and discussing science.

Finally, I would like to thank and acknowledge my thesis writing buddy, workout partner, and hiking companion, all manifested as my partner for life, Dr. Marilyn Vera Chang, who made the whole graduate school and life experience quite a bit more fun and enjoyable than it would have been without her.

Table of Contents

Abstract	ii
Acknowledgments.....	iii
Table of Contents	iv
List of Tables	ix
List of Figures	x
List of Abbreviations	xii
Chapter 1: Introduction	1
1.1 Enzyme engineering.....	1
1.2 Computational enzyme redesign	2
1.2.1 CPD Methods.....	3
1.2.2 Addressing protein flexibility	13
1.2.3 Computational enzyme design vs. redesign	17
1.3 Thesis rationale	21
1.3.1 Multi-substrate specificity	21
1.3.2 Conformational changes in enzymes	23
1.3.3 Selecting suitable test cases	24
1.4 Aminotransferases.....	25
1.4.1 Aminotransferase structures.....	26
1.4.2 Mechanism.....	28

1.4.3	Assaying ATs.....	29
1.4.4	Dual specificity of ATs.....	30
1.4.5	Domain movement during catalysis.....	32
1.5	Thesis objectives	33
Chapter 2: Rotamer optimization calculations to recapitulate aminotransferase active site configurations with different rotamer libraries		35
2.1	Statement of contribution.....	35
2.2	Introduction	35
2.3	Methods.....	36
2.4	Results and Discussion.....	37
2.4.1	Structure selection.....	37
2.4.2	Description of active sites.....	39
2.4.3	Recovery of active-site residue rotamer configurations	41
2.4.4	Recovery of substrate rotamers.....	47
2.4.5	Targeted substrate rotamer library	49
2.4.6	Modeling substrates with the Targeted library	51
2.4.7	New substrate on a new backbone	53
2.5	Conclusions	59
Chapter 3: Computational design of multi-substrate enzyme specificity		61
3.1	Statement of contribution.....	61

3.2	Introduction	61
3.3	Results and discussion.....	62
3.3.1	Contribution of each chemical state to prediction accuracy	76
3.3.2	Comparison to other methods	79
3.4	Conclusions	80
3.5	Materials and methods	80
3.5.1	Selection of the intermediate PLP states to model.....	80
3.5.2	Structure Preparation	81
3.5.3	MD ensemble generation	82
3.5.4	Computational details	82
3.5.5	Construction of position matrices	87
3.5.6	Library design	89
3.5.7	Saturation mutagenesis	89
3.5.8	Mutant Library Construction	90
3.5.9	Screening.....	91
3.5.10	Steady-state kinetics.....	91
Chapter 4: Computational design of conformational equilibrium to modulate the substrate specificity of an enzyme		93
4.1	Statement of contribution	93
4.2	Introduction	93

4.3	Results	96
4.3.1	Domain movement analysis and residue selection	96
4.3.2	Ensemble Preparation	97
4.3.3	Computational Design	99
4.3.4	Screening.....	103
4.3.5	Steady-state kinetics of AAT mutants.	110
4.4	Discussion	113
4.4.1	Library characteristics.....	113
4.4.2	Future experiments to validate results.	115
4.4.3	Importance of including multiple conformational states.	116
4.4.4	Comparison to the literature.....	118
4.5	Conclusions	118
4.6	Methods.....	119
4.6.1	Computational Protein Design details	119
4.6.2	Mutagenesis	119
4.6.3	Screening.....	120
4.6.4	Steady-state kinetics of AAT mutants	121
Chapter 5: Summary and perspectives.....		122
5.1	Summary	122
5.2	Perspectives.....	123

References	131
------------------	-----

List of Tables

<i>Table 2.1. The RMSD of rotamer optimized aminotransferase designed structures</i>	<i>46</i>
<i>Table 2.2. Number of rotamers.....</i>	<i>55</i>
<i>Table 3.1. Position Probability Matrix of HIS state.</i>	<i>67</i>
<i>Table 3.2. Position Probability Matrix of GLU state.....</i>	<i>68</i>
<i>Table 3.3 Position Probability Matrix of PLP state.....</i>	<i>69</i>
<i>Table 3.4. Position Probability Matrix of HisGluPLP state.</i>	<i>72</i>
<i>Table 3.5. Apparent kinetic parameters of E. coli BCAT and its mutants</i>	<i>73</i>
<i>Table 3.6. Apparent kinetic parameters of additional BCAT mutants</i>	<i>74</i>
<i>Table 3.7. Apparent kinetic parameters of E. coli BCAT and its mutants for transamination of leucine with α-ketoglutarate as an amino acceptor.....</i>	<i>74</i>
<i>Table 3.8. Minimum library size required to predict BCAT mutants</i>	<i>77</i>
<i>Table 3.9. BCAT residues optimized by computational design</i>	<i>86</i>
<i>Table 3.10. Correlation of residue probabilities between sub-designs.</i>	<i>88</i>
<i>Table 4.1. Domain motion data for E. coli AAT.....</i>	<i>97</i>
<i>Table 4.2. Structural information of the Open and Closed 50-member PertMin ensembles.</i>	<i>98</i>
<i>Table 4.3. The Both, Closed, and Open AAT combinatorial libraries.</i>	<i>100</i>
<i>Table 4.4. Computed ensemble scores for the Open and Closed state of the WT E. coli AAT sequence,.....</i>	<i>100</i>
<i>Table 4.5. Computed ensemble scores for the Open and Closed state of variants of the Closed library.....</i>	<i>101</i>
<i>Table 4.6. Computed ensemble scores for the Open and Closed state of variants of the Both library.....</i>	<i>102</i>
<i>Table 4.7. Computed ensemble scores for the Open and Closed state of variants of the Open library.....</i>	<i>103</i>
<i>Table 4.8. Screening results of the 2-concentration assay.....</i>	<i>106</i>
<i>Table 4.9. Computed scores and apparent kinetic parameters of AAT variants.....</i>	<i>111</i>

List of Figures

Figure 1.1. Computational protein design.	3
Figure 1.2. Flow diagram of the FASTER algorithm.	10
Figure 1.3. Multistate approaches to computational protein design.	14
Figure 1.4. Input templates for examples of possible multistate experiments.	16
Figure 1.5. Ligand placement and matching algorithms used in computational enzyme design.	19
Figure 1.6. Active sites of multi-substrate enzymes.	23
Figure 1.7. Transamination reaction catalyzed by α -aminotransferases.	26
Figure 1.8. Fold Type I and IV aminotransferases.	27
Figure 1.9. Mechanism for the half reaction of aminotransferases.	29
Figure 1.10. Quinonoid intermediate.	29
Figure 1.11. GDH coupled enzyme assay.	30
Figure 1.12. Active site of <i>P. denitrificans</i> AroAT complexes	31
Figure 1.13. Active site of <i>E. coli</i> BCAT in complex with various substrates.	32
Figure 1.14. AAT domain movement.	33
Figure 2.1. <i>E. coli</i> BCAT active site.	40
Figure 2.2. <i>E. coli</i> AAT active site.	41
Figure 2.3. Comparison of BCAT crystal and designed structures.	43
Figure 2.4. Comparison of AAT crystallographic and designed structures.	45
Figure 2.5. Hydrogen-bond networks of AAT designed structures.	48
Figure 2.6. Workflow of the Targeted rotamer library generation.	51
Figure 2.7. Substrate-bound PLP complexes.	54
Figure 2.8. Rotamer library coverage plots for substrates	56
Figure 2.9. Rotamer prediction of Glu ^{PLP} bound to PLP as an external aldimine	58
Figure 3.1. BCAT transamination bi bi ping pong reaction	63
Figure 3.2. Active site of <i>E. coli</i> BCAT.	65

<i>Figure 3.3. Comparison of position energy matrices generated using different chemical states.</i>	69
<i>Figure 3.4. Effect of mutations shown in CPD models on the GLU, HIS, and PLP chemical states.</i>	70
<i>Figure 3.5. Sequence optimization scheme for a single chemical state.</i>	85
<i>Figure 3.6. SDS-PAGE of wild type BCAT and the top 4 BCAT mutants</i>	92
<i>Figure 4.1. Comparison of the different conformational forms of wild type (WT) AAT and the AAT hexamutant (HEX).</i>	94
<i>Figure 4.2. Representative kinetic traces of the 2-concentration GDH screening assay.</i>	108
<i>Figure 5.1. States adopted by an enzyme as it undergoes its catalytic cycle.</i>	127
<i>Figure 5.2. Hypothetical consequences of transition state stabilization by computational enzyme design.</i>	127
<i>Figure 5.3. Multistate computational enzyme design.</i>	128
<i>Figure 5.4. Potential energy surfaces of hypothetical enzymes.</i>	128

List of Abbreviations

Å	Ångstrom
AAT	aspartate aminotransferase
AroAT	aromatic amino acid aminotransferase
ATs	aminotransferases
bbdep	Dunbrack 2002 backbone dependent rotamer library
bbind	Dunbrack 2002 backbone independent rotamer library
BCAT	branched chain amino acid aminotransferase
BMEC	backbone-determined minimum energy configuration
CLEARSS	Combinatorial Libraries Emphasizing And Reflecting Scored Sequences
CPD	computational protein design
DAAT	D-amino acid aminotransferases
deg	degree
dPR	double perturbation and relaxation
E.C.	Enzyme commission
elec	electrostatics
FASTER	Fast and Accurate Side-Chain Topology and Energy Refinement
GDH	L-glutamate dehydrogenase
GMEC	global minimum energy configuration
H-bond	hydrogen bond
HEX	<i>E. coli</i> AAT hexamutant (V39L, K41Y, T47I, N69L, T109S and N297S)
HspAT	histidinol phosphate aminotransferase
iBr	iterative batch relaxation
IMEC	intermediate minimum energy configuration
ISM	Iterative saturation mutagenesis
MCMEC	Monte-Carlo-determined minimum energy configuration
MD	molecular dynamics
MLS	Minimum library size
MOE	Molecular Operating Environment
MSA	Multistate analysis
MSD	Multistate design
NAD ⁺	β-nicotinamide adenine dinucleotide, oxidized form

NADH	β -nicotinamide adenine dinucleotide, reduced form
Ni-NTA	nickel-nitrilotriacetic acid
NMR	nuclear magnetic resonance
PCC	Pearson correlation coefficient
PDB	Protein Data Bank
PDB ID	Protein data bank identification
PEM	position energy matrix
PertMin	coordinate perturbation followed by energy minimization
PID	position
PLP	pyridoxal 5' phosphate
PPM	position probability matrix
PQM	Position cumulative partition function Matrix
RMSD	Root mean square deviation
RMSD ₀	Root mean square deviation from the input structure
RMSG	Root mean square gradient
SC ^{bbdep} Sub ^{bbind}	Side chain rotamers are optimized with bbdep library, substrate side chain is optimized with bbind library
SC ^{bbdep} Sub ^{fix}	Side chain rotamers are optimized with bbdep library, substrate side chain is kept fixed
SC ^{bbind} Sub ^{bbind}	Side chain rotamers are optimized with bbind library, substrate side chain is optimized with bbind library
SC ^{bbind} Sub ^{fix}	Side chain rotamers are optimized with bbind library, substrate side chain is kept fixed
SC ^{bbind} Sub ^{sub}	Side chain rotamers are optimized with bbind library, substrate side chain is optimized with Targeted substrate rotamer library
SC ^{fix} Sub ^{bbind}	Side chain rotamers are kept fixed, substrate side chain is optimized with bbind library
SC ^{fix} Sub ^{sub}	Side chain rotamers are kept fixed, substrate side chain is optimized with Targeted substrate rotamer library
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
solv	solvation
sPR	single perturbation and relaxation
SRC	Spearman rank coefficient
ss	secondary structure propensities

SSD	single state design
TAT	tyrosine aminotransferase
U	Enzymatic unit (mg min^{-1})
vdw	van der Waals
WT	wild-type

Chapter 1: Introduction

Some of the language and figures in this chapter were adapted from a book chapter coauthored with Olivier Gagnon and Dr. Roberto A. Chica:

St-Jacques AD, Gagnon O & Chica RA (2018) Computational Enzyme Design: Successes, Challenges and Future Directions. In: Williams G & Hall M (Ed.) Modern Biocatalysis: Advances Towards Synthetic Biological Systems (pp. 88-116). Cambridge, UK. Royal Society of Chemistry.

1.1 Enzyme engineering

The ability of enzymes to catalyze chemical reactions with high rate acceleration ($k_{cat}/k_{uncat} = 10^7\text{--}10^{26}$)¹⁻² and unmatched selectivity makes them highly attractive for use as biocatalysts in a large number of industrial syntheses. However, the finite repertoire of naturally-occurring enzyme activities and the fact that few enzymes inherently catalyze desired reactions under industrially-relevant conditions restrict their applicability in many industrial processes. To make enzymes practical for industrial syntheses, it is often necessary to tailor their catalytic activity to the intended application using protein engineering. Protein engineering techniques such as rational design and directed evolution have yielded many improved enzymatic activities,³⁻⁷ yet the creation of new enzymes that possess desired catalytic properties remains very challenging due to our incomplete understanding of the structural determinants of enzyme catalysis and the necessity for high-throughput screening of large mutant libraries. Thus, the development of a broadly-applicable computational method for the design of enzymes that can catalyze any desired chemical reaction with high efficiency under defined conditions would bolster the use of industrial biocatalysis while opening the door to biocatalytic processes that are currently inaccessible using natural enzymes.

1.2 Computational enzyme redesign

Computational enzyme redesign is the alteration of pre-existing enzyme activity or substrate specificity using computational protein design (CPD) methodologies.⁸ CPD algorithms have been developed over the past 20 years⁹⁻¹⁰ to facilitate the identification of protein sequences that can adopt a specific three-dimensional structure and possess a desired property such as increased stability,¹¹ improved binding affinity,¹² or altered specificity.¹³ Most CPD algorithms consist of three steps (Figure 1.1A): (i) a side-chain placement step where discrete side-chain rotamers¹⁴ are threaded onto specified positions on a predefined protein backbone template, (ii) an energy calculation where interaction energies between pairs of rotamers and between each rotamer and the backbone template are computed using a potential energy function,^{9, 15-18} and (iii) a sequence optimization step where combinations of rotamers are optimized using a search algorithm¹⁹⁻²⁴ that explores both rotamer and sequence space in order to identify optimal sequences. In the context of enzyme redesign, a substrate molecule can be included in the active site to identify mutant sequences able to form energetically favourable interactions with the substrate that are predicted to allow productive binding.²⁵ At the conclusion of this process, a list of sequences along with their corresponding rotamer-optimized structures are generated (Figure 1.1B), and each sequence is ranked according to an energy value that reflects its predicted stability on the target protein template structure.

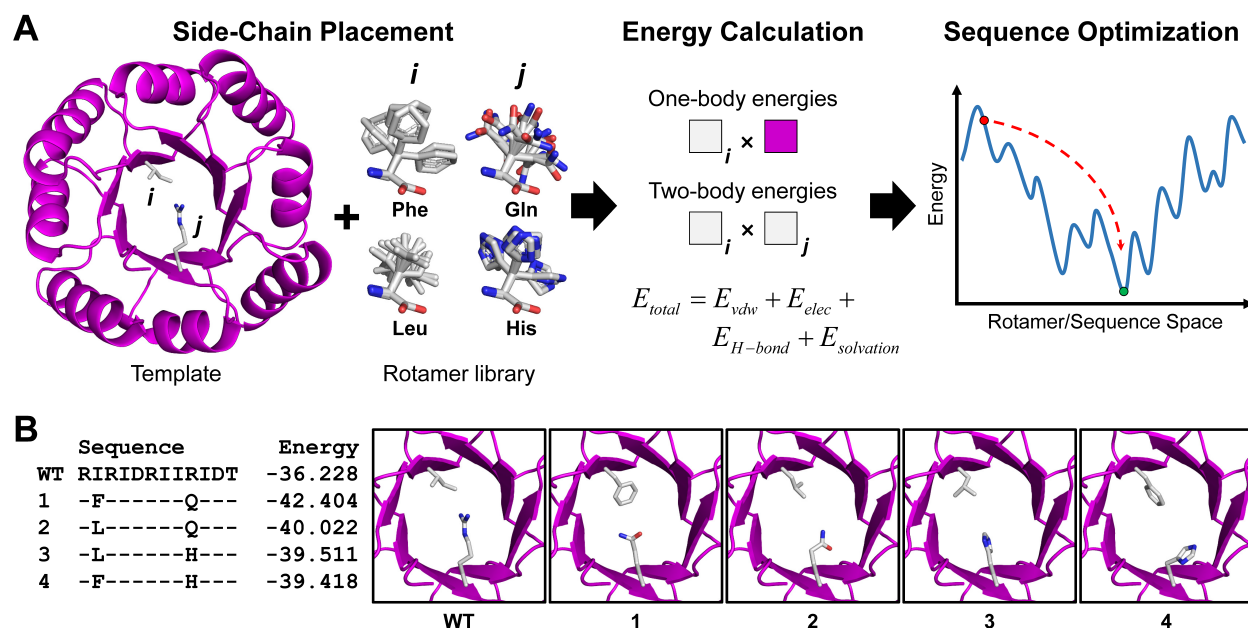


Figure 1.1. Computational protein design. (A) The general methodology of computational protein design (CPD) consists of three steps: (1) a side-chain placement step where amino acids at user-defined positions (*i* and *j*) in the protein template are replaced by a library of rotamers for each allowed amino-acid type (Phe and Leu, or Gln and His at positions *i* and *j*, respectively), (2) an energy calculation where interaction energies between pairs of rotamers (two-body energies) and between each rotamer and the backbone template (one-body energies) are computed using a potential energy function that can include van der Waals (vdw), electrostatics (elec), hydrogen bonding, and solvation terms, and (3) a sequence optimization step where a search algorithm is used to optimize combinations of rotamers in order to identify optimal sequences. (B) The CPD procedure returns a ranked list of scored sequences along with their corresponding rotamer-optimized structures. For clarity, side chains of residues at non-designed positions are omitted. Each sequence is ranked according to an energy value that reflects its predicted stability on the target protein template structure. WT indicates the wild-type sequence.

1.2.1 CPD Methods

1.2.1.1 Template preparation

Prior to the first step in CPD, a protein backbone template must be retrieved and prepared. Experimental atomic resolution structures that have been solved using X-ray crystallography or nuclear magnetic resonance (NMR) can be retrieved from databases such as the worldwide Protein Data Bank (PDB).²⁶ Homology modelling can be used to construct a backbone template in instances where experimental structural data is not available for a target protein. Once a template has been selected and inspected for missing atoms, hydrogens can be added to the structure using a variety of methods such as Reduce²⁷ and Protonate3D.²⁸ A new rotamer can be threaded onto the

backbone template in cases where there are atoms missing for a residue and loops can be modeled if missing.²⁹ After hydrogens and missing residues are added, the template is then minimized to prevent van der Waals clashes and reduce any strain. For further discussion, we are defining the template to contain the protein backbone and the side chain residues not included in rotamer optimization calculations.

1.2.1.2 Rotamer libraries

Rotamers are discrete sets of allowed conformations that correspond to local minima on potential energy maps of organic molecules such as amino acids.¹⁴ Rotamers vary from one another by rotations around side chain torsion (χ) angles, but not the backbone. Rotamer libraries are often derived from statistical analysis of side chain conformations of proteins and can be backbone-dependent or backbone-independent.^{14, 30-32} Backbone-dependent rotamer libraries differ from backbone-independent ones in that the dihedral angles of rotamers and their frequencies depend on the local backbone conformation, which is not the case for backbone-independent rotamer libraries. Backbone-dependent rotamer libraries have been shown to more accurately predict side-chain conformations in protein design calculations than backbone-independent libraries.^{30, 33-34} However, amino acid residues that are in specialized environments such as an enzymatic active site may adopt a rotamer conformation that is sparsely populated for a given (Φ , Ψ) bin and thus its rotamer conformation may not be represented in the backbone-dependent rotamer library, as will be demonstrated in Chapter 2. In these cases, it may be beneficial to use backbone-independent rotamer libraries since they contain more rotamers, which are compiled from all backbone conformations observed in proteins. In this thesis, the Dunbrack 2002 backbone-dependent (bbdep) and backbone-independent (bbind) rotamer libraries^{14, 30} with expansions of ± 1 standard deviation around χ_1 and χ_2 torsion angles are used to model the side chain of amino acid residues.

To construct the Dunbrack 2002 rotamer libraries, the rotamer frequencies of side chains in over 500 protein chains deposited in the wwPDB with 2.0 Å resolution or better were counted. Next, the prior estimate of rotamer probabilities for each (Φ , Ψ) bin were combined with the raw counts of the rotamers overlapping in $20^\circ \times 20^\circ$ bins using a Bayesian formalism, where the prior estimate of rotamer probabilities come from the product of Φ and Ψ dependencies. The backbone-dependent rotamer library contains the rotamer frequencies and their mean dihedral angles and their standard deviations on a $10^\circ \times 10^\circ$ grid of the backbone angles while the backbone-independent rotamer library is not on a grid of backbone dihedral angles.

1.2.1.3 Energy Functions

In this thesis, computational design was performed using the PHOENIX protein design software,³⁵⁻³⁷ which uses a physics-based energy function to measure the interaction energies between pairs of rotamers (two-body energies) and between each rotamer and the backbone template (one-body energies) to compute the rotamer configuration score. The potential energy function (Equation 1.1) includes van der Waals (vdw), electrostatics (elec), hydrogen bonding (H-bond), solvation (solv), and secondary structure propensities (ss) terms. Because the template is kept fixed during rotamer optimization, its internal energy remains constant during energy calculations and is thus not computed. Additionally, the internal energy of rotamers are not computed because the rotamer geometry is considered to be optimal in the rigid rotamer approximation.

$$E_{total} = E_{vdw} + E_{elec} + E_{H-bond} + E_{solv} + E_{ss}$$

Equation 1.1. PHOENIX energy function.

The van der Waals term, E_{vdw} , is computed using a Lennard-Jones 12-6 potential with radii and well depth parameters from the Dreiding II force field.³⁸ The Lennard-Jones 12-6 potential is

shown in Equation 1.2, where D_0 is the energy well depth, R is the interatomic distance, R_0 is the geometric mean of the radii of the two atoms, and α is the van der Waals radius scaling factor. The first term, which is proportional to $(1/R)^{12}$, is the short-range repulsive component and the second term, which is proportional to $-(1/R)^6$, is the long-range attractive component.

$$E_{vdw} = D_0 \left[\left(\frac{\alpha \cdot R_0}{R} \right)^{12} - 2 \left(\frac{\alpha \cdot R_0}{R} \right)^6 \right]$$

Equation 1.2. Lennard-Jones 12-6 van der Waals potential.

The electrostatics potential, E_{elec} , between two atoms is computed by approximating the two atoms as point charges and using Coulomb's Law, shown in Equation 1.3, where q and q' are the charges on atom 1 and atom 2, ϵ is the dielectric constant, and R is the interatomic distance. The PARSE set of atomic charges³⁹ are used for the electrostatic and solvation energy calculations.

$$E_{elec} = \frac{q \cdot q'}{\epsilon \cdot R}$$

Equation 1.3. Coulomb's Law equation.

The hydrogen bond potential scoring function⁴⁰ used in this thesis is defined in Equation 1.4, where D_0 is the hydrogen bond well depth, R is the interatomic distance between the hydrogen bond donor and acceptor heavy atoms, R_0 is the hydrogen bond equilibrium distance between the donor and acceptor, and $F(\theta, \Phi, \Psi)$ is an angle dependent term, which depends on the hybridization state of donor and acceptor atoms.⁴⁰ In this thesis, the recommended values for D_0 (8 kcal mol⁻¹) and R_0 (2.8 Å) were used.

$$E_{H-bond} = D_0 \left[5 \left(\frac{R_0}{R} \right)^{12} - 6 \left(\frac{R_0}{R} \right)^6 \right] F(\theta, \Phi, \Psi)$$

Equation 1.4. Hydrogen bond potential.

Since modeling explicit solvent in CPD is complicated by the large computational cost of surveying a large range of water configurations,⁴¹ water is modeled using an implicit solvation

potential in PHOENIX. The implicit solvation potential can be calculated using a surface area-based model^{9, 42-43} or an occlusion-based model.^{37, 40, 44} In the surface area-based model, the solvation energy is proportional to the solvent-accessible surface area of amino acid side chains, where the interactions with water are treated as favourable for polar residues and unfavourable for nonpolar residues (i.e., an energy benefit is given to polar exposure and nonpolar burial while an energy penalty is given to nonpolar exposure and polar burial). Similar to the surface area-based model, burial of nonpolar and exposure of polar groups are considered beneficial while burial of polar and exposure of nonpolar groups are penalized in the occlusion-based model. However, rather than using the solvent-accessible surface area of the side chain, in the occlusion-based model the atoms are scored according to the percentage of their volumes that are occluded by other atoms in the protein structure. The occlusion model is computationally less expensive than the solvent-accessible surface area since the occlusion model does not require the solvent accessible surface area of each amino acid to be computed prior to calculating the implicit solvation term. For this reason, we used the occlusion-based model in this thesis.

The last energy term of the PHOENIX energy function is the secondary structure amino acid phi-psi propensities term,⁴⁵ which is a statistical function given in Equation 1.5, where E_{ss} is the energy for the secondary structure propensity, N_{ss} is the scaling factor (5.25), and $P_{x,y}$ is the propensity of observing an amino acid type x in Φ/Ψ bin y , where the bin size is $10^\circ \times 10^\circ$. $P_{x,y}$ is defined by Equation 1.6, where $N_{x,y}$ is the count of observing amino acid type x in Φ/Ψ bin y , $N_{x,any}$ is the count of observing amino acid type x in any Φ/Ψ bin, $N_{any,y}$ is the count of observing any amino acid type in Φ/Ψ bin y , and $N_{any,any}$ is the count of observing any amino acid type in any Φ/Ψ bin. The denominator scales the term between positions but does not affect the relative

preference for amino acids since for any position y , the denominator is constant in a fixed backbone design.

$$E_{ss} = -N_{ss} \log(P_{x,y})$$

Equation 1.5. Energy of the secondary structure propensity.

$$P_{x,y} = \frac{N_{x,y}/N_{x,any}}{N_{any,y}/N_{any,any}}$$

Equation 1.6. Amino acid secondary structure propensity.

As described above, energy functions used in CPD do not include entropy terms, preventing the calculation of free energies. Instead, these energy functions compute potential energies. This is done because free energy calculations would be too expensive when optimizing sequences across large sequences spaces. Therefore, it is important to note that differences in the calculated potential energies between sequences are more important for analysis than absolute values.

1.2.1.4 Sequence optimization algorithms

During the sequence optimization step of CPD, rotamer configurations are optimized to find the global minimum energy configuration (GMEC). The potential energy of a sequence with a specific rotamer configuration is computed using Equation 1.7, where E_{global} is the sequence potential energy, $E_{template}$ is the template potential energy, $E(r_i)$ is the one-body interaction energy of residue i in rotameric state r_i with the template and $E(r_i \cdot r_j)$ is the two-body pairwise interaction energy of residue i in rotameric state r_i with the side chain rotamers of other amino acids that are also being optimized.

$$E_{global} = E_{template} + \sum_{i=1}^N E(r_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^N E(r_i \cdot r_j)$$

Equation 1.7. Potential energy of a sequence.

To identify favourable rotamer combinations, deterministic or stochastic search algorithms⁴⁶ can be used. Deterministic search algorithms used in separate CPD simulations will result in identical solutions if the same set of parameters are used. In contrast, stochastic algorithms rely on probabilistic trajectories, which means the solutions are not always consistent and rely on a random number generator seed. While some deterministic algorithms like self-consistent mean field result in the same solution but do not guarantee the GMEC, there are other deterministic search algorithms that can guarantee the GMEC such as Dead-End-Elimination,¹⁹ Dead-End-Elimination combined with the A* algorithm,⁴⁷ Cost Function Networks algorithms,⁴⁸ Integer Linear Programming solvers,⁴⁹ and Markov Random Fields solvers.⁵⁰ However, since the rotamer optimization problem is NP-hard,⁵¹ these deterministic and provable algorithms are extremely time consuming and do not converge within a reasonable amount of time for sufficiently complex design problems.⁴⁶ Alternatively, stochastic algorithms such as the genetic algorithm,⁵² Monte Carlo combined with simulated annealing,^{24, 46} and Fast and Accurate Side-Chain Topology and Energy Refinement (FASTER)²² all converge towards low energy sequences rapidly, but at the cost of no longer guarantying the GMEC.

One of the most efficient stochastic algorithms used in CPD is the FASTER algorithm (Figure 1.2). The FASTER optimization trajectory is executed by following five steps. First, the backbone-derived minimum energy configuration (BMEC) is generated by selecting the rotamer at each position with the most favourable one-body rotamer-backbone interaction energy, which is used as input for the first cycle of iterative batch relaxation (iBR). iBR is executed in the same way the BMEC was generated, except two-body rotamer-rotamer interactions are included: the

intermediate minimum energy configuration (IMEC) is generated by selecting the rotamer at each position that interacts most favourably with the backbone and input rotamer configurations. The rotamers at every position are simultaneously updated after all positions have been considered at the end of each iBR cycle. iBR is repeated using the IMEC as input until convergence. Next, conditional iBR, which is the same as iBR except new rotamers are only accepted with 80% probability, is executed and repeated until convergence. The fourth step is single perturbation and relaxation (sPR), where a single position is randomly picked and its rotamer is changed to another. Then, the rest of the positions are allowed to relax in a single iBR cycle. The resulting IMEC is accepted if it has the lowest energy found so far and the sPR is repeated until convergence. The final step is double perturbation and relaxation (dPR), which is executed the same as sPR, except a pair of rotamers at two different positions are perturbed together.

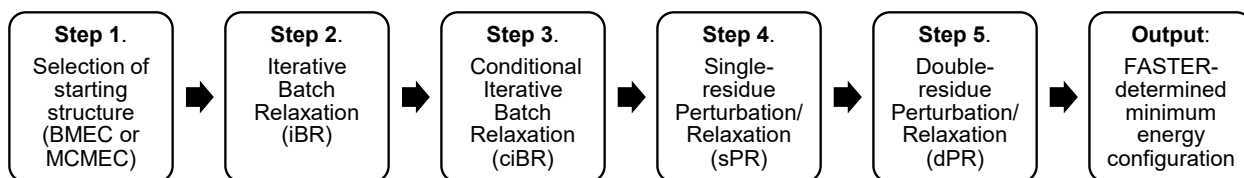


Figure 1.2. Flow diagram of the FASTER algorithm. The BMEC is the backbone-determined minimum energy configuration, the MCMEC is the Monte-Carlo-determined minimum energy configuration.

In this thesis, we used the improved version of FASTER developed by Allen et al.²³ to optimize sequences. Rather than generating the BMEC, the starting configuration is generated using a short Monte Carlo trajectory starting from a random configuration, which results in a significantly lower energy rotamer configuration than the BMEC. Additionally, relaxation of residues during sPR and dPR cycles are limited to only positions that strongly interact with the perturbed rotamer(s). These two changes improved the efficiency of the original FASTER protocol by up to 2 orders of magnitude with no loss of accuracy.

1.2.1.5 Library design

To determine the success of a CPD calculations, it is important to experimentally test the resulting top-ranked sequences. Sequences can be tested individually, but this limits the number that can be tested because of the incremental expense and labor associated with the synthesis and characterization of each sequence individually.³⁶ Alternatively, a combinatorial mutant library that is representative of the best scoring sequences can be built for roughly the same cost and effort as a single sequence and tested with a high-throughput experimental assay. There are various ways to design combinatorial libraries. Hayes et al.⁵³ constructed a frequency table for each amino acid type found at each designed position in the list of low energy sequences generated by CPD. They then limited the library to amino acids found at each position more frequently than a designated frequency cutoff. Rather than applying a frequency cutoff, Mena and Daugherty⁵⁴ designed libraries that include the most frequent mutations, but are limited to amino acids that can be encoded together using degenerate codons. Treynor et al.⁵⁵ took a different approach and scored the interactions between sets of amino acids at various positions during the rotamer optimization step of CPD to find the global minimum combinatorial library, rather than using individual sequence scores during rotamer optimization to find the GMEC.

In this thesis, the Combinatorial Libraries Emphasizing And Reflecting Scored Sequences (CLEARSS) method developed by Allen et al.³⁶ was used. This method was used because it explicitly considers the CPD energies, allows for a library size range to be specified, and for amino acid sets to be included or excluded from the final library. To design a protein library after sequence optimization of n positions to all 20 proteinogenic amino acids, the CLEARSS³⁶ library algorithm constructs a $20 \times n$ position cumulative partition function matrix (PQM) of each amino acid at all n designed positions (Equation 1.8). The Boltzmann weight, w , is calculated for each

sequence (Equation 1.9), where E is the score of the sequence, k_B is the Boltzmann constant, and T is the simulation temperature. The cumulative partition function, q , of an amino acid, aa , at position n , is the sum of w of all sequences where the identity of the residue at position n is the amino acid in question ($resi_n = aa$ in Equation 1.10). For a library size configuration, which is the specific number of amino acids at each designed protein position, the optimal amino acids at each position are those with the highest absolute partition function values for each position. All library size configurations that lead to a combinatorial library within a target size range are scored by taking the sum of all partition functions of the chosen amino acid sets over all positions, and subsequently ranking them. In doing so, the final combinatorial library contains the amino acids predicted to have the most stabilizing effect at each designed position. It is important to note that instantiation of each individual sequence included in the combinatorial library will not necessarily yield the exact same list of top-ranked sequences found by CPD. Nevertheless, many of the combinatorial library sequences will have similar potential energy to the top-ranked sequences found by CPD.

$$PQM = \begin{bmatrix} q_{A,1} & q_{A,2} & \cdots & q_{A,n-1} & q_{A,n} \\ q_{C,1} & q_{C,2} & \cdots & q_{C,n-1} & q_{C,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ q_{W,1} & q_{W,2} & \cdots & q_{W,n-1} & q_{W,n} \\ q_{Y,1} & q_{Y,2} & \cdots & q_{Y,n-1} & q_{Y,n} \end{bmatrix}$$

Equation 1.8. The Position Partition function matrix, PQM.

$$w = e^{-E/k_B T}$$

Equation 1.9. The Boltzmann Weight.

$$q_{aa,n} = \sum_{i=1}^S w_i [resi_n = aa]$$

Equation 1.10. The cumulative partition function.

1.2.2 Addressing protein flexibility

The CPD method described thus far has been performed on a single, fixed backbone template, which we call single-state design (SSD). However, even in cases where SSD has been successful, the artefacts caused by the rigid rotamer and fixed protein backbone approximations make these designs prone to false negative predictions since sequences are rejected when they cause small steric clashes even though such clashes would be relieved if the backbone was allowed to move slightly.⁵⁶ Strategies such as using softer repulsive potential energy terms,^{40, 57-58} continuous rotamer optimization⁵⁹ or flexible backbone design⁶⁰⁻⁶² can be incorporated into SSD calculations to address these artefacts. Alternatively, sequences can be optimized on a conformational ensemble of multiple protein backbones that mimic protein flexibility (called the positive state) using a positive multistate approach, which has been shown to improve accuracy of protein stability predictions.⁵⁶ To further improve the accuracy of mutant stability prediction, an additional conformational ensemble that represents non-native protein folds (called the negative state) can be incorporated in an approach called negative design,⁶³ which predicts sequences that stabilize positive states while simultaneously destabilizing negative states. In addition to improving accuracy for stability prediction, multistate approaches can be used for other protein design tasks that require different states, such as the design of conformational switches,⁶⁴ oligomeric association specificity,⁶⁵ and antibiotic resistance.⁶⁶

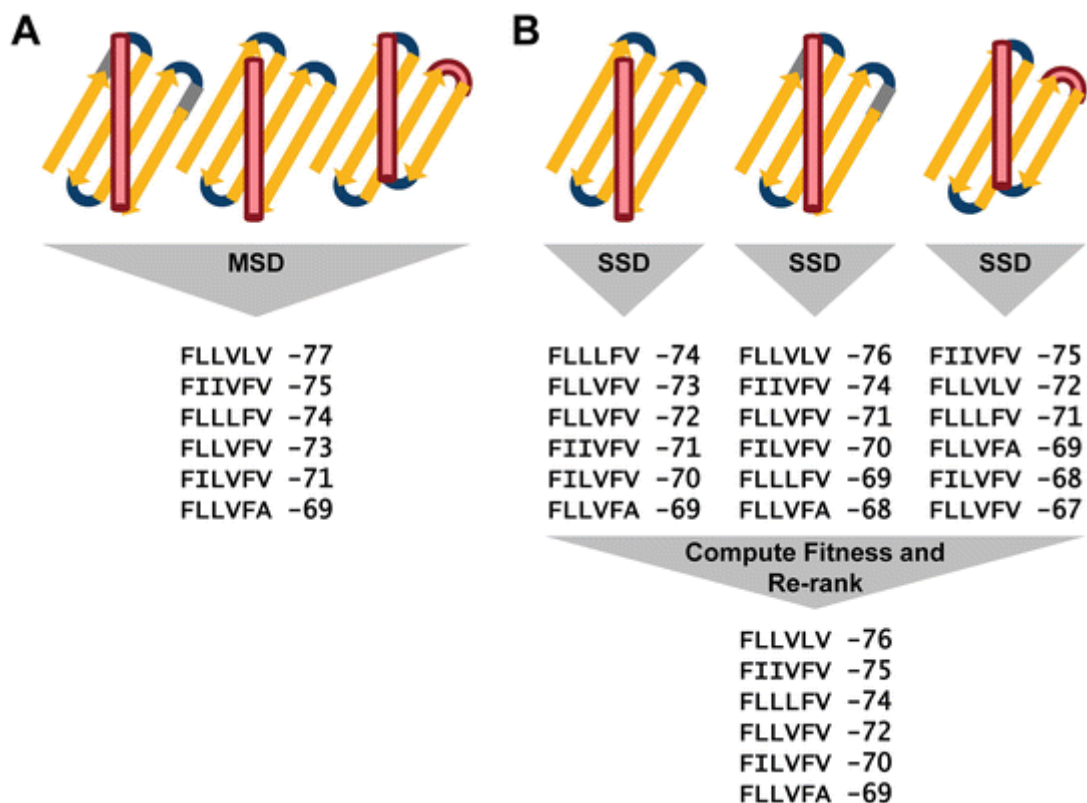


Figure 1.3. Multistate approaches to computational protein design. (A) In multistate design (MSD), sequence optimization is guided by the energetic contributions of multiple protein structures simultaneously. Thus, sequence optimization and scoring are performed concertedly, resulting in a list of sequences that are ranked according to their predicted stability across the ensemble of three structures. (B) In multistate analysis (MSA), multiple independent SSD calculations are performed in parallel using alternate backbones as input templates. The sequence scores obtained from each SSD calculation are combined post-CPD into a fitness value for each sequence across the ensemble of three structures. Sequences are then re-ranked based on their fitness values, generating a new ranked list of scored sequences. Reprinted by permission from Springer Nature: Humana Press, New York, NY, Methods in Molecular Biology. Davey J.A., Chica R.A. (2017) Multistate Computational Protein Design with Backbone Ensembles. In: Samish I. (eds) Computational Protein Design. Methods in Molecular Biology, vol 1529. Humana Press, New York, NY., © Springer Science+Business Media New York 2017

In the multistate approach called multistate design (MSD),⁶⁷⁻⁶⁸ the sequence optimization trajectory is guided by a fitness function that considers energy contributions of mutations across multiple protein structures simultaneously to generate a ranked list of sequences (Figure 1.3a). During sequence optimization, rotamers for a specific sequence are optimized on each backbone template (called states), enabling the energy of that sequence to be calculated for each state. Next, the energies of the different states are combined using a fitness function to produce a single “ensemble score” that reflects the predicted stability of that sequence across the ensemble. This ensemble score is also the “fitness value” in designs that contain only one ensemble. However, many designs can include multiple ensembles (Figure 1.4). For instance, if two structural ensembles representing distinct conformational states of the protein (e.g., folded and unfolded states, open and closed conformations, etc.), we call each ensemble different “conformational states.” Alternatively, if two ensembles contain different bound small molecules (i.e. substrates) or represent apo and holo states that are not significantly different conformationally, we call each ensemble different “chemical states.” In designs that include multiple conformational or chemical states, the fitness value can be computed by combining the ensemble scores using any desired fitness function. The disadvantages of MSD are that the fitness function must be known *a priori* and the number of states that can be modeled is limited by computational resources, as each state (i.e., backbone template) requires its own dedicated computer core.

