NEW TOOLS AND APPROACHES FOR COMPUTATIONAL PROTEIN DESIGN

By

KRISTIN MARIE BLACKLOCK

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Programs in Quantitative Biomedicine and Chemistry and Chemical Biology

Written under the direction of

Dr. Sagar D. Khare

And approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2019

ABSTRACT OF THE DISSERTATION

New Tools and Approaches for Computational Protein Design

by KRISTIN MARIE BLACKLOCK

Dissertation Director:

Sagar D. Khare, Ph.D.

This dissertation describes the development, benchmarking, validation, and application of computational methods, mostly written within the Rosetta suite of macromolecular modeling software, for the design and/or analysis of proteins.

The introductory chapter brings the reader into the historical context of computational protein design, while detailing important concepts referenced in subsequent chapters.

The second chapter is comprised of a benchmarking study of "LooDo", a computational algorithm for the design of novel nested-domain proteins, which are proteins where one domain is inserted into another. This algorithm, named after its primary sampling method of loop-directed domain placement, was shown to be able to recapitulate native domain orientations for a benchmark set of five nested-domain proteins, as well as recapitulate domain-domain interface sequences and rank native versus nonnative domain combinations highly.

In the next chapter, to improve the therapeutic ratio of the yeast cytosine deaminase (yCD)/5-fluorocytosine (5FC) directed-enzyme prodrug therapy system for targeted chemotherapy, we hypothesized that light-induced structural changes in yCD via bifunctional azobenzene derivative cross-linking would allow for the design of a photoswitchable yCD enzyme. Using generalizable computational design methods and experimental validation, we present one such design that allowed for a roughly 2-fold increase in activity towards cytosine under UV versus blue light stimuli.

The fourth chapter presents a study in which the Rosetta and Amber energy functions are systematically compared by their performance in two structural evaluation tests and afterwards combined to increase the overall performance over both individually. The minimum-sum-of-ranks method employed in this chapter reduces the RMSD of the selected decoy by 1Å in 14 cases for the ff14SBonlySC energy function in Amber and 13 cases for the current Rosetta energy function, REF2015, in a large decoy discrimination benchmark test.

The final chapter investigates bioisosteric alternatives to Axitinib in order to reduce the metabolic vulnerability of the heteroaryl thioether group. Using QM calculations, Rosetta docking protocols, and Amber molecular dynamics simulations, this study computationally evaluates four proposed structures by their predicted behaviors within the VEGFR2 kinase and ABL1 T315I gatekeeper mutant kinase binding pockets.

Together, the work herein represents a collection of developments in the fields of computational biology and protein design.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my advisor, Dr. Sagar D. Khare, for his inspiration and mentorship throughout my graduate career. Without his guidance, encouragement, and support, the studies presented here (and many others not presented!) would not have been possible.

Alongside my advisor, I would like to extend a heartfelt thank you to my committee members, Dr. David Case, Dr. Gaetano Montelione, Dr. Vikas Nanda, and Dr. G. Andrew Woolley, for their insightful comments, assistance in steering my scientific development, and generosity in sharing their time and resources with me.

I am grateful to the many members of the CBMB/QB administration who helped me navigate through the business side of graduate school, and who were always willing to lend me their advice in matters of science and life, including Dr. Gail Ferstandig-Arnold, Luz Fajardo, Michelle Sanghera, and Kenneth Dalenberg.

I would also like to thank our endlessly patient postdoc, Dr. Brahm Yachnin, for his constant help throughout the latter years of my graduate career. I greatly appreciate everything you have taught me, not only in your role as a mentor of technical skills but also as a role model for resilience and thoughtfulness.

Thank you to my many collaborators and mentors throughout Rutgers who helped me learn new techniques and expand my scientific experiences. I would especially like to thank Dr. G.V.T. Swapna for helping me learn fluorine NMR, Dr. Conceicao Minetti and Dr. David Remeta for helping me with circular dichroism, Dr. Hai Nguyen for his support with

AMBER MD simulations and his pytraj expertise, and Dr. Spencer Knapp and Robert Barrows for including me in the Axitinib bioisosteres project.

I am thankful for my labmates and friends, who have not only helped me develop as a computational biologist and biochemist, but who also supported me mentally and emotionally throughout graduate school, including Dr. Manasi Pethe, Dr. Lu Yang, Dr. Aliza Grossman, Dr. Nancy Hernandez, William Hansen, Elliott Dolan, and Dmitri Zorine.

And finally, I want to thank my husband, Dr. Patrick Nosker, for every brainstorming session, every push in the right direction, and every day you made sure I could see the bigger picture.

DEDICATION

To my eternal cheerleaders, my parents,

Kathryn Marie Blacklock and Arthur Scott Blacklock

And stepparents,

Elmer Frank Elkins and Mary Melinda "Lindy" Blacklock

To my supportive older brother,

Arthur Ryan Blacklock

To my pillars, my grandparents,

Jim and Ann McKown

To my beloved husband and best friend,

Patrick Leonard Nosker

To my peanut butter,

Nancy Estela Hernandez

TABLE OF CONTENTS

ABSTRACT OF THE DISSERTATION ii

ACKNOWLEDGEMENTS iv

DEDICATION vi

TABLE OF CONTENTS vii

LIST OF TABLES x

LIST OF ILLUSTRATIONS xi

LIST OF ABBREVIATIONS xiv

Chapter 1: Introduction 1

1.1 Proteins: Sequence/Structure/Function and the Folding Funnel 1

1.2 Challenges and Successes of Computational Protein Design 2

1.3 A Brief Overview of Key Scoring Methods 4

1.4 A Brief Overview of Key Sampling Methods 10

1.5 References 18

Chapter 2: A Computational Method for the Design of Nested Proteins by Loop-Directed Domain Insertion 25

2.1 Preface 25

2.2 Abstract 25

2.3 Introduction 26

2.4 Materials and Methods 29

2.5 Results 35

2.6 Discussion 52

2.7 Conclusion 54

2.8 References 55

Chapter 3: Computational Design of a Photo-controlled Cytosine Deaminase....60

3.1 Preface....60

3.2 Abstract....60

3.3 Introduction....61

3.4 Results and Discussion....62

3.5 Conclusion....76

3.6 Methods:....77

3.7 References....88

Chapter 4: Systematic comparison of Amber and Rosetta energy functions for protein structure evaluation....90

4.1 Preface....90

4.2 Abstract....90

4.3 Introduction....91

4.4 Methods....94

4.5 Results....100

4.6 Discussion....118

4.7 Supplementary Figures....120

4.8 Supplementary Tables....167

4.9 Supplementary Software....182

4.10 References....198

Chapter 5: Computational assessment of thioether isosteres....202

5.1 Preface....202

5.2 Abstract....202

5.3 Introduction....203

5.4 Results and discussion....206

5.5 Conclusions....229

5.6 Methods....230

5.7 References 234

LIST OF TABLES

Table 1-1: Summary of terms in REF15 for proteins. 6

Table 2-1: Benchmark set of two-domain protein complexes. 35

Table 2-2: Benchmark set of two-domain protein complexes, in more detail. 36

Table 2-3: Benchmark results for native domain combinations. 39

Table 2-4: Quantification of funnel-shaped distributions found in Figure **2-3**. 40

Table 2-5: Results of combining each insert domain with each parent domain. 43

Table 2-6: Effect of decreasing number of catalytic residue constraints originating from the insert domain on rank of native insert-parent domain combination. 46

Table 2-7: Effect of decreasing the molten region on lowest insert domain RMSD in the benchmark set. 48

Table 2-8: Subset of 10 PDBs from the Gray set. 52

Table 3-1: yCD Variants. 64

Table 3-2: Michaelis-Menten kinetics for yCD-Opt1 under UV or blue light condtions. 74

Table 4-1: *B* metric, false minima, and model selection summary comparisons for Amber ff14SBonlySC, Rosetta talaris2014, and Rosetta REF2015 energy functions. 102

Table 4-2: *B*-metric values for Rosetta talaris2014, Rosetta REF2015, and Amber ff14sbonlysc for the Decoy Discrimination benchmark set. 167

Table 4-3: Values of RMSD for Min-sum-selected, Rosetta-selected, and Amber-selected models, and minimum RMSD model in the set (All). 174

Table 4-4: *B*-metric values for Rosetta talaris2014, Rosetta REF2015, and Amber ff14SBonlySC for the Loop Modeling benchmark set. 180

Table 5-1: Calculated structural parameters for diphenyl thioether (3a) and proposed isosteres 4a–25a. 209

Table 5-2: Calculated intramolecular N–N' distances for bis(pyridin-4-yl) thioether (3b) and proposed isosteres 4b–25b, and intermolecular N'-N' distances in overlap of each of 4b–25b with 3b. 213

Table 5-3: Calculated distances and energies for Axitinib and four isosteres in two protein binding sites. 222

LIST OF ILLUSTRATIONS

Figure 2-1: Flowchart and cartoon diagram of the design pipeline, including the LooDo algorithm.....30

Figure 2-2: Cartoon representations of the five benchmark proteins.....37

Figure 2-3: Graphs of (A) Total Energy, (B) FNat, and (C) Interface Energy versus Insert Domain RMSD.....40

Figure 2-4: Energy distributions among the set of 500 decoys containing the top 100 decoys from each of the five insert-parent domain combination sets with the same parent domain.....44

Figure 2-5: Graphs of Total Energy (A), FNat (B), and Interface Energy (C) versus Insert Domain RMSD for the native 1WM1 insert-parent combination set when LooDo Hits are filtered with 0, 1, or 2 residues included in the insert domain catalytic constraints.....45

Figure 2-6: Energy distributions among the set of 500 decoys containing the top 100 decoys from each of the five insert-parent domain combination sets with the 1WM1 parent domain.....46

Figure 2-7: Decreasing the molten region during fragment picking shifts the distribution of the lowest insert domain RMSD toward more near-native conformations in the Gray set. 48

Figure 2-8: LooDo analyses for five PDBs from the Gray set where non-catalytic conformational constraints were used during the Matching stage show a lack of FNat or Interface Energy funnels.....50

Figure 2-9: LooDo analyses for ten PDBs from the Gray set where non-catalytic conformational constraints were used during the Matching stage show a lack of FNat or Interface Energy funnels.....52

Figure 2-10: Total Energy versus Insert Domain RMSD for structure prediction benchmark. 54

Figure 3-1: Computational design of ABDM-labeled yCD-dC3 variants.....63

Figure 3-2: *Cis*-to-*trans* simulations of yCD-Opt1 and yCD-Opt2.....66

Figure 3-3: Denaturing, reducing SDS-PAGE Gels showing labeling patterns of azobenzene derivatives conjugated to yCD variants.....68

Figure 3-4: Post-crosslinking analysis via free-cysteine PEGylation.....69

Figure 3-5: yCD-Opt1 Labeling with ABDM.....71

Figure 3-6: Bound ABDM characterization and optical control of cross-linked yCD-Opt1 cytosine deaminase activity.....73

Figure 3-7: yCD-Opt1 activity assays.....74

Figure 3-8: CD Thermal Melt Spectra.....76

Figure 3-9: ABDM internal degrees of freedom.....78

Figure 3-10: ABDM-cysteine geometric degrees of freedom. 80

Figure 4-1: (A-C) Energy landscapes for 2QY7, 1T2I, and 1SEN respectively using Rosetta talaris2014 (salmon; left), Rosetta REF2015 (purple; middle) and Amber ff14SBonlySC (turquoise; right) energy functions. 98

Figure 4-2: Scatterplots to depict general performance of the scoring functions in decoy discrimination. 101

Figure 4-3: Contact maps and representative structure comparisons for 1SEN, 2QY7, and 1T2I. 105

Figure 4-4: Per-residue and per-score-term propensity of energy functions toward false minima. 107

Figure 4-5: Minimum-sum model selection.....111

Figure 4-6: Loop modeling benchmark.....114

Figure 4-7: Comparison between the lowest energy decoy and lowest RMSD decoy for 1CYO. 115

Figure 4-8: Comparison between the lowest-energy decoy and lowest-RMSD decoy for 1MSC. 117

Figure 4-9: Energy landscapes for all decoy discrimination proteins. 120

Figure 4-10: Plot of RMSD for the decoy selected by various criteria 156

Figure 4-11: Energy landscapes for all loop modeling proteins. 157

Figure 5-1: Representative aryl and heteroaryl thioethers from the medicinal chemistry literature.6 204

Figure 5-2: Comparison of the calculated structural parameters of diphenyl thioether (1) and diphenyl ether (2). 206

Figure 5-3: Proposed isosteres of diphenyl thioether (3a).....208

Figure 5-4: One-ring overlap models of selected isosteres (labeled), each with bis(pyrid-4-yl) thioether (3b)..... 212

Figure 5-5: Comparison of the calculated structural parameters of bis(pyridin-4-yl) thioether (3b, left) and bis(pyrimidin-4-yl) thioether (3c, right). 215

Figure 5-6: Chart4: Electrostatic potential diagrams for diphenyl thioether (3a), bis(pyrid-4-yl) thioether (3b), bis(pyrimid-4-yl) thioether (3c), and three isosteres. 217