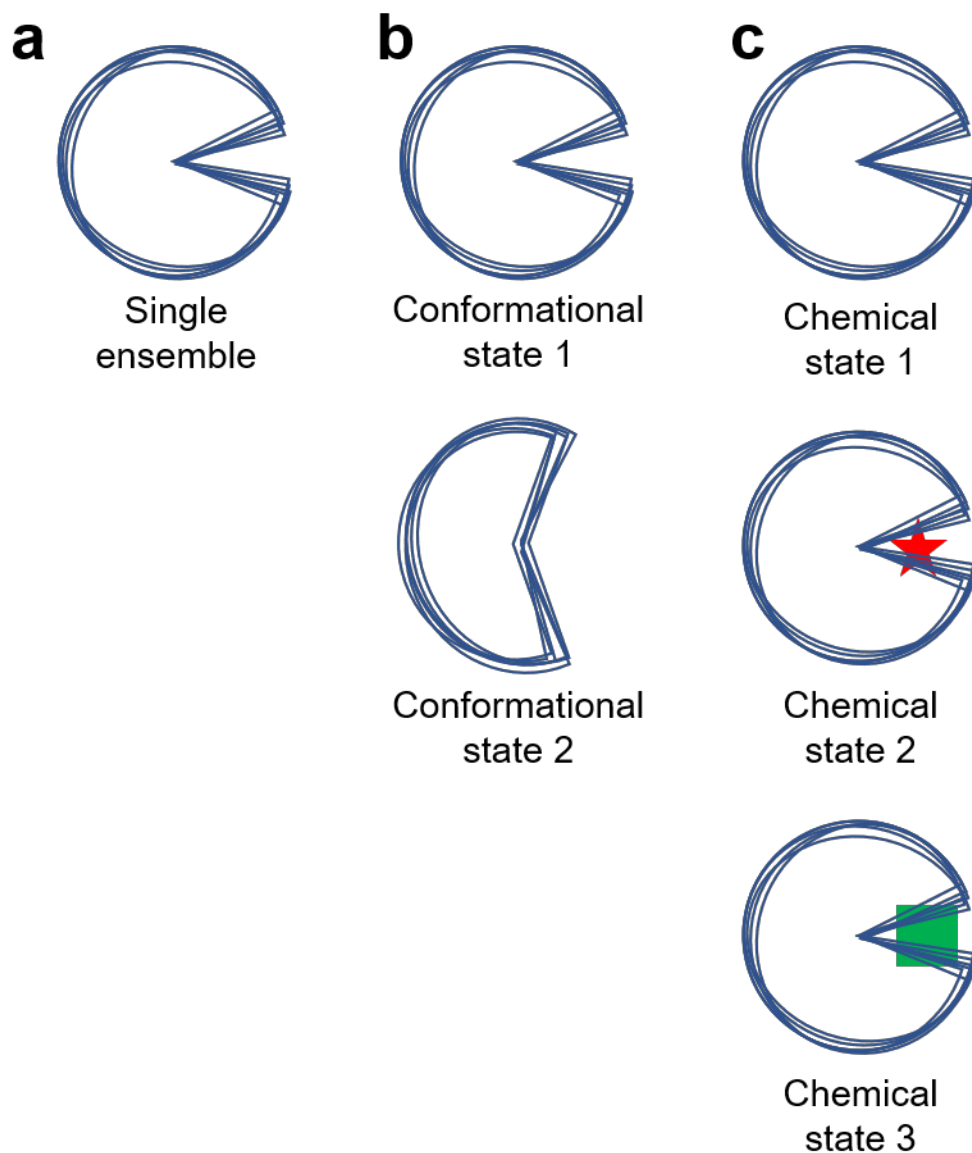


Figure 1.4. Input templates for examples of possible multistate experiments. (a) Ensemble of multiple backbone templates. (b) Two ensembles of two distinct conformational states of a protein. (c) Three ensembles of three separate chemical states that differ by the substrate bound at the active site.

An alternate approach to MSD that does not display these disadvantages is multistate analysis (MSA).⁶⁹ Unlike MSD, where sequences are optimized across all states simultaneously, multiple SSD calculations are performed on each protein template individually in MSA (Figure 1.3). The ranked list of sequences from each SSD calculation are then combined using the fitness function to build a single ranked list of sequences. By optimizing sequences on each backbone individually, MSA requires fewer computational resources than MSD and the fitness function can be optimized following sequence optimization calculations, allowing for multiple design hypotheses to be tested. However, since sequence optimization in MSA is not guided by multiple states, the explored sequence space by stochastic search algorithms can be constrained to local sequence space minima. This can result in little or no overlap in the list of top-ranked sequences obtained from each individual SSD calculation, which will prevent application of MSA.

In this thesis, MSA will be used since many different fitness functions and chemical or conformational states are tested for design calculations. To avoid searching sequence spaces constrained to local minima, all sequences within the design specifications will be rotamer optimized on each protein template individually.

1.2.3 Computational enzyme design vs. redesign

In computational enzyme design, CPD algorithms are used to create a catalytic site in a protein scaffold that does not natively bind the target substrate and catalyze the desired reaction. This is a very challenging design problem because the CPD procedure must simultaneously create a binding pocket specific to the desired substrate(s), introduce the required catalytic machinery for the chemical reaction of interest, and preserve the overall structure and stability of the template protein fold. From a calculation perspective, computational enzyme design differs from standard CPD methodologies by the inclusion of an additional step where a user-defined active site

comprising one or more catalytic side chains and a high-energy reaction intermediate such as a transition state is placed onto a pre-existing protein scaffold according to a set of predefined geometrical constraints postulated to be necessary for the designed chemical reaction. Active site construction begins by identifying positions in the protein scaffold where all necessary catalytic contacts can be made to a transition state derived from quantum mechanics calculations. This can be done using either a ligand placement⁷⁰ or matching⁷¹ algorithm (Figure 1.5). Once the position of catalytic residues has been determined, amino acids at neighbouring positions in the active site are designed to optimize substrate recognition and binding in a step called packing. Stabilization of the transition state via additional non-bonding interactions introduced during the packing step is crucial to ensure that the designed protein can perform catalysis, as transition state stabilization is one of the key features of enzyme catalysis. Sequences predicted to efficiently stabilize the transition state are then experimentally validated. It should be noted that CPD methods optimize sequence stability, and not catalytic activity. However, by identifying sequences predicted to efficiently stabilize the transition state of a chemical reaction of interest, catalytic activity can be inferred.

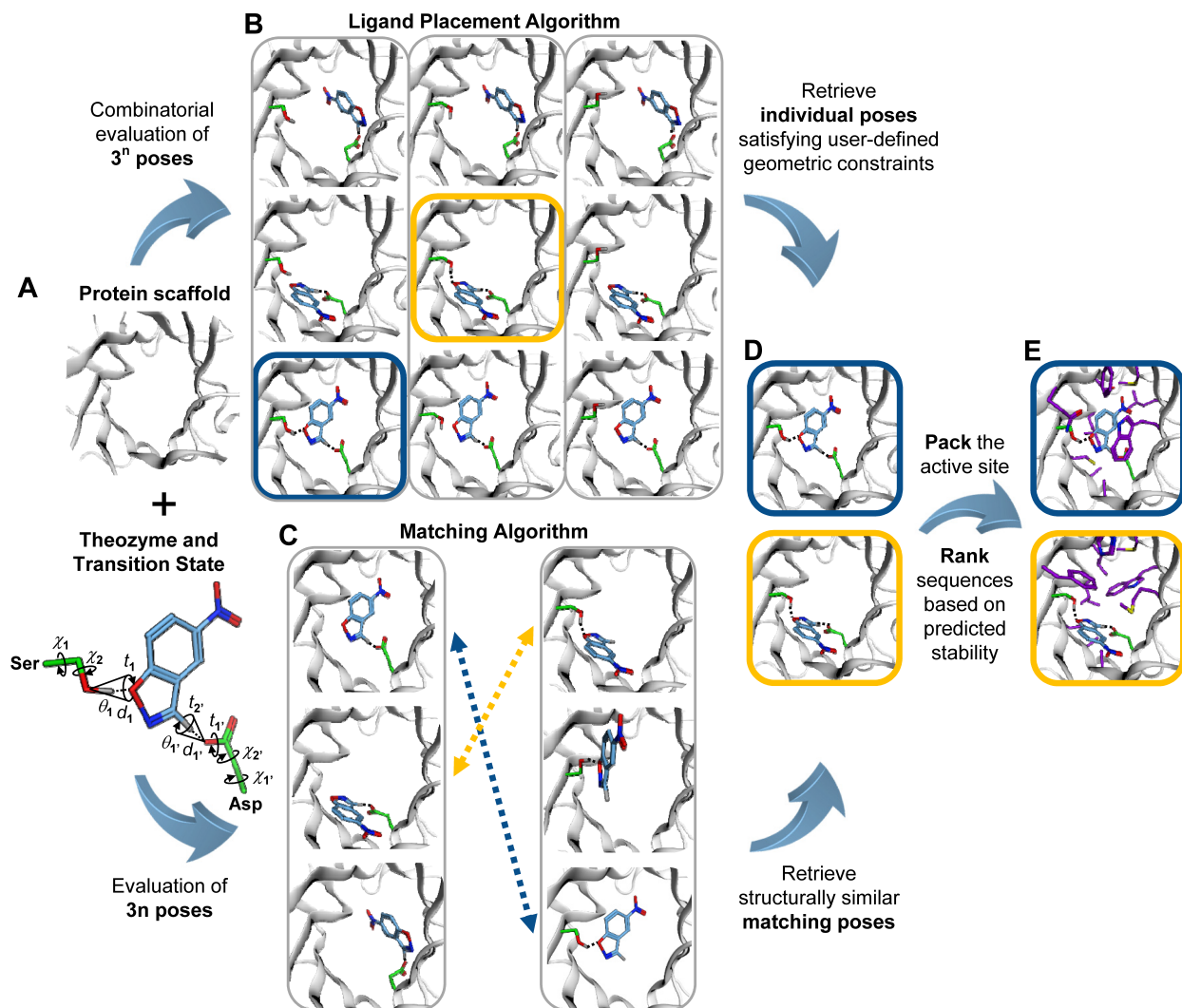


Figure 1.5. Ligand placement and matching algorithms used in computational enzyme design. (A) A protein scaffold and a theozyme (green) with its associated transition state (TS, cyan) are required as inputs for either algorithm. Although the theozyme and TS are derived from quantum mechanics calculations, several theozyme rotamers and TS poses are sampled during calculation to compensate for the use of a fixed protein scaffold during design. Theozyme rotamers are generated according to user-defined torsional degrees of freedom (χ) while TS poses are generated from each theozyme rotamer using translational (d), rotational (θ), and torsional (t) degrees of freedom to maintain user-defined catalytic contacts (dashed lines). The various theozyme rotamers and TS poses are placed inside the active-site pocket of the protein scaffold using either a ligand placement (B) or matching (C) algorithm. For clarity purposes, only three rotamers each of the designed catalytic residues Asp and Ser, as well as a single TS binding pose are shown. In ligand placement (B), the single TS binding pose is generated relative to each of the three catalytic Asp rotamers (grey boxes). For each of these Asp and TS combinations, the three Ser rotamers are tested to identify which one(s) form(s) user-defined catalytic contacts (dashed lines). Active sites containing TS poses that form both catalytic contacts (blue and yellow boxes) are optimized via a packing step, and are ranked according to their predicted stability. In this example, the blue active site ranks better than the yellow one. In matching (C), each Asp and Ser rotamer of the theozyme

is placed in the active site, and the TS pose is built off of each to create the user-defined catalytic contact. The coordinates of each rotamer and TS pose combination are saved in a hash map corresponding to the catalytic contact formed. Structurally similar TS poses forming different catalytic contacts are matched (blue and yellow arrows). Following matching, the active sites are optimized by a packing step and ranked. In this toy example, both algorithms yield the same results, however the matching algorithm is more computationally efficient given that it scales linearly ($3 \text{ rotamers} \times 2 \text{ catalytic residues} = 6 \text{ active sites}$) instead of exponentially ($3^2 = 9 \text{ active sites}$), as is the case in ligand placement.

However, the success rate of the computational enzyme design procedure remains low and even when computational enzyme design yields active biocatalysts, catalytic efficiencies (k_{cat}/K_M) obtained are orders of magnitude lower than those of natural enzymes.¹ These modest catalytic efficiencies are mostly due to k_{cat} values that are 1–6 orders of magnitude lower than those of natural enzymes,^{1, 72} and not to Michaelis constants (K_M), which fall within the range of those found in natural biocatalysts. This result demonstrates the ability of computational enzyme design to create active sites capable of productively binding a substrate but that this does not ensure highly efficient catalysis. Indeed, enzymatic reaction rates are determined by the rate-limiting step of the complete catalytic process, which may not necessarily be the chemical transformation that is modeled during design. For example, it has been shown that for many enzymes, product release⁷³ or conformational exchange between various sub-states during the catalytic cycle⁷⁴ are rate-limiting. It is thus possible that such processes slow down the rate of reactions catalyzed by *de novo* enzymes.

Therefore, in cases where an enzyme catalyzes a reaction of interest but does not bind a target substrate, it is often preferred to redesign a pre-existing active site of the enzyme rather than to create a *de novo* enzyme. By redesigning a pre-existing enzyme, one can benefit from the fact that catalytic contacts and the conformational exchanges between sub-states during the catalytic cycle have already been fine-tuned by evolution. Thus, the design problem is reduced to predicting

mutations to bind the target substrate without disrupting the pre-existing catalytic machinery and mechanism in any significant way.

1.3 Thesis rationale

1.3.1 Multi-substrate specificity

Computational enzyme redesign has been used to alter the substrate specificity of enzymes for a variety of applications, which include the change of the cofactor specificity of a xylose reductase to accept the non-native cofactor NADH rather than the native cofactor NADPH;⁷⁵ the redesign of a galactose 6-oxidase to become a glucose 6-oxidase for use in a proposed pathway to convert D-glucose to D-glucaric acid;⁷⁶ the redesign of the substrate specificity of a serine endopeptidase to display proteolytic activity towards α -gliadin, an immunogenic peptide implicated in celiac disease;⁷⁷ and even to redesign a benzaldehyde lyase to catalyze the coupling of formaldehyde into dihydroxyacetone, called the formose reaction, a reaction not yet found in nature.⁷⁸

It is worth noting that for these SSD examples, the redesign of an enzyme to bind a new target substrate resulted in significant loss of activity towards the native substrate, an unsurprising result given that mutations that stabilize binding to the target substrate may no longer be optimal for productive binding of the native substrate. In the applications described above, specificity towards the native substrate was either not required to be maintained or undesired. However, many enzymes require multiple substrates to complete their catalytic cycle.⁷⁹ For clarity, these enzymes will be denoted as “multi-substrate” enzymes. When redesigning a multi-substrate enzyme, mutations designed to change specificity for one substrate may be detrimental to the binding of a second substrate. Particularly, when redesigning enzymes that catalyze reactions via a one-site bi bi ping-pong (e.g. aminotransferases,⁸⁰⁻⁸¹ transketolases,⁸²⁻⁸³ and lipases⁸⁴), both substrates bind in

the same location of the active site (Figure 1.6b). Therefore, engineering specificity for a target substrate in the place of one substrate could greatly diminish the binding of the second substrate and render the enzyme inactive. To avoid this, we believe these enzymes should be treated as multi-chemical state systems, where all substrates required for a catalytic turnover are modeled. Multi-chemical state approaches have been developed to redesign substrate specificity in multi-substrate enzymes,^{66, 85-86} in both positive and negative designs, but in these cases, the substrates required for a complete catalytic turnover do not bind in the same position within the active site. In Chapter 3 of this thesis, we develop multistate methods to redesign multi-substrate enzymes that catalyze reactions via a one-site bi bi ping pong mechanism.

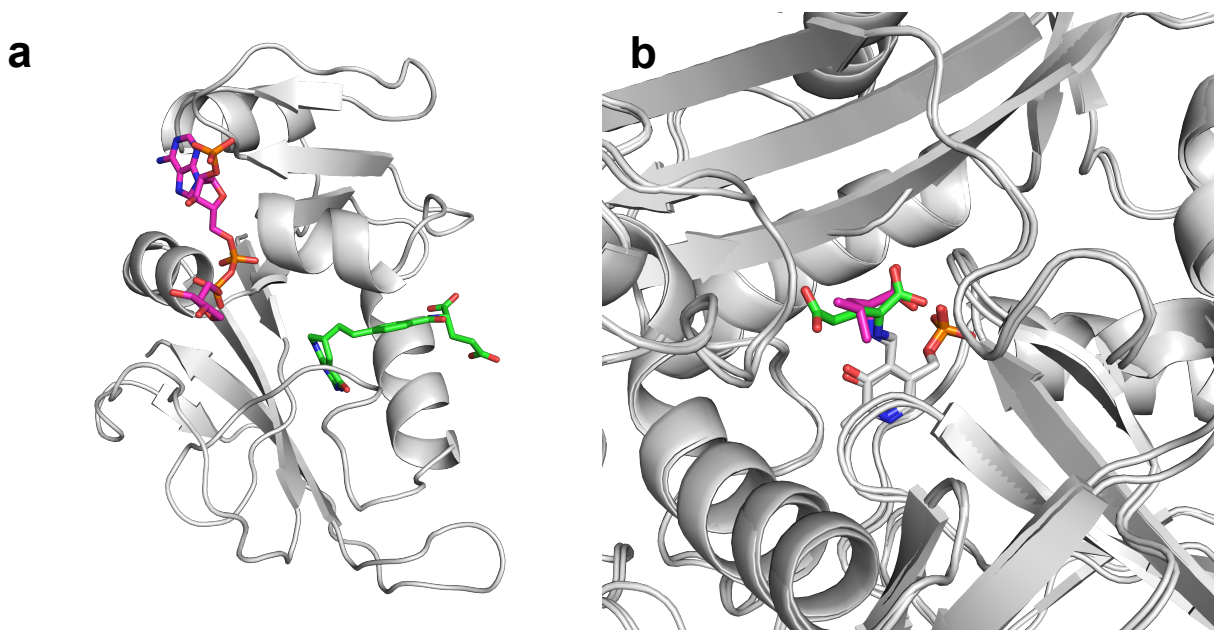


Figure 1.6. Active sites of multi-substrate enzymes where both substrates required for a complete catalytic turnover (a) bind in different active sites or (b) bind in the same active site at the same place. (a) *E. coli* dihydrofolate reductase complexed with reduced adenine di-nucleotide phosphate (pink sticks) and 7,8-dihydrofolate (green sticks) (PDB code: 1RX6⁸⁷). (b) *E. coli* Branched Chain Amino acid Aminotransferase (BCAT) complexed with glutamate (green sticks, PDB: code 1IYE⁸⁸) superimposed with *E. coli* BCAT complexed with a leucine analog (pink sticks, PDB code: 1IIL⁸⁹).

1.3.2 Conformational changes in enzymes

Conformational changes between intermediate sub-states throughout a catalytic cycle are common to many enzymes such as ribonuclease,⁹⁰⁻⁹¹ dihydrofolate reductase,⁹² chymotrypsin,⁹³⁻⁹⁴ adenylate kinases,⁷⁴ aspartate aminotransferase,^{81, 95-97} and T4 lysozyme.⁹⁸ In most proteins, the binding of ligands is coupled to conformational changes,⁹⁹⁻¹⁰⁰ which can be driven by conformational selection, where the ligand binds to a sparsely populated protein conformation. Conformational changes can also be triggered by the binding of a ligand to the open conformational state in an induced fit mechanism to give the preferred ligand-bound closed conformation. Conformational changes can enhance substrate binding in some enzymes⁹⁹ and can induce strain in the catalytic groups to enhance catalysis in others.⁹⁵ Additionally, conformational changes allow the enzyme to create a different environment for the active site. Water can be expelled from the active site, which creates a low dielectric medium where electrostatic

interactions are enhanced and the pKa of ionizable groups are significantly altered.¹⁰¹ The reaction takes place in the closed state followed by a conformational change to the open state, where the product is released.

These conformational changes are directly linked to catalysis and there is evidence that many enzyme rates are determined by the rate of conformational movement.¹⁰⁰ For instance, a single active site second shell mutation inverted the equilibrium between two catalytically essential states in proline isomerase CypA and decreased interconversion rate between the two sub-states, which led to a reduction in catalysis.¹⁰² Two second shell mutations identified via directed evolution partially rescued the enzyme activity of this single mutant by accelerating the interconversion rate between the two sub-states.¹⁰³ This correlation between the interconversion rate of the two catalytically essential sub-states and the enzyme activity highlights the importance of conformational dynamics to enzyme catalysis.

A major obstacle in using CPD to design enzymes with high activities is the ability to determine a priori which enzyme sub-states are functionally important and should be stabilized. Additionally, it is a challenge to determine which positions to mutate to modulate the equilibrium between sub-state populations for optimal enzyme activity. In Chapter 4, we develop a multistate computational framework to identify and design positions for the modulation of conformational equilibria of enzymes to affect substrate specificity and activity.

1.3.3 Selecting suitable test cases

To develop multistate methods for the redesign of multi-substrate enzyme specificity and conformational equilibrium, we needed to select enzymes to use as test cases. We were looking for enzymes with the following features: they catalyze reactions that are industrially relevant; they can be characterized using a high-throughput assay; and they have crystal structures available in

the PDB. To redesign multi-substrate enzyme specificity and conformational equilibrium, we also required enzymes that catalyze reactions using a one-site bi bi ping pong mechanism, and that undergo conformational changes throughout the reaction mechanism. For both projects, we selected aminotransferases as model enzymes.

1.4 Aminotransferases

Aminotransferases (ATs), also known as transaminases, (EC 2.6.1.X) are pyridoxal 5'-phosphate (PLP) dependent enzymes that reversibly catalyze the transamination reaction, which is the transfer of an amino group from an amino donor to an amino acceptor. There are α -aminotransferases, which convert α -amino and α -keto acids (Figure 1.7); ω -aminotransferases, which can accept substrates with a distal carboxylate group; and amine transaminases, which accept substrates that lack a carboxylate group. Additionally, the exquisite stereoselectivity, environmental friendliness, and compatibility with other enzymatic or chemical systems of ATs have made these enzyme important biocatalysts for the synthesis of chiral primary amines,^{80, 104-105} which are important building blocks for bioactive natural products, pharmaceuticals, fine chemicals and agrochemicals.¹⁰⁶⁻¹⁰⁷

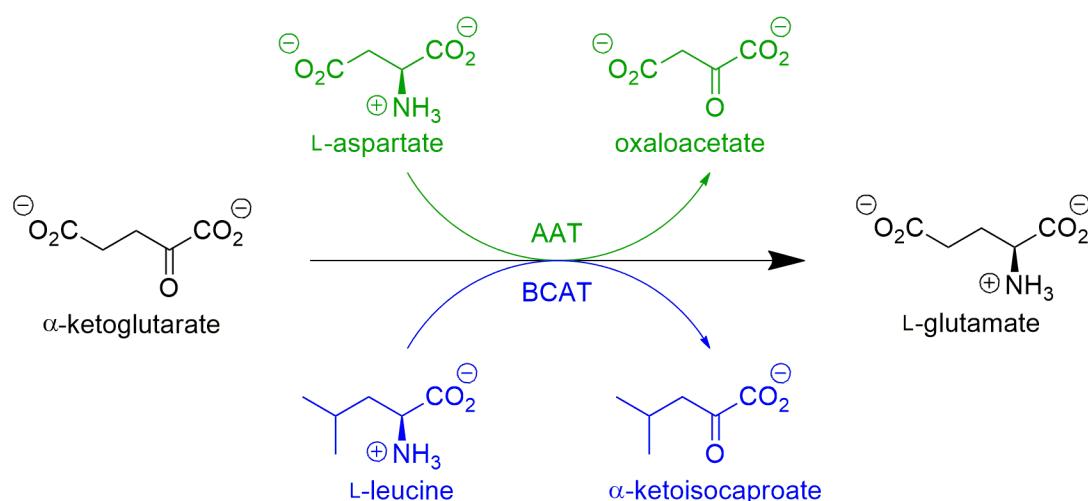


Figure 1.7. Transamination reaction catalyzed by α -aminotransferases. In green: the transamination reaction catalyzed by aspartate aminotransferase (AAT); in blue: the transamination reaction catalyzed by branched-chain amino acid aminotransferase (BCAT).

1.4.1 Aminotransferase structures

ATs adopt Fold Types I and IV of the PLP-dependent enzyme folds.^{81, 108} In both folds, the ATs are obligate homodimers with active sites that are composed of residues from both chains. Most aminotransferases belong to the Fold Type I (Figure 1.8a), with the exception of D-amino acid aminotransferases (DAAT), branched chain L-amino acid aminotransferases (BCAT), and (R)-amine transaminases, which adopt Fold Type IV folds (Figure 1.8b).¹⁰⁷⁻¹⁰⁹ When an amino acid substrate binds to Fold Type I ATs, such as aspartate aminotransferase (AAT), it is oriented so that its α -carboxylate group lies proximal to the PLP phenolate group and its side chain lies proximal to the PLP phosphate group (Figure 1.8c). To react with and synthesize L-amino acids, the catalytic lysine of Fold Type I ATs is positioned on the *si* face of the PLP while the *re* face is solvent exposed (Figure 1.8d). With ATs that belong to the Fold Type IV, the catalytic lysine is positioned on the other side of the PLP. For DAAT, the orientation of the substrate/PLP complex and the catalytic lysine is essentially a mirror image of the Fold Type I active sites, where the substrate α -carboxylate group and side chain are positioned proximal to the PLP phenolate and phosphate groups, respectively. Consequently, the repositioned catalytic lysine is now on the *re*

face while the *si* face is solvent exposed, which causes DAAT to react with and synthesize D-amino acids. In addition to the DAAT family of ATs, there are also BCATs, which react with branched chain L-amino acids, in this Fold Type. To react with and synthesize L-amino acids, the substrate orientation is inverted relative to DAAT and Fold Type I ATs (Figure 1.8e), where the α -carboxylate group lies proximal to the PLP phosphate group and the substrate side chain lies proximal to the PLP phenolate group.

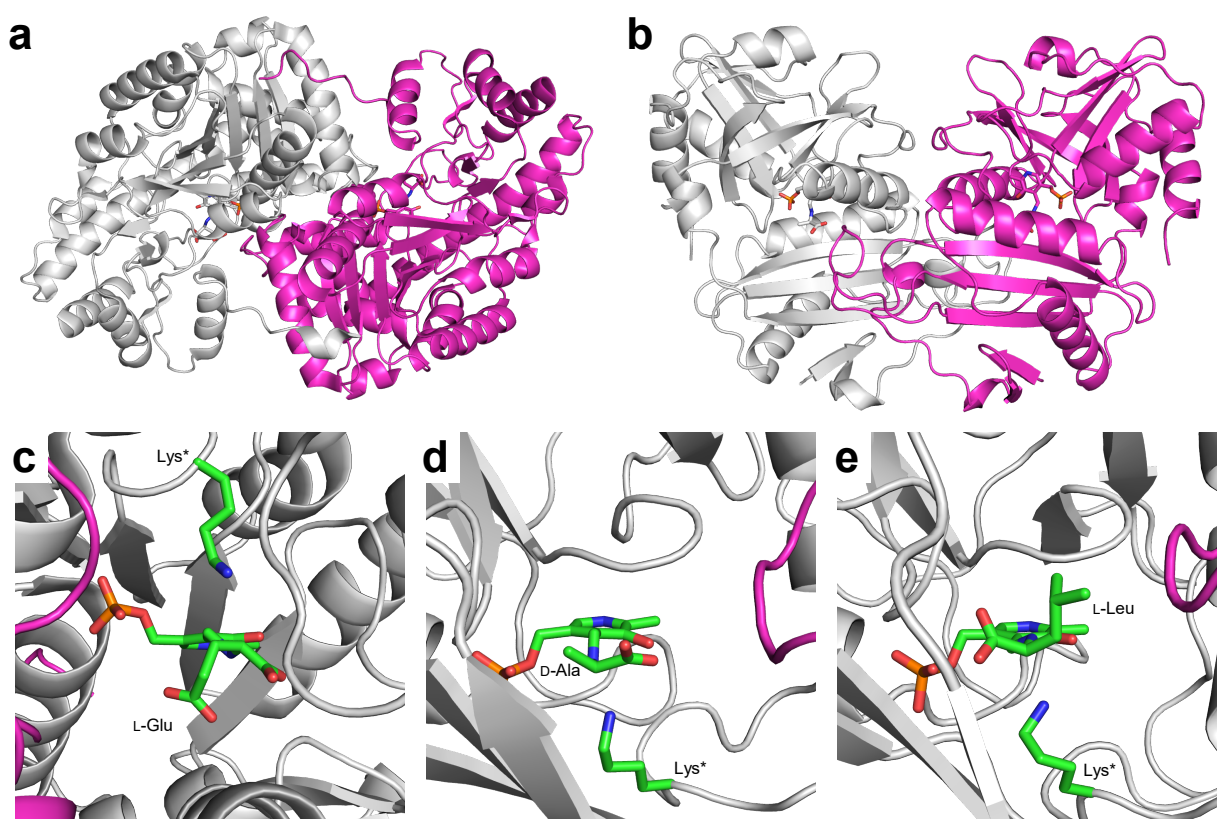


Figure 1.8. Fold Type I and IV aminotransferases. (a) Crystal structure of *E. coli* AAT, which adopts a type I fold (PDB ID: 1X28). (b) Crystal structure of *Bacillus* sp. YM-1 DAAT, which adopts a type IV fold (PDB ID: 3DAA). (c) Active site of *E. coli* AAT. (d) Active site of *Bacillus* sp. YM-1 DAAT. (e) Active site of *E. coli* BCAT (PDB ID: 1I1L). Chain A is shown in white, chain B is shown in magenta, and both the catalytic lysine (Lys*) and the substrate-PLP complex are shown as green sticks. The substrate is labelled next to its side chain.

1.4.2 Mechanism

The transamination reaction of aminotransferases follows a PLP-dependent one-site bi bi ping mechanism that comprises various intermediate PLP states along the reaction coordinate (Figure 1.9). The unliganded aminotransferase contains a PLP cofactor that forms a Schiff base with the catalytic lysine in the active site, to form the internal aldimine state. The amino donor, an amino acid, will enter the active site and form a Michaelis complex, which undergoes transaldimination to yield the external aldimine and release the lysine. Then, the external aldimine undergoes a lysine-mediated 1,3-prototropic shift to form the ketimine. A water hydrolyzes the imine bond to form the carbinolamine, which is then hydrolyzed further to release an α -keto acid and form the pyridoxamine phosphate PLP state. The pyridoxamine phosphate then binds the amino acceptor co-substrate, an α -keto acid, and the reverse reaction happens to release an amino acid, regenerate the internal aldimine and restart the catalytic cycle. Conflicting reports disagree in whether the concerted 1,3-prototropic shift happens in all aminotransferases, or if the H_α of the substrate is shifted to the C_{4A} carbon on PLP via two distinct steps and the quinonoid intermediate (Figure 1.10). However, the quinonoid appears to be an intermediate off of the reaction pathway in many aminotransferases, including *E. coli* AAT and *Mycobacterium tuberculosis* BCAT.¹¹⁰⁻¹¹²

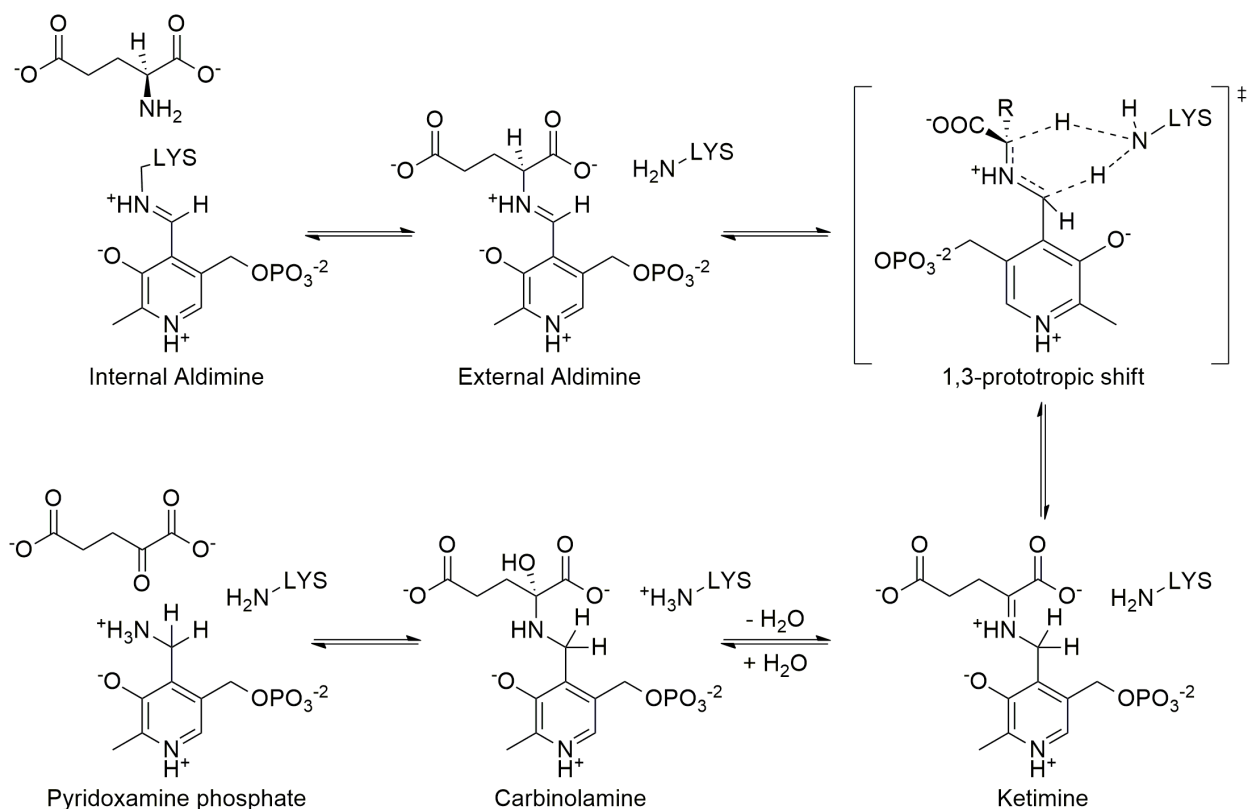


Figure 1.9. Mechanism for the half reaction of aminotransferases.

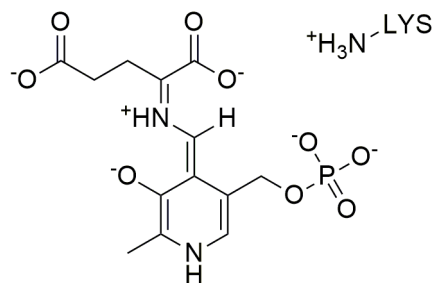


Figure 1.10. Quinonoid intermediate.

1.4.3 Assaying ATs

Aminotransferase activity can be measured using a variety of enzymatic assays such as radiometric methods,¹¹³ chromatography,¹¹⁴ capillary electrophoresis coupled with mass spectrometry,¹¹⁵ amino acid oxidase (e.g., L-glutamate¹¹⁶, D-amino acid¹¹⁷ and L-amino acid¹¹⁷ oxidases) coupled assays, and dehydrogenase (e.g., L-glutamate,¹¹⁸ L-leucine,¹¹⁹ and L-malate¹²⁰ dehydrogenases) coupled assays. In this thesis, aminotransferase activity was measured using the L-glutamate dehydrogenase (GDH) assay. The GDH assay was selected because it can be used to

assay any of the >160 currently known D- or L-amino acid specific aminotransferases that use the α -ketoglutarate/glutamate (D- or L-glutamate) substrate/product pair. Additionally, the GDH assay is convenient since it is continuous and colorimetric, and because bovine GDH is commercially available.

In the GDH assay, the L-glutamate produced by the AT transamination reaction is oxidatively deaminated by the dehydrogenase in the presence of NAD^+ to produce α -ketoglutarate and NADH (Figure 1.11), where the increasing NADH concentration can be conveniently monitored at 340 nm ($\epsilon = 6220 \text{ M}^{-1} \text{ cm}^{-1}$). An additional advantage of the GDH assay is that it has been developed for the high-throughput screening of aminotransferase mutant libraries in clarified cell lysates,¹²¹ which will be useful for testing computationally designed libraries.

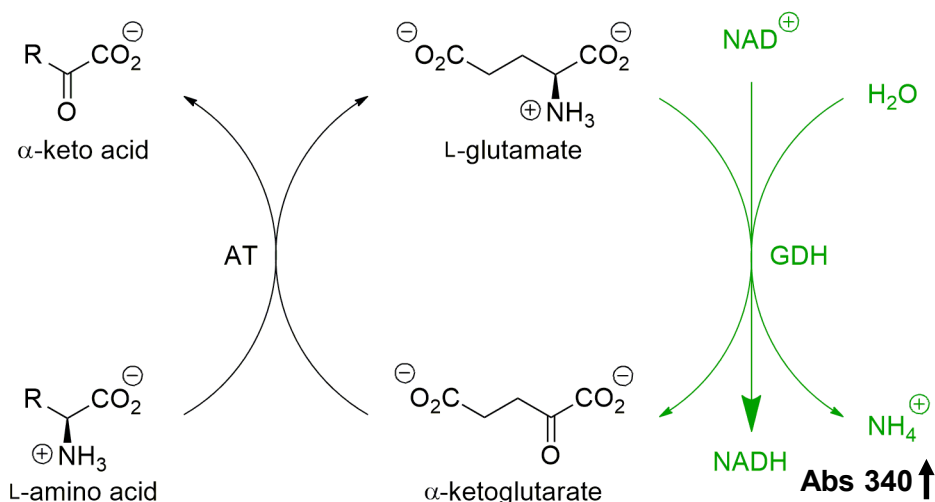


Figure 1.11. GDH coupled enzyme assay. The transamination reaction catalyzed by L- α -aminotransferases (AT) produces L-glutamate, which is oxidatively deaminated by GDH (green) in the presence of NAD^+ . The coupled reaction generates NADH, which results in an absorbance increase at 340 nm.

1.4.4 Dual specificity of ATs

The nature of the transamination reaction necessitates that ATs must be able to bind two amino acids with side chains of different shapes and properties in the same active site. Consequently, different aminotransferases have developed various strategies to display this dual

specificity. Most aminotransferases use rearrangement of the hydrogen bonding and salt bridging network to bind both substrates in the same site. An example is with *Paracoccus denitrificans* aromatic amino acid aminotransferase (AroAT), which reacts with L-glutamate and aromatic L-amino acids.¹²²⁻¹²³ Upon binding the acidic substrate analog maleate, an arginine (R292) forms a bidentate hydrogen bond/ion pair with what would be the side chain carboxylate group (Figure 1.12a). Alternatively, the arginine reorients to point out of the active site when the aromatic substrate analog 3-phenylpropionate binds (Figure 1.12b). To accommodate this arginine switch, there is also a large change in the hydrogen bonding network due to the different arginine orientation.

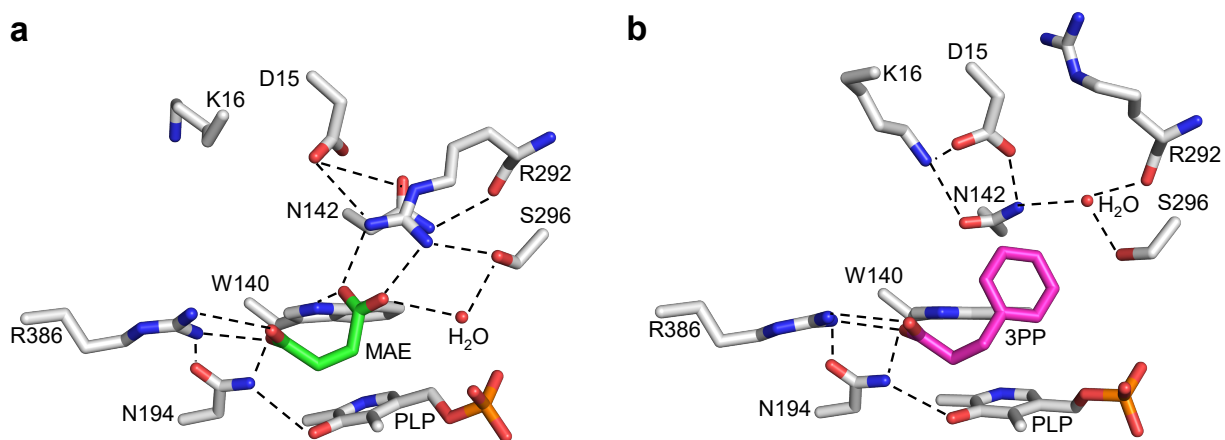


Figure 1.12. Active site of *P. denitrificans* AroAT complexes showing the hydrogen bond and salt bridge interactions with (a) maleate, MAE, (PDB ID: 1AY5) and (b) 3-phenylpropionate, 3PP, (PDB ID: 1AY8). Backbone atoms of side chain residues are not shown, except for R292. The active site residues are labelled with 1-letter amino acid code at the C $_{\alpha}$. Selected hydrogen bonds are shown as black dotted lines.

To bind hydrophobic and acidic side chains, the BCAT active site also undergoes hydrogen bonding rearrangement, but to a lesser extent than what is observed in AroAT. The BCAT active site is hydrophobic to accommodate the branched-chain amino acids (L-leucine, L-isoleucine, and L-valine) and contains four hydrophilic residues to bind the glutamate γ -carboxylate group (Figure 1.13).^{88, 108} Because of the simplicity of the dual substrate recognition utilized by BCAT, we chose

BCAT as our model enzyme in Chapter 3 to develop methods to engineer multi-substrate enzyme specificity.

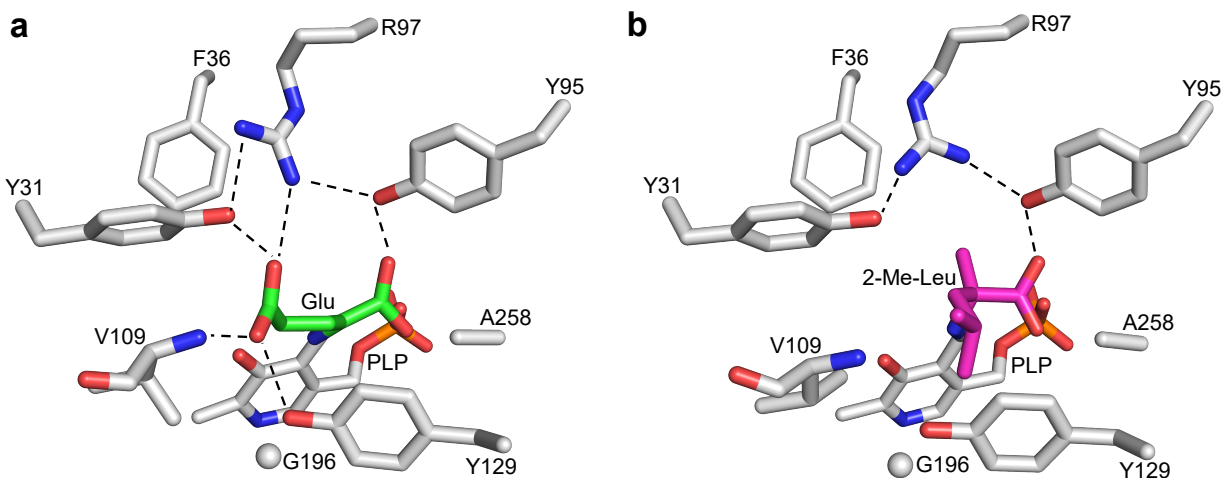


Figure 1.13. Active site of *E. coli* BCAT in complex with various substrates. Hydrogen bond and salt bridge interactions with (a) glutamate (PDB ID: 1IYE) and (b) the L-leucine analog 2-methyl-L-leucine (PDB ID: 1I1L) covalently bound to PLP are shown as dashed lines. Backbone atoms of active site residues are not shown. The active site residues are labelled with 1-letter amino acid code at the C α . The C α of glycine is shown as a sphere.