Figure 5-7: Proposed Axitinib isosteres..... 219

Figure 5-8: Axitinib shown in a PDB rendering in the respective binding sites of ABL1 (top) and VEGFR (bottom)..... 220

Figure 5-9: Implicit molecular dynamics simulations of Abl kinase with Axitinib (first column), CP-Ax (middle column), and FE-Ax (third column)..... 226

Figure 5-10: Implicit molecular dynamics simulations of the VEGFR2 kinase with Axitinib (first column), FE-Ax (middle column), and DMS-Ax (third column)..... 228

Figure 5-11: Molecular dynamics simulation analyses of CP-Ax, FE-Ax, DFS-Ax, and DMS-Ax in the Abl Kinase (left side) and VEGFR (right side) binding pockets..... 229

Figure 5-12: Electron density from total SCF density (isoval = 0.0001)..... 231

LIST OF ABBREVIATIONS

5FC	:	5-fluorocytosine
5FU	:	5-fluorouracil
5FW	:	5-fluorotryptophan
Å	:	Angstrom
ABDM	:	4,4'-bis(maleimido)azobenzene
AMBER	:	Assisted Model Building with Energy Refinement
CADRES	:	Collaborative Assessment and Development of Rosetta Energetics and Sampling
CPD	:	Computational Protein Design
CST	:	Constraint (used in context as CST Score, which is a score term reflecting the penalty associated with a conformation's lack of upholding a set of constraints)
FNat	:	Fraction of Native Amino Acid Identities
FRET	:	Förster Resonance Energy Transfer
GB	:	Generalized Born
KIC	:	Kinematic Loop Closure
LooDo	:	Loop-Directed Domain Insertion
MC	:	Monte Carlo
MD	:	Molecular Dynamics
NMR	:	Nuclear Magnetic Resonance
PDB	:	Protein Databank
QM	:	Quantum Mechanics
RMSD	:	Root-Mean-Square Deviation
SAXS	:	Small-Angle X-ray Scattering
VDW	:	Van der Waals
yCD	:	Yeast Cytosine Deaminase

Chapter 1: Introduction

1.1 Proteins: Sequence/Structure/Function and the Folding Funnel

Nearly all biological processes are supported by the vast and diverse repertoire of proteins and their associated sets of complex functions. These functions, including cellular structure and organization, storage, defense, transport, signaling, sensing, and catalysis, are intimately dependent on the *shape* of each protein involved in these processes[1,2]. A protein's shape, more specifically, reflects the ability of its one-dimensional primary sequence of amino acids to adopt a specific thermodynamic ensemble of three-dimensional structures, where one structure—the "native" structure—is usually considered to be representative of both the ensemble and/or the functional conformation[3].

Although environmental factors and helper molecules can also play a role in the final form of a protein's three-dimensional, or "tertiary", structure, it is clear from many studies that structure—and therefore function—is strongly correlated (to a first approximation) with the one-dimensional amino acid sequence[4]. Remarkably, the amino acid sequence for almost all proteins is built by chaining together monomers from the same set of roughly 20 amino acid building blocks in different configurations, yet still a great amount of multifaceted information is passed on to each three-dimensional structure. The nature of this transition from the one-dimensional world of sequences to the three-dimensional world of macromolecules (also known as protein "folding") has captivated biochemists and molecular biophysicists for more than 50 years, beginning with Anfinsen's hypothesis that a protein in its native state will adopt a structure with the lowest possible energy[5,6]. This, as well as Levinthal's question as to how proteins can rapidly fold into their native structures despite their polypeptide chains encountering a huge landscape

of possible unstructured conformations[7], eventually led to the development of one popular representation of a protein's folding pathway, or energy landscape: the "folding funnel". This model posits that proteins fold along one or more paths whereby random thermal motions mostly drive a protein's structure incrementally downwards in energy and towards the native structure located at the free-energy minimum[8–11].

Hence, understanding both the sequence/structure relationship of proteins and the energetic forces underlying their structural transitions will be important factors in building a holistic understanding of proteins and their functions. Once this deep understanding has been attained, it will be possible to apply any generalizable rules towards the design of tailored proteins from scratch, which can potentially be endowed with any number of desired properties useful for the advancement of biomolecular research. This challenge to comprehend and model such complex biological puzzles is being tackled by the field of computational protein design (CPD), a field that not only aims to critically test our current understanding of protein sequence/structure/function relationships, but also aims to apply our understanding to create useful and novel proteins that are custom-made for specific tasks.

1.2 Challenges and Successes of Computational Protein Design

For many reasons, it may be impossible or prohibitively time-consuming to experimentally determine the three-dimensional structures of proteins of interest. In fact, only a tiny fraction of the known sequences in the Universal Protein Resource (UniProt)[12], have been experimentally characterized[13]. Therefore, to aid in the formulation of testable hypotheses and further research efforts regarding these proteins, it is often required to construct models of their structures using tools from CPD. In some CPD challenges, one's goal may be to predict the structure of a protein given an amino acid sequence (the

structure prediction problem). In other cases, the three-dimensional structure may be known and used as a template for probing the effects of focused sets of mutations (template-based design), or the backbone conformation may also be a designed element such that one's goal is to *create* both the sequence and the structure "from scratch" (*de novo* protein design). In these last two scenarios, the structure prediction challenge is inverted: instead, one must find amino acid sequences that fold into the given/designed structure (the inverse folding problem)[14].

If both the structure prediction and inverse folding problems are posed on the grounds of the folding funnel hypothesis, then the probability of finding suitable answers for either approach requires (at least) two components: (1) a reasonably accurate scoring system for computing the quality of a protein conformation, and (2) methods for adequately sampling the landscape of possible protein sequences and conformations. Encouragingly, several computational methods and algorithms for structure prediction and inverse folding problems have resulted in a series of remarkable successes, often producing designs that were validated to atomic-level accuracy when compared to experimentally-determined structures. These include, but are by no means limited to, the development of *de novo* proteins and peptides with novel folds[15–17], novel enzymes[18–22], high-affinity ligand-binding proteins[23–25], nanocages and self-assembling superstructures[26–30], and thermostablilized proteins[31–34].

The challenges faced by endeavors to find appropriate energy functions and sampling methods are numerous, but fortunately considerable progress on both fronts has been made in recent decades. The following sections provide a brief introduction to the scoring and sampling problems, followed by descriptions of key methods utilized by the subsequent chapters of this dissertation.

1.3 A Brief Overview of Key Scoring Methods

Most methods for CPD make use of a scoring function to differentiate between "good" and "bad" conformations that are sampled during or after the modeling protocol. In both structure prediction and inverse folding simulations, scoring functions are often used to direct sampling towards lower energy states, to evaluate the goodness of fit between a sequence and a structure, and to select one or more native-like conformations from a pool of discrete conformations ("decoys") that are computed by various sampling methods. Most scoring functions aim to accurately condense the information regarding a protein's conformation and environmental interactions into a single number or score such that it becomes possible to discriminate between native, folded but non-native, and unfolded conformations due to large differences in the scores or rankings of each structure.

There are (at least) four different types of scoring functions, and different scoring approaches can be comprised entirely of one type or a combination of multiple types[35]:

- Physics-based scoring functions, which encompass force-field-based scoring functions along with the addition of solvation energy terms, attempt to model physical forces explicitly,
- Empirically-derived scoring functions are those comprised of a weighted sum of contributions of individual terms, which may not necessarily be of the same form as those in physics-based scoring functions and may be composed entirely from scratch. The functional forms of empirical scoring functions are usually more intuitive interpretations of important energetic factors and the weights are usually determined through regression analysis.
- Knowledge-based scoring functions utilize statistical pairwise potentials derived from a knowledge base, such as the PDB, to classify certain

interactions/conformations as energetically favorable or unfavorable by comparing the frequency of the current state to some reference state, i.e. a random distribution.

- Descriptor-based scoring functions derive statistical models from molecular features, such as element-element distance counts or number of hydrogen bonds, and use machine-learning for variable selection. This type of scoring function differs from empirical models in that, depending on the machine-learning technique used, descriptor-based models may not adopt a linear functional form.

As subsequent chapters in this dissertation utilize two popular protein modeling programs, Rosetta and Amber, further review will be focused on the functional forms and underlying meanings of the scoring functions used by these programs.

1.3.1 The Rosetta Energy Function

The Rosetta scoring function (or energy function) is a hybrid linear-weighted sum of both physics-based energy terms and knowledge-based potentials. The current functional form of the default Rosetta energy function, Rosetta Energy Function 2015 (REF15)[36], shown in **Eq. (1)**, approximates the energy of a conformation (ΔE_{total}) by summing over the weighted (w_i) values of 19 energy terms (E_i) that are calculated as functions of degrees of freedom and chemical identities (Θ_i, aa_i):

$$\Delta E_{total} = \sum_i w_i E_i(\Theta_i, aa_i) \tag{1}$$

Table 1-1 provides a succinct snapshot of each energy term and its meaning. These 19 terms are discussed in detail below and can be divided into two categories: one-body and two-body interactions.

Table 1-1: Summary of terms in REF15 for proteins.(Reprinted with permission from Alford. et. al.[36])

term	description	weight	units
fa_atr	attractive energy between two atoms on different residues separated by a distance d	1.0	kcal/mol
fa_rep	repulsive energy between two atoms on different residues separated by a distance d	0.55	kcal/mol
fa_intra_rep	repulsive energy between two atoms on the same residue separated by a distance d	0.005	kcal/mol
fa_sol	Gaussian exclusion implicit solvation energy between protein atoms in different residues	1.0	kcal/mol
lk_ball_wtd	orientation-dependent solvation of polar atoms assuming ideal water geometry	1.0	kcal/mol
fa_intra_sol	Gaussian exclusion implicit solvation energy between protein atoms in the same residue	1.0	kcal/mol
fa_elec	energy of interaction between two nonbonded charged atoms separated by a distance d	1.0	kcal/mol
hbond_lr_bb	energy of short-range hydrogen bonds	1.0	kcal/mol
hbond_sr_bb	energy of long-range hydrogen bonds	1.0	kcal/mol
hbond_bb_sc	energy of backbone–side-chain hydrogen bonds	1.0	kcal/mol
hbond_sc	energy of side-chain–side-chain hydrogen bonds	1.0	kcal/mol
dslf_fa13	energy of disulfide bridges	1.25	kcal/mol
rama_prepro	probability of backbone ϕ, ψ angles given the amino acid type	(0.45 kcal/mol)/kT	kT
p_aa_pp	probability of amino acid identity given backbone ϕ, ψ angles	(0.4 kcal/mol)/kT	kT
fa_dun	probability that a chosen rotamer is native-like given backbone ϕ, ψ angles	(0.7 kcal/mol)/kT	kT
omega	backbone-dependent penalty for cis ω dihedrals that deviate from 0° and trans ω dihedrals that deviate from 180°	(0.6 kcal/mol)/AU	AU[a]
pro_close	penalty for an open proline ring and proline ω bonding energy	(1.25 kcal/mol)/AU	AU
yhh_planarity	sinusoidal penalty for nonplanar tyrosine χ_3 dihedral angle	(0.625 kcal/mol)/AU	AU
ref	reference energies for amino acid types	(1.0 kcal/mol)/AU	AU

[a]AU = arbitrary units.

1.3.1.1 One-body Interactions

Rosetta uses several knowledge-based potentials to evaluate the "energies" of backbone and side chain torsion angles. These terms rely on converting the probabilities of finding certain torsional features from large-scale analyses of high-resolution crystal structures into Boltzmann-weighted energies.

The `rama_prepro` term is related to the probability of finding the current backbone ϕ and ψ values given the amino acid identity, and the energy term is summed over all amino acids (residues). The `p_aa_pp` term, on the other hand, is related to the

likelihood of substituting an amino acid at a given position with the given ϕ and ψ values. This energy is also summed over all residues.

Rosetta discretizes amino acid side chains into conformations with specific χ angles (rotamers), and these rotamers are evaluated based on their frequencies, means, and standard deviations found within the Dunbrack 2010 backbone-dependent rotamer library. The energy term reflecting these rotamer probabilities is called the `fa_dun` term, which again is summed over all residues.

The `omega` term is a special-case torsion energy that penalizes peptide bonds for deviating away from values of 0° or 180°. The energy of each residue's omega value is evaluated using backbone-dependent probabilities, and the entire `omega` energy term is summed over all residues. Another special-case energy, the `pro_close` term, imposes closure of the side chain ring for proline residues, which is necessary due to the way Rosetta represents chains as kinematic trees. REF15 also penalizes conformations of tyrosine rotamers that do not have a planar hydroxyl hydrogen with the `yhh_planarity` term.

1.3.1.2 Two-body Interactions

The REF15 energy function represents short-range attractive and repulsive Van der Waals (VDW) forces between neighboring nonbonded atoms as a decomposition of the Lennard-Jones 6-12 potential into two terms, `fa_atr` and `fa_rep`. The inclusion of a third VDW term, `fa_intra_rep`, which has the same functional form as `fa_rep` but a much smaller weighting factor, is necessary to better represent intraresidue clashes.