1.4.5 Domain movement during catalysis

Structural studies have shown that *E. coli* AAT undergoes conformational changes during transamination between L-aspartate and α -ketoglutarate to form oxaloacetate and L-glutamate.^{81, 96-97, 111} The conformational change is due to the small domain moving towards the large domain (Figure 1.14), which closes the active site and shields the substrate from solvent during catalysis. This conformational change is common to other Fold Type I ATs including *P. denitrificans* AroAT¹²² and *Thermus thermophilus* HB8 glutamine AT.¹²⁴ The Fold Type IV *E. coli* BCAT also undergoes a conformational change, but unlike Fold Type I ATs, BCAT does not undergo domain movement. Instead, upon binding substrate, the interdomain loop becomes ordered to close the active site and shield it from solvent.⁸⁸⁻⁸⁹

The conformational equilibrium between the open and closed forms of *E. coli* AAT is known to affect substrate specificity.¹²⁵⁻¹²⁶ In Chapter 4, we exploit this relationship between

conformational equilibrium and substrate specificity, and use AAT as a test case for the development of multistate methods for the modulation of conformational equilibrium of an enzyme.

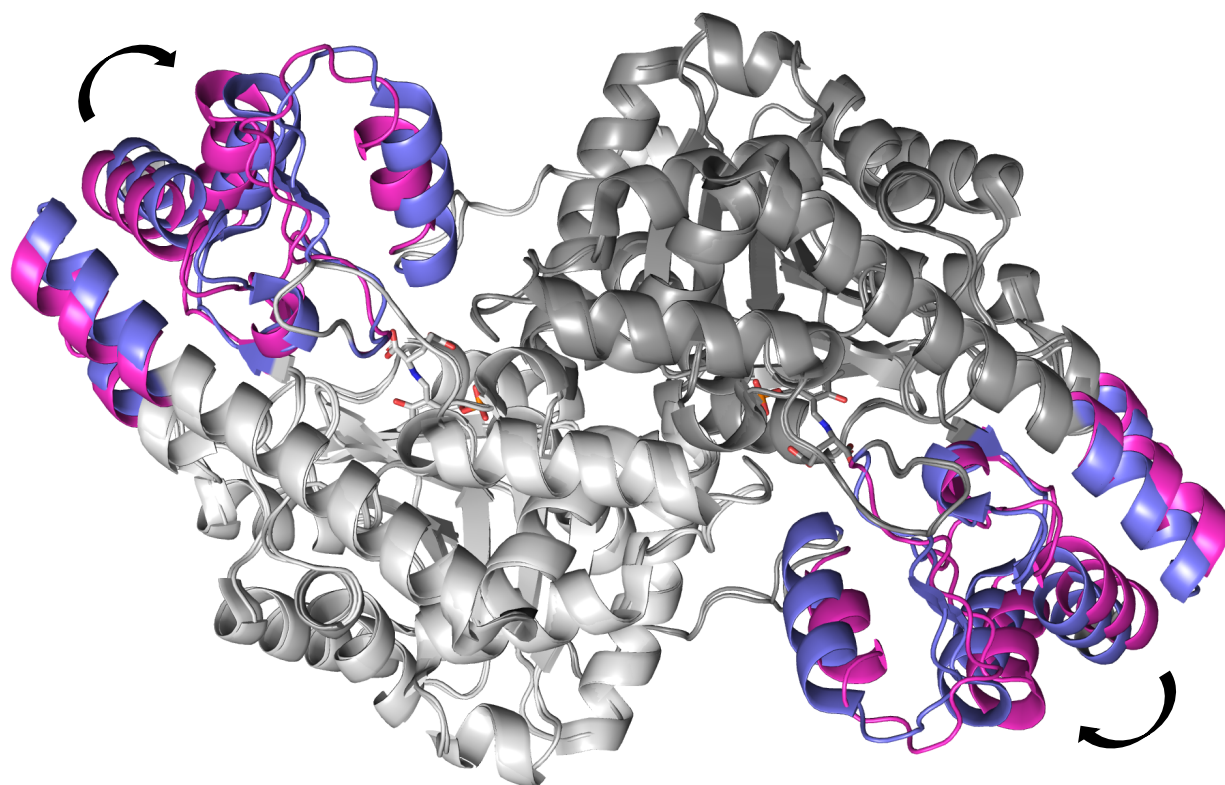


Figure 1.14. AAT domain movement. Superposition of *E. coli* AAT crystal structures in the open (PDB: 1ARS) form and closed (PDB: 1X28) form aligned to the static domains (white and grey). Moving domain is represented by pink (Open) and blue (Closed). PLP is shown as sticks. The static domain of Chain A is in white and the static domain of Chain B is in grey. The moving domains of Chain A and Chain B are on the left and right, respectively. The curved arrows demonstrate the direction of domain movement upon substrate binding.

1.5 Thesis objectives

In Chapter 1, we discussed the application of computational enzyme redesign for engineering enzymes and the necessity of treating enzymes as multistate systems. We emphasized that all substrates required for catalysis should be modeled to successfully redesign multi-substrate enzymes. In Chapter 3, we used BCAT as a test case to develop multistate methods for the redesign

of multi-substrate enzyme specificity. For the redesign of the BCAT active site (Chapter 3), rotamer libraries were required to model the substrate side chain in the external aldimine state. However, these are not available. In Chapter 2, described a methodology to generate a substrate rotamer library for any small molecule. We also discussed the importance of conformational dynamics for enzyme catalysis. In Chapter 4, we used AAT as a test case to develop multistate computational protein design methods for the modulation of conformational equilibrium of an enzyme to affect catalytic activity and substrate specificity. In Chapter 5, we discussed the implications of the developed methodologies described in this thesis, how it impacts new enzyme (re)designs, and future outlooks for the field of computational enzyme (re)design.

Chapter 2: Rotamer optimization calculations to recapitulate aminotransferase active site configurations with different rotamer libraries

Antony D. St-Jacques & Roberto A. Chica

2.1 Statement of contribution

Antony D. St-Jacques developed the rotamer library generation method, performed the CPD calculations, and prepared the chapter. Dr. Roberto A. Chica helped with the design of this study and editing.

2.2 Introduction

Rotamers are discrete sets of allowed conformations that correspond to local minima on potential energy maps of organic molecules such as amino acids.^{14, 30} Rotamers vary from one another by rotations around side chain torsion (χ) angles, but not the backbone. Rotamer libraries that contain rotamers for each type of proteinogenic residue are used for CPD, structure determination, and structure prediction to approximate side-chain amino acid flexibility. Rotamer libraries are often derived from statistical analysis of side-chain conformations of proteins and can be backbone-independent or backbone-dependent.^{14, 30-32} As these rotamer libraries are derived from amino-acid conformers in protein structures, they may not be suitable to model amino acids that are not bound to protein backbones such as amino acids covalently bound to PLP in aminotransferase (AT) active sites.

The objective of Chapter 3 is to use CPD to redesign the active site of *Escherichia coli* Branched Chain Aminotransferase (BCAT) to catalyze the transamination of L-histidine using α -ketoglutarate as an amino acceptor. This design involves modeling the side chains of both active-site residues and amino-acid substrates (e.g., L-glutamate and L-histidine) covalently bound to PLP. Since the backbone atoms of the substrates bound to PLP in the external aldimine and

ketimine states are different from those of a protein backbone, it is unclear how well the rotamer libraries derived from protein crystal structures can be used to model the substrates. To gain confidence that we could successfully redesign the active site of any aminotransferase for altered substrate specificity, our goal was to test the ability of different rotamer libraries to recapitulate the side-chain conformations of residues and substrates observed in crystal structures of aminotransferase active sites.

2.3 Methods

Throughout this thesis, we are interested in repacking the active-site residues of aminotransferases that interact with the side chains of various amino acid substrates. To verify the effect of various rotamer libraries on the accuracy of side-chain repacking, calculations were performed in which side-chain rotamers were optimized for only active-site residues or substrates, or for both. Rotamer combinations were evaluated using a four-term potential energy function consisting of a van der Waals term from the Dreiding II force field³⁸ with atomic radii scaled by 0.9, a direction sensitive hydrogen-bond term with well depth at 8.0 kcal mol⁻¹,⁴⁰ an electrostatic energy term modeled using Coulomb's law with a distance dependent dielectric of 10, and an occlusion-based solvation penalty term.³⁷ Fast and Accurate Side-Chain Topology and Energy Refinement (FASTER)²³ was used to optimize the rotamer configuration of the active sites without altering the amino acid identity. Importantly, the crystallographic conformer for each designed active-site residue or substrate was not allowed in these calculations to avoid biasing the calculation outcome to the crystallographic conformation.

To keep discussion of these designs clear and concise, each rotamer optimization design will use the nomenclature of SC^xSub^y. *x* states the rotamer library used to model the active-site side chains (SC), which can be “bbdep”, “bbind”, or “fix”, where “bbdep” and “bbind” refer to the

Dunbrack 2002 backbone dependent (bbdep) and independent (bbind) rotamer libraries (with expansions of ± 1 standard deviation around χ_1 and χ_2), respectively, and “fix” means the side chains were kept in their crystallographic configuration and not altered during the calculation. Likewise, y states the rotamer library used to model the substrate side chain (Sub), which can be the bbind rotamer library (bbind), our new Targeted substrate rotamer library (sub), or the substrate can be kept fixed in its crystallographic configuration (fix). The crystallographic rotamer was not allowed to be sampled in rotamer optimization calculations, unless the rotamer library was specified to be fixed. As an example, a design labelled $SC^{bbind}Sub^{fix}$ would indicate active-site side-chain rotamer optimization with the Dunbrack 2002 backbone independent rotamer library while the substrate was kept fixed in its crystallographic configuration. The structure that was used as input to a rotamer optimization design will be referred to as the crystallographic structure and the structure that resulted from a rotamer optimization design will be referred to as the designed structure.

2.4 Results and Discussion

2.4.1 Structure selection

Aminotransferases adopt two different general folds, termed Fold Type I and IV of the PLP-dependent enzymes.⁸¹ We selected one aminotransferase from each fold type to evaluate whether the different backbone templates would affect the outcome of our prediction of rotamers for active-site residues and substrates. *E. coli* BCAT adopts a Fold Type IV conformation, and we selected *E. coli* AAT to represent a Fold Type I aminotransferase. As an amino acceptor, α -ketoglutarate is converted by AAT to L-glutamate using L-aspartate as an amino donor, while BCAT does the same but with branched-chain amino acids (e.g., L-leucine) as amino donors.

Therefore, crystal structures of these amino acceptors and donors covalently bound to PLP were selected for each aminotransferase for active-site recapitulation.

For BCAT active-site recapitulation designs, we selected a crystal structure containing 2-methyl-L-leucine covalently bound to PLP (PDB ID: 1I1L⁸⁹) as an external aldimine complex analog and a crystal structure containing L-glutamate covalently bound to PLP as a ketimine complex (PDB ID: 1IYE⁸⁸). For AAT active-site recapitulation designs, we selected a crystal structure containing 2-methyl-L-aspartate covalently bound to PLP as an external aldimine complex analog (PDB ID: 1ART¹²⁷) and a crystal structure containing L-glutamate covalently bound to PLP as a ketimine complex (PDB ID: 1X28⁹⁷). When discussing the substrate or PLP moiety of the substrate-PLP complexes, the substrate moiety will be referred to as the amino acid form (i.e., Glu to represent the glutamate/ α -ketoglutarate-bound substrate moiety), the word “moiety” will be left out for brevity, and the substrates can be identified by the superscript PLP (i.e., Glu^{PLP}, Leu^{PLP}, Asp^{PLP}).

The four crystal structures (i.e., 1I1L, 1IYE, 1ART, 1X28) were obtained from the PDB and subsequently prepared (e.g., hydrogens were added, missing residues were modeled) using the software package Molecular Operating Environment (MOE).¹²⁸ The AMBER99 forcefield¹²⁹ was used with a combined explicit and implicit reaction field solvent model. The 2-methyl group of the 2-methyl-L-leucine and 2-methyl-L-aspartate moieties were replaced with a hydrogen atom to form external aldimine complexes. Hydrogens were added using the Protonate 3D utility within MOE,¹²⁸ and the structures were solvated in a rectangular box of water with counter-ions (Na⁺ and Cl⁻) under periodic boundary conditions with a box cutoff of 6 Å once protonation states and atomic hybridizations of PLP-substrate were verified to be either external aldimines or ketimines. The protonation state and hydrogen bonds of PLP were adjusted to be consistent with NMR

studies.¹³⁰ The pyridine and imine nitrogens were protonated, while all phosphate oxygens and the phenol oxygen were deprotonated. The ϵ -amino group of the catalytic lysine is modeled as neutral to remain consistent with the catalytic cycle. The energy of prepared structures was minimized using conjugate gradient minimization to a gradient below 7 kcal mol⁻¹ Å⁻¹. Water molecules and counter-ions were removed to yield the input crystallographic structures, which are named using a [substrate]-[aminotransferase] notation, where the Glu-BCAT, Leu-BCAT, Glu-AAT, and Asp-AAT structures refer to prepared structures derived from the 1IYE, 1I1L, 1X28, and 1ART PDB structures, respectively.

2.4.2 Description of active sites

To identify the best combination of rotamer libraries to use for modelling active-site residues and substrates in our design calculations, we investigated the ability of our active-site recapitulation designs to rescue the crystallographic active-site configuration and specific hydrogen bonds within the active site. Geometric criteria for determining the presence or absence of hydrogen bonds are adapted from McDonald & Thornton.¹³¹ A hydrogen bond is present if the hydrogen bond donor and acceptor distance is < 3.9 Å; the hydrogen and hydrogen bond acceptor distance is < 2.5 Å; the angle between the hydrogen bond donor, hydrogen, and hydrogen bond acceptor is > 90 °; the angle between the hydrogen, hydrogen bond acceptor, and an atom covalently bound to the acceptor is > 90 °; and the angle between the hydrogen bond donor, acceptor, and an atom covalently bound to the acceptor is > 90 °. To select important hydrogen bonds to consider within designs, we inspected the active site of the crystallographic structures.

Within the BCAT active site (Figure 2.1), the γ -carboxylate group of Glu^{PLP} forms hydrogen bonds to the side chains of Y31, R97, and Y129, as well as the backbone amide of V109. Also, R97 is involved in a hydrogen bonding network between Glu^{PLP}, Y31, and Y95, where the

latter residue forms a hydrogen bond to the α -carboxylate group of Glu^{PLP}. A hydrogen bond network between Y31, Y95, and R97 is also present in the Leu-BCAT structure with Leu^{PLP}. When optimizing the side-chain rotamers of BCAT active-site residues, we also included an additional six residues (Y31, R97, V109, W126, Y129, and E197), which are included in the CPD design calculations performed in Chapter 3.

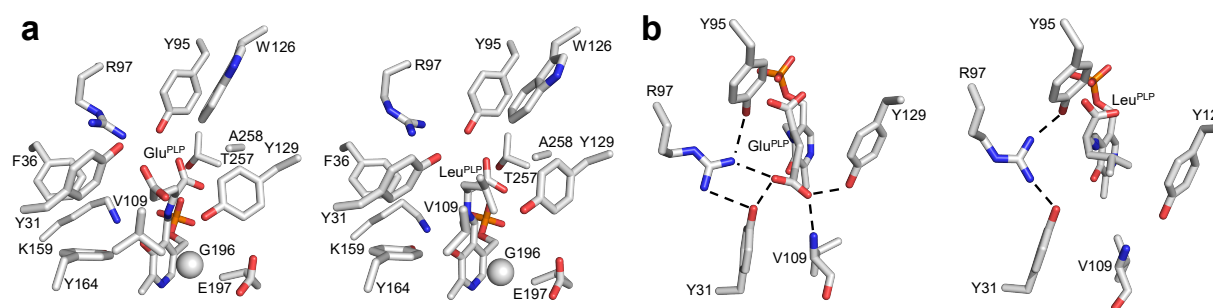


Figure 2.1. *E. coli* BCAT active site. In both panels, the Glu-BCAT and Leu-BCAT structures are on the left and right, respectively. (a) Residues within 3 Å of the substrate moiety of the PLP complex and residue E197 are shown as sticks. The C α of G196 is shown as a sphere. (b) Selected hydrogen bonds between R97 or the Glu^{PLP} γ -carboxylate group and active-site residues are shown as black dashed lines. Backbone atoms of side chain residues are not shown except for V109 in panel (b). The active site residues are labelled with 1-letter amino acid code at the C α .

Within the AAT active site (Figure 2.2), R292 and W140 can form hydrogen bonds with the γ -carboxylate group of Glu^{PLP} or with the β -carboxylate group of Asp^{PLP}. They were thus included in our rotamer-optimization calculations. AAT active site residues with atoms within 3 Å from the side chains of either the Glu or Asp substrates (I17, L18, I37, Y70, W140, R292) were also included in rotamer optimization calculation, along with active site residues T109, S296, and N297, which were included because they are active site residues that have been shown to be important for substrate specificity⁵ (e.g., T109, N297) or form hydrogen bonds with R292 (e.g., S296).

Since our goal is to evaluate the ability of our rotamer-optimization procedure using different rotamer libraries to recapitulate the crystallographic side-chain conformation of the

substrates, catalytic lysine (residue 159 in BCAT; residue 258 in AAT) and active-site residues that hydrogen bond with the α -carboxylate group of substrates were kept in their crystallographic configurations.

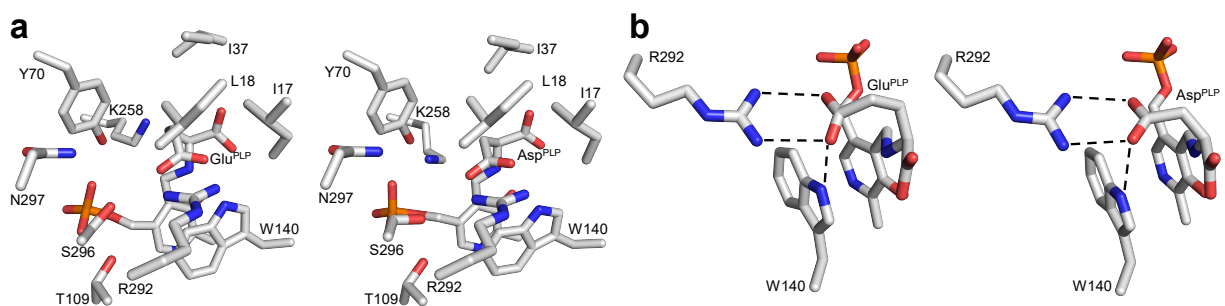


Figure 2.2. *E. coli* AAT active site. In both panels, the Glu-AAT and Asp-AAT structures are on the left and right, respectively. (a) Residues within 3 Å of the side chain of the substrate moiety of the PLP complex and active site residues N297, S296, and T109 are shown as sticks. (b) Selected hydrogen bonds between the substrate side chain carboxylate groups and active site residues in the crystallographic structures. The active site residues are labelled with 1-letter amino acid code at the C $_{\alpha}$. The active site residues are labelled with 1-letter amino acid code at the C $_{\alpha}$. Selected hydrogen bonds are shown as black dotted lines.

2.4.3 Recovery of active-site residue rotamer configurations

Although backbone-dependent rotamer libraries have been shown to more accurately predict side-chain conformations in protein design calculations than backbone-independent libraries,^{30, 33-34} it was unclear which rotamer library would be more accurate for recapitulating substrate rotamers in the aminotransferase active sites. In order to select which rotamer library (bbdep or bbind) to use to model the active-site residues, we first performed SC^{bbdep}Sub^{fix} and SC^{bbind}Sub^{fix} rotamer optimization designs to recapitulate the active-site residue conformations of the Glu-BCAT, Leu-BCAT, Glu-AAT, and Asp-AAT crystallographic structures.

On both BCAT structures, SC^{bbind}Sub^{fix} rotamer optimization designs (Figure 2.3b) resulted in designed structures that contained rotamer configurations with lower all atom active-site side chain RMSD_{SC} (Table 2.1) to the crystallographic structure than SC^{bbdep}Sub^{fix} designs (Figure 2.3a), where the _{SC} subscript denotes that only the atoms of rotamer-optimized active-site

residue were included in RMSD calculations. Notably, the hydrogen bond network between Y31, Y95, and R97 present in the crystallographic structures (Figure 2.1b) was absent in the designed structures that resulted from the $SC^{bbdep}Sub^{fix}$ designs in the presence of both Leu^{PLP} and Glu^{PLP} substrates (Figure 2.3c). Instead, R97 adopted rotamers that were only able to hydrogen bond to either the Glu^{PLP} γ -carboxylate group in the Glu-BCAT designed structure or to the phenolic hydroxyl group of Y31 in the Leu-BCAT designed structure. Rotamer optimization with the *bbdep* library was unable to recapitulate the R97 crystallographic conformer because there is no arginine rotamer within the correct $(\Phi \times \Psi)$ bins ($\Phi = -120^\circ$, $\Psi = +90^\circ$ for Glu-BCAT, $\Phi = -120^\circ$, $\Psi = +100^\circ$ for Leu-BCAT) that is similar enough to the crystallographic R97 conformers, which prevents placement of the guanidinium group in the same position. Due to the specialized environment of active sites, it was not unexpected that an active site residue (i.e., R97) would adopt a rotamer configuration that is rare for its $(\Phi \times \Psi)$ bin, and thus a rotamer that is similar to the rare rotamer configuration would not be included in the *bbdep* library. On the other hand, the *bbind* library contained rotamers similar to the R97 crystallographic conformers. Thus, the designed structures that resulted from the $SC^{bbind}Sub^{fix}$ rotamer optimization designs contained hydrogen bond networks between Y31, Y95, and R97 (Figure 2.3d), although the rotamer-optimized R97 guanidinium group in the Leu-BCAT designed structure was rotated approximately 60° along an axis through the C_ζ atom that is perpendicular to the guanidinium group relative to the R97 crystallographic conformer. Additionally, the χ_1 angle of W126 in the Glu-BCAT designed structure that resulted from the $SC^{bbind}Sub^{fix}$ design was rotated 134.3° relative to the W126 crystallographic rotamer.

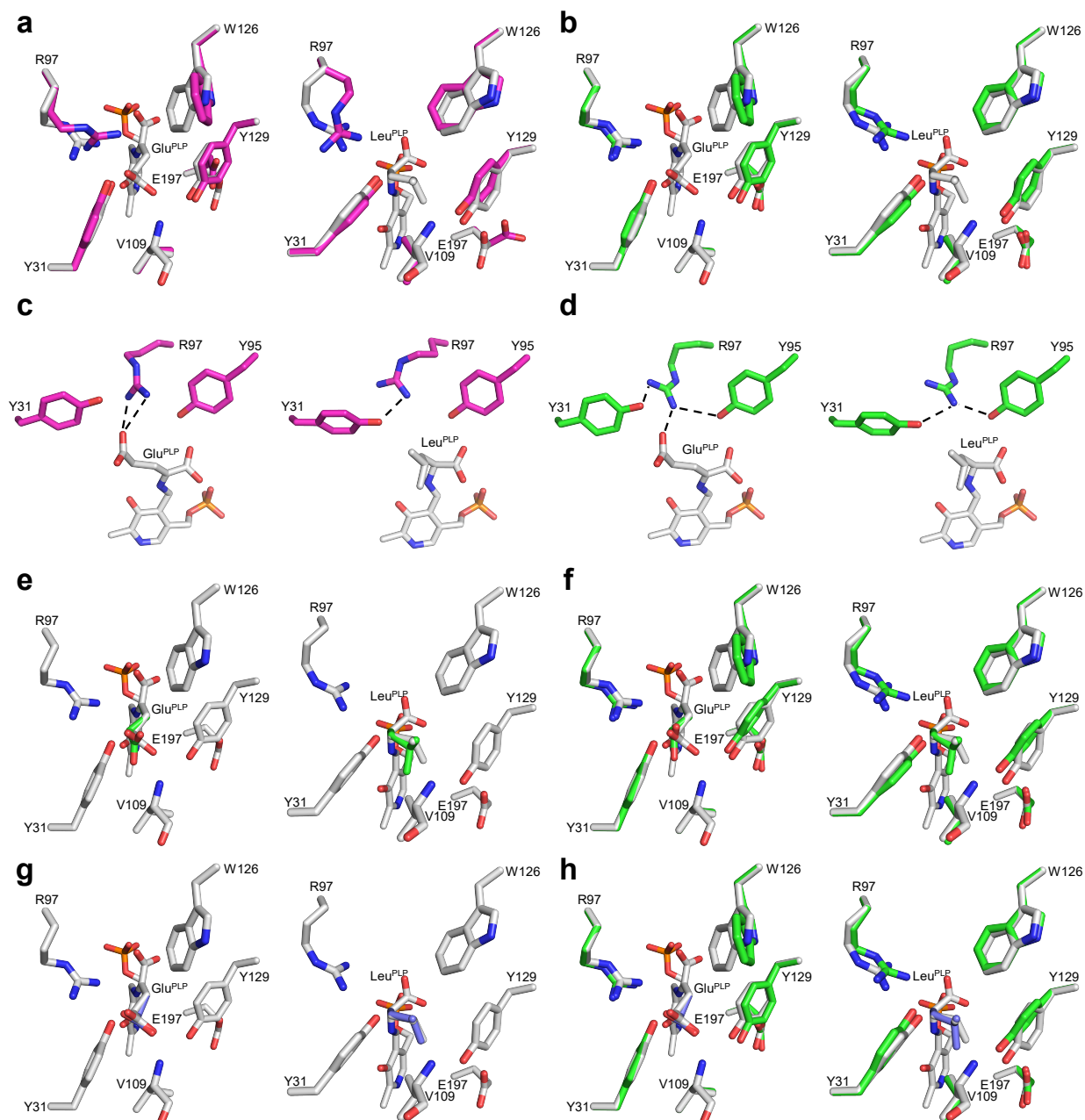


Figure 2.3. Comparison of BCAT crystal and designed structures. (a) $SC^{\text{bbdep}}\text{Sub}^{\text{fix}}$, (b) $SC^{\text{bbind}}\text{Sub}^{\text{fix}}$, (e) $SC^{\text{fix}}\text{Sub}^{\text{bbind}}$, (f) $SC^{\text{bbind}}\text{Sub}^{\text{bbind}}$, (g) $SC^{\text{fix}}\text{Sub}^{\text{sub}}$, and (h) $SC^{\text{bbind}}\text{Sub}^{\text{sub}}$ rotamer optimization designs. Residues that were fixed during design are represented as white sticks, and side chains optimized with either the bbdep, bbind or Targeted rotamer libraries are represented as pink, green or blue sticks, respectively. Panels (c) and (d) display R97 hydrogen bond networks from the $SC^{\text{bbdep}}\text{Sub}^{\text{fix}}$ and $SC^{\text{bbind}}\text{Sub}^{\text{fix}}$ rotamer optimization designed structures, respectively. In all panels, structures on the left and right are the Glu-PLP ketimine complex (Glu^{PLP}) modeled on the Glu-BCAT template structure, or the Leu-PLP external aldimine complex (Leu^{PLP}) on the Leu-BCAT template structure, respectively. Backbone atoms of side chain residues are not shown

except for V109. The active site residues are labelled with 1-letter amino acid code at the C α . Selected hydrogen bonds are shown as black dashed lines.

For the recapitulation of active-site residue rotamer configurations in the AAT active site, both the SC^{bbdep}Sub^{fix} and SC^{bbind}Sub^{fix} rotamer-optimization designs yielded hydrogen bonds between the side chain carboxylate group of the substrates (i.e., Glu^{PLP} and Asp^{PLP}) and R292. Interestingly, the Asp-AAT designed structures that resulted from both rotamer-optimization designs contained side-chain configurations that had identical side-chain RMSD_{SC} from the crystallographic configurations, although the rotamers selected are different (Figure 2.4a,b). Similarly, rotamer optimization of the Glu-AAT active site side-chain rotamers with the bbdep library resulted in a comparable (within 0.01 Å) side chain RMSD_{SC} to the designed structure that resulted from the use of the bbind library (Table 2.1).

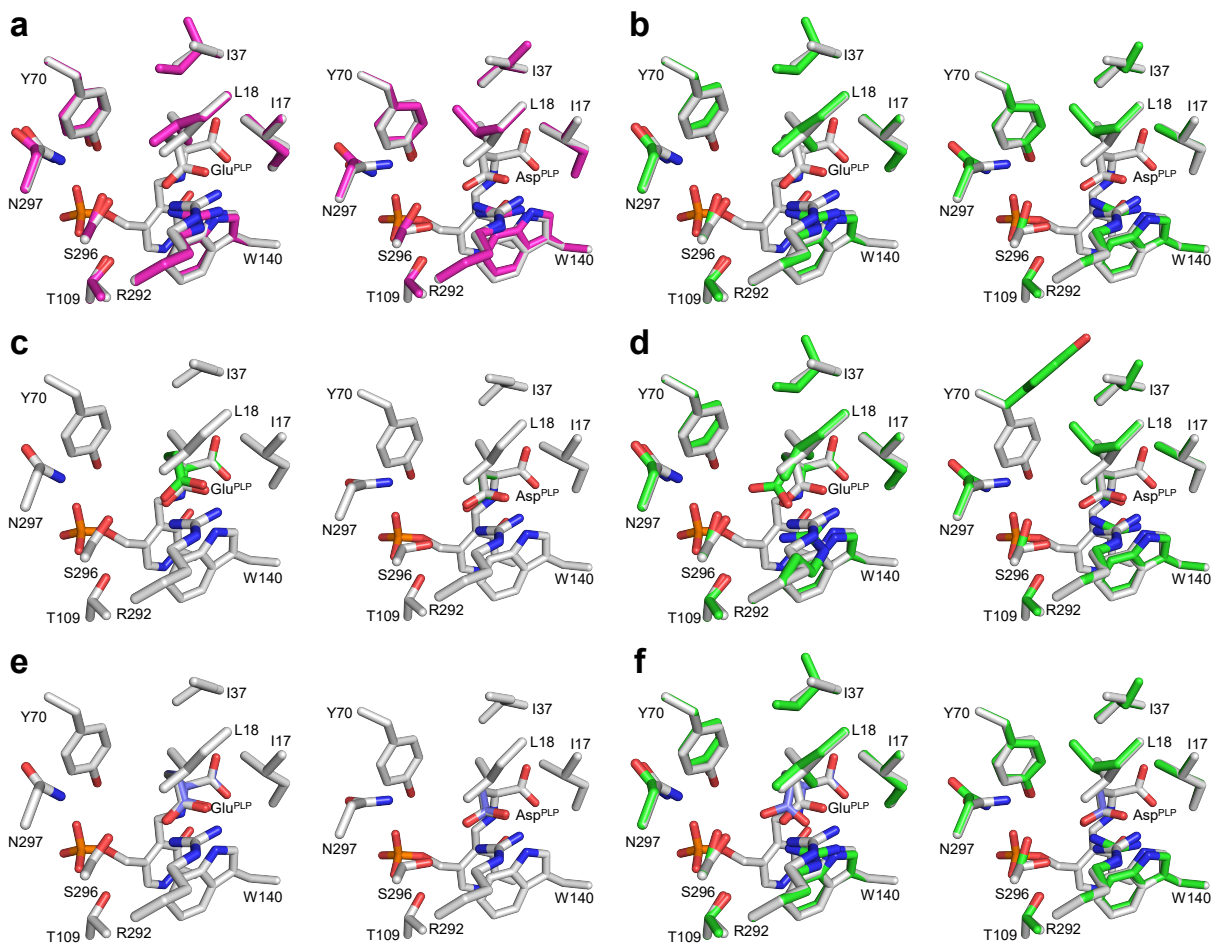


Figure 2.4. Comparison of AAT crystallographic and designed structures. (a) $SC^{bbdep}Sub^{fix}$, (b) $SC^{bbind}Sub^{fix}$, (c) $SC^{fix}Sub^{bbind}$, (d) $SC^{bbind}Sub^{bbind}$, (e) $SC^{fix}Sub^{sub}$, and (f) $SC^{bbind}Sub^{sub}$ rotamer optimization designs, superimposed with the crystallographic structures. Residues that were fixed during design are represented as white sticks, and side chains optimized with either the bbdep, bbind or Targeted rotamer libraries are represented as pink, green or blue sticks, respectively. In all panels, structures on the left and right are the Glu-PLP ketimine complex (Glu^{PLP}) modeled on the Glu-AAT template structure, or the Asp-PLP external aldimine complex (Asp^{PLP}) on the Asp-AAT template structure, respectively. Backbone atoms of side chain residues are not shown. The active site residues are labelled with 1-letter amino acid code at the C_{α} .

Our results indicate that rotamer optimization designs with the bbind library to model the active site side-chains of both BCAT templates result in a lower side-chain $RMSD_{SC}$ from their respective crystallographic structures and that rotamer optimization designs using both rotamer libraries perform similarly to recapitulate the active site rotamer configurations of AAT. Thus, for all further designs, active site residues were modeled with the bbind library.

Table 2.1. The RMSD of rotamer optimized aminotransferase designed structures from their respective crystallographic structures.

Structure	Rotamer Library		RMSD (Å)		
	SC [*]	Sub [†]	SC ^{**}	Sub ^{††}	All ^{‡‡}
Leu-BCAT	bbdep [‡]	fix [§]	1.61	0.00	1.49
	bbind [¶]	fix	1.17	0.00	1.09
	fix	bbind	0.00	0.85	0.32
	bbind	bbind	1.20	0.85	1.16
	fix	sub [#]	0.00	0.21	0.08
	bbind	sub	0.82	0.52	0.79
Glu-BCAT	bbdep	fix	1.89	0.00	1.79
	bbind	fix	1.68	0.00	1.58
	fix	bbind	0.00	0.54	0.18
	bbind	bbind	1.71	0.50	1.62
	fix	sub	0.00	0.36	0.12
	bbind	sub	1.66	0.36	1.57
Asp-AAT	bbdep [‡]	fix [§]	1.13	0.00	1.09
	bbind [¶]	fix	1.13	0.00	1.09
	fix	bbind	0.00	0.10	0.03
	bbind	bbind	2.08	0.10	2.00
	fix	sub [#]	0.00	0.04	0.01
	bbind	sub	1.13	0.05	1.09
Glu-AAT	bbdep	fix	1.56	0.00	1.49
	bbind	fix	1.57	0.00	1.50
	fix	bbind	0.00	0.19	0.06
	bbind	bbind	1.64	0.78	1.58
	fix	sub	0.00	0.08	0.02
	bbind	sub	1.57	0.47	1.51

^{*}Rotamer library used to model protein side chains. [†]Rotamer library used to model substrate side chain. [‡]Dunbrack 2002 backbone-dependent library with expansions of ± 1 standard deviation around χ_1 and χ_2 . [§]The conformers were fixed. [¶]Dunbrack 2002 backbone-independent library with expansions of ± 1 standard deviation around χ_1 and χ_2 . [#]Targeted substrate rotamer library prepared in this manuscript. ^{||}The difference in RMSD is due to the Tyr70 conformer change. Without Tyr70, the RMSD of both instances of this symbol are 1.15 Å. ^{**}Only the atoms of rotamer-optimized active-site residue were included in RMSD calculations. ^{††}Only substrate atoms were included in RMSD calculations. ^{‡‡}Rotamer-optimized active-site residue atoms and substrate atoms were included in RMSD calculations.

2.4.4 Recovery of substrate rotamers

Next, we tested how well our rotamer-optimization procedure using the bbind library could recapitulate the substrate side-chain conformer in SC^{fix}Sub^{bbind} designs with both Glu^{PLP} and Leu^{PLP} for BCAT, and Glu^{PLP} and Asp^{PLP} for AAT.

In the BCAT designed structures that resulted from the SC^{fix}Sub^{bbind} designs (Figure 2.3e), χ angles of the Leu^{PLP} rotamer ($\chi_1 = +57.9^\circ$, $\chi_2 = +149.8^\circ$) were within 25° of the crystallographic conformer dihedral angles ($\chi_1 = +82.1^\circ$, $\chi_2 = +171.7^\circ$). Likewise, the χ_1 and χ_2 angles of the Glu^{PLP} rotamer ($\chi_1 = -55.7^\circ$, $\chi_2 = +179.3^\circ$) were within 20° of the corresponding crystallographic rotamer dihedral angles ($\chi_1 = -38.1^\circ$, $\chi_2 = -171.3^\circ$). However, the χ_3 angle of the Glu^{PLP} rotamer ($\chi_3 = +56.0^\circ$) was rotated 36.4° relative to the crystallographic conformer ($\chi_3 = +92.4^\circ$), which prevented the γ -carboxylate from making the hydrogen bond to Y129 that is formed with the crystallographic configuration. We also ran SC^{bbind}Sub^{bbind} designs to test if different side-chain rotamer combinations of active-site residue could allow Glu^{PLP} to make the hydrogen bond with Y129 that is present in the crystallographic structure (Figure 2.3f), but the γ -carboxylate group of Glu^{PLP} in the designed structure that resulted from the SC^{bbind}Sub^{bbind} design also did not form this hydrogen bond.

For AAT, the side-chain rotamer configurations of both the Asp^{PLP} and Glu^{PLP} substrates were optimized on their respective backbone templates in SC^{fix}Sub^{bbind} designs (Figure 2.4c). The rotamer optimization designs yielded designed structures which contained substrate rotamers that formed hydrogen bonds between the side-chain carboxylate group and the active-site residues R292 and W140 (Figure 2.4). The Asp-AAT designed structure that resulted from the SC^{bbind}Sub^{bbind} design also contained hydrogen bonds between the substrate Asp^{PLP}, R292, and W140 (Figure 2.4d). Additionally, the Asp^{PLP} rotamer of the SC^{bbind}Sub^{bbind} designed structure

displayed the same substrate RMSD_{sub} from the crystallographic Asp^{PLP} rotamer as the Asp^{PLP} rotamer in the $\text{SC}^{\text{fix}}\text{Sub}^{\text{bbind}}$ designed structure (Table 2.1), where the $_{\text{sub}}$ subscript denotes that only substrate atoms were included in RMSD calculations. Because the χ_1 torsion of Y70 switched from gauche (–) in the crystallographic structure to gauche (+) in the Asp-AAT $\text{SC}^{\text{bbind}}\text{Sub}^{\text{bbind}}$ designed structure, the all atom RMSD_{All} of all active-site residues and the substrate is 2.00 Å, where the $_{\text{All}}$ subscript denotes that rotamer-optimized active-site residue atoms and substrate atoms were included in RMSD calculations. Excluding the Y70 from the RMSD_{All} calculation lowered the RMSD_{All} to 1.15 Å.

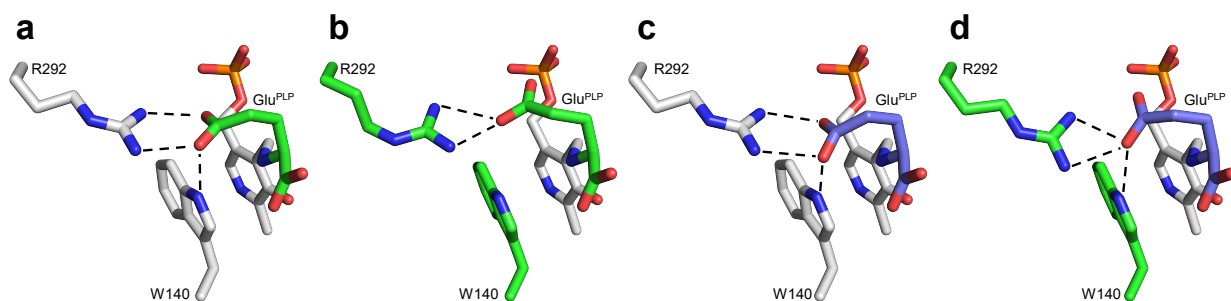


Figure 2.5. Hydrogen-bond networks of AAT designed structures. Selected hydrogen bonds between the substrate side-chain carboxylate group and active-site residues from (a) $\text{SC}^{\text{fix}}\text{Sub}^{\text{bbind}}$, (b) $\text{SC}^{\text{bbind}}\text{Sub}^{\text{bbind}}$, (c) $\text{SC}^{\text{fix}}\text{Sub}^{\text{sub}}$, and (d) $\text{SC}^{\text{bbind}}\text{Sub}^{\text{sub}}$ rotamer-optimization designs. Residues fixed during design are represented as white sticks, and side chains optimized with either the bbdep, bbind or Targeted rotamer libraries are represented as pink, green or blue sticks, respectively. Backbone atoms of active-site residues are not shown. Active-site residues are labelled with 1-letter amino acid code at the C_α . Selected hydrogen bonds are shown as black dotted lines.

The rotamer configuration of the Glu-AAT designed structure yielded from the $\text{SC}^{\text{bbind}}\text{Sub}^{\text{bbind}}$ rotamer-optimization design displayed RMSD_{All} and RMSD_{sub} from the crystallographic rotamer configuration of 1.58 Å and 0.78 Å, respectively (Table 2.1). Importantly, this RMSD_{sub} value is 0.59 Å higher than the equivalent value obtained for the designed structure yielded from the $\text{SC}^{\text{fix}}\text{Sub}^{\text{bbind}}$ rotamer-optimization design, suggesting that the crystallographic rotamers of the active-site residues help to select a more accurate rotamer for the substrate.

Although there are still hydrogen bonds between the γ -carboxylate group of Glu^{PLP} and R292 in the designed structure yielded from the SC^{bbind}Sub^{bbind} design, the hydrogen bond configuration is different: in the crystallographic and SC^{fix}Sub^{bbind} designed structures, the two O _{ϵ} atoms of Glu^{PLP} form hydrogen bonds to the two guanidinium N _{η} atoms of R292 (Figure 2.1b, Figure 2.5a). In contrast, only one of the O _{ϵ} atoms of Glu^{PLP} in the SC^{bbind}Sub^{bbind} designed structure makes hydrogen bonds with both guanidinium N _{η} of R292 (Figure 2.5b). Additionally, the hydrogen bond of the Glu^{PLP} γ -carboxylate group to W140 is not conserved in this designed structure.

2.4.5 Targeted substrate rotamer library

Since we could not consistently and accurately model Glu^{PLP} in either BCAT or AAT using the bbdep or bbind rotamer libraries, we decided to generate our own substrate rotamer library. One option was to use a library of rotamers that have every torsion angle combination in 10 ° increments instead of using the Dunbrack rotamer libraries, which are derived from deposited protein crystal structures. However, the computational cost to evaluate every torsion combination would be high because of the resulting large number of possible rotamers (23,328 rotamers total for glutamate). Also, because internal rotamer energies are not included in CPD energy calculations, highly strained nonphysical conformations could be chosen over physically relevant substrate conformations, which severely reduces computational accuracy.¹⁴ Instead, either torsion terms should be included in energy calculations, which would require altering the energy function, or libraries that only contain low energy, unstrained rotamers should be used.

Since torsion terms are not included in the CPD calculations performed in this thesis, we developed a work flow to (1) generate a library of rotamers for a small molecule, and (2) select low energy rotamers to include in a library (Figure 2.6). To build rotamer libraries for the substrate-PLP complexes in this manuscript, we extracted the coordinates of the entire substrate-PLP

complex and built the rotamers by rotating the substrate side-chain torsion angles. Instead of building rotamers of all dihedral angle combinations, the number of rotamers in the library were reduced by only building those for every canonical dihedral minima, as described in the Dreiding forcefield³⁸ (e.g., $\pm 60^\circ$, 180° for sp^3 - sp^3 bonds), with a deviation of $\pm 10^\circ$ and 20° . Then, the internal energy was calculated *in vacuo* using MOE with the previously described forcefield. The inclusion of PLP in the internal energy calculations allowed the rejection of rotamers that would clash with the PLP. Next, the Boltzmann probability of each rotamer was calculated using Equation 2.1, where, for a given rotamer, i , P_i is the probability, E_i is the internal energy, k_B is the Boltzmann constant, T is the temperature, and n is the number of all rotamers evaluated. The top 95 % of rotamers¹³² ranked by probability were included in what we call the Targeted rotamer library.

$$P_i = \frac{e^{-E_i/k_B T}}{\sum_{i=1}^n e^{-E_i/k_B T}}$$

Equation 2.1. Boltzmann probability.

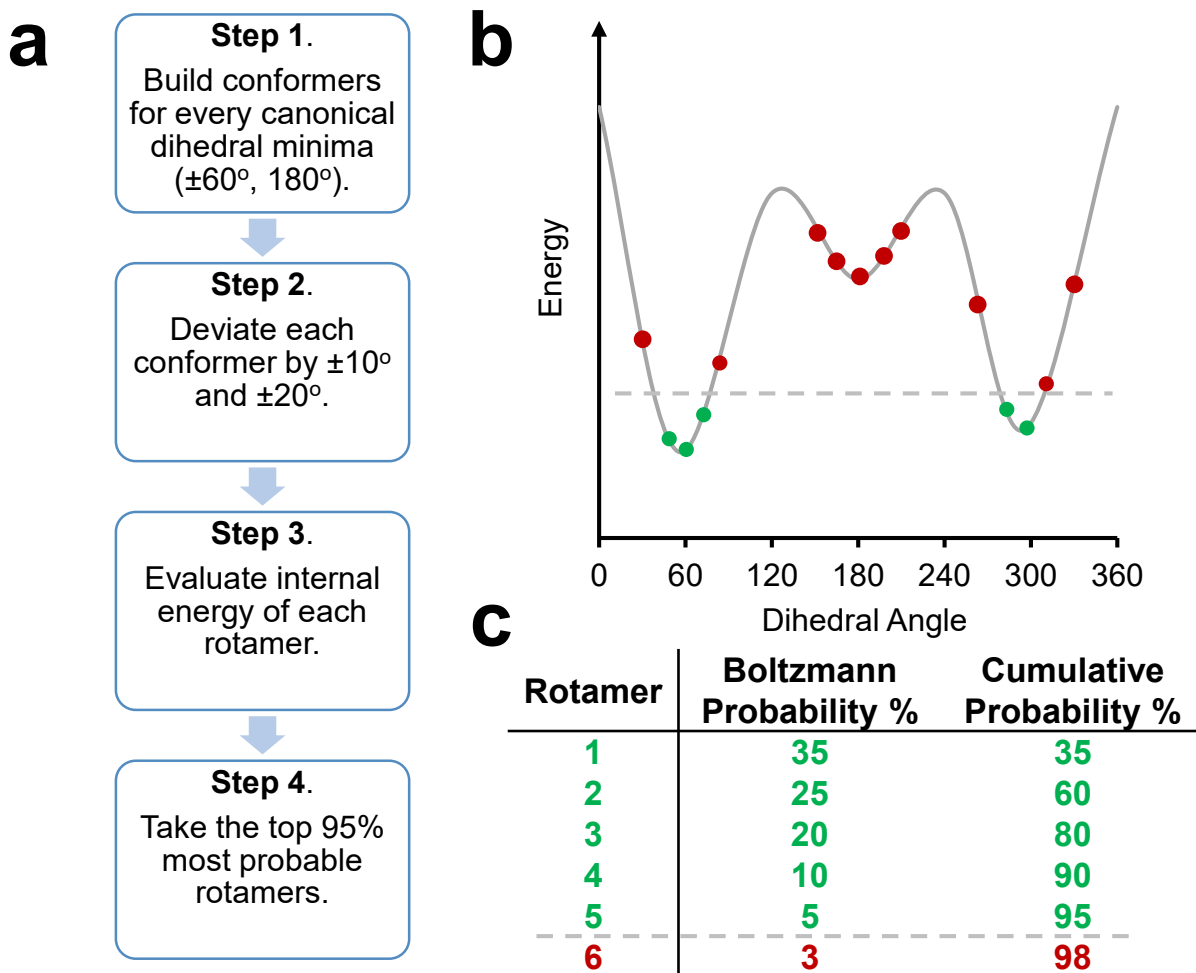


Figure 2.6. Workflow of the Targeted rotamer library generation. (a) Method flow chart. The potential energy diagram (b) is the result of steps 2 and 3, and accepted rotamers (green) are selected based on the top 95 % most probable rotamers based on cumulative probability (c). Rejected rotamers (red) are not included in the rotamer library.

2.4.6 Modeling substrates with the Targeted library

To test the Targeted substrate rotamer library, we performed $SC^{\text{fix}}\text{Sub}^{\text{sub}}$ and $SC^{\text{bbind}}\text{Sub}^{\text{sub}}$ rotamer-optimization designs, which are homologous to the designs we used previously to test how well the bbind library models the substrate conformers, except that the Targeted rotamer library was instead used to model the substrates.

In the BCAT designed structures that resulted from the $SC^{\text{fix}}\text{Sub}^{\text{sub}}$ designs (Figure 2.3g), both the Leu^{PLP} and Glu^{PLP} rotamers had a lower RMSD_{sub} to the crystallographic conformers than

in the SC^{fix}Sub^{bbind} designed structures (Table 2.1). Likewise, in designed structures yielded from SC^{bbind}Sub^{sub} rotamer optimization designs (Figure 2.3. Comparison of BCAT crystal and designed structures.h), the substrate rotamers displayed a lower substrate RMSD_{sub} (Table 2.1) to the crystallographic conformers than in SC^{bbind}Sub^{bbind} designed structures. Importantly, the designed structures yielded from both the SC^{fix}Sub^{sub} and SC^{bbind}Sub^{sub} designs contained a hydrogen bond network between the γ -carboxylate group of Glu^{PLP}, Y31, R97, V109, and Y129.

As was observed for BCAT, SC^{fix}Sub^{sub} designs on AAT with both substrates (Figure 2.4e) resulted in designed structures that displayed a lower all atom substrate RMSD_{sub} than those obtained from the SC^{fix}Sub^{bbind} designs. When active-site residues are also optimized in the SC^{bbind}Sub^{sub} design for Asp^{PLP} (Figure 2.4f), the designed structure displayed an RMSD_{All} (1.15 Å) from the crystallographic configuration that is the same as the SC^{bbind}Sub^{bbind} designed structure if the Y70 residue is not included in the RMSD calculations.

Although the Glu-AAT designed structure yielded from the SC^{bbind}Sub^{sub} design displayed a slightly lower all-atom RMSD_{All} than the SC^{bbind}Sub^{sub} designed structure, the Glu^{PLP} from the SC^{bbind}Sub^{sub} design (Figure 2.5d) also hydrogen bonded via only one O _{ϵ} to both guanidinium N _{η} of R292, as was observed in the SC^{bbind}Sub^{bbind} designed structure. However, the rotamer Glu^{PLP} adopted in the SC^{bbind}Sub^{sub} designed structure allowed the γ -carboxylate group of Glu^{PLP} to form a hydrogen bond to W140, which was not present in the SC^{bbind}Sub^{bbind} designed structure. Therefore, as was the case in the BCAT designs, modeling the AAT substrate side chains with the Targeted rotamer library yielded designed structures with substrate rotamers that are closer to the crystallographic conformers than rotamer-optimization designs that use the bbind library to model the substrate side chain.

Designed structures from SC^{bbind}Sub^{sub} designs on the Glu-BCAT and Asp-AAT templates both resulted in substrate rotamers that formed the hydrogen bonds observed in the crystal structures. However, the SC^{bbind}Sub^{sub} design on the Glu-AAT template structure yielded a designed structure where the Glu^{PLP} rotamer χ_2 torsion angle rotated the γ -carboxylate group by 20 ° towards the PLP phosphate group (Figure 2.5d) even though the Targeted library contains a rotamer that is structurally similar to the one found in the crystal structure (Figure 2.5c). In the crystal structure from which Glu-AAT is derived (1X28), there is a water molecule between the Glu^{PLP} γ -carboxylate group and the PLP phosphate group. However, in our implementation of CPD, occlusion-based implicit solvation terms³⁷ are used in calculations instead of explicit water. This left a void between the phosphate and γ -carboxylate groups, which changed the conformational energy landscape of the enzyme system and allowed the γ -carboxylate group to move into the empty space. This consequence of excluding explicit water molecules from CPD calculations is also apparent in the designed structure that resulted from the SC^{bbind}Sub^{bbind} design of Glu^{PLP} in the AAT active site, where the Glu^{PLP} γ -carboxylate group also moved to fill the empty space next to the PLP phosphate group (Figure 2.5b). Although solvated rotamer libraries have been developed,¹³³ they have not outperformed other methods. Therefore, it remains a significant challenge to incorporate explicit water into CPD calculations.

2.4.7 New substrate on a new backbone

Satisfied that the Targeted substrate rotamer library was sufficient to recapitulate the crystallographic substrate conformers in the active sites of BCAT and AAT in SC^{fix}Sub^{sub} designs, another Targeted substrate rotamer library was built on the Leu-BCAT crystallographic structure for the purpose of redesigning BCAT to be more active with L-histidine (see Chapter 3). We constructed Targeted rotamer libraries for the substrates Glu, Leu, and the tautomers of His, HID and HIE, where the 3rd letter indicates which nitrogen (δ or ϵ) is protonated. A rotamer from the

bbind library of each substrate, which contained specific chi dihedral angles that extend side chains away from the PLP, was threaded onto the PLP in the Leu-BCAT crystallographic structure to form four individual external aldimine complexes, which were used as templates to generate rotamers by rotating the side-chain torsion (χ) angles outlined in Figure 2.7a-d.

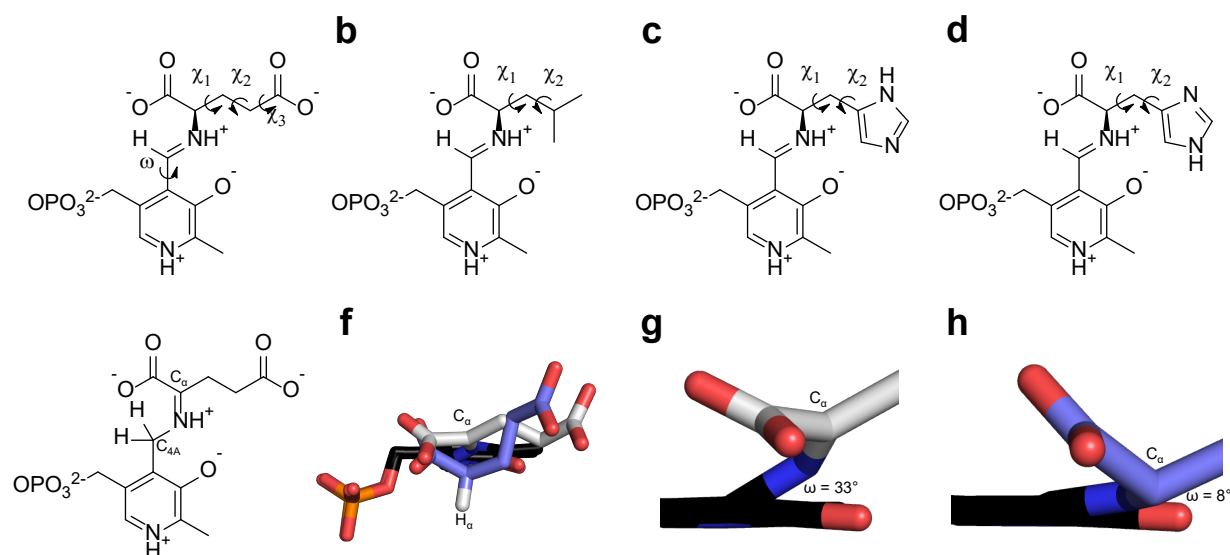


Figure 2.7. Substrate-bound PLP complexes. The torsion angles of amino acids (a) Glu, (b) Leu, (c) HID, and (d) HIE, bound to PLP as external aldimines. The Glu-bound ketimine is shown in panel (e), where the atoms C_α and C_{4A} are labelled. The Glu-bound ketimine (white) and Glu-bound external aldimine (blue) model are overlaid onto each other in panel (f), and the PLP is coloured black. The different ω -angles, as defined in panel (a), of the ketimine and external aldimine are illustrated in panels (g) and (h), respectively.

We also constructed what we call the Global rotamer library, where rotamers are built for every combination of every torsion angle in 10° increments instead of being built around canonical minima such as for the Targeted library (Figure 2.6). For example, 36 conformers would be built for valine and 1,296 conformers would be built for leucine if only the amino acid side-chains were rotated. As with the construction of the Targeted library, the internal energy of all these conformers was calculated *in vacuo* in the presence of the PLP moiety and the top 95 % most probable rotamers were incorporated into the Global rotamer library.

The Targeted library contained 25–52 % fewer rotamers than the Global library, and 54 % less to 68% more rotamers than the corresponding bbind library (Table 2.2). However, the number of rotamers for these substrates were at least 6-fold less than if every dihedral combination across 10 ° increments would have been used.

Table 2.2. Number of rotamers included in the backbone-independent (bbind), “Global”, and “Targeted” libraries built on BCAT and AAT.

Template	Substrate	10 °	bbind	Global	Targeted
Glu-AAT	Glu	23,328	243	NC	261
Asp-AAT	Asp	648	81	NC	52
Glu-BCAT	Glu	23,328	243	NC	1762
Leu-BCAT	Leu	1,296	81	92	44
Leu-BCAT	Glu	23,328	243	1,614	764
Leu-BCAT	HID	1,296	81	209	150
Leu-BCAT	HIE	1,296	81	191	137

10 ° = a rotamer library that contains the maximum number of possible rotamers if side-chain torsion angles are sampled combinatorially at every 10 ° increment.

NC = Not constructed.

We surveyed the distribution of χ_1 and χ_2 angles for the Leu, Glu, HID, and HIE rotamers in the bbind, Targeted, and Global rotamer libraries (Figure 2.8a-d), where the Targeted and Global rotamer libraries were constructed on the external aldimine complex from the Leu-BCAT backbone template. In general, the Global and Targeted libraries contained rotamers that cluster around the same χ_1 and χ_2 combinations as rotamers present in the bbind library, with the Targeted library containing a smaller number of rotamers at each torsion combination than the Global library. Specific combinations present in the bbind libraries are absent in both the Targeted and Global libraries due to steric clashes of the substrate side chain with the additional oxygen in the α -carboxylate group that is not present in a peptide backbone.

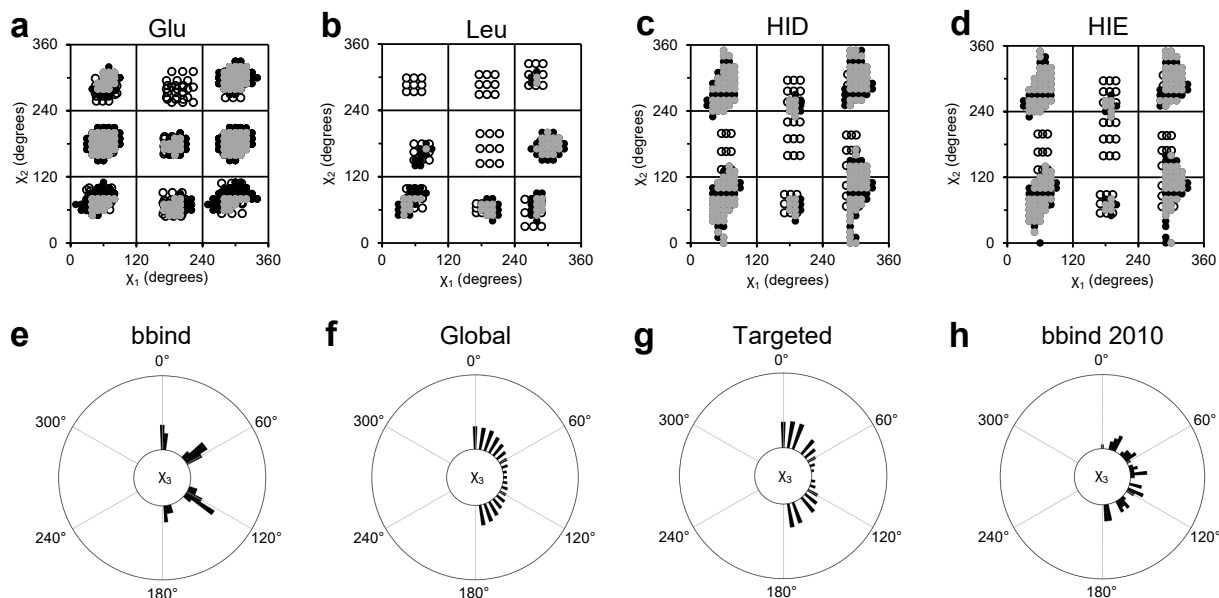


Figure 2.8. Rotamer library coverage plots for substrates bound to PLP as external aldimines built on the Leu-BCAT PLP template structure. The distribution of χ_1 and χ_2 torsion angles are shown for the rotamers of (a) Glu, (b) Leu, and His tautomers, (c) HID and (d) HIE. Interpolated χ torsions in panels a-d are shown as white, black, or grey circles, which refer to rotamers from the Dunbrack backbone-independent, "Global," and "Targeted" libraries, respectively. The population of χ_3 torsion angles for Glu for (e) the Dunbrack 2002 backbone-independent, (f) "Global", (g) "Targeted", and (h) Dunbrack 2010 backbone-independent libraries are shown as polar histograms. The height is the relative population, where the inner circle represents 0 % population and the outer circle represents 33 % population.

Next, we looked at the distribution of χ_3 angles of the glutamate substrate across the γ -carboxylate group in the different rotamer libraries (Figure 2.8e-g). Because of the symmetry of the carboxylate group, angles are only shown up to 180 ° and χ_3 angles measured over 180 ° are remeasured with the other O_ϵ atom for incorporation into the polar histogram. The bbind library treated the γ -carboxylate group χ_3 angles as rotameric, where the χ_3 angles are distributed across the three different sp^3 - sp^2 dihedral angle minima for carboxylate groups (0 °, +60 °, +120 °, 180°). Conversely, the Global and Targeted libraries contain rotamers with χ_3 angles that are distributed between 0 ° and 180 °, which is similar to the nonrotameric distribution of the Glu χ_3 angles observed in the Dunbrack 2010 backbone-independent library³¹ (Figure 2.8h). The success of the Targeted library when modeling Glu^{PLP} in the designs discussed in the previous section is likely

due to treating the χ_3 angle as nonrotameric. Additionally, the increased distribution of χ_3 angles is likely the reason for the approximately 7- or 3-fold increase in the number of Glu^{PLP} rotamers included in the Global and Targeted libraries relative to the bbind library, respectively. Since both libraries cover approximately the same torsion angle distributions, we selected the Targeted library for design since it has a smaller number of rotamers, which will be important to increase computational tractability in Chapter 3.

Lastly, we ran rotamer-optimization designs on the Leu-BCAT active site to test how well we could predict the Glu^{PLP} carboxylate hydrogen bonds present in the Glu-BCAT crystallographic structure using both the bbind and Targeted rotamer libraries (Figure 2.9). Both the SC^{fix}Sub^{bbind} and the SC^{bbind}Sub^{bbind} designs on the Leu-BCAT template with the Glu^{PLP} substrate yielded designed structures that did not contain a hydrogen bond network between the γ -carboxylate of the Glu^{PLP} side-chain, Y31, R97, V109, and Y129 (Figure 2.9a,b). On the other hand, the SC^{fix}Sub^{sub} and the SC^{bbind}Sub^{sub} designs on the Leu-BCAT template with the Glu^{PLP} substrate yielded designed structures that contain Glu^{PLP} rotamers where the γ -carboxylate is positioned in the same orientation as the γ -carboxylate of the Glu^{PLP} crystallographic conformer in the Glu-BCAT structure (Figure 2.9c,d). Additionally, the designed structure that resulted from the SC^{bbind}Sub^{sub} design on the Leu-BCAT template with the Glu^{PLP} substrate contained a Glu^{PLP} rotamer where the γ -carboxylate group of Glu^{PLP} forms hydrogen bonds to Y31 and R97, but not to Y129 nor the backbone amide of V109 (Figure 2.9d).

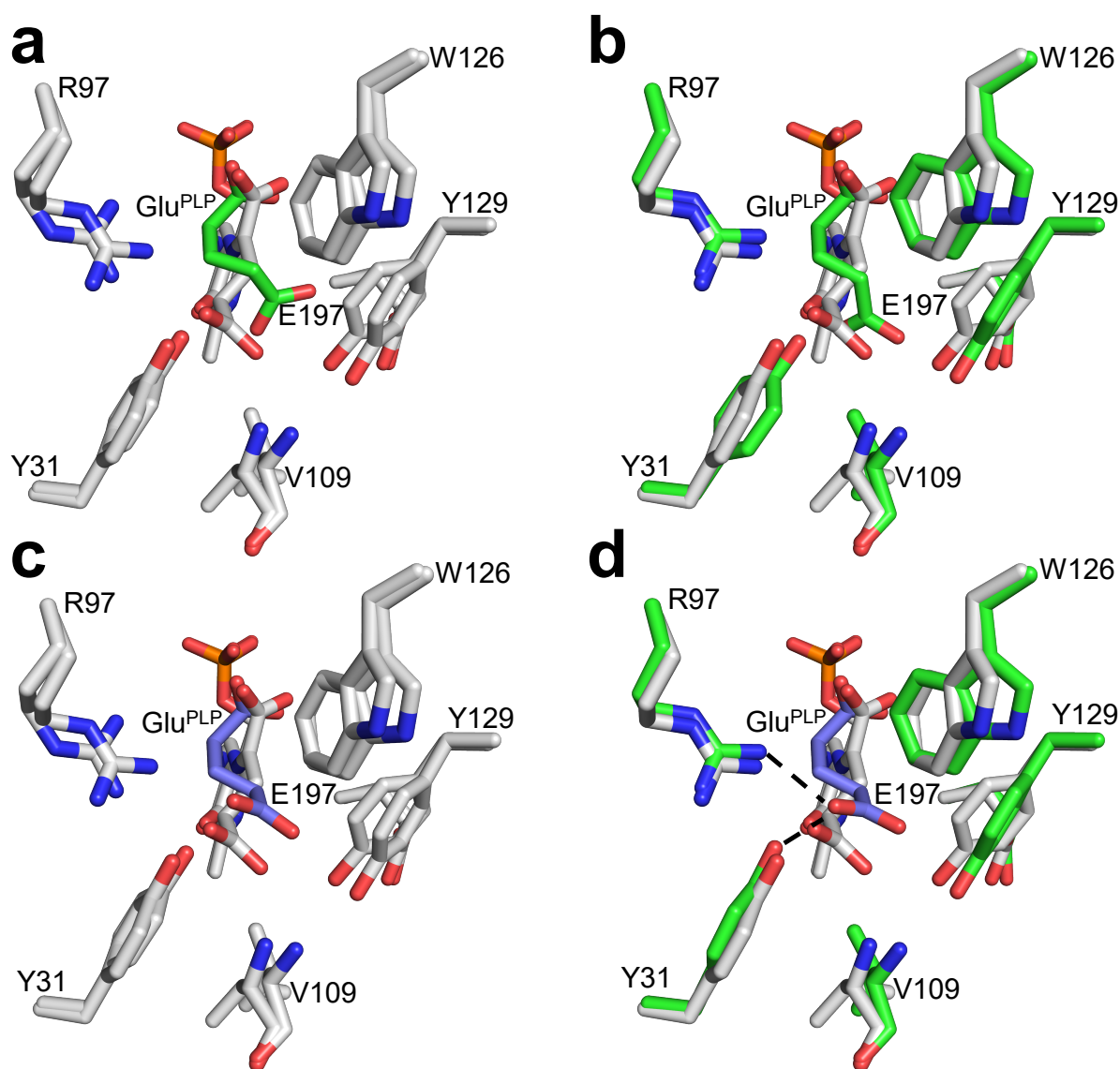


Figure 2.9. Rotamer prediction of Glu^{PLP} bound to PLP as an external aldimine on Leu-BCAT crystal structure. Designed structures obtained from (a) SC^{fix}Sub^{bbind}, (b) SC^{bbind}Sub^{bbind}, (c) SC^{fix}Sub^{sub} and (d) SC^{bbind}Sub^{sub} designs of Glu^{PLP} on Leu-BCAT are superimposed on the crystal structure (white). Residues fixed during design are represented as white sticks, and side-chains optimized with either the bbind or Targeted rotamer libraries are represented as green or blue sticks, respectively. Backbone atoms of side-chain residues are not shown, except for V109. The active site residues are labelled with 1-letter amino acid code at the C_α. Selected hydrogen bonds are shown as black dotted lines.

Glu^{PLP} rotamers are unable to adopt the same conformation on the Leu-BCAT structure as in the Glu-BCAT structure because of the structural difference between the external aldimine and

the ketimine (Figure 2.7a,e). In the external aldimine, the C_α is sp^3 and is bonded to H_α . However, in the ketimine, the C_α is sp^2 and the H_α atom has been moved to C_{4A} (labelled in Figure 2.7e). Additionally, the ω angle (Figure 2.7a) differs by 25° between the two PLP forms (Figure 2.7g,h). These differences are apparent in Figure 2.7f, where the Glu^{PLP} -bound external aldimine and ketimine complexes of the designed structures that resulted from the $SC^{bbind}Sub^{sub}$ designs are superimposed. The torsion angles of the crystallographic ketimine Glu^{PLP} side chain are ($\chi_1 = -38.1^\circ$, $\chi_2 = -171.3^\circ$, $\chi_3 = +92.4^\circ$) and the torsion angles of the optimized external aldimine Glu^{PLP} side-chain rotamer are ($\chi_1 = +70.0^\circ$, $\chi_2 = +160.0^\circ$, $\chi_3 = +172.4^\circ$) or ($\chi_1 = +70.0^\circ$, $\chi_2 = +160.0^\circ$, $\chi_3 = +180.0^\circ$) for the $SC^{fix}Sub^{sub}$ or $SC^{bbind}Sub^{sub}$ designed structures, respectively. In addition, we treat each PLP state as a rigid moiety during CPD even though the aldimine and ketimine both contain single bonds capable of rotation and can translate and rotate in the active site. Conformational ensembles can be constructed to help alleviate the artefacts of the rigid PLP approximation, in addition to alleviating the artefacts of the rigid rotamer and fixed backbone approximations,⁵⁶ which will be investigated in Chapters 3 and 4.

2.5 Conclusions

In this chapter, we tested the ability of rotamer-optimization calculations with different rotamer libraries to recapitulate the side-chain conformation of active-site residues and substrates present in aminotransferase crystal structures. First, we showed that rotamer-optimization calculations with the bbind library could replicate active-site residue rotamer configurations better than with the bbdep rotamer library. Next, we showed that the bbind library was also sufficiently accurate to model the Glu^{PLP}/Asp^{PLP} and Leu^{PLP} substrates during rotamer optimization designs on AAT and BCAT with fixed active-site residues. This was not the case for the Glu^{PLP} substrate in the BCAT active site. Since properly modeling the Glu^{PLP} side chain in the BCAT active site is

important in Chapter 3, we built our own Targeted rotamer library to model the substrate side chains. Modeling the Glu^{PLP} side chains with the Targeted rotamer library in rotamer-optimization designs allowed us to recover the Glu^{PLP} side chain in the Glu-BCAT active site and all hydrogen bonds involved with the Glu^{PLP} γ -carboxylate group. The work-flow developed in this chapter can now be used to generate a rotamer library for any small molecule thereby opening the door to more accurate design of numerous enzymes.

Chapter 3: Computational design of multi-substrate enzyme specificity

Antony D. St-Jacques, Marie-Ève Eyahpaise & Roberto A. Chica *

3.1 Statement of contribution

Marie-Ève Eyahpaise helped prepare the combinatorial DNA library and sequenced BCAT mutants. All CPD calculations and all other experiments were performed by Antony D. St-Jacques. Experimental design was devised by Antony D. St-Jacques and Dr. Roberto A. Chica. The chapter was written by Antony D. St-Jacques and edited by Dr. Roberto A. Chica.

3.2 Introduction

The creation of enzymes displaying desired substrate specificity is an important objective of enzyme engineering. To achieve this goal, computational protein design (CPD) can be used to identify sequences that can fulfill interactions required to productively bind a desired substrate.¹³ Commonly used CPD protocols find optimal sequences in the context of a single state, for example an enzyme structure with a single substrate bound at its active site.^{77, 134} However, many enzymes that are important for biocatalytic applications (e.g., aminotransferases,⁸⁰⁻⁸¹ transketolases,⁸²⁻⁸³ and lipases⁸⁴) follow one-site bi bi ping pong mechanisms, which necessitate that two or more substrates bind to the same active site at different points in the reaction for complete catalytic turnover.¹³⁵ Thus, the computational design of multi-substrate enzyme specificity requires the ability to optimize sequences in the context of multiple chemical states (i.e., multiple substrates) because mutations designed to enable the productive binding of one substrate may be detrimental to the binding of a second substrate.¹³⁶

Previously, multistate approaches to CPD have been developed to enable the optimization of sequences across multiple chemical states.¹³⁷ These approaches have been used to design

transcriptional regulators¹³⁸ and enzymes^{25, 66, 86} that can bind a new target molecule. In these examples, sequences were optimized to bind a single target small molecule ligand, and the alternate chemical state was either the unbound state, so as to predict binding affinity,^{25, 86} or an inhibitor-bound state, so as to identify mutations that confer inhibitor resistance while maintaining catalytic activity.⁶⁶ However, multistate CPD has not yet been used to predict sequences that can bind multiple desired ligands, preventing the redesign of multi-substrate enzyme specificity. Here, we develop and experimentally-validate a multistate CPD methodology that can optimize enzyme sequences for productive binding of any number of target substrates.

3.3 Results and discussion

As a test case for the development of this methodology, we selected *Escherichia coli* branched-chain amino acid aminotransferase (BCAT), a biocatalyst that has been used for the synthesis of many unnatural amino acids.¹³⁹⁻¹⁴¹ BCAT is a pyridoxal 5'-phosphate (PLP) dependent enzyme that catalyzes the reversible transamination of α -ketoglutarate with branched-chain amino acids (leucine, isoleucine, and valine). We picked this enzyme because it follows a one-site bi bi ping-pong mechanism^{110, 142} and exhibits dual specificity⁸⁸ towards substrates containing functional groups of different shapes and properties (Figure 3.1). Furthermore, crystal structures of substrate-bound BCAT are available,⁸⁸⁻⁸⁹ and these show that the side chains of active-site residues interacting with the substrates are not directly involved in catalysis, which should enable the redesign of its specificity without affecting the catalytic machinery.

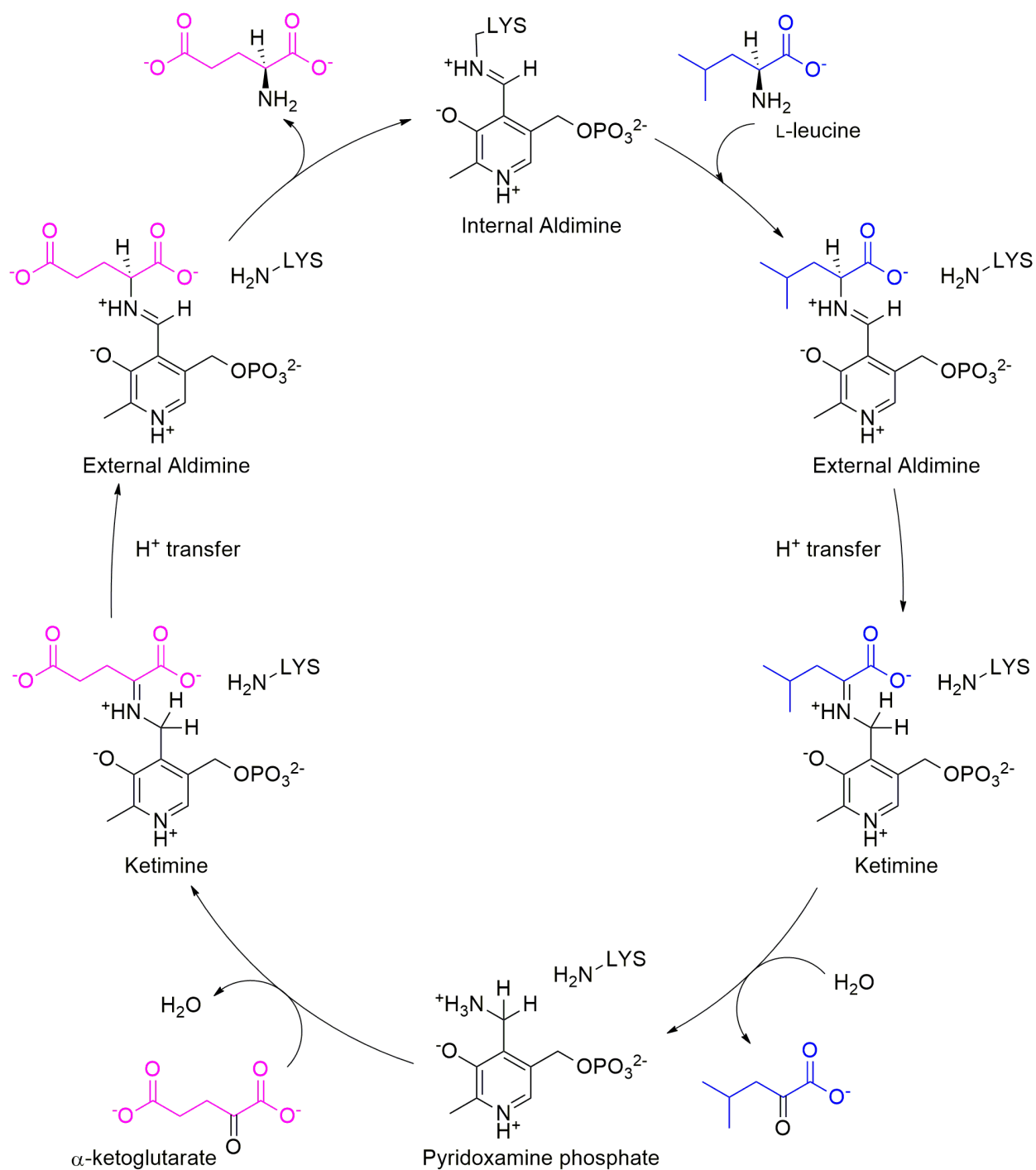


Figure 3.1. BCAT transamination bi bi ping pong reaction of L-leucine (blue) as an amino donor and α -ketoglutarate (pink) as an amino acceptor.

As our target substrates, we picked histidine and α -ketoglutarate as amino donor and acceptor, respectively. Histidine was selected because it is a poor substrate of BCAT (< 1% activity relative to leucine),¹¹⁵ and is chemically distinct from both glutamate and leucine, the native substrates of the enzyme. We selected α -ketoglutarate since its product, glutamate, can be readily detected using a coupled-enzyme assay that we previously developed to screen aminotransferase mutant libraries in 96-well plate format.¹²¹

Analysis of the BCAT crystal structure bound to the 2-methylleucine analog (PDB ID: 1IIL)⁸⁹ enabled us to identify residues that are involved in substrate binding. These residues (Figure 3.2) have side chains that point towards the leucine analog, with atoms within 3 Å of the β -isopropyl group of the substrate. During design, these six residues were allowed to sample side-chain rotamers of all proteinogenic amino acids except proline and lysine, representing a total sequence space of 3.4×10^7 sequences (18⁶). Side-chain rotamers of neighbouring residues (positions R97, F124, E197) were also optimized but their identity was not allowed to vary.

To ensure that designed sequences are active, our design procedure must be able to identify active site mutations that allow productive binding of the two target substrates while not significantly destabilizing the enzyme fold. Therefore, amino acids at the six design positions were optimized in the context of the histidine- (HIS) and glutamate-bound (GLU) external aldimines (Figure 3.2). We also included the substrate-free internal aldimine state (PLP) to assess the effect of mutations on protein stability in the absence of substrate.

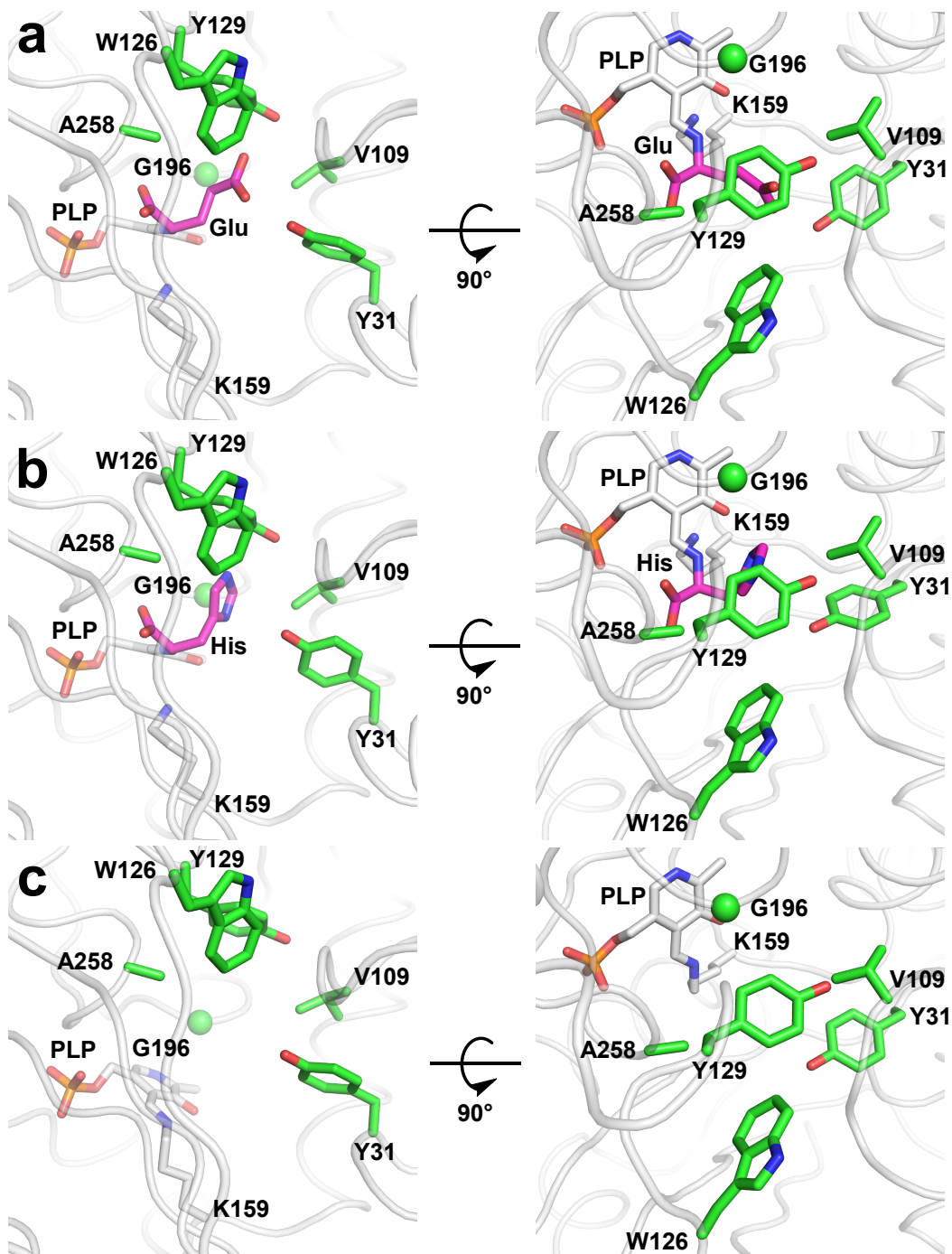


Figure 3.2. Active site of *E. coli* BCAT containing (a) glutamate- and (b) histidine-bound external aldimines, or the (c) internal aldimine chemical states. The PLP cofactor and the catalytic lysine are coloured white, the substrates are coloured magenta, and the residues selected for design are coloured green, which are labelled by their one-letter amino acid code and residue number next to their C α . The G196 C α atom is shown as a sphere. In the internal aldimine, the PLP is covalently bound to the catalytic lysine, K159.