Electrostatic interactions described by the `fa_elec` term arise from a modified version of Coulomb's law, which includes several adjustments for simulations of biomolecules.

Rosetta represents the solvent surrounding biomolecules as bulk water, rather than computationally expensive explicit water molecules, using the Lazaridis-Karplus implicit Gaussian exclusion model. Two solvation terms, `fa_sol` and `lk_ball_wtd`, model solvation energies as isotropic (assumes bulk water is uniformly distributed around atoms) and anisotropic (models the orientation-dependent solvation of polar atoms), respectively. The `fa_intra_sol` term has the same functional form as `fa_sol` but is the intraresidue equivalent.

The hydrogen bonding terms `hbond_sr_bb`, `hbond_lr_bb`, `hbond_bb_sc`, and `hbond_sc`, are evaluated using `fa_elec` and an empirically fit functional form to mimic hydrogen bond orientation preferences seen in distributions of polar contacts from high-resolution crystal structures.

For disulfide bonds, the distance, angles, and torsions across the covalent bond between two cysteine Sγ atoms are treated explicitly with the `dslf_fa13` term. This energy term uses an orientation-dependent model derived from analyzing large sets of intraprotein disulfide bonds.

1.3.2 The AMBER Force Field

The Assisted Model Building with Energy Refinement (AMBER)[37] family of physics-based scoring functions (commonly referred to as force fields) has been used extensively in molecular dynamics simulations of proteins. A simplistic version of these force fields uses a generic harmonic functional form shown in **Eq. (2)**:

$$V(r^N) = \sum_{bonds} k_b(l - l_0)^2 + \sum_{angles} k_a(\theta - \theta_0)^2 + \sum_{torsions} \sum_{n} \frac{1}{2} V_n[1 + \cos(n\omega - \gamma)] + \sum_{j=1}^{N-1} \sum_{i=j=1}^{N} f_{ij} \left\{ \epsilon_{ij} \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\} \quad \textbf{(2)}$$

The **first term** is a harmonic bond-stretching force modeled after an ideal spring equation and controls the length of covalent bonds during simulations. The **second term**, the bond-angle potential or angle-bending term, which is also of the form of a spring equation, is designed to reproduce bond geometry in molecules based on the hybridization of atomic orbitals. The force constants k_b and k_a and the reference structural parameters l_0 and θ_0 are typically derived from experiment or from *ab initio* quantum mechanics (QM) calculations. The force constants are usually further optimized after they are initially parameterized using subsequent methods in order to reproduce experimental and *ab initio* vibrational frequencies[38].

The **third term** in **Eq. (2)** represents the energy for twisting a bond and is arguably the most important of the intramolecular terms in a force field, as torsion parameters play a crucial role in simulating local structures of proteins[39–41]. This energy term is modeled using a Fourier series, where V_n are the torsional rotation force constants and ω is the current torsion angle. γ is a phase angle for the torsional angle parameters, and n is the multiplicity (i.e., the number of minima in the energy *versus* angle graph from torsion angles of 0° to

360°)[40,42]. Parameter sets for this energy term are often derived from QM calculations of dipeptides and/or tetrapeptides, where small empirical adjustments can be made afterwards to improve the reproduction of experimental data[41].

The fourth term captures the non-bonded energy between all pairs of atoms that are in different molecules or in the same molecule but separated by at least three bonds. These through-space interactions do not depend on a specific bonding relationship between the atoms. The non-bonded term is commonly decomposed into the VDW component (modeled using the 6-12 Lennard-Jones equation where ϵ_{ij} is the VDW well depth, r_{0ij} is the distance where the potential is at a minimum, and r_{ij} is the distance between atoms i and j) and an electrostatic component (modeled as a Coulomb potential where ϵ_0 is the permittivity of free space, $q_{i,j}$ are atomic charges, r_{ij} is the distance between atoms i and j)[40]. Additionally, most AMBER force fields apply a special scaling factor to the interactions of atoms separated by exactly three bonds, i.e. 1-4 interactions, aimed to reduce the repulsion energy caused by the 6-12 Lennard-Jones potential and achieve better agreement between simulations and experimental data[43,44].

1.4 A Brief Overview of Key Sampling Methods

The sampling problem in CPD is of fundamental interest due to the necessity that one must be able to *find* low-energy (and therefore near-native) structure(s) in order to predict or design the desired structure or function. Methods of sampling protein conformations are therefore faced with the daunting challenges of (1) reducing the intractable degrees of freedom associated with modeling large biomolecules, (2) generating large sets or long time-based trajectories of conformations in reasonable

configurations, and (3) performing the search within a reasonable amount of time. A review of key techniques related to the content of this dissertation will be highlighted below.

1.4.1 Reducing degrees of freedom

It is infeasible to computationally sample every degree of freedom available to a given protein structure and its environment even with current computing power. Therefore, the search space must be reduced by selectively simplifying certain degrees of freedom to balance accuracy and tractability. In doing so, one hopes to restrict the overwhelmingly large sampling space in such a way that a small and enumerable set of conformations (hopefully including the native conformation or near-native) can then be ranked using a scoring function to pick the most probable final model[45]. Here, we detail methods commonly employed to reduce the degrees of freedom in structure prediction and protein design approaches.

1.4.1.1 Solvent degrees of freedom

Modelling the solvent surrounding a protein is essential for the accurate description of its molecular mechanics, as proteins have evolved to fold and function in environments containing predominantly water and other molecules[46]. While it is possible to include explicit water molecules and ions surrounding the protein in certain simulation techniques, it can become painstakingly slow to compute the motions and interactions of each solvent molecule with the protein and with each other when systems become quite large. Therefore, commonly used techniques to reduce the solvent degrees of freedom include using coarse representations of solvent molecules (pseudoatoms) or treating the solvent as a crudely modeled dielectric continuum (implicit solvent).

Subsequent chapters in this dissertation utilize the continuum representation of solvent for molecular dynamics simulations, namely the Generalized Born (GB) implicit solvent model, for its computational speedup despite its tradeoffs with accuracy. Fortunately, however, recent years have seen significant improvements in accuracy for modeling proteins and small molecules with the GB model as implemented in Amber[47,48]. Briefly, this model, which is an approximation of the more accurate but more computationally-expensive Poisson-Boltzman (PB) method, models solute atoms as a set of charged spheres whose internal dielectric constant is lower than that of the external solvent environment[49].

1.4.1.2 Protein degrees of freedom:

Coarse-graining polypeptide components: Proteins are objects with significant internal degrees of freedom as they are often comprised of long chains of linked amino acids that can easily surpass the upper threshold of a tractable number of atoms for all-atom simulations. For some CPD approaches, coarse-graining the representation of all or parts of the protein may lead to a significant computational speed-up while still allowing reasonable reconstructions of protein conformations. In these coarse-grained models, the protein backbone (or main chain; the longest chain of covalently bonded atoms) and amino acid sidechains (the chemical moieties attached to the main chain that differentiate one amino acid from another) can be reduced to pseudoatoms that model the average position of several atoms as a single sphere[50].

Constrain "stiff" degrees of freedom: The degrees of freedom associated with bond lengths and angles are considered "stiff" degrees of freedom, as the force constants required to produce fluctuations in these values are relatively large compared to the force constants for rotating about torsion angles and sampling non-bonded atom energies ("soft"

degrees of freedom)[51]. Thus, it is common to fix or highly constrain the covalent bond lengths and covalent bond angles during conformational sampling using statistics to exclude conformations that have values not seen in most proteins.

Flexible *versus* rigid template backbone: Proteins commonly exhibit varying degrees of backbone flexibility as suggested by experimental methods such as NMR, and therefore a higher resolution computational design approach may incorporate alternative backbone configurations within the same procedure. Most current design methods, however, use the backbone information from a single protein structure as the design template, which, although may rest on a highly questionable assumption[52], greatly allows for a reduction in the degrees of freedom during sampling protocols.

1.4.1.3 Eliminating subsets of search space with experimental constraints

In some protein design and structure prediction situations, sparse experimental data or *a priori* knowledge may be available regarding aspects of the final target structure. These data may be incorporated into conformation and sequence sampling methods via constraints to confine the search space to include only structures or sequences that are consistent with the experimental observations. Specific examples of successful integrative modeling approaches include the incorporation of experimental data from a wide array of methods, including nuclear magnetic resonance (NMR) techniques[53,54], chemical modification footprinting with mass spectrometry[55,56], Forster Resonance Energy Transfer (FRET) experiments[57], small-angle X-ray scattering (SAXS)[58], studies of catalytic mechanisms[18], sequence evolution comparisons[59], or a complex combination of many methods[60].

1.4.2 Methods of sampling: how to generate sets of decoys and/or trajectories of conformational pathways

Generating protein conformations that will be evaluated by the scoring function is fundamental to protein design and structure prediction. Here, we review the two most widely used methods for ensemble generation and conformational sampling: molecular dynamics (MD) as used by the Amber[37] package of programs, and the Metropolis Monte Carlo (MC) algorithm as used by Rosetta.

1.4.2.1 Molecular Dynamics

Molecular dynamics is a method that simulates the motions of proteins using equations from classical dynamics and is a particularly useful technique for protein structure prediction[48,61,62] or refining low-resolution template-based structure predictions[63]. Although Amber MD simulations do not directly include methods for protein design, the dynamic molecular interactions sampled by MD simulations can be used to guide protein design efforts in many ways, such as by revealing regions of possible instability, identifying positions of functional residues, or providing models of transition state conformations along a folding pathway that may be useful targets for protein design[64].

In molecular dynamics simulations, new conformations of the system are generated by repeatedly applying Newton's equations of motion to all atoms simultaneously. Calculating the forces acting on the position and velocity of individual atoms over a small time-step allows for the step-by-step determination of the new atomic positions and velocities of the next conformation. Repeated iterations of this process eventually generate a time-based trajectory of continuous protein motions.[65]

Molecular dynamics simulations offer an incredibly detailed approach to studying the time-dependent motions of biological elements at the atomic and molecular level. This not only heightens our understanding of important processes taking place dynamically but can also serve as a bridge between protein structures and macroscopic kinetic data[66]. Drawbacks of MD include the potential inability to sample *many* conformations during a simulation due to large energetic barriers that must be crossed, and the inability to model protein motions on *large* timescales due to the computational complexity of the system.

1.4.2.2 Metropolis Monte Carlo Algorithm

The Metropolis Monte Carlo algorithm is a stochastic search method often used in protein design. At its core, this algorithm applies random perturbations to the degrees of freedom in a system, compares the energy of the new conformation to the previous conformation, and accepts the transition if the energy of the new conformation is lower (higher-energy transitions are accepted with certain probabilities).[67]

Because MC simulations do not rely on solving Newton's equations of motion, one advantage of MC approaches is that developers of sampling protocols are free to use clever tools to generate trial configurations with significantly higher computational efficiency than MD simulations. One disadvantage of this method, however, is the inability to gather information on protein dynamics as there is no time dimension between steps, which precludes the ability of using this method for studying time-dependent phenomena.[68]

Many protocols in Rosetta use a version of the Metropolis Monte Carlo algorithm to sample the conformational landscape of an input system. In these protocols, MC offers a computationally inexpensive method that works well with Rosetta's treatment of the solvent as bulk water, which allows for large-scale movements of the protein without

leading to energetically unfavorable overlaps of protein and solvent atoms.[69] For brevity, we describe only a few Rosetta sampling protocols below that are pertinent to the remaining chapters in this dissertation.

1.4.2.3 Sampling Backbone Degrees of Freedom

Sampling alternative protein backbone conformations is often necessary to capture the essential fluctuations around a "native" structure separated by small energy differences that can give rise to a variety of protein functions[70]. Although many protocols exist in Rosetta for sampling backbone degrees of freedom (backrub[71], fragment insertion[72,73], loop closure with cyclic coordinate descent[74]), we will focus on a particular loop closure algorithm used in the following chapters. The kinematic loop closure (KIC) protocol[75,76] and its generalized implementation (GenKIC[17]) in Rosetta analytically determines all accessible conformations of a chain of atoms, subject to certain constraints. Briefly, this protocol requires the user to define three pivot points, which are the start, middle, and ending atoms of the covalently-contiguous chain, and the degrees of freedom that will be allowed to be perturbed during the simulation (such as dihedral angles, bond angles, and bond lengths). After these perturbations have been made to the non-pivot atoms, the two resulting segments have different internal conformations but retain well-defined rigid-body transforms between the pivot points. From here, it is possible to apply inexpensive matrix math to solve for the torsion angles adjacent to each pivot atom that enforce the chain of atoms to be closed with ideal bond geometry. The GenKIC and KIC protocols are often wrapped into an MC protocol that includes a minimization step for the reconstructed full-length segment.