Using a 100-member ensemble of backbone templates generated to approximate conformational flexibility in BCAT (section 3.5.3), we performed multistate CPD to optimize sequences on each individual chemical state (section 3.5.4). It is important to note that each chemical state comprises a native ensemble of 100 backbone templates of unbound, histidine-bound, or glutamate-bound BCAT. From every ranked list of scored sequences obtained from each chemical state calculation, we generated a table of amino-acid probabilities at each designed position based on the Boltzmann distribution (section 3.5.5). These probabilities reflect the predicted ability of each amino-acid type at every design position to productively bind the target substrates (HIS and GLU) or stabilize the enzyme fold (PLP).

The tables of probabilities generated for individual chemical states (Table 3.1-3) are displayed as sequence logos in Figure 3.3. As expected, each sequence logo is unique, demonstrating that mutations that stabilize one chemical state may have a detrimental or neutral effect in another chemical state. For example, in the GLU state, the asparagine mutation is favoured at position 109 (95.1%), but this mutation has a probability of 11.3% in the HIS state and $< 0.1\%$ in the PLP state. This result is unsurprising given that arginine forms stronger favourable interactions with glutamate than with histidine or the internal aldimine (Figure 3.4). Similarly, histidine at position 196 has a high probability in the PLP state ($> 99.9\%$) because it forms stabilizing interactions with the active-site pocket. However, H196 has low probability in the GLU and HIS states ($< 0.1\%$) because the active site residues adopted a less energetically favourable rotamer configuration to avoid clashes between the glutamate and histidine substrates with the H193 imidazole side chain. On the other hand, high-probability amino acids at some positions are very similar across all chemical states (e.g., positions 31 and 126). This result suggests that those residues are important for stability of the BCAT fold because they have a stabilizing effect

independent of the presence, absence, or identity of the substrate. In light of these results, we hypothesized that positions where high-probability amino acids vary significantly across chemical states are likely the ones that are important for controlling the enzyme's substrate specificity, and should be the focus of our multi-chemical state design.

Table 3.1. Position Probability Matrix of HIS state.

HIS	31	109	126	129	196	258
A	< 0.1	0.2	< 0.1	< 0.1	6.1	< 0.1
C	< 0.1	2.0	< 0.1	< 0.1	76.6	< 0.1
D	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
E	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
F	34.6	< 0.1	< 0.1	99.1	< 0.1	< 0.1
G	< 0.1	< 0.1	< 0.1	< 0.1	13.5	< 0.1
H	< 0.1	0.8	< 0.1	0.7	< 0.1	< 0.1
I	< 0.1	23.1	< 0.1	< 0.1	< 0.1	< 0.1
L	< 0.1	1.2	< 0.1	< 0.1	< 0.1	< 0.1
M	19.7	53.5	< 0.1	< 0.1	< 0.1	95.9
N	< 0.1	11.3	< 0.1	< 0.1	< 0.1	< 0.1
Q	< 0.1	7.9	< 0.1	< 0.1	< 0.1	< 0.1
R	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
S	< 0.1	< 0.1	< 0.1	< 0.1	3.6	0.7
T	< 0.1	< 0.1	< 0.1	< 0.1	0.2	3.4
V	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
W	< 0.1	< 0.1	> 99.9	< 0.1	< 0.1	< 0.1
Y	45.6	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1

Table 3.2. Position Probability Matrix of GLU state.

GLU	31	109	126	129	196	258
A	< 0.1	< 0.1	< 0.1	< 0.1	3.1	< 0.1
C	< 0.1	0.4	< 0.1	< 0.1	88.9	0.8
D	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
E	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
F	13.3	< 0.1	< 0.1	39.9	< 0.1	< 0.1
G	< 0.1	< 0.1	< 0.1	< 0.1	4.5	< 0.1
H	< 0.1	1.6	< 0.1	0.1	< 0.1	< 0.1
I	< 0.1	1.5	< 0.1	< 0.1	< 0.1	< 0.1
L	< 0.1	0.1	< 0.1	< 0.1	< 0.1	< 0.1
M	40.5	0.9	< 0.1	< 0.1	< 0.1	3.9
N	0.2	95.1	< 0.1	< 0.1	< 0.1	< 0.1
Q	< 0.1	0.2	< 0.1	< 0.1	< 0.1	< 0.1
R	< 0.1	< 0.1	< 0.1	60.0	< 0.1	< 0.1
S	< 0.1	< 0.1	< 0.1	< 0.1	1.7	76.1
T	< 0.1	< 0.1	< 0.1	< 0.1	1.8	19.2
V	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
W	< 0.1	< 0.1	> 99.9	< 0.1	< 0.1	< 0.1
Y	46.0	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1

Table 3.3 Position Probability Matrix of PLP state.

PLP	31	109	126	129	196	258
A	< 0.1	< 0.1	< 0.1	0.2	< 0.1	< 0.1
C	< 0.1	< 0.1	0.3	0.8	< 0.1	< 0.1
D	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
E	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
F	57.4	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
G	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
H	< 0.1	< 0.1	< 0.1	0.1	> 99.9	< 0.1
I	< 0.1	27.0	< 0.1	< 0.1	< 0.1	< 0.1
L	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
M	< 0.1	3.9	< 0.1	0.4	< 0.1	< 0.1
N	< 0.1	< 0.1	< 0.1	0.5	< 0.1	86.9
Q	< 0.1	0.7	< 0.1	64.8	< 0.1	12.8
R	< 0.1	0.3	< 0.1	2.5	< 0.1	< 0.1
S	< 0.1	0.6	0.2	30.6	< 0.1	0.3
T	0.2	37.4	< 0.1	< 0.1	< 0.1	< 0.1
V	< 0.1	30.0	< 0.1	< 0.1	< 0.1	< 0.1
W	< 0.1	< 0.1	99.3	< 0.1	< 0.1	< 0.1
Y	42.4	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1

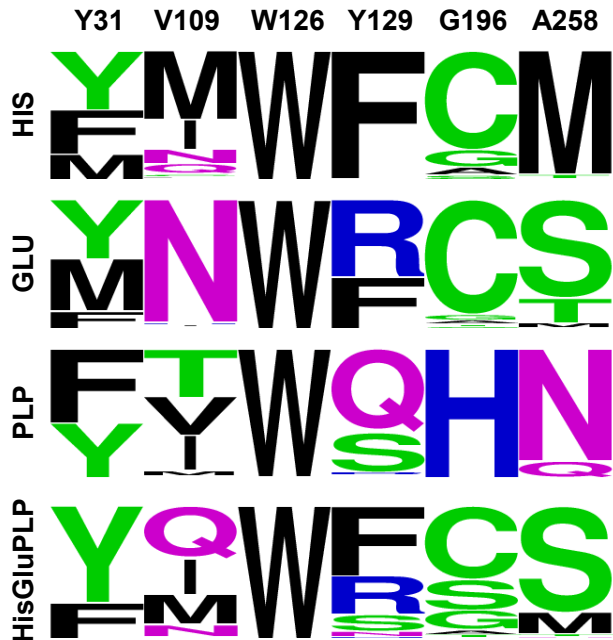


Figure 3.3. Comparison of position energy matrices generated using different chemical states. The height of the letter corresponds to the probability of the amino acid at that position from the position probability matrix. Amino acids with less than 1% probability are not included.

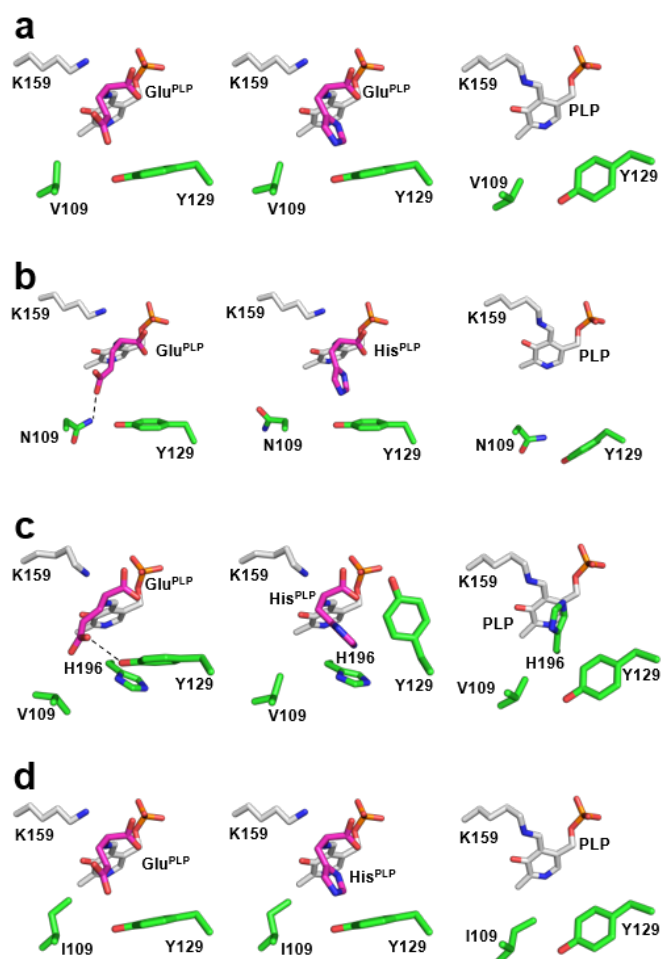


Figure 3.4. Effect of mutations shown in CPD models on the GLU, HIS, and PLP chemical states. (a) wild type, (b) V109N variant, (c) G196H variant, (d) V109I variant. The PLP cofactor and the catalytic lysine are coloured white, the substrates are coloured magenta, and the residues that have been rotamer-optimized are coloured green, which are labelled by their one-letter amino acid code and residue number next to their C_α. In the internal aldimine, the PLP is covalently bound to the catalytic lysine, K159.

To design BCAT variants that can catalyze the transamination of histidine with α -ketoglutarate, we evaluated mutations in the context of all three chemical states. This was done by first averaging the energies obtained for each BCAT sequence on all chemical states, and then using these average energies to generate a table of probabilities for the HisGluPLP state. As expected, the multi-chemical state sequence logo (Figure 3.3) was unique but contained many of the high-probability amino acids from each individual chemical state. Interestingly, most positions included a combination of the highest probability amino acids from more than one chemical state (e.g., positions 31, 126, and 129), suggesting that those positions are not important for substrate specificity. However, some positions (e.g., positions 109 and 258) favoured mostly the high-probability amino acids from only one of the chemical states. For example, all amino acids having a probability > 1% in the HisGluPLP state at position 109 are the same as in the HIS state. This result suggests that these mutations may be beneficial for productive binding of one of the substrates (in this case histidine) without being destabilizing to the enzyme fold or detrimental for binding of the other substrate (i.e. glutamate).

Table 3.4. Position Probability Matrix of HisGluPLP state.

HisGluPLP	31	109	126	129	196	258
A	< 0.1	< 0.1	< 0.1	< 0.1	3.6	< 0.1
C	< 0.1	0.5	< 0.1	< 0.1	54.9	0.8
D	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
E	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
F	27.4	< 0.1	< 0.1	52.2	< 0.1	< 0.1
G	< 0.1	< 0.1	< 0.1	< 0.1	14.7	< 0.1
H	< 0.1	< 0.1	< 0.1	1.5	< 0.1	< 0.1
I	< 0.1	26.6	< 0.1	< 0.1	< 0.1	< 0.1
L	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
M	< 0.1	22.3	< 0.1	< 0.1	< 0.1	16.3
N	< 0.1	9.0	< 0.1	4.1	1.2	1.1
Q	< 0.1	39.5	< 0.1	1.0	< 0.1	< 0.1
R	< 0.1	< 0.1	< 0.1	28.7	< 0.1	< 0.1
S	< 0.1	< 0.1	< 0.1	12.5	22.9	79.5
T	< 0.1	0.7	< 0.1	< 0.1	2.7	2.3
V	< 0.1	1.0	< 0.1	< 0.1	< 0.1	< 0.1
W	< 0.1	< 0.1	> 99.9	< 0.1	< 0.1	< 0.1
Y	72.5	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1

Next, we used the probability table of the HisGluPLP state (Table 3.4) to generate a combinatorial library with the CLEARSS method³⁶ that includes the top amino acids that are predicted to productively bind both histidine and glutamate without significantly destabilizing the enzyme, as well as the wild-type residue at that position. This yielded a 32-member library containing I/M/Q/V, F/Y, C/G, and A/S at positions 109, 129, 196, and 258, respectively. Screening of this library for enhanced transamination activity with histidine (> 300% wild-type activity) yielded four active mutants containing one or two mutations (Table 3.5).

Table 3.5. Apparent kinetic parameters of *E. coli* BCAT and its mutants for transamination of histidine with α -ketoglutarate as an amino acceptor

Enzyme	K_M (mM)	k_{cat} (s^{-1})	k_{cat} / K_M ($M^{-1} s^{-1}$)	Relative k_{cat} / K_M
Wild type	16 ± 2	0.022 ± 0.002	1.4 ± 0.2	1
V109I/G196C	5.2 ± 0.3	0.332 ± 0.007	64 ± 4	46
V109I	3.5 ± 0.4	0.161 ± 0.006	46 ± 6	33
V109I/Y129F	3.3 ± 0.2	0.0414 ± 0.0008	12.4 ± 0.7	9
V109M/G196C	36 ± 4	0.28 ± 0.02	8 ± 1	6

For all reactions, the acceptor substrate was 0.5 mM α -ketoglutarate
All reactions were performed in triplicate (n = 3, $\pm$ s.d.)

The most active variant (V109I/G196C) displayed an apparent k_{cat}/K_M for histidine (k_{cat}/K_M^{His}) of $64 \pm 4 M^{-1} s^{-1}$, which is approximately 50-fold higher than the corresponding value for the wild type. Interestingly, although the binding of the substrate was targeted in design, the K_M of this variant was only reduced 3-fold relative to wild-type (WT) while k_{cat}^{His} increased 15-fold, even though our computational procedure does not explicitly design catalytic rates. The V109I mutation that is present in 3 of the top 4 BCAT mutants results in a 7-fold increase in k_{cat}^{His} and a 5-fold decrease in K_M^{His} relative to WT whereas the G196C mutation results in a 5-fold higher k_{cat}^{His} and a 2.5-fold higher K_M^{His} relative to WT (Table 3.6). These results suggest that the V109I and G196C mutations act synergistically to increase the k_{cat}/K_M^{His} in the most active variant. As is the case for V109I/G196C, the k_{cat}^{His} for two of the other three most active mutants was increased more than the K_M^{His} was decreased.

Table 3.6. Apparent kinetic parameters of additional BCAT mutants from the screened 32-member library for transamination of histidine with α -ketoglutarate as an amino acceptor

Enzyme	K_M (mM)	k_{cat} (s ⁻¹)	k_{cat} / K_M (M ⁻¹ s ⁻¹)	Relative k_{cat} / K_M
Wild type	16 ± 2	0.022 ± 0.002	1.4 ± 0.2	1
G196C	41 ± 3	0.103 ± 0.005	2.5 ± 0.2	1.78
V109M/Y129F	6 ± 0.3	0.0046 ± 0.0001	0.77 ± 0.04	0.55
V109M	48 ± 6	0.030 ± 0.003	0.6 ± 0.1	0.4
Y129F	9.8 ± 0.9	0.0042 ± 0.0002	0.43 ± 0.05	0.31
V109M/Y129F/G196C/A258S	14 ± 5	0.004 ± 0.0008	0.3 ± 0.1	0.21
V109I/Y129V ^a	1.7 ± 0.1	0.118 ± 0.003	68 ± 6	49
V109I/Y129I ^a	1.61 ± 0.08	0.0324 ± 0.0005	20 ± 1	14

^a BCAT mutants identified from saturation mutagenesis.

For all reactions, the acceptor substrate was 0.5 mM α -ketoglutarate

All reactions were performed in triplicate (n = 3, ± s.d.)

Table 3.7. Apparent kinetic parameters of *E. coli* BCAT and its mutants for transamination of leucine with α -ketoglutarate as an amino acceptor

Enzyme	Leucine ^a			α -ketoglutarate ^b			
	K_M (mM)	k_{cat} (s ⁻¹)	k_{cat} / K_M (M ⁻¹ s ⁻¹)	K_M (mM)	K_i^c (mM)	k_{cat} (s ⁻¹)	k_{cat} / K_M (M ⁻¹ s ⁻¹)
Wild type	0.106 ± 0.006	1.19 ± 0.02	11200 ± 700	0.060 ± 0.004	1.3 ± 0.1	1.28 ± 0.04	21000 ± 2000
V109I/G196C	0.150 ± 0.005	1.042 ± 0.007	6900 ± 200	0.075 ± 0.006	3.0 ± 0.3	0.64 ± 0.02	8600 ± 800
V109I	0.084 ± 0.006	2.24 ± 0.03	27000 ± 2000	0.035 ± 0.004	1.9 ± 0.2	0.33 ± 0.01	9000 ± 1000
V109I/Y129F	0.14 ± 0.02	0.213 ± 0.005	1500 ± 200	0.21 ± 0.01	2.9 ± 0.2	0.266 ± 0.007	1240 ± 80
V109M/G196C	0.175 ± 0.008	0.927 ± 0.009	5300 ± 200	0.039 ± 0.005	2.3 ± 0.3	0.205 ± 0.009	5200 ± 700

^a For all reactions, the acceptor substrate was 0.5 mM α -ketoglutarate

^b For all reactions, the donor substrate was 2 mM leucine

^c Substrate inhibition constant

All reactions were performed in triplicate (n = 3, ± s.d.)

To evaluate the effect of the mutations on the native activity, we determined the apparent kinetic parameters for both leucine and α -ketoglutarate. The k_{cat}/K_M for leucine (k_{cat}/K_M^{Leu}) was decreased for 3 out of 4 mutants, with the only exception being the V109I mutant (Table 3.7), whose k_{cat}/K_M^{Leu} is 2.4-fold higher than that of the WT. This increase in specificity for leucine is mostly due to an approximately two-fold higher k_{cat}^{Leu} , which suggests that the V109I mutation increases turnover rates for both amino donor substrates (L-histidine and L-leucine). However, in

all cases, the k_{cat}/K_M for α -ketoglutarate was decreased (2- to 17-fold), which is mostly due to decreases in k_{cat} for this acceptor substrate (2- to 6-fold). The lowered K_M for the two amino acids and α -ketoglutarate observed for the V109I mutant suggests that this mutation plays a role in improving productive binding of these substrates.

In contrast, the V109I/G196C mutant displayed an increase in K_M^{Leu} and $K_M^{\alpha\text{-KG}}$ of 40% and 25% relative to WT, respectively. Despite the fact that the V109I/G196C mutant is approximately 100-fold more active towards the native substrate leucine than towards the target substrate histidine, its increased specificity (k_{cat}/K_M) towards histidine is accompanied by a significantly decreased specificity towards the native substrates leucine and α -ketoglutarate. Although a large decrease in specificity towards native substrates may be desirable for some applications, it is likely that negative design will be required to generate a biocatalyst that is more active towards a new target substrate than its native substrate, for example if the goal is to generate a histidine aminotransferase from BCAT.

Interestingly, the majority of BCAT variants displaying high transamination activity towards histidine contain the V109I mutation. Indeed, the V109I single mutant displays the second highest activity towards L-histidine of all the BCAT mutants characterized due to an increase in $k_{\text{cat}}^{\text{His}}$ and a decrease in K_M^{His} relative to WT. To gain a better understanding of the functional effect of the V109I mutation, we analyzed the structure of the models derived by CPD for these mutants. The CPD models (Figure 3.4d) reveal that the longer β -ethyl group of I109 adopts a position underneath the side chain of the histidine and glutamate substrates. Possibly, the V109I mutation restricts the conformational flexibility of the substrate side chains, which limits the nonproductive conformations that the side chain can adopt.

Overall, our computational library design procedure was able to design a 32-member library which contained four mutants that displayed a significant increase in $k_{\text{cat}}/K_{\text{M}}$ (6- to 46-fold) towards L-histidine, representing a hit rate of 12.5%.

3.3.1 Contribution of each chemical state to prediction accuracy

In the previous section, our computational procedure with the chemical state HisGluPLP allowed us to design a 32-member that contained mutants that displayed up to a 46-fold increase in $k_{\text{cat}}/K_{\text{M}}$ towards L-histidine. To test if our computational procedure could have done better, we tested our method with different chemical states. Although we cannot say other sequences were not active with L-histidine, we looked at the minimum library size (MLS) required by our library design procedure to design a library that contains the BCAT mutants most active with L-histidine to compare the predictive power of different chemical states (Table 3.8). In this analysis, the application of our library design procedure with a highly predictive chemical state would require a small MLS to include all the BCAT variants with enhanced activity towards L-histidine. For comparison, our library design procedure with the HisGluPLP state would have included 3 of the top 4 BCAT variants in MLS of 24 members.

Table 3.8. Minimum library size required to predict BCAT mutants for a variety of chemical states

Mutant	His	Glu	PLP	HisGlu		HisPLP		HisGluPLP	
				Mean ^a	Max ^a	Mean ^a	Max ^a	Mean ^a	Max ^a
V109I/G196C	24	375	>500	100	36	108	144	24	32
V109I	12	75	36	50	18	20	8	12	32
V109I/Y129F	12	75	260	50	18	50	175	12	>500
V109M/G196C	16	375	>500	36	60	81	192	32	72
V109I/Y129V	>500	>500	>500	>500	>500	170	425	170	>500

^a Mean and Max indicate that the fitness function used during design is the mean or maximum energy across multiple chemical states, respectively.

First, we tested how predictive a chemical state that used a scoring function that took the maximum sequence energy between the HIS, GLU and PLP states (i.e., the HisGluPLP_{Max} state) would have been if were used during our library design procedure. If we had used the HisGluPLP_{Max} state, the MLS to include 3 of the 4 mutants would have risen to 72 members, which suggests that a fitness function that takes the average ensemble score of all the chemical states is better than a fitness function than the maximum sequence energy between the chemical states. Next, we tested chemical states that were only composed of the HIS and GLU states (i.e., HisGlu) or the HIS and PLP states (i.e., HisPLP). The MLS analysis revealed that the HisGluPLP state was more predictive than both the HisGlu and HisPLP states, which suggests that consideration of as many relevant chemical states as possible for the redesign of a multi-substrate enzyme may be beneficial.

In order to assess contributions of each individual chemical state to the HisGluPLP state, we looked at the MLS required for our library design method to predict the 4 BCAT variants for each individual chemical state. Interestingly, library design with the single-chemical state HIS would have included the 4 BCAT variants with MLSs less than or equal to the MLSs for library design with the HisGluPLP state. Alternatively, library design with either the GLU state or PLP

state would have required an MLS at least 3-fold higher than library design with the HisGluPLP state to include any of the 4 BCAT variants.

The low MLS required for library design with the HIS state and the high MLS required for library design with the PLP state suggests that the HIS state likely had the largest contribution towards the success of the HisGluPLP chemical state to predict BCAT variants that display enhanced transamination activity towards L-histidine relative to WT, while the PLP state removed mutations that significantly destabilize the enzyme but form stabilizing nonproductive interactions with substrates. The high MLS of the GLU state suggests that the GLU state predicted mutations that contribute to interactions not necessarily beneficial to either activity, specificity, or stability. Additionally, the HIS state was the only single chemical state where the V109I and G196C mutations both had probabilities > 2% in the PPMs and sequence logos (Figure 3.3), which also indicates that the HIS state contributed the most towards the HisGluPLP state for the prediction of the V109I/G196C double mutant.

The success of the HIS chemical state is likely because L-histidine is a poor substrate for WT BCAT and because α -ketoglutarate is a natural substrate of WT BCAT. Therefore, although mutations that are beneficial for the transamination of L-histidine may be detrimental to L-leucine and α -ketoglutarate transamination, the BCAT active site was optimized for its natural substrates enough that the transamination of L-leucine and α -ketoglutarate still proceeds more efficiently than for L-histidine. Although the success of the single HIS state suggests that the consideration of multiple chemical states was not required for the redesign of BCAT to catalyze the transamination of L-histidine and α -ketoglutarate, the methods now exist to incorporate multiple chemical states for applications that do require it.

3.3.2 Comparison to other methods

A strategy that has been used to engineer proteins is iterative saturation mutagenesis (ISM).¹⁴³⁻¹⁴⁵ ISM is a rational design approach consisting of iterative rounds of saturation mutagenesis at user-defined sites in an enzyme structure. After each round of saturation mutagenesis, the gene of the most improved variant then serves as a template for another round of saturation mutagenesis at the respective other sites. This process is continued until a mutant variant displays the desired trait. Application of ISM would have required saturation mutagenesis of the six designed positions, which means 324 colonies ($6 \text{ positions} \times 18 \text{ amino acids} \times 3 \text{ to ensure } 95\% \text{ sequence coverage}^{146}$) would have been required to be screened in the first round to identify the V109I mutant alone. Additional rounds of saturation mutagenesis of the remaining 5 positions with the V109I variant (an additional 270 variants to screen) would have been required to identify the V109I/G196C double mutant. For comparison, our library design procedure with either the HisGluPLP chemical state or the HIS chemical state would have predicted a 24-member library that included both the V109I/G196C double mutant and the V109I single mutant, which would have required only 72 variants to have been screened to ensure 95% sequence coverage¹⁴⁶ of the library.

Another common protein engineering strategy is directed evolution,¹⁴⁷ which mimics evolution by artificial selection. Random mutations are introduced to an existing protein. The progeny proteins are screened for an improved trait. The improved proteins then serve as parents for the next round of mutation and this process is continued until a mutant variant displays the desired trait. The library design we developed was also significantly more efficient than directed evolution to change substrate specificity, for which the frequency of beneficial mutations in random libraries is on the order of 0.1%.¹⁴⁸ For example, Yano et al.¹⁴⁹ engineered aspartate aminotransferase to be $>10^5$ times more active towards branched chain amino acids with 5 rounds

of directed evolution with 10^6 - 10^7 variants per round, which is on average a 10-fold increase in activity per round. In another example, in a second round of directed evolution, Savile et al.¹⁵⁰ screened 2,880 (R)-selective transaminase (ATA-117) mutants to identify a variant that displayed a 75-fold increase in activity relative to the parent enzyme towards the amination of prositagliptin ketone to provide enantiopure sitagliptin, which is a hit rate of 0.03%. In contrast, our computational procedure was able to yield four mutants that displayed a large increase in k_{cat}/K_M (6- to 46-fold) from a library that contained just 32 variants, which represents a hit rate of 12.5%.

3.4 Conclusions

The multi-chemical state CPD framework presented here incorporates multiple chemical states into the redesign of enzymes that require multiple substrates to undergo a complete reaction mechanism. The method helps to generate mutant libraries that are rich in diversity and are enriched in functional proteins. Here, we used the framework to redesign the BCAT active site to enhance its ability to catalyze the transamination of L-histidine using α -ketoglutarate as an amino acceptor. Our best BCAT variant from a library of 32 members displayed a L-histidine transamination activity 50-fold higher than WT BCAT. If desired, this variant could be improved through additional rounds of design or directed evolution. Although the inclusion of multiple chemical states was not necessary in this application, the method developed here can be used in applications where multiple chemical states are required.

3.5 Materials and methods

3.5.1 Selection of the intermediate PLP states to model.

The aminotransferase reaction mechanism comprises several intermediate PLP states. Ideally, every PLP state along the reaction coordinate would be included in our computational procedure to ensure that designed sequences retain catalytic efficiency. However, a BCAT

structure has not been solved for every intermediate PLP state along the reaction coordinate. The external aldimine state was selected as the template for design since a BCAT crystal structure containing this state is available.⁸⁹ Additionally, the amino-acid moiety in the external aldimine state is structurally similar to an amino-acid residue on a peptide chain, with the sp^3 C_α atom bound to carboxylate and amino groups via single bonds. This allowed us to parameterize substrates with the van der Waals and partial charge parameters used for its amino acid residue counterpart in the PHOENIX force field. Since the enzyme must also be stable in the absence of substrate, the substrate-free internal aldimine was included in our design procedure.

3.5.2 Structure Preparation

The BCAT crystal structures 1I1L⁸⁹ and 1I1K⁸⁹ were obtained from the PDB. The BCAT structures crystallized as a homo trimer. The symexp function of PyMol (DeLano Scientific) was used to reconstruct the hexamer structure. Then, the coordinates of the dimer containing the A chain from the original crystal structure was extracted and used for further structure preparation. The 1I1L structure contains the substrate analogue 2-methylleucine covalently bound to PLP in the external aldimine form. The 2-methyl group was removed and replaced with a hydrogen atom to form a leucine-bound external aldimine.

Although the 1I1K crystal structure was the open conformational state of BCAT containing the internal aldimine, it could not be used as a template in computational calculations because electron density of the interdomain loop (P125 to E136) that hangs over the active site entrance is missing. To prepare an internal aldimine model, the coordinates of the PLP, Y164, Y31, and the catalytic K159 were taken from the 1I1K structure and transplanted into the 1I1L structure. This structure is used to model the closed internal aldimine structure.

All structures have been prepared using the software package Molecular Operating Environment (MOE).¹²⁸ The AMBER99 force field¹²⁹ was used with a combined explicit and implicit reaction field solvent model as implemented by MOE. Hydrogens were added using the Protonate 3D utility, and the structures were solvated with counter ions (Na^+ and Cl^-) in a rectangular box of water under periodic boundary conditions with a box cutoff of 6 Å once protonation states and atomic hybridizations of PLP-substrate were verified to be external or internal aldimines. The protonation state and hydrogen bonds of PLP were adjusted to be consistent with NMR studies:¹³⁰ the pyridine and imine nitrogen atoms were protonated, while all phosphate oxygens and the phenol oxygen were deprotonated. The ϵ -amino group of the catalytic lysine was modeled as neutral to remain consistent with the catalytic cycle.

3.5.3 MD ensemble generation

The prepared leucine-bound external aldimine and internal aldimine structures were minimized by conjugate gradient to a root mean square gradient (RMSG) of below $0.1 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$. These two seed structures were heated to 300 K over a course of 10 ps and equilibrated for 10 ps, which was followed by a 1 ns 300 K NPT production run. 100 structures were extracted from each trajectory at 10 ps intervals and were minimized to a RMSG below $7 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$, a RMSG that gave favourable results to Davey et al.⁵⁶ The 100-member ensemble derived from the leucine-bound external aldimine structure is called MD_{EX} . The deviation from the input structure and diversity of the MD_{EX} ensemble is $1.0 \pm 0.1 \text{ \AA}$ and $0.8 \pm 0.2 \text{ \AA}$, respectively. The 100-member ensemble derived from the internal aldimine is called MD_{IN} , and its deviation from the input structure and diversity of the MD_{IN} ensemble is $0.9 \pm 0.1 \text{ \AA}$ and $0.8 \pm 0.1 \text{ \AA}$, respectively.

3.5.4 Computational details

The backbone-independent Dunbrack rotamer libraries³⁰ with expansions of ± 1 standard deviation around χ_1 and χ_2 were used to model side-chain conformations and the Targeted substrate

rotamer libraries generated in Chapter 2 were used to model the substrate conformations while the PLP moiety of the external or internal aldimine was kept fixed. Both neutral tautomers of histidine, HID and HIE, where the δ - and ϵ -imidazole nitrogen is protonated, respectively, were included when modeling this substrate. Rotamer combinations were evaluated using a four-term potential energy function consisting of a van der Waals term from the Dreiding II force field³⁸ with atomic radii scaled by 0.9, a direction sensitive hydrogen-bond term with well depth at 8.0 kcal mol⁻¹,⁴⁰ an electrostatic energy term modeled using Coulomb's law with a distance dependent dielectric of 10, and an occlusion-based solvation penalty term.³⁷ Fast and Accurate Side-Chain Topology and Energy Refinement (FASTER)^{22-23, 67} was used to optimize sequences for single- and multistate designs.

In order to develop a multistate method for the design of multi-substrate enzyme specificity, we need to compare the score values for each sequence across many chemical states (e.g., HIS, GLU, PLP). However, the sequence space that we propose to search is very large, which makes it intractable to instantiate every possible sequence on every chemical state. On the other hand, sequences could be simultaneously optimized in the context of every chemical state using a multistate design framework, but this could result in a sequence space search that is heavily biased by the chemical state that contributes the most to the fitness function used during optimization. Thus, a new multistate method was developed to enable the design of multi-substrate enzyme specificity (Figure 3.5).

In this method, we start by generating backbone ensembles of the enzyme in each of its three chemical states. This is done to approximate the intrinsic flexibility of the protein, which we have shown improves the accuracy of design predictions. Next, the design positions are divided into separate design calculations that we call sub-designs. This is done so that the number of

possible sequences that are to be rotamer-optimized can be reduced to a value that is computationally tractable. Importantly, each sub-design must contain at least one design position in common with each of the other sub-designs (Table 3.9). This overlap in design positions is necessary to ensure that results from each sub-design can be recombined to yield a single set of mutations predicted to satisfy the design objective.

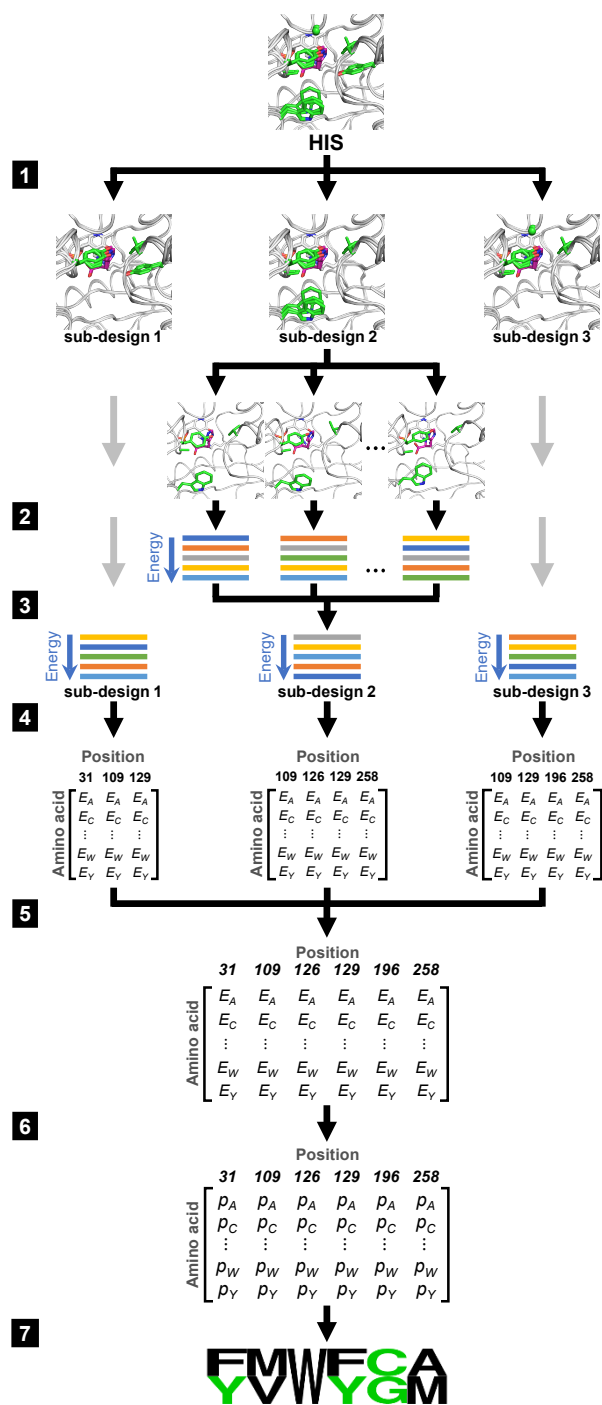


Figure 3.5. Sequence optimization scheme for a single chemical state. Step 1. The design is divided into sub-designs to allow a full search of sequence space. Step 2. Each ensemble member is rotamer-optimized in the context of each sub-design to yield a list of scored sequences for each ensemble member. Step 3. The Boltzmann weighted average is taken over all ensemble members to create a single list of scored sequences for each sub-design. Step 4. The lists of scored sequences are converted to position energy tables for each sub-design. Step 5. The sub-design position energy matrices are combined into a final position energy matrix. Step 6. The position energy matrix is

converted into a position probability matrix. Step 7. The position probability matrix is used to design a combinatorial library of a target size.

Table 3.9. BCAT residues optimized by computational design

Sub-design	Designed positions ^a	Floated positions ^b	Total sequence space
1	Y31, V109, Y129	R97, W126, E197	5,832 (18 ³)
2	V109, W126, Y129, A258	R97, F124, E197	104,976 (18 ⁴)
3	V109, Y129, G196, A258	Y31, R97, W126, E197	104,976 (18 ⁴)

^a All proteinogenic amino acids with the exception of proline and lysine were allowed at these positions.

^b Residues whose rotameric configuration but not identity was allowed to vary

Each sub-design consists of multiple independent single-state calculations performed in parallel. During these calculations, rotamers are optimized on each individual backbone from the ensemble for every chemical state, which results in a ranked list of scored sequences obtained from each ensemble member. The sequence scores obtained from each single-state calculation are combined post-CPD into a fitness value for each sequence across the ensemble. Sequences are then re-ranked based on their fitness values, generating a new ranked list of scored sequences for each sub-design.

To represent the energies of mutations across the different sub-designs, these sequence lists are then used to construct a position energy matrix (PEM) for each sub-design. A PEM is a table of amino acid energies versus position, where each energy reflects the stability of an amino acid at that position. These sub-design PEMs are then combined into a single PEM representing the energy of each amino acid substitution at every position for the design for every chemical state. The PEM can be converted to a position probability matrix (PPM), which is used to generate the final combinatorial library.

3.5.5 Construction of position matrices

A $20 \times n$ Position cumulative partition function Matrix (PQM) of each amino acid at all n designed positions (Equation 3.1) is constructed. The Boltzmann weight, w , is calculated for each sequence (Equation 3.2), where E is the score of the sequence, k_B is the Boltzmann constant, and T is the simulation temperature. The cumulative partition function, q , of an amino acid, aa , at position n , is the sum of w of all sequences where the identity of the residue at position n is the amino acid in question ($resi_n = aa$ in Equation 3.3). A Position Energy Matrix (PEM) is constructed by using Equation 3.4 to convert the cumulative partition function, q , of amino acid aa at position n to energy, E , where k_B is the Boltzmann constant, and T is the simulation temperature.

$$PQM = \begin{bmatrix} q_{A,1} & q_{A,2} & \cdots & q_{A,n-1} & q_{A,n} \\ q_{C,1} & q_{C,2} & \cdots & q_{C,n-1} & q_{C,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ q_{W,1} & q_{W,2} & \cdots & q_{W,n-1} & q_{W,n} \\ q_{Y,1} & q_{Y,2} & \cdots & q_{Y,n-1} & q_{Y,n} \end{bmatrix}$$

Equation 3.1. The Position Partition function matrix, PQM.

$$w = e^{-E/k_B T}$$

Equation 3.2. The Boltzmann Weight.

$$q_{aa,n} = \sum_{i=1}^S w_i [resi_n = aa]$$

Equation 3.3. The cumulative partition function.

$$E = -k_B T \cdot \ln(q_{aa,n})$$

Equation 3.4. The Boltzmann energy.

We were interested in choosing a method to combine the PEMs of the 3 sub-designs in a way that closely reflects the PEMs of all three sub-designs. There are 3 central positions (109, 129, 258) that are included within more than one sub-design. The PEM column from different sub-designs corresponding to the same position were combined by taking the worst energy (max), the mean energy, and the best energy (min). The mean energy had on average the highest correlation (Table 3.10) to the individual sub-design PEM columns for all chemical states. Therefore, the sub-designs were combined by taking the mean energies in PEMs for positions shared across different sub-designs.

Table 3.10. Correlation of residue probabilities between sub-designs. Average Pearson correlation coefficient (PCC) and average Spearman rank coefficient (SRC) of residue probabilities assigned to positions (PID) 109, 129, 258 between sub-designs and the combination from taking the Max, Mean and Min of the sub-designs. PCCs are the average over all chemical states and all sub-designs.

PID	PCC			SRC		
	Max	Mean	Min	Max	Mean	Min
109	0.944	0.966	0.950	0.909	0.929	0.909
129	0.717	0.777	0.661	0.724	0.766	0.674
258	0.893	0.926	0.895	0.890	0.920	0.894

The Position Probability Matrix (PPM) can be constructed using the PEM by converting the PEM back into a PQM. Each column is divided by its sum, Q_n , to produce a PPM (Equation 3.5), which consists of the probabilities of amino acids to be present in a specific position of a sequence.

$$PPM = PQM \div \begin{bmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_{n-1} \\ Q_n \end{bmatrix} = \begin{bmatrix} p_{A,1} & p_{A,2} & \cdots & p_{A,n-1} & p_{A,n} \\ p_{C,1} & p_{C,2} & \cdots & p_{C,n-1} & p_{C,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{W,1} & p_{W,2} & \cdots & p_{W,n-1} & p_{W,n} \\ p_{Y,1} & p_{Y,2} & \cdots & p_{Y,n-1} & p_{Y,n} \end{bmatrix}$$

Equation 3.5. Position Probability Matrix.

3.5.6 Library design

The PPM is used as input for library design. For a library size configuration, which is the specific number of amino acids at each position in the protein, the optimal amino acids at each position are those with the largest probabilities for each position. All library size configurations that lead to a combinatorial library within a target size range are scored by taking the sum of all partition functions of the chosen amino acid sets over all positions, and subsequently ranked.

3.5.7 Saturation mutagenesis

Saturation mutagenesis libraries were made for 5 active-site residues that directly interact with the side chain of the α -amino acid substrate (Y31, R97, V109, Y129, G196). The four primers containing one of the four degenerate codons NDT (encoding 12 amino acids: C, D, F, G, H, I, L, N, R, S, V, Y), VMA (encoding 6 amino acids: A, E, K, P, Q, T), ATG (M), and TGG (W) were synthesized independently and mixed at a ratio of 12:6:1:1 to give a primer solution containing 20 unique codons in equal amounts that encode for all 20 proteinogenic amino acids.¹⁵¹ All saturation mutagenesis libraries were screened for enhanced activity for histidine using the same protocol described for the library screening in section 3.5.9.

The V109I mutation was the only mutation that displayed a high activity towards histidine relative to WT, so we tested an additional saturation library of nearby residue Y129 with the V109I mutation using the same protocol as above. The double mutants V109I/Y129V and V109I/Y129I

displayed 50- and 15-fold higher catalytic efficiency towards L-histidine than wild type, the former having a $k_{\text{cat}}/K_{\text{M}}^{\text{His}}$ within error of the most active mutant from the designed library. Interestingly, these two Y129 mutations are scored unfavourably in PPMs.

3.5.8 Mutant Library Construction

Codon-optimized *E. coli* BCAT was obtained from Integrated DNA Technologies and subcloned into pET11-a (Novagen) via *NdeI/BamHI*. The plasmid was transformed into *E. coli* BL21-Gold(DE3) cells (Stratagene). To construct the 32-member combinatorial BCAT mutant library, mutagenesis was performed as described by Vallejo et al.¹⁵² using VentR DNA polymerase. Inspired by Tang et al.,¹⁵¹ multiple degenerate codons were used in order to eliminate redundant codons and minimize uneven distributions of amino acids in the library. Following gel purification (Omega Biotek), the amplicons were recombined by overlap extension PCR, resulting in the final gene. The gene was gel purified and digested with the restriction enzymes *NdeI/BamHI*. Following another gel purification, the digested fragment was ligated into the expression vector pET11-a with T4 DNA ligase. The resulting expression vectors containing the aminotransferase genes was then transformed into *E. coli* BL21-Gold(DE3) gold cells. Colonies were picked into individual wells of V96 MicroWell polypropylene plates (Nunc) containing 300 μL of medium (LB with 100 $\mu\text{g mL}^{-1}$ ampicillin). Wild-type BCAT in pET11-a vector, pET11-a vector containing no BCAT, and the BCAT V109I/Y129V double mutant in pET11-a vector were picked into two wells each as controls. These plates were covered with a sterile breathable membrane (VWR) and incubated overnight at 37 °C with shaking. These mother plates were used to inoculate Sterile Nunc V96 MicroWell plates containing 300 μL of Overnight Express Instant TB media (Novagen) supplemented with ampicillin per well, incubated overnight at 37 °C with shaking. The cells were washed and lysed to prepare the clarified lysates as described by Walton et al.¹²¹ except 5 U mL^{-1} of benzonase nuclease (EMD) was used instead of 25 U mL^{-1} in the lysis buffer.

3.5.9 Screening

Screening assays were performed in 200- μ L reactions containing final concentrations of 2 mM histidine, 0.5 mM α -ketoglutarate, 5 mM NAD^+ , 16 μ M PLP, and 1 U of GDH in 100 mM potassium phosphate buffer (pH 8.0, 37 °C). 96-well plates (Greiner Bio-One) containing the standard reaction mixture were preincubated at 37 °C for 5 min prior to initiation of the reaction by the addition of 20 μ L of clarified cell lysates. Enzyme reactions were monitored by measuring absorbance of NADH at 340 nm ($\epsilon = 6220 \text{ M}^{-1} \text{ cm}^{-1}$) every 12 seconds for 30 minutes. Path lengths for each well were calculated ratiometrically using the difference in absorbance of potassium phosphate buffer at 900 and 998 nm. The slopes of the kinetic traces were used to measure reaction rates.

3.5.10 Steady-state kinetics

The selected mutants were expressed and purified to at least 95 % purity as described by Mironov et al.¹¹⁵ SDS-PAGE (Figure 3.6) was performed to ensure the mutants were indeed BCAT variants (MW = 35 kDa per monomer). All kinetic assays were performed in triplicate. To measure the steady-state kinetics for histidine and leucine with the BCAT mutants, GDH assays were performed by varying the amino acid concentration up to 20 mM in the presence of 0.5 mM α -ketoglutarate, 5 mM NAD^+ , 16 μ M PLP, 1 U GDH and approximately 10 mU of aminotransferase in 100 mM potassium phosphate buffer (pH 8, 37 °C). To measure the steady-state kinetics for α -ketoglutarate with the BCAT mutants, the α -ketoglutarate concentration was varied up to 4 mM while the leucine concentration was kept constant at 2 mM. The linear phases of the kinetic traces were used to measure the initial reaction rates. The initial reaction rates at different amino acid concentrations were fit to the Michaelis-Menten equation (Equation 3.6) using the GraphPad Prism software to compute apparent steady-state kinetic parameters, k_{cat} and K_{M} , where v is the initial reaction rate, $[E]$ is the total enzyme concentration, and $[S]$ is the substrate

concentration. Since the BCAT mutants displayed substrate inhibition with α -ketoglutarate, the initial reaction rates were fit to the substrate inhibition equation to determine the apparent steady state kinetic parameters k_{cat} and K_M as well as the inhibition constant K_I (Equation 3.7) for that substrate.

$$v = \frac{[S] \cdot k_{\text{cat}} \cdot [E]}{[S] + K_M}$$

Equation 3.6. Michaelis-Menten equation.

$$v = \frac{[S] \cdot k_{\text{cat}} \cdot [E]}{K_M + [S] \cdot \left(1 + \frac{[S]}{K_I}\right)}$$

Equation 3.7. Michaelis-Menten equation with substrate inhibition.

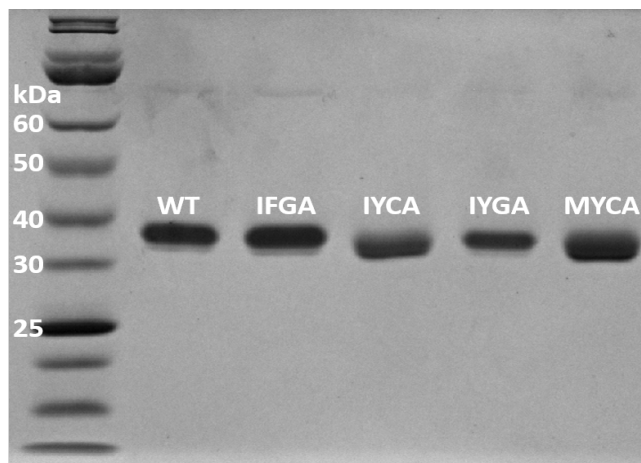


Figure 3.6. SDS-PAGE of wild type BCAT and the top 4 BCAT mutants , where each letter refers to residue at positions 109, 129, 196, and 258. The marker on the left is the P7704S protein marker (New England Biolabs).

Chapter 4: Computational design of conformational equilibrium to modulate the substrate specificity of an enzyme

Antony D. St-Jacques, Safwat T. Khan, Roberto A. Chica

4.1 Statement of contribution

Safwat T. Khan prepared the combinatorial DNA libraries and sequenced AAT mutants. All CPD calculations and all other experiments were performed by Antony D. St-Jacques. Experiments were designed by Antony D. St-Jacques and Dr. Roberto A. Chica. The chapter was written by Antony D. St-Jacques and edited by Dr. Roberto A. Chica.

4.2 Introduction

The conformational change *Escherichia coli* aspartate aminotransferase (AAT) undergoes during transamination of L-aspartate with α -ketoglutarate to form oxaloacetate and L-glutamate is well documented (Figure 4.1a,b).^{81, 96-97} AAT is an obligate homodimer and each subunit is composed of a static large domain and a moving small domain. In solution, AAT is in equilibrium between its open and closed conformations, where the small domain approaches the large domain and packs together the residues at the entrance of the active site. Aspartate binds to the open AAT form and induces a conformational change to the closed form in which the transamination reaction occurs. AAT switches back to the open state and releases the product in a partially rate-determining step.¹¹¹ In the case of the reverse reaction, glutamate also binds AAT in the open state, however, the domain closure is induced during the transaldimination step to form the external aldimine rather than during formation of the noncovalent Michaelis complex with the enzyme.⁹⁷

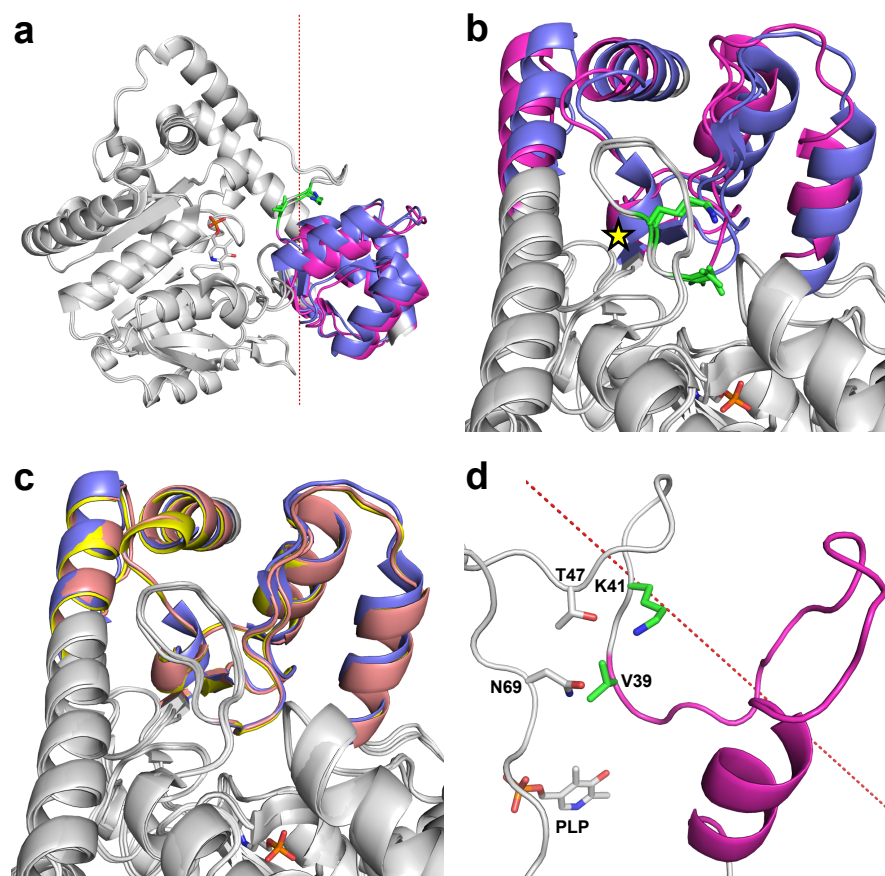


Figure 4.1. Comparison of the different conformational forms of wild type (WT) AAT and the AAT hexamutant (HEX). The static domain is shown in white. Moving domain is represented by pink (Open) and blue (Closed). PLP and bending residues (green) are shown as sticks. The red dotted line is the axis of rotation. (a) Superposition of AAT crystal structures in the open (PDB ID: 1ARS) and closed (PDB ID: 1X28) forms aligned to the static domains. (b) Top view of panel a. The star is the axis of rotation. (c) A superposition of the crystal structures of the closed WT AAT structure from panel b and the crystal structure of the AAT hexamutant with (light pink) and without (yellow) the ligand hydrocinnamic acid (not shown) bound to the active site (PDB ID: 1AHX and 1AHE, respectively). (d) Designed residues (39, 41, 47, and 69) and PLP are shown as sticks on the crystal structure of the open form (PDB ID: 1ARS). Residues and backbones in white, pink, and green are part of the static domain, moving domain, and bending residues, respectively. Residues are labelled next to the C α by their one-letter amino acid code and residue number.

E. coli tyrosine aminotransferase (TAT) is structurally homologous to AAT (43% and 72% sequence identity and similarity, respectively) yet catalyzes the transamination of α -ketoglutarate with L-aromatic amino acids such as L-tyrosine and L-phenylalanine, which are poor AAT substrates.⁵ Previously, AAT was engineered to display TAT activity by identifying residues

important for binding aromatic amino acids through multiple sequence alignment,⁵ rational design,¹⁵³ or bioinformatics,¹²⁶ and mutating these to their corresponding amino acids found in TAT.

For example, Onuffer et al.⁵ used the sequence alignment of AAT with TAT¹⁵⁴⁻¹⁵⁵ to identify mutations required to increase the activity of AAT towards phenylalanine. This analysis resulted in a hexamutant (HEX) containing the V39L, K41Y, T47I, N69L, T109S, and N297S mutations, which together decrease the K_M for phenylalanine (K_M^{Phe}) by 58-fold and increase overall activity towards this non-native substrate by 50-fold. Unlike wild-type (WT) AAT, which crystallizes in the open form in the absence of substrate, HEX crystallized in the closed conformation in both the presence and absence of substrate in the active site (Figure 4.1c). This result suggests that the conformational equilibrium of the open and closed states for HEX has shifted towards the closed state relative to WT. In a later study, Addington et al.¹²⁶ used a bioinformatics method to convert AAT to TAT and the conformational equilibrium of their AAT variant was also found to have been shifted to prefer the closed state, which suggests that a correlation between conformational equilibrium and affinity for phenylalanine exists in AAT. This raises the possibility that we could use computational protein design (CPD) methods to predict mutations that can modulate the conformational equilibrium of AAT in order to control its specificity towards substrates. Additionally, we could validate the designs to modulate the conformational equilibrium of AAT with kinetic assays.

Conformational changes are a common feature of enzymatic mechanisms¹⁵⁶⁻¹⁵⁸ and therefore controlling conformational changes is an important objective of enzyme engineering. In this work, we aim to develop a multistate CPD method to modulate the conformational equilibrium of an enzyme, and in doing so, alter its substrate specificity. As a case study, we exploit the

apparent link between the conformational equilibrium of AAT and its activity towards the non-native substrate phenylalanine. We postulate that the conformational equilibrium of AAT can be shifted by using multistate analysis (MSA) to predict mutations that will preferentially stabilize one conformer over the other, leading to an associated change in K_M for phenylalanine.

4.3 Results

To design a shift in conformational equilibrium between the open and closed conformational states of AAT, we followed a 6-step process: Identification of residues important for domain movement; ensemble preparation; sequence optimization of each chemical state; multichemical state analysis; library design, and a 2-concentration screening of the mutants.

4.3.1 Domain movement analysis and residue selection

To identify residues important for domain motion in AAT, the crystal structures of the open (PDB ID: 1ARS¹²⁷) and closed forms (PDB ID: 1ART¹²⁷ and 1X28⁹⁷) were used as input for domain movement analysis using the DynDom1D algorithm.¹⁵⁹⁻¹⁶² Three independent analyses with DynDom1D (Table 4.1) revealed a “fixed” domain and a “moving” domain (Figure 4.1a,b), which contained between 295–302 and 80–89 residues, respectively. The moving domain was determined to rotate $7.3 \pm 0.6^\circ$ about the hinge axis from the fixed domain. Although the 3 analyses did not give exactly the same results due to differences in the atomic coordinates of the three closed structures, 97% of residues were assigned to the same domain (i.e., fixed domain, moving domain, unassigned) in all three analyses. Most of the residues that were not assigned to the same domain across all three analyses were hinge-bending residues that are situated at a transition between the fixed and moving domains and are involved in the domain motions. Because of their involvement with domain motions, the hinge-bending residues were determined to be good candidates for design to modulate the preference for one conformation over another. Nine hinge

bending-residues were identified to be involved in domain motion in all three analyses. On the 1ARS PDB structure, the STRIDE¹⁶³ algorithm determined residues 39-41 and 357-359 have coil secondary structure, which is ideal since coils are more flexible than other secondary structure elements. Residues 39 and 41 were selected for design because they are the only pair of hinge-bending residues that are on an interdomain loop with side chains that are situated between the moving and fixed domains (Figure 4.1b,d). Residues 47 and 69 were also included in designs since they are the two closest residues to residues 39 and 41 that (1) have coil secondary structure and (2) do not form a lid over the active site.

Table 4.1. Domain motion data for *E. coli* AAT between the open conformation (PDB ID: 1ARS) and closed conformation from DynDom1D analysis

	PDB ID of closed conformation		
	1ART	1X28, chain A	1X28, chain B
Fixed Domain (residues)	35, 40-336, 351-354, 357-360, 387	35, 40-334, 351-360, 387-390	40-334, 353-360
Moving Domain (residues)	17-34, 36-39, 337-350, 355-356, 361-386, 388-406	17-34, 36-39, 335-350, 361-386, 391-406	15-39, 335-352, 361-406
Unassigned residues*	1-16, 407-408	1-16, 407-408	1-14, 407-408
Angle of rotation (deg)	7.1	6.8	7.9
Translation along axis (Å)	-0.3	-0.4	-0.4
Residues involved in motion	33-36, 39-41, 335-337, 350-361, 386-388	34-36, 39-41, 334-335, 350-351, 357-362, 386-391	39-42, 333-335, 352-353, 356-362
Residues involved in motion common to both chains	39-41,** 335, 357-359,** 360-361		

*Not in the fixed domain or the moving domain

** Bold residues were determined by the STRIDE¹⁶³ algorithm to have a coil secondary structure

4.3.2 Ensemble Preparation

Backbone ensembles of the open and closed states of AAT were generated in order to mimic backbone flexibility and address the fixed backbone approximation in CPD calculations, which has been shown to improve accuracy of protein stability predictions.⁵⁶ The crystal structures 1ARS¹²⁷ and 1X28⁹⁷ from the PDB²⁶ were selected as templates for the open and closed states of

E. coli AAT, respectively. The glutamate-bound ketimine complex and catalytic lysine (K258) in the 1X28 structure were replaced with the coordinates for the K258 and PLP from 1ARS so that both the structural models of the open and closed forms were in the internal aldimine state. Use of the internal aldimines eliminates potential biases in the generated ensemble that can arise from the presence of substrates in the AAT active site. Each structure was separately prepared using the software package Molecular Operating Environment (MOE).¹²⁸ The AMBER99 force field¹²⁹ was used with a combined explicit and implicit reaction field solvent model as implemented by MOE. Hydrogens were added using the Protonate 3D utility, and the structures were solvated in a rectangular box of water under periodic boundary conditions with a box cutoff of 6 Å once the protonation states and atomic hybridizations of the PLP internal aldimine were corrected to reflect NMR data.¹³⁰ The two structures were minimized to a gradient below 7 kcal mol⁻¹ Å⁻¹ and used as input to generate two 50-member ensembles (structural information in Table 4.2) using the PertMin⁵⁶ method. The ensemble derived from the 1ARS crystal structure (i.e., the Open ensemble) and the ensemble derived from the 1X28 crystal structure (i.e., the Closed ensemble) were used as conformational ensembles to model the open and closed states of AAT, respectively.

Table 4.2. Structural information of the Open and Closed 50-member PertMin ensembles.

	Open	Closed
RMSG ^a (kcal·mol ⁻¹ ·Å ⁻¹)	0.63 ± 0.02	0.63 ± 0.04
RMSD ₀ ^b (Å)	0.51 ± 0.02	0.51 ± 0.02
Diversity ^c (Å)	0.29 ± 0.03	0.32 ± 0.03

^a Root mean square gradient.

^b Backbone root mean square deviation from the input structure.

^c Average backbone root mean square deviation between members of the same ensemble.

4.3.3 Computational Design

The target hinge-bending residues (39 and 41) and nearby residues (47 and 69), shown in Figure 4.1d, were modeled to all proteinogenic amino acids combinations except proline on all ensemble members of both the Closed and Open ensembles (130,321 sequences $\times$ 50 backbones $\times$ 2 conformational state = 13,032,100 individual scores). Nearby residues at positions 34, 37, 70, 67, and 263 were also rotamer-optimized but their identity was not allowed to vary. The ensemble score of each sequence for a conformational state (Open or Closed) was computed as the Boltzmann weighted average (300 K) across the individual scores of each backbone member of the ensemble.

To design mutants that stabilize preferentially the closed state, sequence fitness values were computed as the difference between the Closed ensemble score and the Open ensemble score. To avoid favouring sequences that have high energy ensemble scores in both conformational states but have a favourable fitness value, sequences with a Closed ensemble score below the 75th percentile were discarded. Likewise, to design mutants that stabilize preferentially the open state, sequence fitness values were computed as the difference between the Open ensemble score and the Closed ensemble score. To avoid favouring sequences that have high energy ensemble scores in both conformational states but have a favourable fitness value, sequences with an Open ensemble score below the 75th percentile were discarded. To design mutants that equally occupy the open and closed states, the sequences fitness values were computed as the average ensemble score between the two conformational states. The sequence fitness values were then used as input to the CLEARSS³⁶ library design algorithm (300 K) to predict libraries of AAT mutants that will prefer the closed state, open state, or both states, which are called the “Closed,” “Open,” and “Both” libraries, respectively (Table 4.3). Individual ensemble scores for WT sequence and the sequences of these three libraries are listed in Table 4.4-7, respectively. These three libraries were

constructed for experimental screening. AAT mutants are named as the concatenation of the 1-letter code for the amino acids at positions 39, 41, 47, and 69 (e.g. WT is called VKTN).

Table 4.3. The Both, Closed, and Open AAT combinatorial libraries.

Residue	WT	Closed	Both	Open
39	V	ILVW	IMV	ACQS
41	K	FY	FY	I
47	T	I	CN	F
69	N	FLTY	AHS	ACFHMSWY

Table 4.4. Computed ensemble scores for the Open and Closed state of the WT *E. coli* AAT sequence, as well as the energy difference. Scores are in kcal mol⁻¹.

V	K	T	N	Close	Open	ΔE_{C-O}
-	-	-	-	-358.0	-372.2	14.2

Table 4.5. Computed ensemble scores for the Open and Closed state of variants of the Closed library , as well as the energy difference. Scores are in kcal mol⁻¹.

Rank	V K T N	Close	Open	ΔE_{c-o}
1	I F I F	-358.9	-272.9	-86.0
2	L F I L	-354.6	-277.2	-77.4
3	L Y I L	-350.2	-273.4	-76.8
4	I F I L	-353.3	-280.0	-73.3
5	I F I Y	-357.3	-291.6	-65.7
6	- F I Y	-357.8	-300.0	-57.8
7	- F I L	-356.6	-298.8	-57.8
8	I Y I F	-306.9	-250.2	-56.7
9	- Y I L	-350.7	-295.4	-55.3
10	W Y I L	-236.1	-181.2	-55.0
11	- Y I Y	-349.5	-294.6	-54.9
12	- Y I F	-352.8	-302.4	-50.4
13	I F I T	-361.7	-313.4	-48.3
14	I Y I L	-301.5	-253.4	-48.2
15	- F I F	-361.9	-315.7	-46.2
16	- F I T	-366.1	-320.6	-45.5
17	- Y I T	-355.8	-315.2	-40.6
18	I Y I Y	-304.5	-273.2	-31.2
19	L F I F	-321.3	-297.2	-24.2
20	L Y I F	-316.6	-295.2	-21.4
21	L F I Y	-334.7	-320.3	-14.4
22	L F I T	-353.9	-341.9	-12.0
23	L Y I Y	-329.8	-318.3	-11.5
24	I Y I T	-306.4	-294.9	-11.5
25	L Y I T	-348.9	-339.5	-9.5
26	W F I L	-240.7	-260.2	19.5
27	W Y I F	-202.5	-266.4	63.9
28	W F I F	-207.3	-271.6	64.2
29	W Y I T	-237.1	-310.1	73.0
30	W F I T	-242.9	-315.9	73.1
31	W Y I Y	-215.7	-289.9	74.1
32	W F I Y	-220.7	-295.1	74.4

Table 4.6. Computed ensemble scores for the Open and Closed state of variants of the Both library, as well as the energy difference. Scores are in kcal mol⁻¹.

Rank	V K T N	Close	Open	ΔE_{c-o}
1	- Y N S	-394.7	-390.4	-4.3
2	- Y N A	-390.8	-386.9	-3.9
3	- Y N H	-391.9	-388.4	-3.5
4	I Y N H	-390.3	-387.1	-3.3
5	I Y C H	-386.6	-383.5	-3.1
6	- Y C S	-390.8	-388.6	-2.2
7	I Y N A	-389.0	-387.2	-1.8
8	- Y C H	-388.0	-386.3	-1.7
9	- Y C A	-386.8	-385.3	-1.5
10	I Y N S	-392.0	-390.7	-1.3
11	I F N H	-387.7	-388.5	0.8
12	I F C H	-384.1	-385.1	1.1
13	- F N S	-390.6	-391.8	1.2
14	- F N A	-386.8	-388.3	1.5
15	I Y C A	-385.1	-386.8	1.7
16	I Y C S	-388.0	-390.1	2.1
17	- F N H	-388.6	-391.6	3.0
18	I F N A	-385.2	-388.6	3.4
19	I F N S	-388.5	-392.1	3.7
20	- F C S	-385.6	-389.7	4.2
21	- F C A	-381.7	-386.3	4.6
22	- F C H	-384.8	-389.5	4.7
23	M Y N A	-382.8	-388.7	5.9
24	M Y N S	-386.8	-392.8	6.1
25	I F C A	-381.3	-387.8	6.5
26	I F C S	-384.5	-391.2	6.7
27	M Y C S	-383.1	-393.2	10.1
28	M Y C A	-379.0	-389.4	10.4
29	M Y N H	-375.2	-385.9	10.7
30	M Y C H	-371.5	-382.2	10.7
31	M F N A	-378.8	-390.3	11.4
32	M F N S	-382.8	-394.5	11.7
33	M F C H	-368.5	-383.9	15.3
34	M F C S	-379.1	-394.6	15.5
35	M F C A	-375.0	-390.8	15.7
36	M F N H	-371.4	-387.8	16.4

Table 4.7. Computed ensemble scores for the Open and Closed state of variants of the Open library, as well as the energy difference. Scores are in kcal mol⁻¹.

Rank	V K T N	Close	Open	ΔE_{c-o}
1	A I F F	-231.5	-359.9	128.4
2	A I F W	-231.1	-359.4	128.2
3	C I F M	-239.9	-367.5	127.6
4	A I F Y	-231.2	-358.8	127.6
5	A I F S	-237.2	-364.4	127.2
6	A I F M	-235.3	-362.4	127.1
7	C I F S	-241.7	-368.5	126.9
8	A I F H	-239.3	-365.5	126.1
9	S I F F	-232.8	-358.6	125.9
10	S I F S	-238.1	-363.8	125.7
11	C I F H	-244.1	-369.4	125.3
12	S I F Y	-232.1	-357.4	125.3
13	S I F M	-236.8	-362.0	125.2
14	A I F C	-235.2	-360.1	124.9
15	A I F A	-237.4	-362.1	124.7
16	C I F C	-240.0	-364.4	124.3
17	C I F A	-242.1	-366.3	124.2
18	C I F W	-238.7	-362.8	124.1
19	S I F H	-240.3	-364.1	123.9
20	S I F C	-236.0	-359.3	123.3
21	S I F A	-238.3	-361.5	123.2
22	C I F F	-236.1	-358.3	122.2
23	S I F W	-236.2	-358.3	122.1
24	Q I F S	-241.9	-363.7	121.8
25	C I F Y	-236.2	-356.9	120.7
26	Q I F A	-242.7	-361.8	119.1
27	Q I F M	-237.6	-354.7	117.1
28	Q I F C	-240.6	-357.2	116.6
29	Q I F H	-244.7	-346.9	102.2
30	Q I F W	-237.1	-333.9	96.8
31	Q I F Y	-237.2	-321.3	84.2
32	Q I F F	-234.6	-315.0	80.4

4.3.4 Screening

The goal of this study is to modulate the equilibrium between the open and closed state of AAT. As the K_M^{Phe} of AAT for phenylalanine is known to decrease with an equilibrium shift towards the closed state, we sought to use a high-throughput assay that can differentiate AAT variants with different K_M^{Phe} . To assay the ability of AAT mutants to catalyze transamination of α -ketoglutarate with phenylalanine, we used the L-glutamate dehydrogenase (GDH) assay developed by Walton et al.¹²¹ Typically, mutants are tested under a single set of conditions (e.g., substrate concentration, enzyme quantity, etc.) when using this assay, which allows selection of

mutants based on reaction rates. This reaction rate, v , can be determined by the apparent kinetic parameters of the enzyme, k_{cat} and K_M , the total enzyme concentration, $[E]$, and the substrate concentration, $[S]$, by using the Michaelis-Menten equation (Equation 4.1).

$$v = \frac{[S] \cdot k_{\text{cat}} \cdot [E]}{[S] + K_M}$$

Equation 4.1. Michaelis Menten equation.

By taking the ratio, r , of the reaction rate at two different substrate concentrations, the terms of the Michaelis-Menten equation can be reduced to r , the K_M , and the two substrate concentrations, $[S]_1$ and $[S]_2$ (Equation 4.2). Equation 4.2 can be rearranged to solve for K_M (Equation 4.3), which leads to the conclusion that screening enzyme activity at two different substrate concentrations can allow mutants with a desired K_M to be identified. However, due to experimental error in measuring reaction rates and the mathematical limitations of Equation 4.2 and Equation 4.3, appropriate substrate concentrations must be selected to have confidence in the K_M measured with the 2-concentration screening protocol. For instance, if $K_M \ll [S]_1$ and $[S]_2$, r would be measured to be ≈ 1 and Equation 4.3 would calculate a K_M of 0, which is impossible. Additionally, if $[S]_1 \approx [S]_2$, r would be ≈ 1 and K_M would also be calculated to be 0. Moreover, if $K_M \gg [S]_1$ and $[S]_2$, then $r \approx [S]_2/[S]_1$. This creates a problem because if $r \approx [S]_2/[S]_1$, Equation 4.3 would calculate a K_M of infinity. Moreover, if r is computed to be $> [S]_2/[S]_1$, K_M will be negative. Therefore, in order to ensure that the K_M of an enzyme can be confidently measured with the 2-concentration screening protocol, it is important to select a pair of substrate concentrations that are similar to the K_M expected to be displayed by the enzyme, and ensure that these concentrations are sufficiently different from one another to maximize r .

$$r = \frac{v_{[S]_2}}{v_{[S]_1}} = \frac{[S]_2 \cdot k_{\text{cat}} \cdot [E]}{[S]_2 + K_M} \cdot \frac{[S]_1 + K_M}{[S]_1 \cdot k_{\text{cat}} \cdot [E]} = \frac{[S]_2 \cdot ([S]_1 + K_M)}{[S]_1 \cdot ([S]_2 + K_M)}$$

Equation 4.2. Ratio, r , of Michaelis-Menten equation at two different substrate concentrations.

$$K_M = \frac{[S]_1 \cdot [S]_2 \cdot (r - 1)}{[S]_2 - [S]_1 \cdot r}$$

Equation 4.3. Calculation of K_M based on Equation 4.2.

AAT mutant libraries were screened in the presence of WT AAT and pET11-a vector containing no AAT as positive and negative controls, respectively. An additional positive control was also included: the VYIL triple mutant (K41Y/T47I/N69L), which we had previously made that displays a K_M^{Phe} 63-fold lower than WT and a k_{cat}/K_M for L-phenylalanine, $k_{\text{cat}}/K_M^{\text{Phe}}$, 20-fold higher than WT. Two positive controls were used because they display different apparent K_M^{Phe} for L-phenylalanine (WT $K_M^{\text{Phe}} = 45 \pm 5$ mM; VYIL $K_M^{\text{Phe}} = 0.71 \pm 0.02$ mM), which allows us to verify whether the 2-concentration screening can be used for the identification of AAT mutants with a desired K_M^{Phe} . The 2 L-phenylalanine concentrations that were selected for the 2-concentration screening assay was 3 and 40 mM. The 3 mM concentration of donor substrate was selected to ensure that only the VYIL positive control is saturated, while the 40 mM concentration was the highest concentration of L-phenylalanine that could be achieved within its solubility limits.

Table 4.8. Screening results of the 2-concentration assay.

	v_{3mM} (mU) ^a	v_{40mM} (mU) ^b	$r_{40mM/3mM}$ ^c	K_M^{Phe} (mM) ^d
WT	4 ± 1	12 ± 3	3 ± 1	8 ± 3
Neg	1 ± 2	1 ± 2	1 ± 6	0
VYIL	7 ± 1	8 ± 2	1.2 ± 0.4	0.7 ± 0.2
Close ^{e,f}	3 ± 3	3 ± 4	1.0 ± 0.8	0
Both ^e	6.6 ± 3	9.6 ± 5.4	1.3 ± 0.5	1.0 ± 0.4
Open ^e	0.9 ± 1.4	1 ± 2	1.3 ± 0.9	1.0 ± 0.7

^a Initial reaction rate at 3 mM L-phenylalanine.

^b Initial reaction rate at 40 mM L-phenylalanine.

^c The ratio of the two reaction rates.

^d K_M^{Phe} derived from the 2-substrate screening.

^e The reaction rates for Close, Both, and Open were obtained by averaging the reaction rates of all mutants of each library that were above 0. The $r_{40mM/3mM}$ for Close, Both, and Open were obtained by averaging the $r_{40mM/3mM}$ of all mutants of each library that displayed reaction rates at both concentrations above 0.

^f WFIY and WYIT were excluded from this analysis of the AAT libraries since they have a Closed ensemble score larger than the Open ensemble score, which implies a preference for the open state rather than the closed state

To evaluate the effect of the different mutations to the AAT hinge residues, we screened clarified cell lysates containing individual mutants from the three designed AAT libraries with 3 or 40 mM L-phenylalanine (Table 4.8). As expected, the positive control VYIL displayed the same reaction rates at both L-phenylalanine concentrations, suggesting that VYIL was saturated at 3 mM L-phenylalanine. Conversely, WT displayed a 3-fold increase in reaction rate at 40 mM relative to 3 mM L-phenylalanine and reached a rate similar to that of VYIL at the same concentration (Figure 4.2). This result suggests that WT was not saturated at 3 mM L-phenylalanine, as expected. To evaluate the ability of the 2-concentration assay to correctly estimate the K_M of enzymes, we compared the apparent K_M^{Phe} for the WT and VYIL positive controls that we measured using standard Michaelis-Menten analysis (Table 4.9) with those obtained by the 2-concentration screening assay (Table 4.8). This comparison revealed that the 2-concentration screening assay was able to correctly determine the K_M^{Phe} of VYIL. However, this result may be due to chance

rather than accuracy because reaction rates of VYIL can deviate by up to 30% when using small amounts of clarified lysate in our microtiter plate assay. Also, the K_M^{Phe} of WT calculated using the 2-concentration screening assay is approximately 6-fold lower than the one determined by Michaelis-Menten analysis, which could be due to interactions with other compounds in the clarified lysates. Nevertheless, the 2-concentration screening method enabled us to correctly determine that K_M^{Phe} of the WT is higher than that of VYIL, suggesting that this method can be used to qualitatively assess K_M^{Phe} for different AAT mutants.

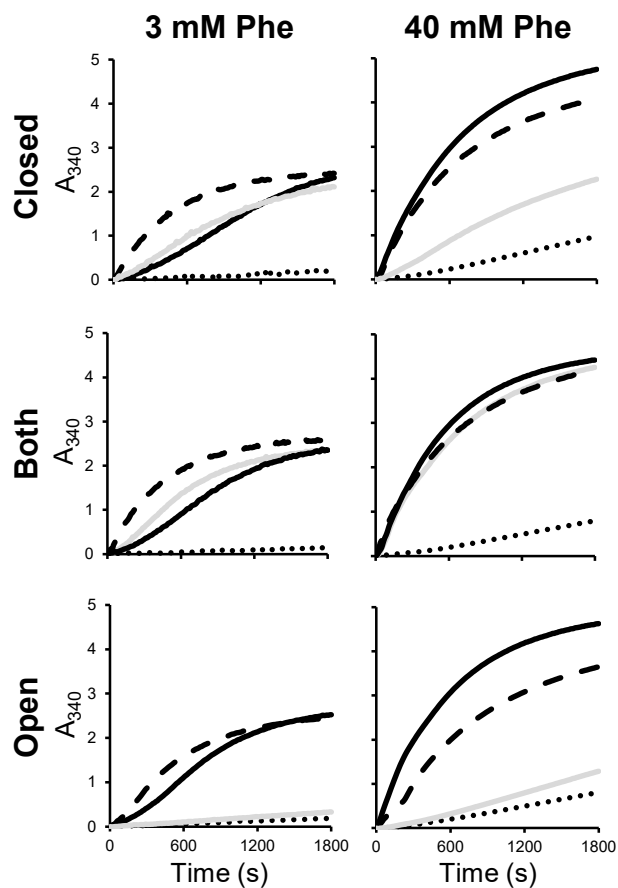


Figure 4.2. Representative kinetic traces of the 2-concentration GDH screening assay. The Closed, Both, and Open libraries were screened for the transamination of α -ketoglutarate with the donor substrate phenylalanine at concentrations of 3 mM (left) and 40 mM (right). The positive controls, WT (solid black) and VYIL (dashed black), and the negative control (dotted black) are shown. AAT mutants IYIT, VFCS, and SIFS from the Closed, Both, and Open libraries, respectively, are shown as grey traces.

Wells from all 3 libraries that displayed high v_{3mM} , v_{40mM} , or $r_{40mM/3mM}$ were sequenced. The data from mutants of each library were averaged together and is listed in Table 4.8 along with the K_M^{Phe} derived from the ratio. Two mutants from the Closed library identified during the screening (WFIY, WYIT) have a Closed ensemble score larger than the Open ensemble score, which implies a preference for the open state rather than the closed state. Therefore, these two mutants were excluded from this analysis of the AAT libraries.

Mutants from the Closed library displayed the lowest $r_{40mM/3mM}$ (1.0 ± 0.8). For example, at 3 mM L-phenylalanine concentration, the IYIT mutant from the Closed library displayed 3-fold less transamination activity than the positive control VYIL and approximately 20% more activity than WT (Figure 4.2). At 40 mM L-phenylalanine concentration, IYIT displayed 40% less activity than it did at 3 mM, which suggests IYIT has a low K_M^{Phe} and may display substrate inhibition.

In general, the Both library contained mutants that display higher transamination reaction rates at both phenylalanine concentrations. The average $r_{40mM/3mM}$ of the Both library was 1.3, which predicts a K_M^{Phe} of 1.0 ± 0.4 mM for the Both library variants. As an example of a mutant from the Both library, the activity of the mutant VFCS from the Both library (4.8 mU) was in between that of WT (3.5 mU) and VYIL (8.8 mU) at 3 mM phenylalanine, but all three AAT variants displayed approximately the same reaction rates at 40 mM phenylalanine (11.6, 13.6, and 12 mU for WT, VYIL, and VFCS, respectively), which indicates that the VFCS mutant has an intermediate K_M^{Phe} between the two positive controls.

Mutants from the Open library had on average the lowest activity at both phenylalanine concentrations and its mean ratio-derived K_M^{Phe} is 1.0 ± 0.7 mM. As an example, SIFS, a mutant from the Open library, displayed between 25- and 61-fold lower transamination activity than WT

and VYIL, but its ratio, $r_{40/3mM}$, was 3.7, which suggested SIFS has a K_M^{Phe} similar to WT ($r_{40/3mM} = 3.1$).

From the results of the 2-concentration screening assay, mutants from the Closed library likely have a low K_M^{Phe} relative to WT, mutants from the Both library appear to be similar to WT, and mutants library from the Open library display low activity towards L-phenylalanine. Next, we performed steady-state kinetics in order to help determine if the differences between the AAT libraries was due to changes in the conformational equilibrium of the AAT variants of each library.

4.3.5 Steady-state kinetics of AAT mutants.

Five AAT mutants from each library that displayed the target phenylalanine activity were kinetically characterized, listed in Table 4.9. Mutants that displayed the highest $r_{40mM/3mM}$ from the Open library were selected for steady-state kinetic characterization since a high $r_{40mM/3mM}$ corresponds to a high K_M^{Phe} . Although a low $r_{40mM/3mM}$ corresponds to a low K_M^{Phe} , many mutants displayed $r_{40mM/3mM}$ lower than 1, which means the rate of reaction was slower at higher substrate concentrations, suggesting substrate inhibition. Since substrate inhibition is not a desired feature for mutants from Closed library, mutants from the Closed library that displayed the highest transamination activity at 3 mM L-phenylalanine and displayed a $r_{40mM/3mM} \approx 1$ were chosen for kinetic characterization. Mutants from the Both library that displayed the highest transamination activity at 40 mM L-phenylalanine were chosen for kinetic characterization since the screening analysis revealed mutants from the Both library displayed relatively high transamination activity at that L-phenylalanine concentration.

Table 4.9. Computed scores and apparent kinetic parameters of AAT variants.