1.4.2.4 Sampling Side Chain Degrees of Freedom

Sidechain conformation prediction is a common problem in protein structure modeling that attempts to select a set of rotamers for a given set of amino acid positions that minimize the total energy of the system. In Rosetta, the primary algorithm for solving for the lowest-energy solution of rotamer combinations is called the *packer*, which can be invoked through a number of protocols. The *packer* uses MC to test random replacements of current rotamers at particular positions with randomly-selected rotamers from a set of allowed rotamers. Two protocols used extensively in this dissertation that rely on the *packer* for energy minimization are the RepackMinimize and FastRelax[77] protocols. The former repacks amino acid sidechains, carries out sidechain minimization, and uses MC to evaluate the change. Here, in the minimization step, a vector is chosen as the descent direction (usually the gradient of the energy function), a step is taken along that vector, and then a new direction is chosen. This march is usually repeated until some level of convergence is found. In similar fashion, the FastRelax protocol applies the *packer* and performs minimization, but does so in iterative cycles of the two while ramping the repulsive weight of the forcefield up and down.

1.4.2.5 Sampling Ligand-Protein Interactions

Modeling the interactions of proteins with ligands can give fundamental insights helpful to the fields of drug discovery and development. Besides the MC-based docking protocols within Rosetta, which use rigid-body perturbations and multi-resolution optimization techniques to find the optimal interaction orientation/conformation between a protein and a ligand[78–80], the Rosetta matching algorithm[81,82] is another tool used extensively in this dissertation that enables the

modeling of proteins with covalently-bound small molecules. Briefly, this algorithm takes user-defined constraints that describe the ideal geometry of the protein-ligand interaction, a user-defined set of positions in the protein from which those constraints may be satisfied, and the input protein and small molecule models. The RosettaMatch protocol then searches through the scaffold protein for positions where a full and complete set of constraints can be realized, within some tolerance. By default, every available position for a given constraint is used to generate "hits", which are models where the desired sidechain is built from that position in the scaffold protein, as well as all allowed sidechain-ligand geometries specified in the constraint file. After all constraints/positions have been built, an enumeration step takes places where combinations of "hits" that place the ligand in the same (+/- some tolerance factor) conformation and/or orientation with respect to the protein are combined into a single output model. The RosettaMatch method can become both time- and memory-intensive as the constraints and ligand rotamers become more finely sampled, but offers the benefit of a completely generalizable method for protein-ligand interaction modeling.

1.5 References

1. Berg, J. M., Tymoczko, J. L. & Stryer, L. Biochemistry. *W H Free.* (2002).

2. Fersht, A. *Structure and mechanism in protein science: A guide to enzyme catalysis and protein folding. Lavoisier.Fr* (1999). doi:10.1016/S0307-4412(99)00114-4

3. Dill, K. a & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **338,** 1042–1046 (2012).

4. Dobson, C. M., Sali, A. & Karplus, M. Protein folding: A perspective from theory and experiment. *Angew. Chemie - Int. Ed.* (1998). doi:10.1002/(SICI)1521-3773(19980420)37:7<868::AID-ANIE868>3.0.CO;2-H

5. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science (80-.).* **181,** 223–230 (1973).

6. HABER, E. & ANFINSEN, C. B. Side-chain interactions governing the pairing of half-cystine residues in ribonuclease. *J. Biol. Chem.* (1962).

7. Levinthal, C. How to fold graciously. *Mössbauer Spectrosc. Biol. Syst. Proc.* (1969). doi:citeulike-article-id:380320

8. Leopold, P. E., Montal, M. & Onuchic, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci.* (1992). doi:10.1073/pnas.89.18.8721

9. Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. Navigating the folding routes. *Science* **267,** 1619–1620 (1995).

10. Dill, K. A. & Chan, H. S. From levinthal to pathways to funnels. *Nature Structural Biology* (1997). doi:10.1038/nsb0197-10

11. Nolting, B. Protein folding kinetics: biophysical methods. in *Quarterly reviews of biophysics* (Springer, 2006). doi:10.1017/S003358351400002X

12. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky092

13. Dorn, M., E Silva, M. B., Buriol, L. S. & Lamb, L. C. Three-dimensional protein structure prediction: Methods and computational strategies. *Computational Biology and Chemistry* (2014). doi:10.1016/j.compbiolchem.2014.10.001

14. Khoury, G. A., Smadbeck, J., Kieslich, C. A. & Floudas, C. A. Protein folding and de novo protein design for biotechnological applications. *Trends in Biotechnology* (2014). doi:10.1016/j.tibtech.2013.10.008

15. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science (80-.).* **302,** 1364–1369 (2003).

16. King, I. C. *et al.* Precise assembly of complex beta sheet topologies from de novo designed building blocks. *Elife* **4,** 3–11 (2015).

17. Bhardwaj, G., Mulligan, V. K., Bahl, C. D. & Gilmore, J. M. Accurate de novo design of hyperstable constrained peptides. *Nature* **538,** 329–335 (2016).

18. Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science* **319,** 1387–1391 (2008).

19. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453,** 190–195 (2008).

20. Khare, S. D. *et al.* Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nat. Chem. Biol.* (2012). doi:10.1038/nchembio.777

21. Siegel, J. B. *et al.* Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* **329,** 309–313 (2010).

22. Dwyer, M. A., Looger, L. L. & Hellinga, H. W. Computational design of a biologically active enzyme. *Science* (2004). doi:10.1126/science.1098432

23. Tinberg, C. E. *et al.* Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* (2013). doi:10.1038/nature12443

24. Yang, W. & Lai, L. Computational design of ligand-binding proteins. *Current Opinion in Structural Biology* (2017). doi:10.1016/j.sbi.2016.11.021

25. Stoddard, B. L. Computational Design of Ligand Binding Proteins. *Methods Mol. Biol.* (2016). doi:10.1007/978-1-4939-3569-7

26. Grigoryan, G. *et al.* Computational design of virus-like protein assemblies on carbon nanotube surfaces. *Science (80-.).* (2011). doi:10.1126/science.1198841

27. Bale, J. B. *et al.* Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science (80-.).* (2016). doi:10.1126/science.aaf8818

28. Padilla, J. E., Colovos, C. & Yeates, T. O. Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proc. Natl. Acad. Sci.* (2001). doi:10.1073/pnas.041614998

29. Fletcher, J. M. *et al.* Self-assembling cages from coiled-coil peptide modules. *Science (80-.).* (2013). doi:10.1126/science.1233936

30. King, N. P. *et al.* Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science (80-.).* (2012). doi:10.1126/science.1219364

31. Moore, E. J., Zorine, D., Hansen, W. A., Khare, S. D. & Fasan, R. Enzyme stabilization via computationally guided protein stapling. *Proc. Natl. Acad. Sci.* (2017). doi:10.1073/pnas.1708907114

32. Bednar, D. *et al.* FireProt: Energy- and Evolution-Based Computational Design of Thermostable Multiple-Point Mutants. *PLoS Comput. Biol.* (2015). doi:10.1371/journal.pcbi.1004556

33. Borgo, B. & Havranek, J. J. Automated selection of stabilizing mutations in designed and natural proteins. *Proceedings of the National Academy of Sciences* **109,** 1494–1499 (2012).

34. Korkegian, A., Black, M. E., Baker, D. & Stoddard, B. L. Computational thermostabilization of an enzyme. *Science* **308,** 857–860 (2005).

35. Liu, J. & Wang, R. Classification of current scoring functions. *J. Chem. Inf. Model.* (2015). doi:10.1021/ci500731a

36. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13,** 3031–3048 (2017).

37. D.A. Case and R.M. Betz and W. Botello-Smith and D.S. Cerutti and T.E. Cheatham and III and T.A. Darden and R.E. Duke and T.J. Giese and H. Gohlke and A.W. Goetz

and N. Homeyer and S. Izadi and P. Janowski and J. Kaus and A. Kovalenko and T.S. Lee and S. AMBER 2018. *University of California, San Francisco* (2018).

38. Wang, J. M., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25,** 1157–1174 (2004).

39. Zgarbová, M. *et al.* Refinement of the Cornell et al. Nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.* (2011). doi:10.1021/ct200162x

40. Cheung, D. L. G. Structures and Properties of Liquid Crystals and Related Molecules from Computer Simulation. (University of Durham, 2002).

41. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11,** 3696–3713 (2015).

42. Leach, A. R. *Molecular Modelling: Principles and Applications*. (Pearson, 2001).

43. Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117,** 5179–5197 (1995).

44. Pang, Y. P. Use of 1-4 interaction scaling factors to control the conformational equilibrium between α-helix and β-strand. *Biochem. Biophys. Res. Commun.* (2015). doi:10.1016/j.bbrc.2014.12.084

45. Keasar, C. & Levitt, M. A novel approach to decoy set generation: Designing a physical energy function having local minima with native structure characteristics. *J. Mol. Biol.* (2003). doi:10.1016/S0022-2836(03)00323-1

46. Kleinjung, J. & Fraternali, F. Design and application of implicit solvent models in biomolecular simulations. *Current Opinion in Structural Biology* (2014). doi:10.1016/j.sbi.2014.04.003

47. Nguyen, H., Roe, D. R. & Simmerling, C. Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* **9,** 2020–2034 (2013).

48. Nguyen, H., Maier, J., Huang, H., Perrone, V. & Simmerling, C. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J. Am. Chem. Soc.* **136,** 13959–13962 (2014).

49. Onufriev, A. The generalized Born model: its foundation, applications, and limitations. *Science (80-.).* (2010). doi:10.1002/9783527629251.ch6

50. Kmiecik, S. *et al.* Coarse-Grained Protein Models and Their Applications. *Chemical Reviews* (2016). doi:10.1021/acs.chemrev.6b00163

51. Yakubovich, A. *Theory of Phase Transitions in Polypeptides and Proteins*. (Springer-Verlag Berlin Heidelberg, 2011). doi:10.1007/978-3-642-22592-5

52. Fung, H. K., Welsh, W. J. & Floudas, C. A. Computational de novo peptide and protein design: Rigid templates versus flexible templates. *Industrial and Engineering*

Chemistry Research (2008). doi:10.1021/ie071286k

53. Bowers, P. M., Strauss, C. E. & Baker, D. De novo protein structure determination using sparse NMR data. *J. Biomol. NMR* (2000). doi:10.1023/A:1026744431105

54. Rohl, C. A. & Baker, D. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc.* (2002). doi:10.1021/ja016880e

55. Mathews, D. H. *et al.* Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci.* (2004). doi:10.1073/pnas.0401799101

56. Aprahamian, M. L., Chea, E. E., Jones, L. M. & Lindert, S. Rosetta Protein Structure Prediction from Hydroxyl Radical Protein Footprinting Mass Spectrometry Data. *Anal. Chem.* (2018). doi:10.1021/acs.analchem.8b01624

57. Dimura, M. *et al.* Quantitative FRET studies and integrative modeling unravel the structure and dynamics of biomolecular systems. *Current Opinion in Structural Biology* (2016). doi:10.1016/j.sbi.2016.11.012

58. Schindler, C. E. M., de Vries, S. J., Sasse, A. & Zacharias, M. SAXS Data Alone can Generate High-Quality Models of Protein-Protein Complexes. *Structure* (2016). doi:10.1016/j.str.2016.06.007

59. Mitra, P. *et al.* An Evolution-Based Approach to De Novo Protein Design and Case Study on Mycobacterium tuberculosis. *PLoS Comput. Biol.* (2013). doi:10.1371/journal.pcbi.1003298

60. Andre, I. Modeling the Structure of Helical Assemblies with Experimental Constraints in Rosetta. in *Protein Complex Assembly* 475–489 (Humana Press, 2018). doi:https://doi.org/10.1007/978-1-4939-7759-8

61. Kato, K., Nakayoshi, T., Fukuyoshi, S., Kurimoto, E. & Oda, A. Validation of molecular dynamics simulations for prediction of three-dimensional structures of small proteins. *Molecules* (2017). doi:10.3390/molecules22101716

62. Shaw, D. E. *et al.* Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* (2008). doi:10.1145/1364782.1364802

63. Jagielska, A., Wroblewska, L. & Skolnick, J. Protein model refinement using an optimized physics-based all-atom force field. *Proc. Natl. Acad. Sci. U. S. A.* (2008). doi:10.1073/pnas.0800054105

64. Childers, M. C. & Daggett, V. Insights from molecular dynamics simulations for computational protein design. *Mol. Syst. Des. Eng.* (2017). doi:10.1039/C6ME00083E

65. Jorgensen, W. L. & Tirado-Rives, J. Monte Carlo vs molecular dynamics for conformational sampling. *J. Phys. Chem.* (1996). doi:10.1021/jp960880x

66. Vlachakis, D., Bencurova, E., Papangelopoulos, N. & Kossida, S. Current state-of-

the-art molecular dynamics methods and applications. *Adv. Protein Chem. Struct. Biol.* (2014). doi:10.1016/B978-0-12-800168-4.00007-X

67. Paquet, E. & Viktor, H. L. Molecular dynamics, monte carlo simulations, and langevin dynamics: A computational review. *BioMed Research International* (2015). doi:10.1155/2015/183918

68. Earl, D. J. & Deem, M. W. Monte Carlo Simulations. in *Molecular Modeling of Proteins* (ed. Andreas, K.) 25–36 (Humana Press, 2008). doi:10.1007/978-1-59745-177-2_2

69. Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H. & Meiler, J. Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry* (2010). doi:10.1021/bi902153g

70. Ollikainen, N., Smith, C. A., Fraser, J. S. & Kortemme, T. Flexible backbone sampling methods to model and design protein alternative conformations. *Methods Enzymol.* **18,** 1199–1216 (2013).

71. Smith, C. A. & Kortemme, T. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. *J. Mol. Biol.* (2008). doi:10.1016/j.jmb.2008.05.023

72. 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.* **268,** 209–25 (1997).

73. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein Structure Prediction Using Rosetta. *Methods in Enzymology* (2004). doi:10.1016/S0076-6879(04)83004-0

74. Canutescu, A. A. & Dunbrack, R. L. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.* (2003). doi:10.1110/ps.0242703

75. Coutsias, E. A., Seok, C., Jacobson, M. P. & Dill, K. A. A kinematic view of loop closure. *J. Comput. Chem.* **25,** 510–528 (2004).

76. Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6,** 551–552 (2009).

77. Tyka, M. D. *et al.* Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405,** 607–618 (2011).

78. Gray, J. J. *et al.* Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* **331,** 281–299 (2003).

79. Wang, C., Bradley, P. & Baker, D. Protein-Protein Docking with Backbone Flexibility. *J. Mol. Biol.* **373,** 503–519 (2007).

80. Chaudhury, S. *et al.* Benchmarking and analysis of protein docking performance in

Rosetta v3.2. *PLoS One* **6,** (2011).

81. Zanghellini, A. *et al.* New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* **15,** 2785–2794 (2006).

82. Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S. & Baker, D. De novo enzyme design using Rosetta3. *PLoS One* **6,** (2011).

Chapter 2: A Computational Method for the Design of Nested Proteins by Loop-Directed Domain Insertion

2.1 Preface

A version of Chapter 2 has been published [Blacklock, K. M., Yang, L., Mulligan, V. K. & Khare, S. D. A Computational Method for the Design of Nested Proteins by Loop-Directed Domain Insertion. *Proteins Struct. Funct. Bioinforma.* (2017). doi:10.1002/prot.25445]. I was the lead investigator, responsible for all major areas of concept formation, data collection and analysis, as well as manuscript composition. Lu Yang and Vikram Mulligan contributed to early stages of concept formation and manuscript edits. Sagar Khare was the supervisory author on this project and was involved throughout the project in concept formation and manuscript composition.

2.2 Abstract

The computational design of novel nested proteins – in which the primary structure of one protein domain (insert) is flanked by the primary structure segments of another (parent) – would enable the generation of multifunctional proteins. Here we present a new algorithm, called Loop-Directed Domain Insertion (LooDo), implemented within the Rosetta software suite, for the purpose of designing nested protein domain combinations connected by flexible linker regions. Conformational space for the insert domain is sampled using large libraries of linker fragments for linker-to-parent domain superimposition followed by insert-to-linker superimposition. The relative positioning of the two domains (treated as rigid bodies) is sampled efficiently by a grid-based, mutual placement compatibility search. The conformations of the loop residues, and the identities

of loop as well as interface residues, are simultaneously optimized using a generalized kinematic loop closure algorithm and Rosetta EnzymeDesign, respectively, to minimize interface energy. The algorithm was found to consistently sample near-native conformations and interface sequences for a benchmark set of structurally similar but functionally divergent domain-inserted enzymes from the α/β hydrolase superfamily, and discriminates well between native and nonnative conformations and sequences, although loop conformations tended to deviate from the native conformations. Furthermore, in cross-domain placement tests, native insert-parent domain combinations were ranked as the best-scoring structures compared to nonnative domain combinations. This algorithm should be broadly applicable to the design of multi-domain protein complexes with any combination of inserted or tandem domain connections.

2.3 Introduction

The majority of prokaryotic and eukaryotic proteins consist of multiple domains combined into a singly connected complex[1]. This trend, driven largely by evolutionary pressures to increase the diversity of molecular functions[2–4], stems from genomic duplication and recombination events that lead to the introduction of large jumps in protein sequence space, which in turn leave the component domains susceptible to radical alterations in topology and function[2]. A particular topological characteristic that has become a prominent feature in a significant number of multi-domain proteins is the utilization of the interface between two or more domains to form binding and/or active sites, such as in the periplasmic binding proteins[5], amino acid dehydrogenases[6], and serine proteases[7].

While most multi-domain proteins are formed by connecting domains through end-to-end linkages, there also exists a non-trivial subset (~10% of multi-domain proteins)

featuring a nested topology in which at least one domain is inserted into another[1]. Similar to the case of end-to-end multi-domain proteins, inserted domains provide tunable functionality to a protein complex by creating domain-domain interfaces capable of developing novel activities and altered substrate specificities. For example, many members from each of the haloacid dehalogenase family of proteins, the A, B, and Y superfamilies of DNA polymerases, and many subclasses of the Flavin-dependent monooxygenases contain at least one inserted domain, which has been shown to both greatly diversify the available substrate space for the parent domain and also to ensure tighter interactions with the substrate[8–10].

Insert-domain proteins differ most notably from their end-to-end counterparts in that more sophisticated functions are made possible through the extensive structural coupling that occurs when one domain is inserted into another. Additionally, insert-domain proteins can also be comprised of nested insertions, two-domain insertions, and three-domain insertions, which can lead to further complexification of domain relationships and functions[1]. Other advantages of this natural strategy for protein architecture elaboration are the formation of more rigid structures with higher stabilities, a decrease in susceptibility to proteolytic degradation in *vivo* [11], and facilitated cooperative folding events[12]. The ability to computationally design proteins with these inherent qualities and a desired intricate functionality at the domain-domain interface, then, would allow for the rapid development of robust biosensors[13], bioswitches[14], bifunctional proteins, "affinity clamps"[15], and reporters of parent domain structure[11]. Notably, an engineered insertable protein (UniRapR) designed by Dokholyan and co-workers has been used for precise spatial and temporal control over the activities of signaling cascades *in vivo*, by insertion into accurately predicted allosteric sites[16,17].

Although some approaches have been developed towards predicting the structure of insert-domain proteins[18], the computational design of such enzyme architectures with active sites located at domain-domain interfaces has not been reported. Compared to the widespread use of domain insertion and recombination for generating new functions both naturally[1,8–10], and synthetically[11,13,14,16,19–24], computational design of protein function has largely been limited to making point mutations with a fixed backbone, using limited sampling of backbone degrees of freedom[25], making small changes in loop conformations[26], or designing structures *de novo* [27–29].

A large-scale survey of fusion events over 30 complete genomes has revealed that the α/β fold participates disproportionately in domain fusions[30]. Since the construction of multi-domain enzymes will rely in part on the compatibility of the component domains, we chose to focus our study on an insert-domain protein family whose parent domain has evolutionarily held a prominent role in accepting insertions[1]. The alpha/beta hydrolase fold, specifically, is one of the largest protein superfamilies in the α/β class of folds, and consists of a diverse group of structurally similar proteins with strikingly widespread functions, which has motivated many protein engineering efforts[31–34]. This modularity arises from easily tolerated insertions, which serve to confer unique activities to enzymes in this class *via* novel active sites at the domain-domain interface[31].

Here, we conceptually decompose the problem of designing insert-domain proteins from existing domains into the two processes: loop design and protein interface design. Both have individually been approached successfully by a number of studies[35–37]. We have developed a computational protein engineering approach that combines the two, consisting of a method for the elaboration of existing protein domains containing functional active sites by introducing new ligand-specific binding/catalytic groups at the domain-

domain interface, while using variable linker segments as guiding connectors for the inserted domain. In this study, we describe an *in silico* benchmark in which we successfully recapitulate the sequences and domain placements of α/β hydrolase insert-domain protein complexes, as well as domain placements for a previously described inserted-domain benchmark (Gray set[18]).

2.4 Materials and Methods

2.4.1 Generation of input structures

A benchmark set of five two-domain protein structures belonging to the α/β hydrolase superfamily was selected from the Protein Data Bank (**Figure 2-1**), and all non-ligand hetero-atoms and water entries were removed prior to analysis. Two eight-residue linker regions separating the insert domain and the parent domain were determined by visual inspection and the four inner residues of each linker region were removed from the structures via PyMOL, thus separating the proteins into insert domains and parent domains. We defined the two residues from each linker region remaining on the termini of the insert domain structure as Insert Attachment Site A and B, where A corresponds to the N-terminal linker region on the insert domain, and B corresponds to the C-terminal linker region on the insert domain. We also defined the two residues from each linker region remaining in the parent domain structure as Parent Insertion Site A and B, where A corresponds to the N-terminal linker region on the parent domain, and B corresponds to the C-terminal linker region on the parent domain. Each parent domain and insert domain was then separately subjected to full-atom relaxation via the RosettaScripts[38] FastRelax[39] protocol in which side-chain coordinate constraints were imposed on each the starting structures. A placement grid was also generated from the native conformation of the insert domain with respect to the parent domain, which consists of a set of voxels corresponding

to the space occupied by the insert domain. The placement grid could be made arbitrarily large, and the results are generally insensitive to the size and shape of the placement grid as long as it was sufficiently large.

2.4.2 LooDo algorithm

The LooDo algorithm (**Figure 2-1:A**) employs an iterative search of rigid-body placements between a parent domain and insert domain to define pairs of linkers that are capable of producing insert domain placements close to one another in three-dimensional space (six-dimensional rotation-translation parameter space).

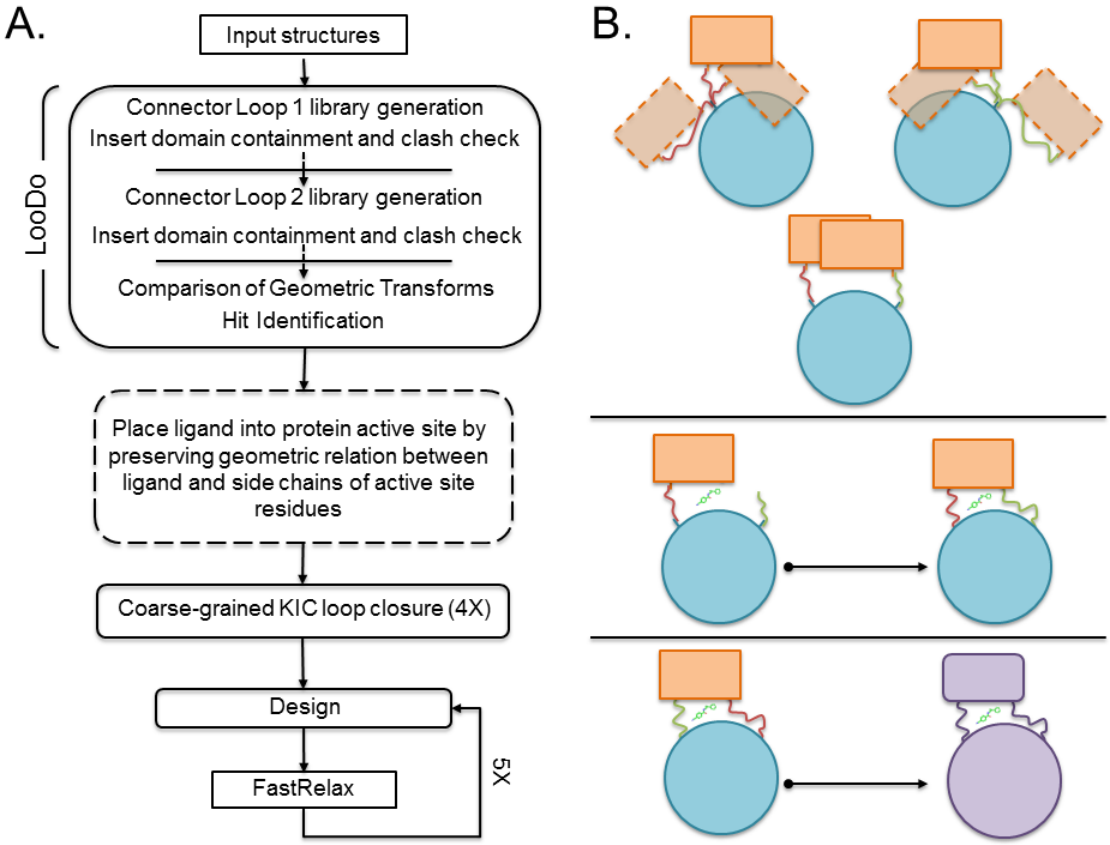


Figure 2-1: Flowchart and cartoon diagram of the design pipeline, including the LooDo algorithm.The input structures are the insert domain, parent domain with insertion site, and placement grid file. The algorithm culls two libraries of fragments of user-defined length from an internal Rosetta database of idealized proteins, and uses these fragments to find placements of the insert domain relative to the parent domain. Placements from both libraries are filtered on two criteria that require the domain to occupy a defined amount of space in

the active site grid and filter any clashing placements. Placements from each library are compared against one another, and any two that are close to one another in space are defined as hits. Optionally, the user can choose to further filter insert domain placements through by requiring the insert domain to be able to interact with the native ligand through native-like geometries. KIC is used to close all four breaks in the peptide chain, followed by five iterations of sequence and structure optimization. (B) Top: Libraries of fragments (red and green lines) are used to place the insert domain (orange rectangle) relative to the parent domain (blue circle) through terminal residue superimposition. Shaded insert domains do not pass either the containment or clash filters required by LooDo. If two insert domain placements are similar to one another, both are considered Hits. Middle: The native ligand-enzyme residue geometry is optionally imposed on all decoys (green stick representation), and the peptide chain is closed using KIC at each end of each fragment. Bottom: Decoys undergo five iterations of the FastRelax and Design protocols in the RosettaScripts framework to find the lowest energy conformation of the decoy (purple complex).