	39 41 47 69				Scores (kcal·mol ⁻¹)			L-aspartate			L-phenylalanine			Relative specificity
	V	K	T	N	Close	Open	ΔE_{C-O}	K_M (mM)	k_{cat} (s ⁻¹)	k_{cat}/K_M (M ⁻¹ ·s ⁻¹)	K_M (mM)	k_{cat} (s ⁻¹)	k_{cat}/K_M (M ⁻¹ ·s ⁻¹)	
Close	-	-	-	-	-358.0	-372.2	14.2	0.082 ± 0.006	6.9 ± 0.1	85000 ± 6000	45 ± 5	33 ± 3	700 ± 100	0.008
	-	Y	I	L	-350.7	-295.4	-55.3	0.019 ± 0.002	0.304 ± 0.006	16000 ± 2000	0.71 ± 0.02	12.06 ± 0.02	16900 ± 500	1.056
	I	Y	I	T	-306.4	-294.9	-11.5	ND	< 0.1	ND	0.35 ± 0.01	3.46 ± 0.01	10000 ± 400	ND
	-	F	I	T	-366.1	-320.6	-45.5	0.12 ± 0.02	0.194 ± 0.007	1600 ± 200	1.03 ± 0.04	7.63 ± 0.04	7400 ± 300	4.625
	-	F	I	Y	-357.8	-300.0	-57.8	0.06 ± 0.003	0.142 ± 0.002	2300 ± 100	1 ± 0.05	5.87 ± 0.05	5900 ± 300	2.565
Both	-	Y	I	T	-355.8	-315.2	-40.6	ND	< 0.1	ND	0.5 ± 0.01	3.86 ± 0.01	7700 ± 200	ND
	-	Y	I	Y	-349.5	-294.6	-54.9	0.016 ± 0.002	0.11 ± 0.002	6800 ± 700	0.91 ± 0.06	5.17 ± 0.06	5700 ± 300	0.838
	I	F	C	A	-381.3	-387.8	6.5	0.033 ± 0.003	0.82 ± 0.02	25000 ± 3000	3.5 ± 0.2	21.5 ± 0.2	6200 ± 400	0.248
	M	F	C	A	-375.0	-390.8	15.7	0.04 ± 0.003	0.306 ± 0.006	7600 ± 600	1.9 ± 0.1	17.7 ± 0.1	9500 ± 500	1.250
	-	F	C	A	-381.7	-386.3	4.6	0.059 ± 0.005	1.33 ± 0.03	23000 ± 2000	6.1 ± 0.4	26 ± 0.4	4300 ± 300	0.187
Open	-	F	C	S	-385.6	-389.7	4.2	0.043 ± 0.004	2.22 ± 0.06	52000 ± 5000	4.3 ± 0.2	26.5 ± 0.2	6200 ± 300	0.119
	-	Y	N	S	-394.7	-390.4	-4.3	0.045 ± 0.003	0.361 ± 0.007	8100 ± 600	1.6 ± 0.1	18.1 ± 0.1	11400 ± 800	1.407
	A	I	F	S	-237.2	-364.4	127.2	0.078 ± 0.007	0.76 ± 0.02	10000 ± 1000	2.3 ± 0.2	0.6 ± 0.2	260 ± 70	0.026
	C	I	F	C	-240.0	-364.4	124.3	0.052 ± 0.005	1.27 ± 0.03	25000 ± 2000	3.8 ± 0.2	3.3 ± 0.2	860 ± 60	0.034
	C	I	F	S	-241.7	-368.5	126.9	0.077 ± 0.008	0.2 ± 0.006	2600 ± 300	2.5 ± 0.2	0.2 ± 0.2	100 ± 60	0.038
	S	I	F	H	-240.3	-364.1	123.9	0.052 ± 0.006	0.61 ± 0.01	12000 ± 1000	2.5 ± 0.1	0.5 ± 0.1	190 ± 40	0.016
	S	I	F	S	-238.1	-363.8	125.7	0.055 ± 0.006	0.77 ± 0.02	14000 ± 2000	2.7 ± 0.2	0.4 ± 0.2	140 ± 60	0.010

ΔE_{C-O} , the energy difference between the Close and Open score

ND, could not be determined

Relative specificity, the $k_{cat}/K_M^{Phe}/k_{cat}/K_M^{Asp}$ of the AAT variant

All reactions were performed in triplicate (n = 3, ± s.d.)

Mutants from the Closed library that were characterized using steady-state kinetics displayed an apparent K_M^{Phe} lower than both WT and mutants from the other libraries. Steady state kinetics revealed the mutants from the Closed library had on average a 59-fold lower apparent K_M^{Phe} than WT. In addition to being a positive control for the 2-substrate screening assay, VYIL is also a member of the Closed library and displays the characteristic relatively low K_M^{Phe} of this library. Alongside the relatively low K_M^{Phe} of the Closed library mutants, the k_{cat} for L-phenylalanine, k_{cat}^{Phe} , was on average 6.3-fold lower than WT. Despite this drop in k_{cat}^{Phe} , the catalytic efficiency of the mutants from the Closed library for phenylalanine was on average 10.5-fold higher than WT. The two mutants from the Closed library with the lowest K_M^{Phe} (IYIT and VYIT) displayed low aspartate transamination activity and Michaelis parameters could not be measured. The aspartate transamination activity was also reduced for the remaining 3 mutants from the Closed library relative to WT, and displayed an average k_{cat}^{Asp} for L-aspartic acid, k_{cat}^{Asp} , and k_{cat}/K_M for L-aspartic acid, k_{cat}/K_M^{Asp} , 46- and 24-fold lower than WT, respectively. Interestingly,

these 3 mutants did not display a trend for the K_M for L-aspartic acid, K_M^{Asp} , and ranged from being twice as high as WT ($120 \pm 20 \mu\text{M}$) to being 4-fold lower than WT ($16 \pm 2 \mu\text{M}$).

Mutants from the Both library displayed a K_M^{Phe} that was on average 4.6-fold higher than mutants from the Closed library, which suggests the conformational equilibrium could be shifted towards the open state relative to the mutants from the Closed library, as designed. The average k_{cat}^{Phe} displayed by the Both mutants was 4.2-fold higher than the Closed library, so the Both library mutants displayed catalytic efficiencies approximately as high as mutants from the Closed library. As with the apparent kinetic parameters for phenylalanine, the k_{cat}^{Asp} and k_{cat}/K_M^{Asp} displayed by the Both mutants were highest among the designed mutants across all libraries despite, on average, a 6.8- and 3.7-fold decrease relative to WT, respectively. Additionally, the K_M^{Asp} of the Both mutants are lower than WT.

Mutants from the Open library were designed to prefer the open conformation more than WT, which could manifest in a poorer affinity for phenylalanine relative to WT. However, the average K_M^{Phe} of the mutants from the Open library was 16-fold lower than WT and similar to the mutants of the Both library. Additionally, the k_{cat}^{Phe} and k_{cat}/K_M^{Phe} of the Open library mutants were the lowest amongst the 3 libraries. For aspartate, the K_M^{Asp} of the mutants from the Open library ranges between 52 and 78 μM , which is similar to the K_M^{Asp} of WT. The mutants from the Open library displayed k_{cat}^{Asp} and k_{cat}/K_M^{Asp} that were on average reduced 9.6- and 6.7-fold relative to WT, respectively.

4.4 Discussion

4.4.1 Library characteristics

Different libraries were designed to predict mutations that will cause mutants from the Closed library to prefer closed conformation, for mutants from the Open library to prefer the open conformation, and for mutants from the Both library to have equally populated open and closed conformations. An apparent K_M^{Phe} relative to WT was used as a proxy for the conformational equilibrium, where a K_M^{Phe} lower than 45 mM would indicate a higher preference for the closed conformation relative to WT.

AAT variants in the Closed library were designed to favour the closed conformation relative to WT, and the 2-concentration screening protocol showed that active mutants from the Closed library have a lower K_M^{Phe} than WT, which was confirmed via steady state kinetic characterization of 5 mutants selected from the Closed library. If the K_M^{Phe} is an accurate proxy, the Closed design successfully predicted mutations that favour the closed state.

During screening, the mutants from the Both library displayed the highest transamination activity at both phenylalanine concentrations but displayed a higher K_M^{Phe} than the mutants from the Closed library. Indeed, steady state kinetic characterization of 5 mutants showed an average K_M^{Phe} lower than WT but still 4.6-fold higher than the mutants from the Closed library. Consistent with the high activity measured during screening, the Both libraries also displayed a high k_{cat}^{Phe} relative to the other libraries. As the energy difference between the Close and Open score of WT is similar to the Both mutants, the Both mutants were expected to show a K_M^{Phe} similar to WT. Therefore, the equilibrium constant may have been shifted towards the closed state more than designed. However, the K_M^{Phe} of the Both mutants suggest the conformational equilibrium was not shifted as far as mutants from the Closed library to prefer the closed state.

The Open library mutants were designed to favour the open conformation more than WT, which we predicted would display a K_M^{Phe} greater than 45 mM. However, the steady-state kinetic characterization of the 5 mutants from the Open library suggest that the conformational equilibrium was shifted towards the closed state to the same extent as the Both mutants, which would suggest the Open library design was unsuccessful.

On the other hand, if all steady-state kinetic parameters of the mutants are considered, a trend can be ascertained that may indicate all three library designs were indeed successful. The mutants from the Open library displayed the lowest k_{cat}^{Phe} and k_{cat}/K_M^{Phe} amongst the libraries, while the mutants from the Closed library displayed the lowest k_{cat}^{Asp} and k_{cat}/K_M^{Asp} . The mutants from the Both library displayed a k_{cat}^{Asp} and a k_{cat}/K_M^{Asp} at least as high as the mutants from the Open library, and a k_{cat}^{Phe} and a k_{cat}/K_M^{Phe} at least as high as the mutants from the Closed library. Additionally, although the mutants from the Open library are less active than WT, they display a relative substrate specificity ($k_{cat}/K_M^{Phe} / k_{cat}/K_M^{Asp}$) between phenylalanine and aspartate (0.010 to 0.038) most similar to WT (0.008), while the mutants from the Closed library generally displayed the highest (0.838 to 4.625) relative substrate specificity towards phenylalanine (Table 4.9). Again, the mutants from the Both library displayed an intermediate relative substrate specificity (0.119 to 1.407) between the relative substrate specificities displayed by the mutants from the Closed library and the mutants from the Open library. This transition in substrate specificity may be a result of an altered conformational equilibrium from mutants in the Closed library that prefer the closed state to mutants in the Open library that prefer the open state, while mutants from the Both library have a relatively balanced conformational equilibrium between the two conformational states.

4.4.2 Future experiments to validate results.

To test if this switch in specificity is because of changes in conformational equilibria between the libraries or another factor, we selected mutants from the Closed and Open libraries for structural study. Among the mutants selected for kinetic characterization, IYIT from the Closed library had the lowest K_M^{Phe} of 0.35 mM, which was 129-fold lower than WT. This mutant was selected for crystallization in the absence of substrate. If IYIT crystallizes in the closed structure without a bound substrate like the HEX mutant, it would add evidence that the conformational equilibrium has been shifted to favour the closed state. To determine the success of the Open library, we selected the SIFS mutant for crystallization in the presence and absence of substrate. SIFS was selected because it had one of the highest K_M^{Phe} of the Open library, it displayed one of the lowest k_{cat}^{Phe} and k_{cat}/K_M^{Phe} , and its ratio of k_{cat}^{Asp} to k_{cat}^{Phe} was highest. If SIFS crystallizes in the open state when the substrate is absent like WT, it would not prove the conformational equilibrium has shifted towards the open state relative to WT, but it would at least indicate the equilibrium has not been unintentionally shifted towards the closed state. If SIFS also crystallizes in the open state when a substrate or inhibitor is bound, it could suggest that the equilibrium has been shifted towards the open conformation.

Although the AAT mutants may crystallize in their target conformations, this does not mean the target conformation is the most populated conformational state in solution. To detect the most populated conformation that the AAT variants occupy in solution, we could use NMR or fluorescent probes. If the WT AAT structures resolves well in an HSQC experiment, 3D triple resonance experiments including HNCO, HNCACB and CBCA(CO)NH spectra can be used to assign peaks. Additionally, we may even be able to measure conformational equilibrium constants with methods such as ZZ-exchange. If the AAT structures do not resolve well in multi-dimensional NMR experiments, we can also consider 1H resonance experiments. The residue W140 is in the

fixed domain but lies in the AAT active site and interacts with the PLP. Previously, Metzler et al.¹⁶⁴ have recorded the ¹H NMR spectra of WT *E. coli* AAT and single point mutants in the 10-18 ppm range and assigned the W140 indole NH proton peak. If the W140 indole NH proton peak is sensitive to the different AAT conformations, this can be used to probe the populated conformations. Alternatively, Picot et al.¹⁶⁵ probed the conformations of mitochondrial chicken AAT by attaching the fluorescent label monobromo(trimethylammonio)bimane to C166, which is H166 in *E. coli* AAT. The fluorescence of the bimane-labelled PLP-bound enzyme was reduced by 40-70 % upon binding ligands, which induced the conformational switch from the open to closed state. If we mutate H166 to a cysteine, we could also probe the open and closed conformations of the AAT mutants, although replacing a histidine residue with a bimane-labelled cysteine may affect the conformational equilibrium, dynamics and even the overall structure.

4.4.3 Importance of including multiple conformational states.

In cases where an enzyme undergoes conformational changes during a catalytic cycle, modulation of the relative stability between the conformational states can be an effective strategy to help fine-tune the binding affinity and catalytic turnover of the enzyme. To accomplish this, multiple conformational states should be simultaneously considered in a MSA scheme.

The design of AAT variants to display a low K_M^{Phe} by stabilizing the closed state without consideration of the open state would not have been successful. Had we only designed a library of 36 members based on the Closed ensemble scores alone (Closed_{only}), the library would have contained 0 mutants from the Closed library and 12 mutants from the Both library, which includes one of the mutants selected for kinetic characterization, VYNS. Although VYNS displayed a K_M^{Phe} 28-fold lower than WT, it was still higher than the K_M^{Phe} displayed by all 5 of the mutants from the Closed library that were kinetically characterized. Additionally, 4 of the 5 mutants from the

Closed library have higher Closed ensemble scores in the closed state than WT (ranked 13230th in the closed state), which means these mutants would have never been included in the Closed_{only} library. In fact, the enzyme with the best K_M^{Phe} , IYIT, had the highest Closed ensemble score (i.e., highest energy) amongst the other mutants from the Closed library. Only by consideration of the open and closed states in multistate analysis could a Closed library that contains mutants which display such a low K_M^{Phe} be designed.

Likewise, if a library was built that only considered the Open ensemble scores (Open_{only}), the library would contain 24 of the 36 mutants from the Both library but no mutants from the Open library. Therefore, if the Open library design is confirmed to be successful, it was only possible due to the inclusion of multiple conformational states.

It is less clear that both conformations were required to design AAT mutants with a small ensemble score difference between the two conformational states. Of the 36 mutants in the Both library, 12 variants would have been included in the Closed_{only} library and 24 variants would have been included in the Open_{only} library. Additionally, 8 mutants from the Both library are common to all three libraries (Both, Closed_{only}, Open_{only}). However, only 3 (VYNS, VFCS, VYCS) of the 24 mutants from the Open_{only} library and 2 (VYNS, VFNA) of the 12 mutants from the Closed_{only} library that were also in the Both library were identified during the 2-substrate screening. Alternatively, 9-10 of the 12 mutants from the Both library that were most active with phenylalanine and identified during screening were unique to the Both library and would not have been included in either single-state libraries, which suggests that inclusion of both conformational states during design may have been beneficial.

4.4.4 Comparison to the literature

Previously, CPD has been used to identify mutations that alter the conformational equilibrium between the ground and excited state of a T4 lysozyme mutant to invert the relative population of the two species.⁹⁸ Also, Michielssens et al.¹⁶⁶ identified single-point mutations that modified conformational equilibria of ubiquitin to control protein-protein binding rather than optimizing its protein-protein binding interface. The work presented here is different in that AAT must undergo conformational changes during its catalytic mechanism, which necessitates that both the open and closed states remain energetically accessible for the AAT variants to maintain catalytic activity. Additionally, we introduced the use of domain movement analysis to identify positions to consider for mutation.

4.5 Conclusions

The MSA procedure presented here predicts mutations at hinge regions of domain movement that could affect the conformational equilibrium of enzymes. Here, we used the procedure to design libraries of AAT mutants that exhibit a preference for either the open conformation, the closed conformation, or both. The 2-concentration screening protocol we developed here and steady state kinetics indicated the designs were successful and further experiments to validate the success of the designs are underway. The method developed here opens the door for the design of any enzyme that undergoes conformational changes.

4.6 Methods

4.6.1 Computational Protein Design details

Computational protein design (CPD) calculations were performed with PHOENIX,^{23, 35, 37} using Fast and Accurate Side-Chain Topology and Energy Refinement (FASTER)²³ to optimize each sequence for every backbone of each conformational state (130,321 sequences $\times$ 50 backbones $\times$ 2 conformational state = 13,032,100 individual scores). The backbone dependent Dunbrack rotamer libraries³⁰ with expansions of ± 1 standard deviation around χ_1 and χ_2 were used to model side-chain conformations. Rotamer combinations were evaluated using a four-term potential energy function consisting of a van der Waals term from the Dreiding II force field¹⁶⁷ with atomic radii scaled by 0.9, a direction sensitive hydrogen-bond term with well depth at 8.0 kcal mol⁻¹,⁴⁰ an electrostatic energy term modeled using Coulomb's law with a distance dependent dielectric of 10, and an occlusion-based solvation penalty term.³⁷ The crystallographic conformers were not allowed to avoid crystallographic bias. In addition, the explicit energies of all rotamers within 5 Å of the optimized residues were also included in pairwise energy interaction calculations. The backbone template energy was not included in the scoring function.

4.6.2 Mutagenesis

The *E. coli* AAT gene cloned into plasmid pET-45b (Novagen) was a generous gift from Michael D. Toney (University of California, Davis). Overlap extension mutagenesis was performed as described by Vallejo et al.¹⁵² using VentR DNA polymerase. Inspired by Tang et al.,¹⁵¹ multiple degenerate codons were used in order to eliminate redundant codons and minimize uneven distributions of amino acids in the library. Following gel purification (Omega Biotek), the amplicons were recombined by overlap extension PCR, resulting in the final gene. The gene was gel purified and digested with the restriction enzymes *NdeI/BamHI*. Following another gel purification, the digested fragment was ligated into the expression vector pET11-a with T4 DNA

ligase. The resulting expression vectors containing the aminotransferase genes was then transformed into *E. coli* BL21-Gold(DE3) gold cells. Colonies were picked into individual wells of V96 MicroWell polypropylene plates (Nunc) containing 300 μ L of medium (LB with 100 μ g mL⁻¹ ampicillin). Wildtype AAT in pet45b vector, a triple mutant AAT (K41Y/T47I/N69L) in pet11-a vector, and pet11-a vector containing no AAT were picked into two wells each as two positive controls and one negative control, respectively.

4.6.3 Screening

The mother plates containing the AAT mutant libraries were used to inoculate sterile Nunc V96 MicroWell plates containing 300 μ L of Overnight Express Instant TB media (Novagen) supplemented with ampicillin per well, which were incubated overnight at 37 °C and then washed and lysed to prepare the clarified lysates as described by Walton et al.¹²¹ except 5 U mL⁻¹ of benzonase was used instead of 25 U mL⁻¹ in the lysis buffer. Screening assays were performed in 200 μ L reactions containing final concentrations of either 3 or 40 mM L-phenylalanine, 0.2 mM α -ketoglutarate, 5 mM NAD⁺, 16 μ M PLP, and 1 U of GDH in 100 mM potassium phosphate buffer (pH 8.0, 37 °C). 96-well plates (Greiner Bio-One) containing the standard reaction mixture were preincubated at 37 °C for 5 min prior to initiation of the reaction by the addition of 10 μ L of clarified cell lysates. Enzyme reactions were monitored with a SpectraMax 384 Plus plate reader (Molecular Devices) by measuring absorbance of NADH at 340 nm ($\epsilon_{340} = 6220 \text{ M}^{-1} \text{ cm}^{-1}$) every 12 seconds for 30 minutes. Pathlengths for each well were calculated ratiometrically using the difference in absorbance of potassium phosphate buffer at 900 and 998 nm. The linear phases of the kinetic traces were used to measure the initial reaction rates. The average background reaction rates plus 10% of the negative controls were subtracted from each well before computing the reaction rates.

4.6.4 Steady-state kinetics of AAT mutants

The mutants were expressed and purified as described by Mironov et al.¹¹⁵ All kinetic assays were performed in triplicate. To measure the steady-state kinetics for L-aspartate and L-phenylalanine with the AAT mutants, GDH assays were performed by varying the amino acid concentration (up to 20 mM for aspartate and 40 mM for phenylalanine) in the presence of 200 μ M α -ketoglutarate, 5 mM NAD^+ , 16 μ M PLP, 1 U GDH and approximately 10 mU of aminotransferase in 100 mM potassium phosphate buffer (pH 8, 37 °C). The linear phases of the kinetic traces were used to measure the initial reaction rates. The initial reaction rates at different amino acid concentrations were fit to the Michaelis Menten equation (Equation 4.1) using the GraphPad Prism software to compute apparent steady state kinetic parameters, k_{cat} and K_M (Table 4.9).

Chapter 5: Summary and perspectives

Some of the language and figures in this chapter were adapted from a book chapter coauthored with Olivier Gagnon and Dr. Roberto A. Chica:

St-Jacques AD, Gagnon O & Chica RA (2018) Computational Enzyme Design: Successes, Challenges and Future Directions. In: Williams G & Hall M (Ed.) Modern Biocatalysis: Advances Towards Synthetic Biological Systems (pp. 88-116). Cambridge, UK. Royal Society of Chemistry.

5.1 Summary

The development of a broadly-applicable computational method for the design of enzymes that can catalyze any chemical reaction with high efficiency and selectivity under defined conditions is a major objective of protein engineers. Standard computational protein design (CPD) protocols find optimal sequences in the context of a single state, for example an enzyme structure with a single substrate bound at its active site. However, many enzymes require multiple substrates and many enzymes undergo conformational changes throughout their catalytic cycle. In this thesis I presented the development and implementation of methodologies (1) to generate rotamer libraries for any small molecule, (2) to redesign the substrate specificity of multi-substrate enzymes and (3) to modulate the conformational equilibrium of an enzyme.

Chapter 2. We determined the Dunbrack 2002 backbone independent rotamer library^{14, 30} (bbind) contained rotamers that could replicate aminotransferase active site residue rotamer configurations better than the rotamers included in the Dunbrack 2002 backbone dependent rotamer library. However, the bbind library was not always sufficient to recapitulate the input amino acid substrate rotamers of the PLP-bound substrates. Therefore, we built our own substrate rotamer library to successfully model the input substrate rotamers, which was used in the following

chapter to model substrates in the BCAT active site. The method developed in this chapter can be used to generate a rotamer library for any small molecule.

Chapter 3. The multi-chemical state CPD framework presented here incorporates multiple chemical states into the redesign of enzymes that require multiple substrates to undergo a complete reaction mechanism. Here, we developed a multi-chemical state CPD framework that incorporates multiple chemical states to redesign the BCAT active site to enhance its ability to catalyze the transamination of L-histidine using α -ketoglutarate as an amino acceptor up to 50-fold higher than WT BCAT.

Chapter 4. We developed a multistate analysis (MSA) framework to identify and design residues important to domain movement in order to modulate the conformational equilibrium of an enzyme and thereby alter its substrate specificity. Domain movement analysis identified residues distal to the active site that act as a hinge to the opening and closing of domains in AAT. These hinge-bending residues were designed using MSA to alter the conformational equilibrium of AAT to predict combinatorial AAT mutant libraries that prefer the open state, the closed state, or to exhibit no preference. Apparent steady-state kinetic parameters of AAT variants from each library suggest the designs were successful and further experiments to validate these designs are underway.

The multistate methods described in this thesis open the door for future CPD projects to redesign enzymes for any application and can also be applied to *de novo* enzyme design.

5.2 Perspectives

The methods presented in this thesis will contribute more tools in the CPD toolbox for enzyme (re)design. However, the success rate of the computational procedures remains low as

many of the failed designs simply do not fold,¹⁶⁸⁻¹⁶⁹ while others are folded but cannot bind the substrate.³⁵ Failed designs can be attributed to several approximations made in the computational methodologies to ensure tractability. For example, the protein backbone is held fixed during design, and translational, rotational, and conformational sampling of the substrate or transition state and amino-acid side chains is discretized, preventing the computational procedure from achieving the sub-ångstrom level of precision that is required for highly efficient catalysis.¹⁷⁰ Additionally, potential energy functions used in computational enzyme design contain inaccuracies in their treatment of non-bonded interactions.^{167, 171-173} Particularly problematic is the accurate description of long-range electrostatics and atom polarization,¹⁷⁴ which are crucial to enzymatic catalysis.¹⁷⁵ These inaccuracies have thus far prevented kinetic parameters of enzymes (*i.e.*, k_{cat} and K_{M}) predicted by CPD to be strongly correlated with experimental values.¹⁷⁶

Even when computational enzyme design yields active biocatalysts, catalytic efficiencies ($k_{\text{cat}}/K_{\text{M}}$) obtained are orders of magnitude lower than those of natural enzymes, especially with respect to *de novo* enzymes.^{1, 177} These modest catalytic efficiencies are mostly due to k_{cat} values that are 1–6 orders of magnitude lower than those of natural enzymes,^{1, 72} and not to Michaelis constants (K_{M}), which fall within the range of those found in natural biocatalysts.¹⁷⁷ This result demonstrates the ability of computational enzyme design to create active sites capable of productively binding a substrate but that this does not ensure highly efficient catalysis. Indeed, enzymatic reaction rates are determined by the rate-limiting step of the complete catalytic process, which may not necessarily be the chemical state that is modeled during design (*i.e.*, the substrate-enzyme complex, or the transition state and its theoretical optimal geometry for stabilization by enzyme functional groups). For example, it has been shown that for many enzymes, product release⁷³ or conformational exchange between various sub-states during the catalytic cycle⁷⁴ are

rate-limiting. It is thus possible that such processes slow down the rate of reactions catalyzed by (re)designed enzymes.

In order to improve the success rate of computational enzyme (re)design, continued development of CPD force fields and sampling methodologies will be vital. Recently developed electrostatic potential energy terms¹⁷⁸ as well as polarizable force fields¹⁷⁹⁻¹⁸⁰ could be incorporated into future computational enzyme design methodologies to better model electron distribution during the catalytic step. Use of flexible backbone design algorithms,^{60-61, 181-182} as well as continuous rotamer^{59, 183} and backbone¹⁸⁴ flexibility may lead to improved sampling of the protein scaffold during design. As well, multistate methods with large structural ensembles^{36, 63} may increase the success rate of computational enzyme design by helping to address the fixed backbone approximation.⁵⁶ In addition to improvements to the computational methodology, it will be crucial for researchers to broaden their selection of residue positions during design. Typically, designed residue positions in the protein are restricted to those in the active site. Yet, distant mutations have been shown to have a large impact on the catalytic properties of many enzymes,¹⁸⁵⁻¹⁸⁶ making their prediction by computational enzyme design essential if we are to create highly active enzymes. However, it remains very difficult to know *a priori* which positions to consider. In Chapter 4, we showed that domain movement analysis can be used to identify residue positions distal to the active site that act as hinges for the modulation of conformational exchange. Other potential computational approaches to identify designable distal sites include a self-consistent mean field entropy-based method,¹⁸⁷ which identifies residue positions in the protein that may be tolerant to mutation, and the Monte Carlo side chain entropy method,¹⁸⁸ which could be used to calculate side-chain fluctuations and their coupling across residues in order to identify distal sites that can affect catalysis.

Ultimately, the (re)design of enzymes with catalytic efficiencies matching those of natural enzymes will likely require a holistic view of enzymatic catalysis where every possible state that the enzyme adopts as it undergoes its catalytic cycle (Figure 5.1) is included in the calculation. This could be achieved by the implementation of multistate approaches^{68, 137, 189-190} in computational enzyme (re)design. In a multistate computational enzyme (re)design process, a minimum of four states should be considered for the creation of biocatalysts catalyzing simple, single-step mechanisms: the free enzyme, the Michaelis complex, the enzyme-bound transition state structure, and the product-bound state. Consideration of these states can ensure that mutations designed to create the active site and stabilize the transition state do not inadvertently result in over-stabilization of the enzyme-product intermediate, leading to decreased rate due to a slow product release step, or over-packing of the transition state, which may reduce catalytic efficiency by creating a high-energy barrier preventing substrate entry and product exit into and from the active site (Figure 5.2). Thus, an ideal multistate computational enzyme (re)design calculation should aim to (i) facilitate substrate entry and binding in the active site, (ii) destabilize the substrate ground state by active site preorganization, (iii) stabilize the transition state, and (iv) destabilize the enzyme-product complex to facilitate product release (Figure 5.3). Achieving this delicate balance between the relative energies of these states could theoretically lead to more catalytically efficient enzymes in the future. The multistate approach developed in Chapter 3 will be essential for the inclusion of all relevant states for enzyme design. Additionally, the rotamer library generation method we developed in Chapter 2 enables more complete descriptions of ligand flexibility when modeling ligand-enzyme complexes.

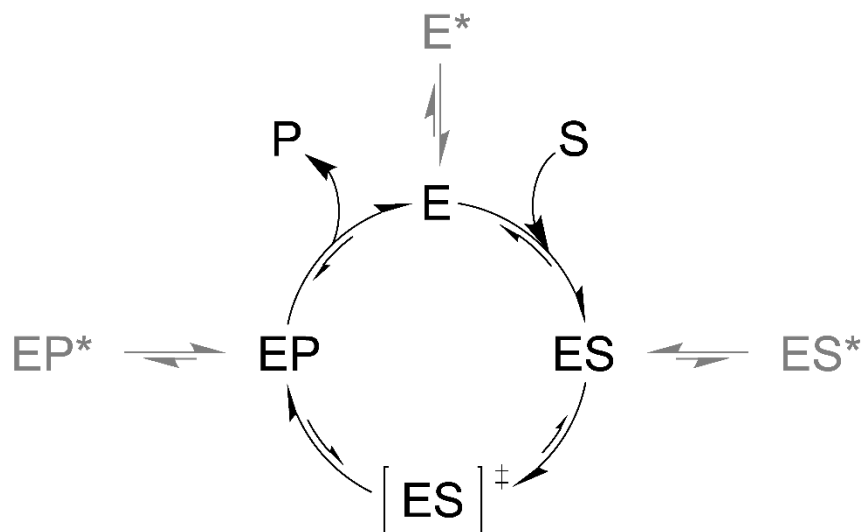


Figure 5.1. States adopted by an enzyme as it undergoes its catalytic cycle. As an enzyme (E) undergoes its catalytic cycle, it must bind the substrate (S) to form the Michaelis complex (ES), leading to catalysis by transition state stabilization ($[ES]^{\ddagger}$). Following the reaction, the enzyme-product complex is formed (EP), and product (P) is released to regenerate the free enzyme. Potential competing states in grey include unproductive conformations of the enzyme (E^*) and unproductive binding poses of the substrate or product to the enzyme (ES^* and EP^* , respectively).

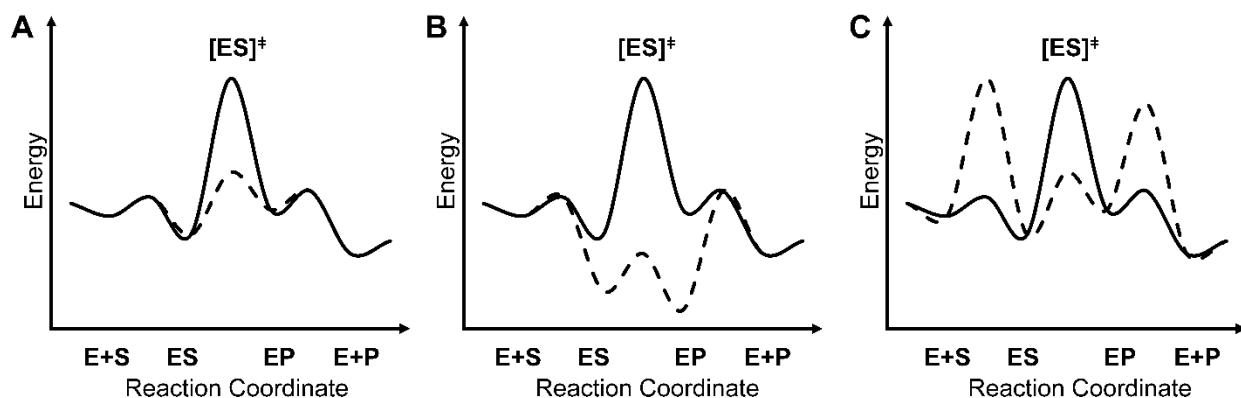


Figure 5.2. Hypothetical consequences of transition state stabilization by computational enzyme design. In all cases, reaction coordinates before and after transition state stabilization are shown as full or dashed lines, respectively. (A) Transition state stabilization by computational enzyme design aims to lower the energy barrier of the rate-limiting step and thus increase the rate of the reaction. However, transition state stabilization may also result in the unintended over-stabilization of the enzyme-substrate and enzyme-product complexes (B), which would result in a rate-limiting product-releasing step that would decrease catalytic efficiency. Alternatively, the computational procedure may result in transition state over-packing (C), which could create a high barrier for substrate entry in the active site or product release and thereby decrease catalytic efficiency.

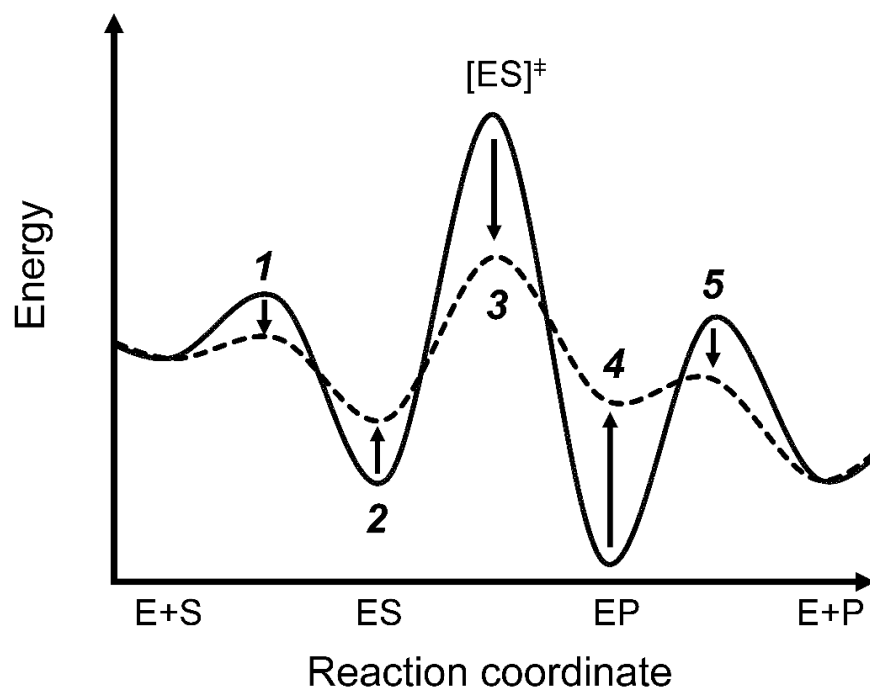


Figure 5.3. Multistate computational enzyme design. The multistate computational design procedure should aim to (1) facilitate substrate entry and binding in the active site, (2) destabilize the substrate ground state by active site preorganization, (3) stabilize the transition state, and (4) destabilize the enzyme-product complex to (5) facilitate product release. Full and dashed lines show theoretical reaction coordinates of low and high efficiency enzymes, respectively.

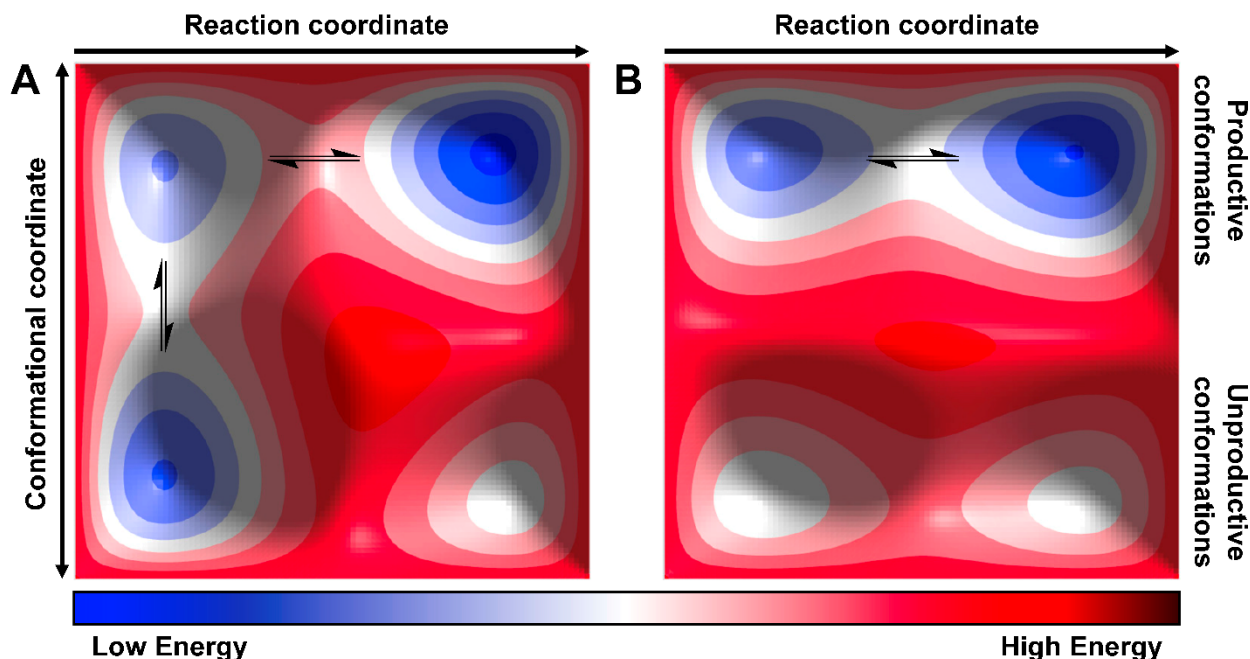


Figure 5.4. Potential energy surfaces of hypothetical enzymes. An enzyme that exchanges between productive and unproductive conformations (A) is expected to be less catalytically efficient than an enzyme that cannot exchange to catalytically unproductive conformations (B).

In addition to the four chemical states described above, alternate conformations of the unbound enzyme need to also be considered, as enzymes are intrinsically dynamic molecules that can sample numerous conformations.^{74, 191} The ability of enzymes to exchange between multiple conformations is vital to catalysis,⁹² as different conformational sub-states are needed to bind the substrate(s), stabilize the transition state, and release the product(s). However, dynamics can also be detrimental to catalysis, as motions that cause sampling of unproductive conformations can limit the rate of enzyme turnover.¹⁹²⁻¹⁹³ For example, an enzyme that exchanges between catalytically productive and unproductive conformations would be less active than an enzyme where unproductive conformations are inaccessible (Figure 5.4). Thus, the multistate computational enzyme design procedure should aim to eliminate sequences predicted to stabilize unproductive conformations, which can be done by negative multistate methods.⁶³ However, the structural features of the conformational sub-states that the enzyme can adopt throughout the catalytic cycle are difficult to define. Fortunately, recent developments in room temperature x-ray crystallography have revealed hidden alternative protein conformations that may be functionally relevant to catalysis, and these can be modeled as discrete structures^{191, 194-195} to be used in computational enzyme design. Alternatively, models of such structures could be obtained using long timescale MD simulations,¹⁹² which are starting to approach the millisecond timescale¹⁹⁶ of protein motions pertinent to catalysis.¹⁹⁷ In cases where multiple biophysically relevant sub-states of an enzyme are available, the framework we developed in Chapter 4 can be used to control the conformational exchange between these sub-states.

Continued development of computational enzyme (re)design methodologies will be essential to realize the on-demand design of enzymes that can catalyze any desired chemical reaction with high efficiency and selectivity under defined conditions. We expect that fulfilment

of this long-term objective will require the design of complete enzymatic catalytic cycles for reactions of interest. This objective may be tackled following the development and validation of more multistate approaches to computational enzyme (re)design, which could open the door to the next phase of biocatalyst design.