To perform this search, two libraries of eight-residue peptide fragments are culled from an internal database of idealized proteins (Vall) within Rosetta[40], where the amino acid sequences and secondary structure sequences of the four terminal residues (two on either end of the fragment) are biased toward the amino acid and secondary structure identities of the two residues in the Parent Insertion Site and Insert Attachment Site.

Each fragment in the library generated by Parent Insertion Site A and Insert Attachment Site A, termed Linker A, is systematically superimposed by two N-terminal residues onto Parent Insertion Site A followed by superimposition of the insert domain by Insert Attachment Site A onto the two C-terminal residues of the library fragment, thus yielding a placement of the insert domain with respect to the parent domain completely directed by Linker A, termed Insert A (**Figure 2-1:B, top panel**). The same procedure is used for the generation of insert domain placements dictated by library fragments generated by Parent Insertion Site B and Insert Attachment Site B (Linker B), to yield a set of Insert B configurations.

Mutually compatible placements, termed Hits, are identified from pairs of insert placements, one superimposed onto Linker A (termed Hit A), the other onto linker B (termed Hit B), that are close to one another and have similar orientation with respect to the parent domain (**Figure 2-1:B, top panel**). Exploiting a rotational/translational description of the insert domain with respect to the parent domain as a hash function to map insert

domain placements to six-dimensional space allows for a more efficient computational comparison between insert domains, which is necessary when dealing with large libraries of domain placements[41].

During the LooDo search, placement Hits generated by the above strategy also must pass three criteria to be considered for further analysis: (1) the residue closest to the center of mass of the insert domain must lie in a user-defined placement grid, (2) the insert domain must not occupy any space in the grid of voxels surrounding the parent domain, and (3) the interface between the insert domain and parent domain must be able to coordinate a ligand molecule at the interface in a native-like geometry. The first of these criteria serves as a filter to decrease the amount of passing decoys by focusing the rigid-body sampling to a smaller, pre-defined region around the parent domain where the insert domain is most likely to be found. The last two of these criteria allow for the simultaneous design of function as well as structure as discussed below.

After both sets of insert domain placements are generated, every Insert B is compared to every Insert A and pairs that are deemed sufficiently close to one another in rotational and translational relation to the parent domain are identified as Insert Pairs (**Figure 2-1:B, top panel**). Both Insertion Attachment Sites in each of the two Insert Pairs are removed, as well as each Parent Insertion Site in the parent domain. Each insert domain conformation in the Insert Pair was then assembled with both the parent domain segments and both Linkers such that the primary structure of the resulting decoy was complete and in the sequential order of parent domain segment 1, Linker A, Insert A or B, Linker B, and parent domain segment 2 (**Figure 2-1:B, middle panel**).

2.4.3 Active site placement using the RosettaMatch algorithm

In order to reduce the number of decoys after the LooDo placement algorithm and to focus further analysis on sets of decoys capable of performing the catalytic activity of the native parent/insert domain combination and conformation, the RosettaMatch algorithm was used to assess each decoy for the ability to bind the native parent domain ligand. In brief, the matcher searches a protein backbone for residue mutations that can accommodate all of the geometrical constraints specific to pre-defined ligand-binding geometry without clashing with the protein backbone or ligand sidechains. After generation of the constraint definitions via visual inspection of the native protein-ligand binding interactions, all residue positions in the parent and insert domains are systematically searched as potential participants of a constellation of backbone positions in which a correct and complete ligand-binding interaction can be realized. The CST energy further discussed here is a measure of how similar the ligand-binding geometry is to the native ligand-binding geometry.

2.4.4 Generalized Kinematic Loop closure

After LooDo-based placement, the assembled decoy protein structures of each Insert Pair undergo iterative loop closure cycles of a generalized kinematic loop closure (KIC)-based algorithm[36,42,43], which is a computationally inexpensive calculation that idealizes the peptide bonds at the four chain breakpoints (two at each end of each linker region introduced into the complex upon assembly) and also serves to increase the diversity of backbone conformations in the loop regions (**Figure 2-1:B, middle panel**). First, quantification of the breakpoints is performed in terms of the spatial separation of the upper and lower residues occurring before and after the protein chain break. Closure is attempted first at the site at which the break is evaluated to be the largest, while rigid body

movements of the insert domain with respect to the parent domain are allowed during minimization steps. The remaining chainbreak points are closed without rigid body movements of the insert domain.

Picking fragments in a backbone-independent fashion from a library of idealized proteins allows for a reduction in the amount of loop sampling necessary while also allowing for a high level of loop diversity. In a structure prediction benchmark set by Gray and co-workers, where the sequences of loops are known, docking the two domains and performing loop modeling has been shown to be successful in recapitulating domain insertion proteins[18]. However, in the more general case when the sequence is not known *a priori*, the possible loop conformational space is too large to sample in a computationally tractable amount of time. Therefore, we have utilized fragments from a library of plausible loop conformations to place the insert domain initially, and through the use of KIC along with structural minimization, we are able increase loop variability within the new environment of the insert and parent domains.

2.4.5 Sequence and structure optimization

The assembled and closed structures were subjected to five cycles of sequence and structure optimization, which entailed amino acid identity and rotamer sampling at each of the protein-ligand interface region, the loop regions, and the loop-neighboring regions, while catalytic residues in the parent and insert domains were kept fixed, followed by structural refinement using the Rosetta FastRelax application within the RosettaScripts framework (**Figure 2-1:B, bottom panel**). Decoys were scored using the talaris2013 score function with a constraints weight of 1.0.

2.5 Results

2.5.1 Loop-directed Domain Insertion Algorithm

In order to test the LooDo algorithm (**Figure 2-1**) on native insert-domain protein structures, five crystal structures of two-domain protein complexes belonging to the α/β hydrolase superfamily and bound to ligands were obtained from the Protein Data Bank[44] and used as the benchmark set (**Table 2-1**). Each PDB-derived structure was separated by its insert domain, two eight-residue linker regions, and noncontiguous parent domain in which two insertion regions were defined as the two residues on either side of the insertion site (Parent Insertion Site A and Parent Insertion Site B) (**Figure 2-2:A-E**). The five parent domains are very similar to one another in secondary structure topology (**Figure 2-2:F**) with only minor differences occurring in 3ANS, 3B12, and 3GZJ that do not affect the overall fold of the parent domain, as seen in the low Cα -RMSDs between each parent domain and the 1WM1 parent domain (**Table 2-2**). Although the general structure of each parent domain is similar, the insertion sites in each parent domain vary greatly in end-to-end distance and solvent accessibility (**Table 2-2**). The insert domains, however, are very different from one another, featuring large size disparities (**Table 2-1**), considerably different secondary structure topologies (**Figure 2-2A-E**), and widely varying insertion site end-to-end distances and solvent accessibilities (**Table 2-2**).

Table 2-1: Benchmark set of two-domain protein complexes.

PDB Accession Code	Resolution (Å)	Enzyme Name	Parent Domain	Insert Domain	Catalytic Residues
1WM1	2.1	*Serratia marcescens* prolyl aminopeptidase	4 - 135 247 - 316	144 - 238	S113, H296, D268, E232, E204
2WUF	1.9	*Mycobacterium tuberculosis* 2-hydroxy-6-oxo-6-phenylhexa-2,4-dienoate hydrolase	7 - 144 219 - 288	153 - 210	W270, N113, N54, R192
3ANS	1.98	*Homo sapiens* epoxide hydrolase	232 - 363 479 - 545	372 - 470	D104, H524, D496, Y466, Y383
3B12	1.2	*Burkholderia* sp. FA1 fluoroacetate dehalogenase	1 - 130 235 - 294	139 - 226	D104, H271, D128, W150, Y149
3GZJ	2.19	*Rauvolfia serpentina* polyneuridine-aldehyde esterase	10 - 112 202 - 263	121 - 193	S87, H244*, D216, Y123

[a]Indicates residue mutated back to WT.

Table 2-2: Benchmark set of two-domain protein complexes, in more detail.

PDB Accession Code	Resolution (A)	Enzyme Name	Parent Domains					Insert Domains				Catalytic Residues	Number of Residues at Parent/Insert Interface
			Residue Delineation	Total Number of Residues	Ca-RMSD to 1WUF Parent Domain (A)	Insertion Site Residue Ca Distance (A)	Summed SASA of Insertion Site Residues (A2)	Residue Delineation	Total Number of Residues	Insertion Site Residue Ca Distance (A)	Summed SASA of Insertion Site Residues (A2)		
1WM1	2.1	*Serratia marcescens* prolyl aminopeptidase	4 - 135 247 - 316	202	0	18.4	216.831	144 - 238	95	8.9	277.914	S113, H296, D268, E232, E204	47
2WUF	1.9	*Mycobacterium tuberculosis* 2-hydroxy-6-oxo-6-phenylhexa-2,4-dienoate hydrolase	7 - 144 219 - 288	208	3.491	18.8	332.654	153 - 210	58	12.4	350.229	W270, N113, N54, R192	52
3ANS	1.98	*Homo sapiens* epoxide hydrolase	232 - 363 479 - 545	199	4.023	5.6	209.708	372 - 470	99	10.3	238.155	D104, H524, D496, Y466, Y383	59
3B12	1.2	*Burkholderia* sp. FA1 fluoroacetate dehalogenase	1 - 130 235 - 294	190	8.127	19.8	216.642	139 - 226	88	15	315.327	D104, H271, D128, W150, H149	76
3GZJ	2.19	*Rauvolfia serpentina* polyneuridine-aldehyde esterase	10 - 112 202 - 263	180	3.599	5.6	241.038	121 - 193	58	15.6	402.826	S87, H244*, D216, Y123	56

[a]Indicates residue mutated back to WT.

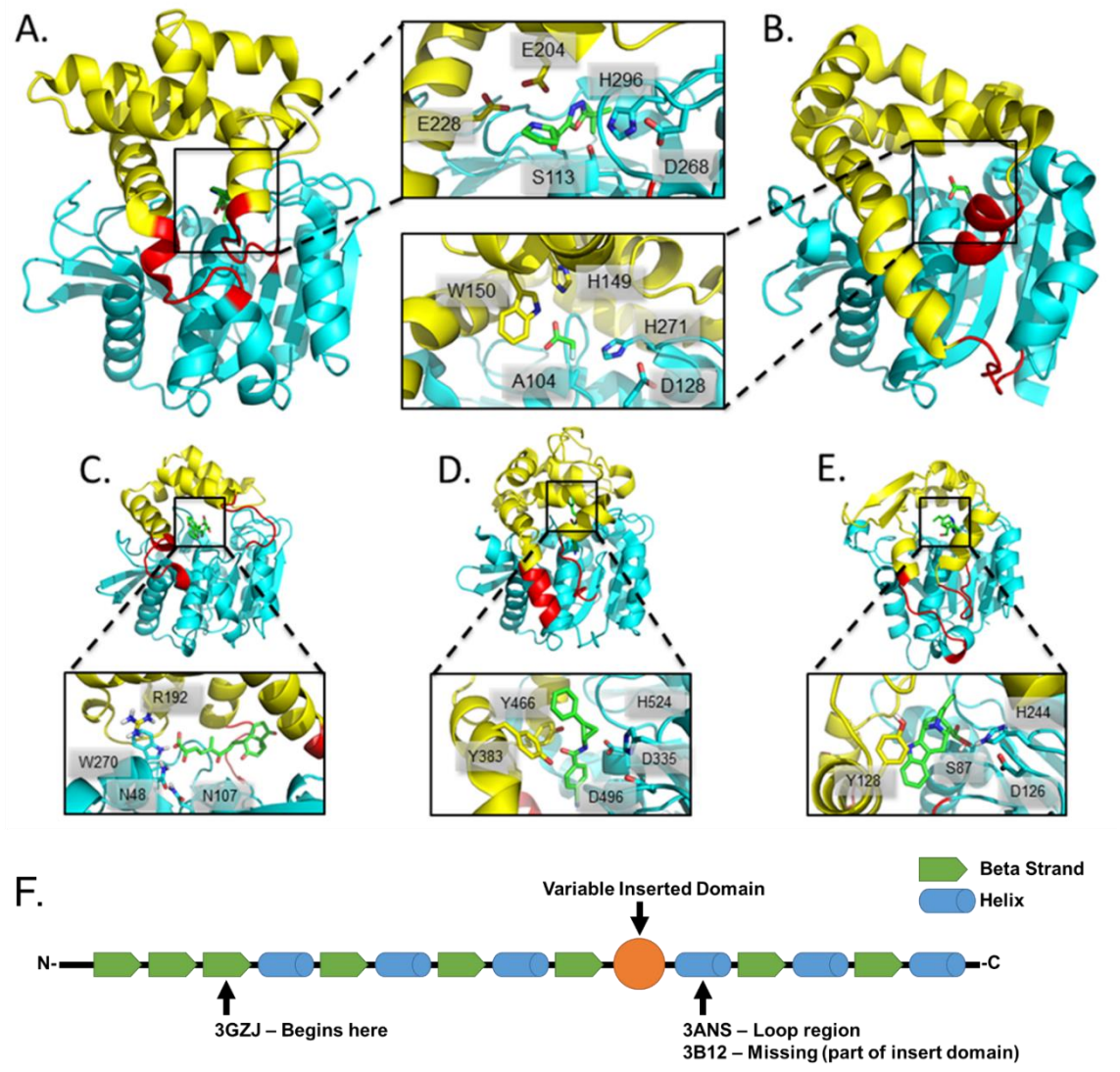


Figure 2-2: Cartoon representations of the five benchmark proteinscolored by parent domain (cyan), insert domain (yellow), and eight-residue linker regions (red). Inserts show the geometric relationship between catalytic residues (labeled, stick representation) and native ligands (green, stick representation). (A) Serratia marcescens prolyl aminopeptidase 45 (PDB 1WM1, chain A). (B) Burkholderia sp. FA1 fluoroacetate dehalogenase 46 (PDB 3B12, chain A). (C) Mycobacterium tuberculosis 2-hydroxy-6-oxo-6-phenylhexa-2,4-dienoate hydrolase 47 (PDB 2WUF, chain B). (D) Homo sapiens epoxide hydrolase 48 (PDB 3ANS, chain A). (E) Rauvolfia serpentina polyneuridine-aldehyde esterase 49 (PDB 3GZJ, chain A). The catalytic H244 residue has been mutated in PyMOL to its wild type identity. (F) A cartoon representation of the secondary structure sequence common to all five of the benchmark parent domains. Beta strands are shown as green arrows, helices are shown as blue cylinders, loop regions are black lines. Relative lengths are not shown to scale. The variable inserted domain, shown as an orange circle, occurs in the same loop region in all benchmark sets. Variations of the secondary structure topology unique to a particular benchmark protein are annotated.

2.5.2 Native insert domain conformations and interface sequences are recapitulated.

In order to assess the ability of the LooDo algorithm to adequately sample near-native geometric space, the native insert domain and native parent domain of each starting protein complex were used as inputs for the protocol outlined in **Figure 2-1**. This generated large sets of decoys consisting of conformations with differing relative orientations of each insert domain with respect to its corresponding parent domain, where the size of each decoy set is reported in **Table 2-3**. All hits generated by LooDo are one of a pair of insert domain orientations (one guided by linker A and one by linker B), so duplications of insert domain placements may arise when one loop generates a hit that is compatible with many other secondary loops. Therefore, we also report the number of unique insert domain placements in **Table 2-3**.

In each of the five benchmark sets, we find good sampling of native insert domain positions both directly after the LooDo algorithm, in which the insert domain is placed with respect to the parent domain, and after the loop closure algorithm. All five benchmark sets contain a decoy with an insert domain placed within 4 Å RMSD to the native placement before loop closure and within 2 Å after loop closure and structural refinement. Thus, adequate sampling of insert domain geometric space can be accomplished through LooDo given a large enough linker library and appropriately positioned placement grid.

Furthermore, we found that linker pairs that were culled from the fragment library and that produced Hit pairs also recapitulate the native linker pair geometries well during initial insert domain placement. However, in later stages of the design process linker pairs deviate to a greater extent from their native conformations, as seen by the increase in lowest linker pair RMSDs from before loop closure to after structural refinement in all five

benchmark sets. The initially low loop RMSDs indicate that supplying the fragment picker with at least half of the native residues of the loop region is sufficient to sample and score highly near-native loop conformations, while the deviations in RMSD in the later stages of design are to be expected when modeling flexible loop regions (**Table 2-1**).

Table 2-3: Benchmark results for native domain combinations.

PDB Accession Code	Number of Hits	Number of Unique Hits	Before Loop Closure		After Loop Closure	
			Lowest Insert Domain RMSD (Å)	Lowest Loop Pair RMSD (Å)	Lowest Insert Domain RMSD (Å)	Lowest Loop Pair RMSD (Å)
1WM1	2,086	881	1.331	0.362	0.84	1.1
2WUF	430	225	3.082	2.118	0.72	2.25
3ANS	7,899	1,435	0.747	0.510	0.89	1.96
3B12	998	437	0.637	1.220	0.86	1.62
3GZJ	10,034	3,153	2.180	0.540	1.197	0.68

We also expect a good design algorithm to rank the native insert domain placement and interface sequence relatively highly compared to more distantly placed insert domains with more divergent interface sequences. Encouragingly, the rank of decoys with native insert domain placements and sequences is high in all cases, as evidenced by the funnel-shaped distributions in **Figure 2-3:A-C**, which have been generated by plots of total energy, the frequency of native amino acid identities at designable positions (FNat), and interface energy versus insert domain RMSD. All three metrics show adequate discrimination between native and nonnative insert domain placements and sequences. PDB 2WUF presents a much flatter landscape compared to the other complexes in the benchmark set. We attribute this to the fact that 2WUF has the smallest insert domain of the five, and is able to find alternate interfaces of equivalent energy and compact conformation.

One can also quantify the funnels in **Figure 2-3**:**A** by counting the number of low-scoring decoys with an insert domain RMSD less than 2 Å or 5 Å from the native insert domain placement. In **Table 2-4**, the set of the top five decoys (ranked by total energy) from each benchmark set contains at least three decoys with insert domain placements

less than 5 Å from the native insert domain placement, and in four of the five benchmark sets the algorithm predicts at least one structure within 2 Å from the native.

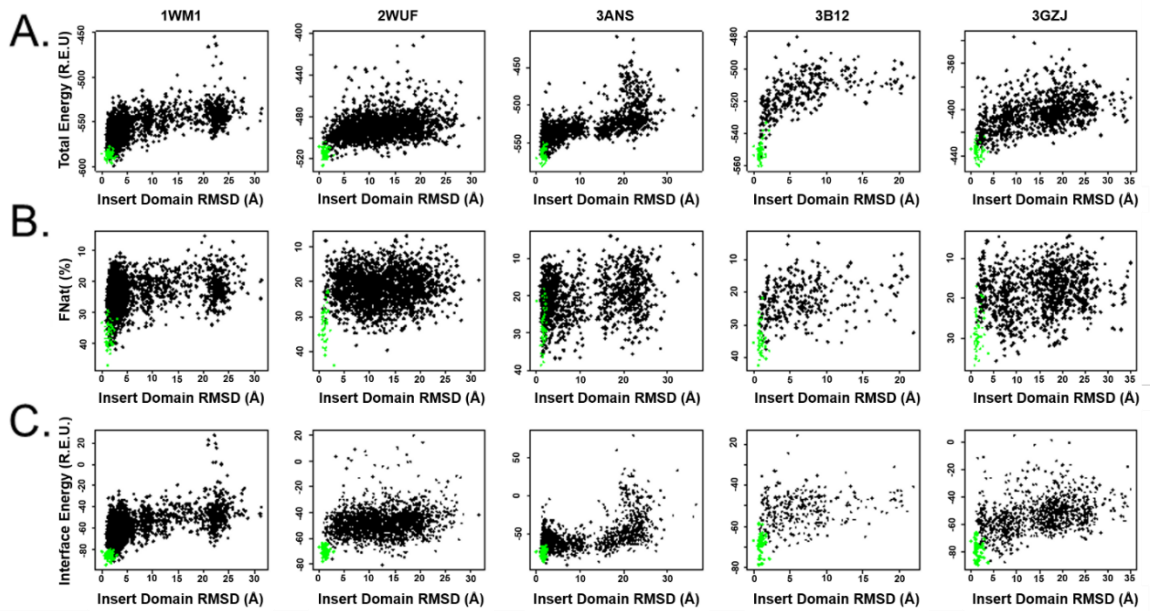


Figure 2-3: Graphs of (A) Total Energy, (B) FNat, and (C) Interface Energy versus Insert Domain RMSDshow best scoring decoys have near-native conformations and sequences. RMSD is calculated over all Cα atoms in the insert domain. Black points represent decoy structures; green points represent 50 native structures subjected to five iterations of FastRelax and Design.

Table 2-4: Quantification of funnel-shaped distributions found in Figure 2-3.

PDB Accession Code	N2	N5	$RMSD_{best}$	$RMSD_{score}$	$FNat_{best}$	$FNat_{score}$	$Rank_{catres}$
1WM1	1	5	1.46 (0.84)	2.35	34.04% (42.00%)	30.19%	1
2WUF	1	4	1.29 (0.72)	2.6	25.49% (39.58%)	22.81%	681
3ANS	2	5	1.38 (0.89)	2.8	25.00% (36.73%)	21.15%	7
3B12	5	5	0.88 (0.86)	1.02	37.50% (37.50%)	30.43%	1
3GZJ	0	3	3.612 (1.197)	5.537	35.71% (35.71%)	35.71%	3

N2, number of the five top-scoring decoy structures that have insert domain placements less than 2Å from the native insert domain placement; N5, number of the five top-scoring decoy structures that have insert domain placements less than 5Å from the native insert domain placement; $RMSD_{best}$, RMSD between the insert domain of the lowest-rmsd decoy within the five top-scoring decoys and the native insert domain; number in parenthesis is the lowest RMSD among all decoys; $RMSD_{score}$, RMSD between the insert domain of the top-scoring decoy and the native insert domain; $FNat_{best}$, frequency of native amino acids at designed positions in the decoy with the highest FNat of the five top-scoring decoys; number in parenthesis is the highest FNat among all decoys; $FNat_{score}$, frequency of native amino acids at designed positions in the top-scoring decoy; $Rank_{catres}$, the rank of the lowest-energy decoy among all decoys sorted by total score that has all catalytic residues placed at native positions.

In addition to native geometry, the algorithm also samples and ranks highly native-like interface amino acid sequences. In quantifying the funnels in **Figure 2-3:B**, we found that frequency of native amino acids at designed positions (FNat), including loop residues, is usually above 35% in low-RMSD decoys. Large-scale analyses of designing amino acids at protein-protein interfaces result in the recovery of ~51% of native residue identities, which represents the upper limit obtainable by current Rosetta energy functions[50–52], and the lower recovery obtained by us likely reflects the two flexible and solvent-exposed loop regions included in the designable residue set.

In order to simultaneously optimize decoys for structure and function, the matcher application[41,53] within Rosetta was also used as a filter for decoy structures after the LooDo placement algorithm and before loop closure in order to focus the structure optimization steps on Hits capable of recapitulating native-like geometric interactions between the active site residues and the native ligand molecule. In sorting each set of decoys by total score, we were able to give a rank to the lowest-energy decoy in which all of the active site residue constraints originating from the insert domain were satisfied by native residue positions in the insert domain ($Rank_{catres}$). This rank was found to be higher in cases where two catalytic residue constraints were required from the insert domain, as in 1WM1, 3ANS, and 3B12, as compared to only one, as in 2WUF and 3GZJ. This result may arise due to the cooperativity of functional constraints found through the requirement of having two or more residues coordinated in a specific conformation for the execution of catalytic function, as opposed to a single residue, whose specific geometric relationship with the ligand molecule could be found in a number of other placements of the insert domain.

2.5.3 Native insert-parent domain combinations are ranked highly compared to nonnative domain combinations.

As our benchmark test involves the recapitulation of the active site geometry (function) along with energy (stability), we also expect a domain insertion design algorithm to be able to rank native parent-insert domain combinations against nonnative parent and insert domain combinations well. To test whether LooDo can discriminate between native and non-native pairings, the insert domains from each of the five input structures were placed using LooDo around each of the five parent domains using our protocol, thus generating decoys for both native- and non-native parent-insert domain combinations. Each decoy in each insert-parent combination set was ranked separately by parent domain interface energy, which corresponds to the sum of the energies of each residue in the parent domain involved in the native protein-protein and protein-ligand interactions, and CST energy, which corresponds to a constraint score in which penalties for nonnative active site ligand-to-side chain interactions are imposed. The CST score is a measure of the deviation from the active site geometry. The top 100 decoys from each of the 5 insert-parent combinations for each parent domain were chosen. The results of this benchmark test reveal that in each set composed of variable insert domains and fixed parent domains, the native combination is always ranked within the top three structures (**Table 2-5**). After performing a sweep of weights to determine the best linear combination score function of CST Energy and Parent Domain Interface Energy, we found that by ranking decoys using CST energy + 5 × Parent Domain Interface Energy, four of the five cross-combination sets ranked the native combination first. Furthermore, in the cross-docked decoy sets, the native-domain combination set always occupies the lowest energy distributions within the distribution of the larger set (**Figure 2-4**). Therefore, discrimination between native and nonnative domain combinations is attainable *via* a score function that includes measures of

both domain–domain interface compatibility and ligand–specific topologies specific to the parent domain.

Table 2-5: Results of combining each insert domain with each parent domain.