References

1. Wolfenden, R.; Snider, M. J., The depth of chemical time and the power of enzymes as catalysts. *Acc Chem Res* **2001**, *34* (12), 938-45.
2. Edwards, D. R.; Lohman, D. C.; Wolfenden, R., Catalytic proficiency: the extreme case of S-O cleaving sulfatases. *J Am Chem Soc* **2012**, *134* (1), 525-31.
3. Gupta, R. D.; Goldsmith, M.; Ashani, Y.; Simo, Y.; Mullokandov, G.; Bar, H.; Ben-David, M.; Leader, H.; Margalit, R.; Silman, I.; Sussman, J. L.; Tawfik, D. S., Directed evolution of hydrolases for prevention of G-type nerve agent intoxication. *Nat Chem Biol* **2011**, *7* (2), 120-5.
4. Bornscheuer, U. T.; Pohl, M., Improved biocatalysts by directed evolution and rational protein design. *Curr Opin Chem Biol* **2001**, *5* (2), 137-43.
5. Onuffer, J. J.; Kirsch, J. F., Redesign of the substrate specificity of Escherichia coli aspartate aminotransferase to that of Escherichia coli tyrosine aminotransferase by homology modeling and site-directed mutagenesis. *Protein Sci* **1995**, *4* (9), 1750-7.
6. Chen, K.; Arnold, F. H., Enzyme Engineering for Nonaqueous Solvents: Random Mutagenesis to Enhance Activity of Subtilisin E in Polar Organic Media. *Bio/Technology* **1991**, *9*, 1073.
7. Arnold, F. H., Design by Directed Evolution. *Accounts Chem. Res.* **1998**, *31* (3), 125-131.
8. Penning, T. M.; Jez, J. M., Enzyme redesign. *Chem Rev* **2001**, *101* (10), 3027-46.
9. Dahiyat, B. I.; Mayo, S. L., Protein design automation. *Protein Sci* **1996**, *5* (5), 895-903.
10. Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D., Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* **1997**, *268* (1), 209-25.
11. Malakauskas, S. M.; Mayo, S. L., Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Biol* **1998**, *5* (6), 470-5.
12. Altman, M. D.; Nalivaika, E. A.; Prabu-Jeyabalan, M.; Schiffer, C. A.; Tidor, B., Computational design and experimental study of tighter binding peptides to an inactivated mutant of HIV-1 protease. *Proteins* **2008**, *70* (3), 678-94.
13. Ashworth, J.; Havranek, J. J.; Duarte, C. M.; Sussman, D.; Monnat, R. J., Jr.; Stoddard, B. L.; Baker, D., Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **2006**, *441* (7093), 656-9.
14. Dunbrack, R. L., Jr., Rotamer libraries in the 21st century. *Curr Opin Struct Biol* **2002**, *12* (4), 431-40.
15. Street, A. G.; Mayo, S. L., Computational protein design. *Structure* **1999**, *7* (5), R105-9.
16. Boas, F. E.; Harbury, P. B., Design of protein-ligand binding based on the molecular-mechanics energy model. *J Mol Biol* **2008**, *380* (2), 415-24.
17. Schmidt Am Busch, M.; Lopes, A.; Mignon, D.; Simonson, T., Computational protein design: software implementation, parameter optimization, and performance of a simple model. *J Comput Chem* **2008**, *29* (7), 1092-102.
18. Rohl, C. A.; Baker, D., De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J Am Chem Soc* **2002**, *124* (11), 2723-9.
19. Desmet, J.; De Maeyer, M.; Hazes, B.; Lasters, I., The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **1992**, *356* (6369), 539-42.
20. Koehl, P.; Delarue, M., Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol* **1994**, *239* (2), 249-75.
21. Koehl, P.; Delarue, M., A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nat Struct Biol* **1995**, *2* (2), 163-70.
22. Desmet, J.; Spriet, J.; Lasters, I., Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **2002**, *48* (1), 31-43.
23. Allen, B. D.; Mayo, S. L., Dramatic performance enhancements for the FASTER optimization algorithm. *J Comput Chem* **2006**, *27* (10), 1071-5.

24. Kuhlman, B.; Baker, D., Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci U S A* **2000**, *97* (19), 10383-8.
25. Lilien, R. H.; Stevens, B. W.; Anderson, A. C.; Donald, B. R., A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the substrate specificity of the gramicidin synthetase a phenylalanine adenylation enzyme. *J Comput Biol* **2005**, *12* (6), 740-61.
26. Berman, H.; Henrick, K.; Nakamura, H.; Markley, J. L., The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* **2007**, *35* (Database issue), D301-3.
27. Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C., Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol* **1999**, *285* (4), 1735-47.
28. Labute, P., Protonate3D: assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins* **2009**, *75* (1), 187-205.
29. Fiser, A., Comparative Protein Structure Modelling. In *From Protein Structure to Function with Bioinformatics*, J. Rigden, D., Ed. Springer Netherlands: Dordrecht, 2017; pp 91-134.
30. Dunbrack, R. L., Jr.; Cohen, F. E., Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* **1997**, *6* (8), 1661-81.
31. Shapovalov, M. V.; Dunbrack, R. L., Jr., A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **2011**, *19* (6), 844-58.
32. Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C., The penultimate rotamer library. *Proteins* **2000**, *40* (3), 389-408.
33. Dunbrack, R. L., Jr.; Karplus, M., Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* **1993**, *230* (2), 543-74.
34. Bower, M. J.; Cohen, F. E.; Dunbrack, R. L., Jr., Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J Mol Biol* **1997**, *267* (5), 1268-82.
35. 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 U S A* **2012**, *109* (10), 3790-5.
36. 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 U S A* **2010**, *107* (46), 19838-43.
37. Chica, R. A.; Moore, M. M.; Allen, B. D.; Mayo, S. L., Generation of longer emission wavelength red fluorescent proteins using computationally designed libraries. *Proc Natl Acad Sci U S A* **2010**, *107* (47), 20257-62.
38. Mayo, S. L.; Olafson, B. D.; Goddard, W. A., Dreiding - a Generic Force-Field for Molecular Simulations. *J Phys Chem* **1990**, *94* (26), 8897-8909.
39. Sitkoff, D.; Sharp, K. A.; Honig, B., Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *J Phys Chem* **1994**, *98* (7), 1978-1988.
40. Dahiyat, B. I.; Mayo, S. L., Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A* **1997**, *94* (19), 10172-7.
41. Lazaridis, T.; Karplus, M., Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol* **1999**, *288* (3), 477-87.
42. Eisenberg, D.; McLachlan, A. D., Solvation energy in protein folding and binding. *Nature* **1986**, *319* (6050), 199-203.
43. Street, A. G.; Mayo, S. L., Pairwise calculation of protein solvent-accessible surface areas. *Fold Des* **1998**, *3* (4), 253-8.
44. Lazaridis, T.; Karplus, M., Effective energy function for proteins in solution. *Proteins* **1999**, *35* (2), 133-52.

45. Shortle, D., Propensities, probabilities, and the Boltzmann hypothesis. *Protein Sci* **2003**, *12* (6), 1298-302.
46. Voigt, C. A.; Gordon, D. B.; Mayo, S. L., Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J Mol Biol* **2000**, *299* (3), 789-803.
47. Leach, A. R.; Lemon, A. P., Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins* **1998**, *33* (2), 227-39.
48. Traore, S.; Allouche, D.; Andre, I.; de Givry, S.; Katsirelos, G.; Schiex, T.; Barbe, S., A new framework for computational protein design through cost function network optimization. *Bioinformatics* **2013**, *29* (17), 2129-36.
49. Kingsford, C. L.; Chazelle, B.; Singh, M., Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics* **2005**, *21* (7), 1028-36.
50. Traore, S.; Allouche, D.; Andre, I.; Schiex, T.; Barbe, S., Deterministic Search Methods for Computational Protein Design. *Methods Mol Biol* **2017**, *1529*, 107-123.
51. Pierce, N. A.; Winfree, E., Protein Design is NP-hard. *Protein Eng Des Sel* **2002**, *15* (10), 779-782.
52. Jones, D. T., De novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci* **1994**, *3* (4), 567-74.
53. Hayes, R. J.; Bentzien, J.; Ary, M. L.; Hwang, M. Y.; Jacinto, J. M.; Vielmetter, J.; Kundu, A.; Dahiyat, B. I., Combining computational and experimental screening for rapid optimization of protein properties. *Proc Natl Acad Sci U S A* **2002**, *99* (25), 15926-31.
54. Mena, M. A.; Daugherty, P. S., Automated design of degenerate codon libraries. *Protein Eng Des Sel* **2005**, *18* (12), 559-61.
55. Treynor, T. P.; Vizcarra, C. L.; Nedelcu, D.; Mayo, S. L., Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proc Natl Acad Sci U S A* **2007**, *104* (1), 48-53.
56. Davey, J. A.; Chica, R. A., Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Proteins* **2014**, *82* (5), 771-84.
57. Grigoryan, G.; Ochoa, A.; Keating, A. E., Computing van der Waals energies in the context of the rotamer approximation. *Proteins* **2007**, *68* (4), 863-78.
58. Kellogg, E. H.; Leaver-Fay, A.; Baker, D., Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **2011**, *79* (3), 830-8.
59. Gainza, P.; Roberts, K. E.; Donald, B. R., Protein design using continuous rotamers. *PLoS Comput Biol* **2012**, *8* (1), e1002335.
60. Havranek, J. J.; Baker, D., Motif-directed flexible backbone design of functional interactions. *Protein Sci* **2009**, *18* (6), 1293-305.
61. Huang, P. S.; Ban, Y. E.; 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.
62. Smith, C. A.; Kortemme, T., Predicting the tolerated sequences for proteins and protein interfaces using RosettaBackrub flexible backbone design. *PLoS One* **2011**, *6* (7), e20451.
63. Davey, J. A.; Damry, A. M.; Euler, C. K.; Goto, N. K.; Chica, R. A., Prediction of Stable Globular Proteins Using Negative Design with Non-native Backbone Ensembles. *Structure* **2015**, *23* (11), 2011-21.
64. Havranek, J. J.; Harbury, P. B., Automated design of specificity in molecular recognition. *Nat Struct Biol* **2003**, *10* (1), 45-52.
65. Ambroggio, X. I.; Kuhlman, B., Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J Am Chem Soc* **2006**, *128* (4), 1154-61.
66. Frey, K. M.; Georgiev, I.; Donald, B. R.; Anderson, A. C., Predicting resistance mutations using protein design algorithms. *Proc Natl Acad Sci U S A* **2010**, *107* (31), 13707-12.
67. Allen, B. D.; Mayo, S. L., An efficient algorithm for multistate protein design based on FASTER. *J Comput Chem* **2010**, *31* (5), 904-16.

68. Leaver-Fay, A.; Jacak, R.; Stranges, P. B.; Kuhlman, B., A generic program for multistate protein design. *PLoS One* **2011**, *6* (7), e20937.
69. Davey, J. A.; Chica, R. A., Multistate Computational Protein Design with Backbone Ensembles. *Methods Mol Biol* **2017**, *1529*, 161-179.
70. 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 U S A* **2006**, *103* (45), 16710-5.
71. Zanghellini, A.; Jiang, L.; Wollacott, A. M.; Cheng, G.; Meiler, J.; Althoff, E. A.; Rothlisberger, D.; Baker, D., New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci* **2006**, *15* (12), 2785-94.
72. Bar-Even, A.; Noor, E.; Savir, Y.; Liebermeister, W.; Davidi, D.; Tawfik, D. S.; Milo, R., The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **2011**, *50* (21), 4402-10.
73. Cleland, W. W., What limits the rate of an enzyme-catalyzed reaction. *Accounts Chem. Res.* **1975**, *8* (5), 145-151.
74. Kerns, S. J.; Agafonov, R. V.; Cho, Y. J.; Pontiggia, F.; Otten, R.; Pachov, D. V.; Kutter, S.; Phung, L. A.; Murphy, P. N.; Thai, V.; Alber, T.; Hagan, M. F.; Kern, D., The energy landscape of adenylate kinase during catalysis. *Nat Struct Mol Biol* **2015**, *22* (2), 124-31.
75. Khoury, G. A.; Fazelinia, H.; Chin, J. W.; Pantazes, R. J.; Cirino, P. C.; Maranas, C. D., Computational design of *Candida boidinii* xylose reductase for altered cofactor specificity. *Protein Sci* **2009**, *18* (10), 2125-38.
76. Lippow, S. M.; Moon, T. S.; Basu, S.; Yoon, S. H.; Li, X.; Chapman, B. A.; Robison, K.; Lipovsek, D.; Prather, K. L., Engineering enzyme specificity using computational design of a defined-sequence library. *Chem Biol* **2010**, *17* (12), 1306-15.
77. Gordon, S. R.; Stanley, E. J.; Wolf, S.; Toland, A.; Wu, S. J.; Hadidi, D.; Mills, J. H.; Baker, D.; Pultz, I. S.; Siegel, J. B., Computational design of an alpha-gliadin peptidase. *J Am Chem Soc* **2012**, *134* (50), 20513-20.
78. Siegel, J. B.; Smith, A. L.; Poust, S.; Wargacki, A. J.; Bar-Even, A.; Louw, C.; Shen, B. W.; Eiben, C. B.; Tran, H. M.; Noor, E.; Gallaher, J. L.; Bale, J.; Yoshikuni, Y.; Gelb, M. H.; Keasling, J. D.; Stoddard, B. L.; Lidstrom, M. E.; Baker, D., Computational protein design enables a novel one-carbon assimilation pathway. *Proc Natl Acad Sci U S A* **2015**, *112* (12), 3704-9.
79. Saraswathi, G.; Panda, T.; Basak, T., A simplified approach to derive Cleland model for enzymatic reactions. *Biotechnol Lett* **2013**, *35* (5), 785-789.
80. Guo, F.; Berglund, P., Transaminase biocatalysis: optimization and application. *Green Chem* **2017**, *19* (2), 333-360.
81. Eliot, A. C.; Kirsch, J. F., Pyridoxal phosphate enzymes: mechanistic, structural, and evolutionary considerations. *Annu Rev Biochem* **2004**, *73*, 383-415.
82. Draths, K. M.; Pompliano, D. L.; Conley, D. L.; Frost, J. W.; Berry, A.; Disbrow, G. L.; Staversky, R. J.; Lievense, J. C., Biocatalytic synthesis of aromatics from D-glucose: the role of transketolase. *J Am Chem Soc* **1992**, *114* (10), 3956-3962.
83. Kremer, A. B.; Egan, R. M.; Sable, H. Z., The active site of transketolase. Two arginine residues are essential for activity. *J Biol Chem* **1980**, *255* (6), 2405-10.
84. Dumont, T.; Barth, D.; Corbier, C.; Branlant, G.; Perrut, M., Enzymatic reaction kinetic: Comparison in an organic solvent and in supercritical carbon dioxide. *Biotechnol Bioeng* **1992**, *40* (2), 329-333.
85. Stevens, B. W.; Lilien, R. H.; Georgiev, I.; Donald, B. R.; Anderson, A. C., Redesigning the PheA domain of gramicidin synthetase leads to a new understanding of the enzyme's mechanism and selectivity. *Biochemistry* **2006**, *45* (51), 15495-504.
86. Chen, C. Y.; Georgiev, I.; Anderson, A. C.; Donald, B. R., Computational structure-based redesign of enzyme activity. *Proc Natl Acad Sci U S A* **2009**, *106* (10), 3764-9.
87. Sawaya, M. R.; Kraut, J., Loop and subdomain movements in the mechanism of *Escherichia coli* dihydrofolate reductase: crystallographic evidence. *Biochemistry* **1997**, *36* (3), 586-603.

88. Goto, M.; Miyahara, I.; Hayashi, H.; Kagamiyama, H.; Hirotsu, K., Crystal structures of branched-chain amino acid aminotransferase complexed with glutamate and glutarate: true reaction intermediate and double substrate recognition of the enzyme. *Biochemistry* **2003**, *42* (13), 3725-33.
89. Okada, K.; Hirotsu, K.; Hayashi, H.; Kagamiyama, H., Structures of Escherichia coli branched-chain amino acid aminotransferase and its complexes with 4-methylvalerate and 2-methylleucine: induced fit and substrate recognition of the enzyme. *Biochemistry* **2001**, *40* (25), 7453-63.
90. Markus, G.; Barnard, E. A.; Castellani, B. A.; Saunders, D., Ligand-induced conformational changes in ribonuclease. *J Biol Chem* **1968**, *243* (15), 4070-6.
91. Loria, J. P.; Berlow, R. B.; Watt, E. D., Characterization of enzyme motions by solution NMR relaxation dispersion. *Acc Chem Res* **2008**, *41* (2), 214-21.
92. Boehr, D. D.; McElheny, D.; Dyson, H. J.; Wright, P. E., The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* **2006**, *313* (5793), 1638-42.
93. Labouesse, B.; Oppenheimer, H. L.; Hess, G. P., Conformational changes accompanying the formation of chymotrypsin-substrate complexes. Evidence for the involvement of an N-terminal α -amino group in the activity and the conformation of the enzyme. *Biochem Biophys Res Commun* **1964**, *14* (4), 318-322.
94. Marzola, P.; Forte, C.; Pinzino, C.; Veracini, C. A., Activity and conformational changes of alpha-chymotrypsin in reverse micelles studied by spin labeling. *FEBS Lett* **1991**, *289* (1), 29-32.
95. Hayashi, H.; Mizuguchi, H.; Miyahara, I.; Nakajima, Y.; Hirotsu, K.; Kagamiyama, H., Conformational Change in Aspartate Aminotransferase on Substrate Binding Induces Strain in the Catalytic Group and Enhances Catalysis. *J Biol Chem* **2003**, *278* (11), 9481-9488.
96. Hayashi, H.; Mizuguchi, H.; Miyahara, I.; Islam, M. M.; Ikushiro, H.; Nakajima, Y.; Hirotsu, K.; Kagamiyama, H., Strain and catalysis in aspartate aminotransferase. *Biochim Biophys Acta* **2003**, *1647* (1-2), 103-9.
97. Islam, M. M.; Goto, M.; Miyahara, I.; Ikushiro, H.; Hirotsu, K.; Hayashi, H., Binding of C5-dicarboxylic substrate to aspartate aminotransferase: implications for the conformational change at the transaldimination step. *Biochemistry* **2005**, *44* (23), 8218-29.
98. Bouvignies, G.; Vallurupalli, P.; Hansen, D. F.; Correia, B. E.; Lange, O.; Bah, A.; Vernon, R. M.; Dahlquist, F. W.; Baker, D.; Kay, L. E., Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. *Nature* **2011**, *477* (7362), 111-4.
99. Gutteridge, A.; Thornton, J., Conformational change in substrate binding, catalysis and product release: an open and shut case? *FEBS Lett* **2004**, *567* (1), 67-73.
100. Hammes, G. G.; Chang, Y. C.; Oas, T. G., Conformational selection or induced fit: a flux description of reaction mechanism. *Proc Natl Acad Sci U S A* **2009**, *106* (33), 13737-41.
101. Hammes, G. G., Multiple conformational changes in enzyme catalysis. *Biochemistry* **2002**, *41* (26), 8221-8.
102. Fraser, J. S.; Clarkson, M. W.; Degnan, S. C.; Erion, R.; Kern, D.; Alber, T., Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **2009**, *462* (7273), 669-73.
103. Otten, R.; Liu, L.; Kenner, L. R.; Clarkson, M. W.; Mavor, D.; Tawfik, D. S.; Kern, D.; Fraser, J. S., Rescue of conformational dynamics in enzyme catalysis by directed evolution. *Nat Commun* **2018**, *9* (1), 1314.
104. Gomm, A.; O'Reilly, E., Transaminases for chiral amine synthesis. *Curr Opin Chem Biol* **2018**, *43*, 106-112.
105. Hwang, B.-Y.; Cho, B.-K.; Yun, H.; Koteswar, K.; Kim, B.-G., Revisit of aminotransferase in the genomic era and its application to biocatalysis. *J Mol Catal B Enzym* **2005**, *37* (1), 47-55.
106. Paul, C. E.; Rodríguez-Mata, M.; Busto, E.; Lavandera, I.; Gotor-Fernández, V.; Gotor, V.; García-Cerrada, S.; Mendiola, J.; de Frutos, Ó.; Collado, I., Transaminases Applied to the Synthesis of High Added-Value Enantiopure Amines. *Org Process Res Dev* **2014**, *18* (6), 788-792.

107. Pavkov-Keller, T.; Strohmeier, G. A.; Diepold, M.; Peeters, W.; Smeets, N.; Schürmann, M.; Gruber, K.; Schwab, H.; Steiner, K., Discovery and structural characterisation of new fold type IV-transaminases exemplify the diversity of this enzyme fold. *Sci Rep* **2016**, *6*, 38183.
108. Hirotsu, K.; Goto, M.; Okamoto, A.; Miyahara, I., Dual substrate recognition of aminotransferases. *Chem Rec* **2005**, *5* (3), 160-172.
109. Percudani, R.; Peracchi, A., The B6 database: a tool for the description and classification of vitamin B6-dependent enzymatic activities and of the corresponding protein families. *BMC Bioinformatics* **2009**, *10*, 273.
110. Amorim Franco, T. M.; Hegde, S.; Blanchard, J. S., Chemical Mechanism of the Branched-Chain Aminotransferase IlvE from Mycobacterium tuberculosis. *Biochemistry* **2016**, *55* (45), 6295-6303.
111. Goldberg, J. M.; Kirsch, J. F., The reaction catalyzed by Escherichia coli aspartate aminotransferase has multiple partially rate-determining steps, while that catalyzed by the Y225F mutant is dominated by ketimine hydrolysis. *Biochemistry* **1996**, *35* (16), 5280-91.
112. Julin, D. A.; Kirsch, J. F., Kinetic isotope effect studies on aspartate aminotransferase: evidence for a concerted 1,3 prototropic shift mechanism for the cytoplasmic isozyme and L-aspartate and dichotomy in mechanism. *Biochemistry* **1989**, *28* (9), 3825-33.
113. Hutson, S. M.; Fenstermacher, D.; Mahar, C., Role of mitochondrial transamination in branched chain amino acid metabolism. *J Biol Chem* **1988**, *263* (8), 3618-25.
114. Midelfort, K. S.; Kumar, R.; Han, S.; Karmilowicz, M. J.; McConnell, K.; Gehlhaar, D. K.; Mistry, A.; Chang, J. S.; Anderson, M.; Villalobos, A.; Minshull, J.; Govindarajan, S.; Wong, J. W., Redesigning and characterizing the substrate specificity and activity of Vibrio fluvialis aminotransferase for the synthesis of imagabalin. *Protein Eng Des Sel* **2013**, *26* (1), 25-33.
115. Mironov, G. G.; St-Jacques, A. D.; Mungham, A.; Eason, M. G.; Chica, R. A.; Berezovski, M. V., Bioanalysis for biocatalysis: multiplexed capillary electrophoresis-mass spectrometry assay for aminotransferase substrate discovery and specificity profiling. *J Am Chem Soc* **2013**, *135* (37), 13728-36.
116. Weinhandl, K.; Winkler, M.; Glieder, A.; Camattari, A., A novel multi-enzymatic high throughput assay for transaminase activity. *Tetrahedron* **2012**, *68* (37), 7586-7590.
117. Hopwood, J.; Truppo, M. D.; Turner, N. J.; Lloyd, R. C., A fast and sensitive assay for measuring the activity and enantioselectivity of transaminases. *Chem Commun (Camb)* **2011**, *47* (2), 773-5.
118. Kagamiyama, H.; Hayashi, H., [11] - Branched-Chain Amino-Acid Aminotransferase of Escherichia coli. In *Methods Enzymol*, Harris, R. A.; Sokatch, J. R., Eds. Academic Press: 2000; Vol. 324, pp 103-113.
119. Cooper, A. J. L.; Conway, M.; Hutson, S. M., A continuous 96-well plate spectrophotometric assay for branched-chain amino acid aminotransferases. *Anal Biochem* **2002**, *308* (1), 100-105.
120. Schadevaldt, P.; Adelmeyer, F., Coupled Enzymatic Assay for Estimation of Branched-Chain-Amino Acid Aminotransferase Activity with 2-Oxo Acid Substrates. *Anal Biochem* **1996**, *238* (1), 65-71.
121. Walton, C. J.; Chica, R. A., A high-throughput assay for screening L- or D-amino acid specific aminotransferase mutant libraries. *Anal Biochem* **2013**, *441* (2), 190-8.
122. Okamoto, A.; Nakai, Y.; Hayashi, H.; Hirotsu, K.; Kagamiyama, H., Crystal structures of Paracoccus denitrificans aromatic amino acid aminotransferase: a substrate recognition site constructed by rearrangement of hydrogen bond network11Edited by R. Huber. *J Mol Biol* **1998**, *280* (3), 443-461.
123. Okamoto, A.; Ishii, S.; Hirotsu, K.; Kagamiyama, H., The active site of Paracoccus denitrificans aromatic amino acid aminotransferase has contrary properties: flexibility and rigidity. *Biochemistry* **1999**, *38* (4), 1176-84.
124. Goto, M.; Omi, R.; Miyahara, I.; Hosono, A.; Mizuguchi, H.; Hayashi, H.; Kagamiyama, H.; Hirotsu, K., Crystal structures of glutamine:phenylpyruvate aminotransferase from Thermus thermophilus HB8: induced fit and substrate recognition. *J Biol Chem* **2004**, *279* (16), 16518-25.

125. Malashkevich, V. N.; Onuffer, J. J.; Kirsch, J. F.; Jansonius, J. N., Alternating arginine-modulated substrate specificity in an engineered tyrosine aminotransferase. *Nat Struct Biol* **1995**, 2 (7), 548-53.
126. Addington, T. A.; Mertz, R. W.; Siegel, J. B.; Thompson, J. M.; Fisher, A. J.; Filkov, V.; Fleischman, N. M.; Suen, A. A.; Zhang, C.; Toney, M. D., Janus: prediction and ranking of mutations required for functional interconversion of enzymes. *J Mol Biol* **2013**, 425 (8), 1378-89.
127. Okamoto, A.; Higuchi, T.; Hirotsu, K.; Kuramitsu, S.; Kagamiyama, H., X-ray crystallographic study of pyridoxal 5'-phosphate-type aspartate aminotransferases from *Escherichia coli* in open and closed form. *J Biochem* **1994**, 116 (1), 95-107.
128. Inc, C. C. G. *Molecular Operating Environment (MOE)*, 1010 Sherbrooke St. West Suite #910, Montreal, QC, Canada, H3A 2R7, 2014.09.
129. Wang, J.; Cieplak, P.; Kollman, P. A., How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem* **2000**, 21 (12), 1049-1074.
130. Limbach, H. H.; Chan-Huot, M.; Sharif, S.; Tolstoy, P. M.; Shenderovich, I. G.; Denisov, G. S.; Toney, M. D., Critical hydrogen bonds and protonation states of pyridoxal 5'-phosphate revealed by NMR. *Biochim Biophys Acta* **2011**, 1814 (11), 1426-37.
131. McDonald, I. K.; Thornton, J. M., Satisfying Hydrogen Bonding Potential in Proteins. *J Mol Biol* **1994**, 238 (5), 777-793.
132. Renfrew, P. D.; Choi, E. J.; Bonneau, R.; Kuhlman, B., Incorporation of noncanonical amino acids into Rosetta and use in computational protein-peptide interface design. *PLoS One* **2012**, 7 (3), e32637.
133. Jiang, L.; Kuhlman, B.; Kortemme, T.; Baker, D., A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins* **2005**, 58 (4), 893-904.
134. Khare, S. D.; Kipnis, Y.; Greisen, P., Jr.; Takeuchi, R.; Ashani, Y.; Goldsmith, M.; Song, Y.; Gallaher, J. L.; Silman, I.; Leader, H.; Sussman, J. L.; Stoddard, B. L.; Tawfik, D. S.; Baker, D., Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nat Chem Biol* **2012**, 8 (3), 294-300.
135. Cleland, W. W., The kinetics of enzyme-catalyzed reactions with two or more substrates or products: I. Nomenclature and rate equations. *Biochim Biophys Acta* **1963**, 67, 104-137.
136. Khersonsky, O.; Roodveldt, C.; Tawfik, D. S., Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr Opin Chem Biol* **2006**, 10 (5), 498-508.
137. Davey, J. A.; Chica, R. A., Multistate approaches in computational protein design. *Protein Sci* **2012**, 21 (9), 1241-52.
138. de los Santos, E. L.; Meyerowitz, J. T.; Mayo, S. L.; Murray, R. M., Engineering Transcriptional Regulator Effector Specificity Using Computational Design and In Vitro Rapid Prototyping: Developing a Vanillin Sensor. *ACS Synth Biol* **2016**, 5 (4), 287-95.
139. Hong, E. y.; Cha, M.; Yun, H.; Kim, B.-G., Asymmetric synthesis of l-tert-leucine and l-3-hydroxyadamantylglycine using branched chain aminotransferase. *J Mol Catal B Enzym* **2010**, 66 (1), 228-233.
140. Taylor, P. P.; Pantaleone, D. P.; Senkpeil, R. F.; Fotheringham, I. G., Novel biosynthetic approaches to the production of unnatural amino acids using transaminases. *Trends Biotechnol* **1998**, 16 (10), 412-418.
141. Wu, S.; Zhou, Y.; Wang, T.; Too, H. P.; Wang, D. I.; Li, Z., Highly regio- and enantioselective multiple oxy- and amino-functionalizations of alkenes by modular cascade biocatalysis. *Nat Commun* **2016**, 7, 11917.
142. Toney, M. D., Aspartate aminotransferase: an old dog teaches new tricks. *Arch Biochem Biophys* **2014**, 544, 119-27.
143. Reetz, M. T.; Carballeira, J. D., Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nature Protoc* **2007**, 2, 891.

144. Reetz, M. T.; Prasad, S.; Carballeira, J. D.; Gumulya, Y.; Bocola, M., Iterative saturation mutagenesis accelerates laboratory evolution of enzyme stereoselectivity: rigorous comparison with traditional methods. *J Am Chem Soc* **2010**, *132* (26), 9144-52.
145. Acevedo-Rocha, C. G.; Hoebeinreich, S.; Reetz, M. T., Iterative saturation mutagenesis: a powerful approach to engineer proteins by systematically simulating Darwinian evolution. *Methods Mol Biol* **2014**, *1179*, 103-28.
146. Patrick, W. M.; Firth, A. E.; Blackburn, J. M., User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Eng* **2003**, *16* (6), 451-7.
147. Arnold, F. H., Directed Evolution: Bringing New Chemistry to Life. *Angew Chem Int Ed* **2017**, *57* (16), 4143-4148.
148. Goldsmith, M.; Tawfik, D. S., Directed enzyme evolution: beyond the low-hanging fruit. *Curr Opin Struct Biol* **2012**, *22* (4), 406-12.
149. Yano, T.; Oue, S.; Kagamiyama, H., Directed evolution of an aspartate aminotransferase with new substrate specificities. *Proc Natl Acad Sci U S A* **1998**, *95* (10), 5511-5.
150. 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), 305-9.
151. Tang, L.; Gao, H.; Zhu, X.; Wang, X.; Zhou, M.; Jiang, R., Construction of "small-intelligent" focused mutagenesis libraries using well-designed combinatorial degenerate primers. *BioTechniques* **2012**, *52* (3), 149-58.
152. Vallejo, A. N.; Pogulis, R. J.; Pease, L. R., PCR Mutagenesis by Overlap Extension and Gene SOE. *CSH Protoc* **2008**, *2008*, pdb prot4861.
153. Hayashi, H.; Kuramitsu, S.; Kagamiyama, H., Replacement of an interdomain residue Val39 of *Escherichia coli* aspartate aminotransferase affects the catalytic competence without altering the substrate specificity of the enzyme. *J Biochem* **1991**, *109* (5), 699-704.
154. Mehta, P. K.; Hale, T. I.; Christen, P., Evolutionary relationships among aminotransferases. Tyrosine aminotransferase, histidinol-phosphate aminotransferase, and aspartate aminotransferase are homologous proteins. *Eur J Biochem* **1989**, *186* (1-2), 249-53.
155. Mehta, P. K.; Hale, T. I.; Christen, P., Aminotransferases: demonstration of homology and division into evolutionary subgroups. *Eur J Biochem* **1993**, *214* (2), 549-61.
156. Hammes, G. G., Mechanism of Enzyme Catalysis. *Nature* **1964**, *204*, 342.
157. Hammes, G. G., Multiple Conformational Changes in Enzyme Catalysis. *Biochemistry* **2002**, *41* (26), 8221-8228.
158. G Careri; P Fasella, a.; Gratton, E., Enzyme Dynamics: The Statistical Physics Approach. *Annu Rev Biophys Bioeng* **1979**, *8* (1), 69-97.
159. Hayward, S.; Kitao, A.; Berendsen, H. J., Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins* **1997**, *27* (3), 425-37.
160. Hayward, S.; Berendsen, H. J., Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. *Proteins* **1998**, *30* (2), 144-54.
161. Hayward, S.; Lee, R. A., Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. *J Mol Graph Model* **2002**, *21* (3), 181-3.
162. Poornam, G. P.; Matsumoto, A.; Ishida, H.; Hayward, S., A method for the analysis of domain movements in large biomolecular complexes. *Proteins* **2009**, *76* (1), 201-12.
163. Frishman, D.; Argos, P., Knowledge-based protein secondary structure assignment. *Proteins* **1995**, *23* (4), 566-79.
164. Metzler, D. E.; Metzler, C. M.; Scott, R. D.; Mollova, E. T.; Kagamiyama, H.; Yano, T.; Kuramitsu, S.; Hayashi, H.; Hirotsu, K.; Miyahara, I., NMR studies of ¹H resonances in the 10-18-ppm range for aspartate aminotransferase from *Escherichia coli*. *J Biol Chem* **1994**, *269* (45), 28027-33.

165. Picot, D.; Sandmeier, E.; Thaller, C.; Vincent, M. G.; Christen, P.; Jansonius, J. N., The open/closed conformational equilibrium of aspartate aminotransferase. Studies in the crystalline state and with a fluorescent probe in solution. *Eur J Biochem* **1991**, *196* (2), 329-41.
166. Michielsens, S.; Peters, J. H.; Ban, D.; Pratihari, S.; Seeliger, D.; Sharma, M.; Giller, K.; Sabo, T. M.; Becker, S.; Lee, D.; Griesinger, C.; de Groot, B. L., A designed conformational shift to control protein binding specificity. *Angew Chem Int Ed Engl* **2014**, *53* (39), 10367-71.
167. Mayo, S. L.; Olafson, B. D.; Goddard, W. A., Dreiding - a Generic Force-Field for Molecular Simulations. *J Phys Chem* **1990**, *94* (26), 8897-8909.
168. Richter, F.; Blomberg, R.; Khare, S. D.; Kiss, G.; Kuzin, A. P.; Smith, A. J.; Gallaher, J.; Pianowski, Z.; Helgeson, R. C.; Grjasnow, A.; Xiao, R.; Seetharaman, J.; Su, M.; Vorobiev, S.; Lew, S.; Forouhar, F.; Kornhaber, G. J.; Hunt, J. F.; Montelione, G. T.; Tong, L.; Houk, K. N.; Hilvert, D.; Baker, D., Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. *J Am Chem Soc* **2012**, *134* (39), 16197-206.
169. Siegel, J. B.; Zanghellini, A.; Lovick, H. M.; Kiss, G.; Lambert, A. R.; St Clair, J. L.; Gallaher, J. L.; Hilvert, D.; Gelb, M. H.; Stoddard, B. L.; Houk, K. N.; Michael, F. E.; Baker, D., Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* **2010**, *329* (5989), 309-13.
170. Sigala, P. A.; Kraut, D. A.; Caaveiro, J. M. M.; Pybus, B.; Ruben, E. A.; Ringe, D.; Petsko, G. A.; Herschlag, D., Testing Geometrical Discrimination within an Enzyme Active Site: Constrained Hydrogen Bonding in the Ketosteroid Isomerase Oxyanion Hole. *J Am Chem Soc* **2008**, *130* (41), 13696-13708.
171. Alvizo, O.; Mayo, S. L., Evaluating and optimizing computational protein design force fields using fixed composition-based negative design. *Proc Natl Acad Sci U S A* **2008**, *105* (34), 12242-7.
172. Song, Y.; Tyka, M.; Leaver-Fay, A.; Thompson, J.; Baker, D., Structure-guided forcefield optimization. *Proteins* **2011**, *79* (6), 1898-909.
173. Leaver-Fay, A.; O'Meara, M. J.; Tyka, M.; Jacak, R.; Song, Y.; Kellogg, E. H.; Thompson, J.; Davis, I. W.; Pache, R. A.; Lyskov, S.; Gray, J. J.; Kortemme, T.; Richardson, J. S.; Havranek, J. J.; Snoeyink, J.; Baker, D.; Kuhlman, B., Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol* **2013**, *523*, 109-43.
174. Vanommeslaeghe, K.; MacKerell, A. D., Jr., CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochim Biophys Acta* **2015**, *1850* (5), 861-71.
175. Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H., Electrostatic basis for enzyme catalysis. *Chem Rev* **2006**, *106* (8), 3210-35.
176. Carlin, D. A.; Caster, R. W.; Wang, X.; Betzenderfer, S. A.; Chen, C. X.; Duong, V. M.; Ryklansky, C. V.; Alpekin, A.; Beaumont, N.; Kapoor, H.; Kim, N.; Mohabbot, H.; Pang, B.; Teel, R.; Whithaus, L.; Tagkopoulos, I.; Siegel, J. B., Kinetic Characterization of 100 Glycoside Hydrolase Mutants Enables the Discovery of Structural Features Correlated with Kinetic Constants. *PLoS One* **2016**, *11* (1), e0147596.
177. Mak, W. S.; Siegel, J. B., Computational enzyme design: transitioning from catalytic proteins to enzymes. *Curr Opin Struct Biol* **2014**, *27*, 87-94.
178. O'Meara, M. J.; Leaver-Fay, A.; Tyka, M. D.; Stein, A.; Houlihan, K.; DiMaio, F.; Bradley, P.; Kortemme, T.; Baker, D.; Snoeyink, J.; Kuhlman, B., Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J Chem Theory Comput* **2015**, *11* (2), 609-22.
179. LuCore, S. D.; Litman, J. M.; Powers, K. T.; Gao, S.; Lynn, A. M.; Tollefson, W. T.; Fenn, T. D.; Washington, M. T.; Schnieders, M. J., Dead-End Elimination with a Polarizable Force Field Repacks PCNA Structures. *Biophys J* **2015**, *109* (4), 816-26.
180. Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A., Jr.; Head-Gordon, M.; Clark, G. N.; Johnson, M. E.; Head-Gordon, T., Current status of the AMOEBA polarizable force field. *J Phys Chem B* **2010**, *114* (8), 2549-64.

181. Friedland, G. D.; Linares, A. J.; Smith, C. A.; Kortemme, T., A simple model of backbone flexibility improves modeling of side-chain conformational variability. *J Mol Biol* **2008**, *380* (4), 757-74.
182. Babor, M.; Mandell, D. J.; Kortemme, T., Assessment of flexible backbone protein design methods for sequence library prediction in the therapeutic antibody Herceptin-HER2 interface. *Protein Sci* **2011**, *20* (6), 1082-9.
183. Roberts, K. E.; Donald, B. R., Improved energy bound accuracy enhances the efficiency of continuous protein design. *Proteins* **2015**, *83* (6), 1151-64.
184. Hallen, M. A.; Keedy, D. A.; Donald, B. R., Dead-end elimination with perturbations (DEEPer): a provable protein design algorithm with continuous sidechain and backbone flexibility. *Proteins* **2013**, *81* (1), 18-39.
185. Gagne, D.; Doucet, N., Structural and functional importance of local and global conformational fluctuations in the RNase A superfamily. *FEBS J* **2013**, *280* (22), 5596-607.
186. Gagne, D.; French, R. L.; Narayanan, C.; Simonovic, M.; Agarwal, P. K.; Doucet, N., Perturbation of the Conformational Dynamics of an Active-Site Loop Alters Enzyme Activity. *Structure* **2015**, *23* (12), 2256-66.
187. Voigt, C. A.; Mayo, S. L.; Arnold, F. H.; Wang, Z. G., Computational method to reduce the search space for directed protein evolution. *Proc Natl Acad Sci U S A* **2001**, *98* (7), 3778-83.
188. Bhowmick, A.; Head-Gordon, T., A monte carlo method for generating side chain structural ensembles. *Structure* **2015**, *23* (1), 44-55.
189. Sevy, A. M.; Jacobs, T. M.; Crowe, J. E., Jr.; Meiler, J., Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences. *PLoS Comput Biol* **2015**, *11* (7), e1004300.
190. Howell, S. C.; Inampudi, K. K.; Bean, D. P.; Wilson, C. J., Understanding thermal adaptation of enzymes through the multistate rational design and stability prediction of 100 adenylate kinases. *Structure* **2014**, *22* (2), 218-29.
191. Keedy, D. A.; Kenner, L. R.; Warkentin, M.; Woldeyes, R. A.; Hopkins, J. B.; Thompson, M. C.; Brewster, A. S.; Van Benschoten, A. H.; Baxter, E. L.; Uervirojnangkoorn, M.; McPhillips, S. E.; Song, J.; Alonso-Mori, R.; Holton, J. M.; Weis, W. I.; Brunger, A. T.; Soltis, S. M.; Lemke, H.; Gonzalez, A.; Sauter, N. K.; Cohen, A. E.; van den Bedem, H.; Thorne, R. E.; Fraser, J. S., Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *Elife* **2015**, *4*.
192. Jimenez-Oses, G.; Osuna, S.; Gao, X.; Sawaya, M. R.; Gilson, L.; Collier, S. J.; Huisman, G. W.; Yeates, T. O.; Tang, Y.; Houk, K. N., The role of distant mutations and allosteric regulation on LovD active site dynamics. *Nat Chem Biol* **2014**, *10* (6), 431-6.
193. Campbell, E.; Kaltenbach, M.; Correy, G. J.; Carr, P. D.; Porebski, B. T.; Livingstone, E. K.; Afriat-Jurnou, L.; Buckle, A. M.; Weik, M.; Hollfelder, F.; Tokuriki, N.; Jackson, C. J., The role of protein dynamics in the evolution of new enzyme function. *Nat Chem Biol* **2016**, *12* (11), 944-950.
194. van den Bedem, H.; Dhanik, A.; Latombe, J. C.; Deacon, A. M., Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers. *Acta Cryst* **2009**, *D65* (Pt 10), 1107-17.
195. Lang, P. T.; Ng, H. L.; Fraser, J. S.; Corn, J. E.; Echols, N.; Sales, M.; Holton, J. M.; Alber, T., Automated electron-density sampling reveals widespread conformational polymorphism in proteins. *Protein Sci* **2010**, *19* (7), 1420-31.
196. Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Shaw, D. E., Picosecond to Millisecond Structural Dynamics in Human Ubiquitin. *J Phys Chem B* **2016**, *120* (33), 8313-20.
197. Bhabha, G.; Lee, J.; Ekiert, D. C.; Gam, J.; Wilson, I. A.; Dyson, H. J.; Benkovic, S. J.; Wright, P. E., A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science* **2011**, *332* (6026), 234-8.