Native PDB ID of Insert Domain	Native PDB ID of Parent Domain	Number of Hits	Number of Unique Hits	Number of Unique Matcher-Filtered Decoys	Number of Closed Structures After KIC	Rank of Lowest Energy Structure (RMSD if native)		
						Parent Domain Interface Energy	CST Energy	CST Energy + 5*Parent Domain Interface Energy
1WM1	1WM1	2,086	881	55	2,508	2 (4.17)	2 (4.77)	1 (4.77)
2WUF	1WM1	15,249	3,016	422	2,680	5	108	115
3ANS	1WM1	4,436	1,403	175	703	8	1	14
3B12	1WM1	10,269	3,307	450	688	1	103	36
3GZJ	1WM1	7,781	2,066	263	2,185	20	113	113
1WM1	2WUF	3,233	1,303	518	1,785	3	35	18
2WUF	2WUF	430	225	214	2,137	1 (7.20)	1 (18.95)	1 (7.20)
3ANS	2WUF	1,492	699	311	675	13	31	13
3B12	2WUF	1,171	582	187	465	28	26	33
3GZJ	2WUF	3,008	1,194	466	3,510	17	42	17
1WM1	3ANS	16,411	4,321	143	1,170	20	1	3
2WUF	3ANS	2,595	1,429	119	255	85	42	60
3ANS	3ANS	7,899	1,435	359	1,317	1 (2.64)	2 (19.98)	1 (19.98)
3B12	3ANS	3,220	1,672	40	225	11	6	4
3GZJ	3ANS	20,572	4,239	171	1,675	7	3	2
1WM1	3B12	1,809	968	77	72	1	1	1
2WUF	3B12	24,553	3,130	368	838	2	11	17
3ANS	3B12	861	504	17	62	165	77	62
3B12	3B12	998	437	78	301	3 (9.8)	2 (3.11)	2 (3.11)
3GZJ	3B12	1,331	640	19	12	237	64	66
1WM1	3GZJ	16,255	4,586	442	1,661	24	4	5
2WUF	3GZJ	15,108	4,128	417	2,400	27	7	11
3ANS	3GZJ	10,393	2,959	160	882	7	34	43
3B12	3GZJ	18,206	5,222	446	938	8	9	14
3GZJ	3GZJ	10,034	3,153	535	3,236	1 (15.117)	1 (17.983)	1 (6.783)

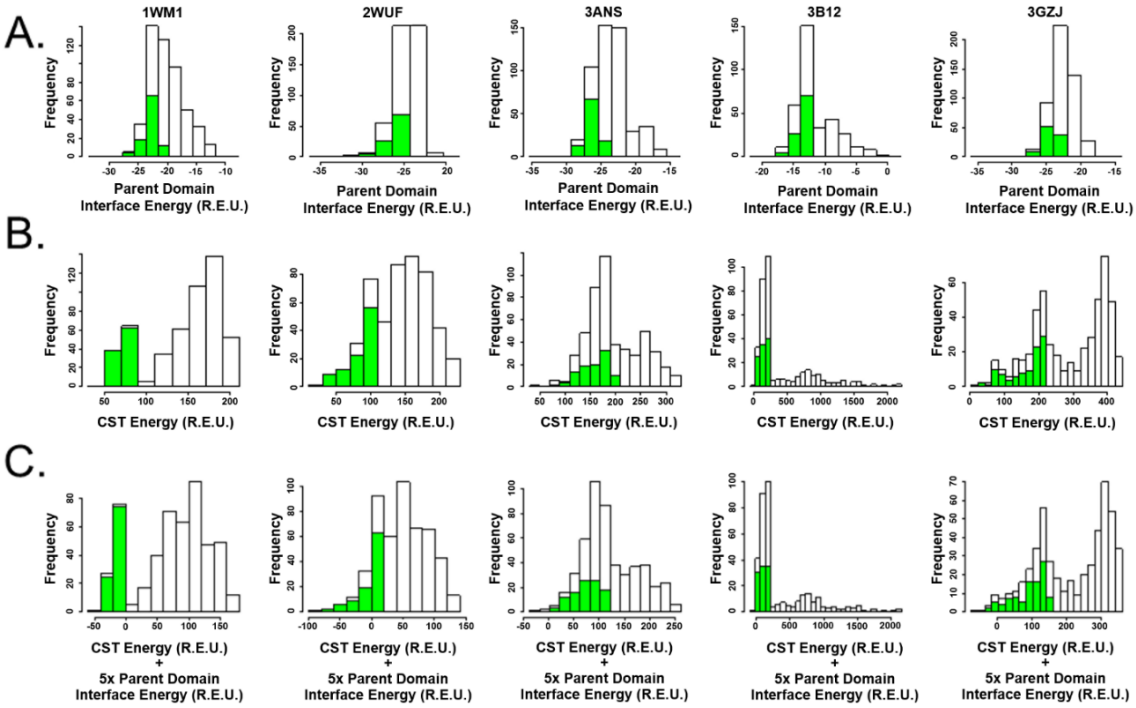


Figure 2-4: Energy distributions among the set of 500 decoys containing the top 100 decoys from each of the five insert-parent domain combination sets with the same parent domain.Decoys from each of the five sets of insert domain, parent domain combination sets were scored using (A) Parent Domain Interface Energy, (B) CST Energy, or (C) a linear combination of the two, and the top 100 decoys from the five sets containing the same parent domain were compiled into a set of 500 decoys. Green bars represent the fraction of the histogram bin composed of native insert-parent combination decoys.

2.5.4 Lowering the number of catalytic constraints on insert domain placement negatively affects discrimination of native-like decoys.

In order to investigate the contribution of the number of catalytic constraints originating from the insert domain on the funnel-shaped distributions in **Figure 2-3**:**A-C**, we repeated the above experiments using PDB 1WM1, for which we originally imposed two catalytic residues constraints in the insert domain (residues E228 and E204), with one (residue E228 only or residue 204 only) and zero insert domain catalytic residue constraints. The removal of catalytic residue constraints causes a reduction in funnel quality across all three metrics of Total Score, FNat, and Interface Energy (**Figure 2-5:A-C**), while the relaxed natives consistently remained the lowest energy points. This effect increases as the number of interface catalytic constraints is reduced. Discrimination of native domain combinations among native- and nonnative-domain combinations also greatly decreases as insert domain catalytic residues constraints are removed (**Figure 2-6**). This decrease in native-domain combination recapitulation and ranking (**Table 2-6**) with lower catalytic geometry requirements may echo the natural tradeoff between function and stability[54]. Protein active sites are optimized such that they do not necessarily adopt the lowest energy configuration, but rather a configuration that is adequate at performing specific functional tasks. Therefore, when constraints are removed from the decoy filtration step and active site residues are allowed to explore lower energy configurations, the funnel-shaped distributions become more flattened, as it is possible for many conformations (both native and nonnative) to be equally low in energy (**Figure 2-6**).

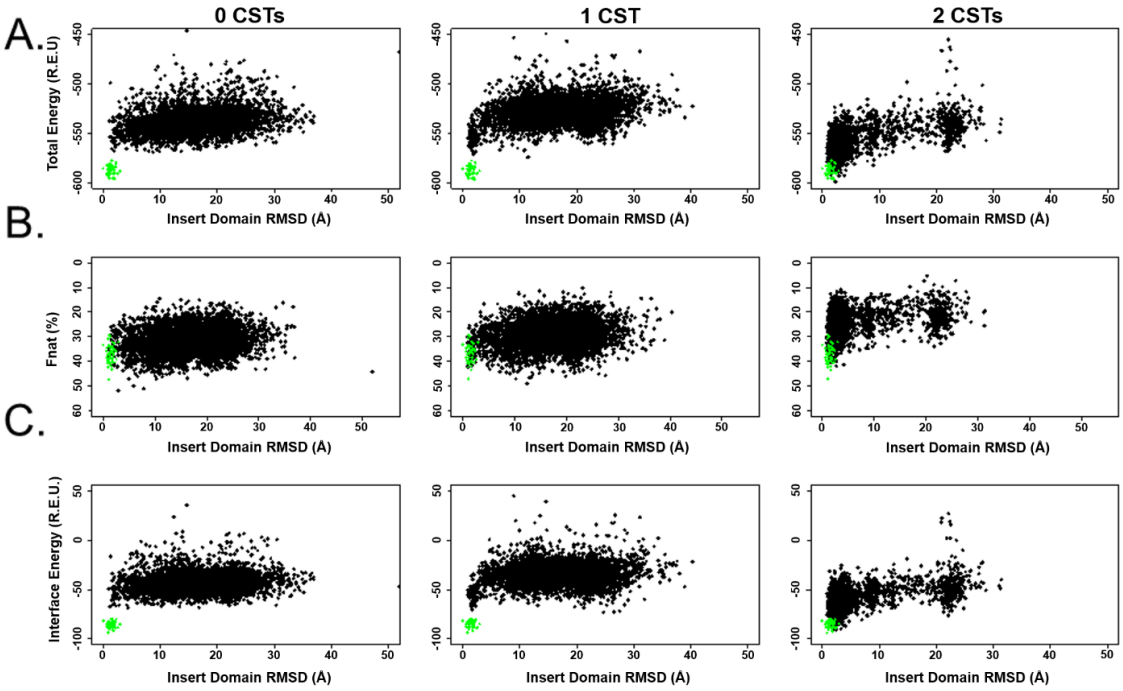


Figure 2-5: Graphs of Total Energy (A), FNat (B), and Interface Energy (C) versus Insert Domain RMSD for the native 1WM1 insert-parent combination set when LooDo Hits are filtered with 0, 1, or 2 residues included in the insert domain catalytic constraints.RMSD is calculated over all Cα atoms in the insert domain. Black points represent decoy structures; green points represent 50 native structures subjected to five rounds of FastRelax and Design.

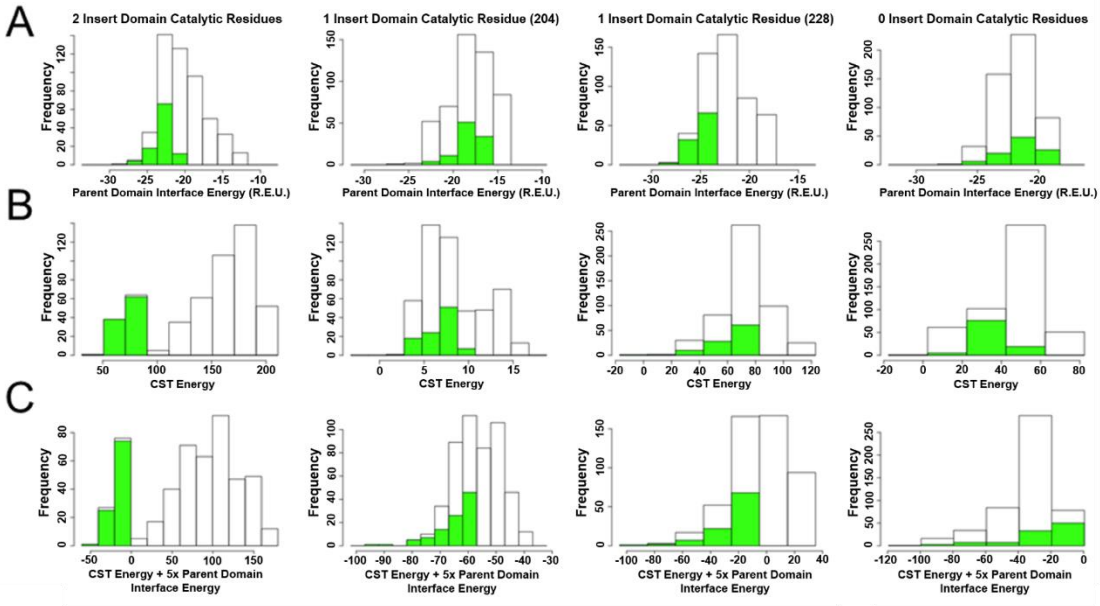


Figure 2-6: Energy distributions among the set of 500 decoys containing the top 100 decoys from each of the five insert-parent domain combination sets with the 1WM1 parent domain.Decoys from each of the five sets of insert domain, parent domain combination sets were scored using (A) Parent Domain Interface Energy, (B) CST Energy, or (C) a linear combination of the two, and the top 100 decoys from the five sets containing the same parent domain were compiled into a set of 500 decoys. Green bars represent the fraction of the histogram bin composed of native insert-parent combination decoys.

Table 2-6: Effect of decreasing number of catalytic residue constraints originating from the insert domain on rank of native insert-parent domain combination.

Number of Insert Domain Catalytic Residues	Native PDB ID of Insert Domain	Native PDB ID of Parent Domain	Number of Unique Matcher Results	Number of Closed Structures After KIC	Rank of Lowest Energy Structure		
					Parent Domain Interface Energy	CST Energy	CST Energy + 5*Parent Domain Interface Energy
2	1WM1	1WM1	55	2,508	2	2	1
	2WUF	1WM1	422	2,680	5	108	115
	3ANS	1WM1	175	703	8	1	14
	3B12	1WM1	450	688	1	103	36
	3GZJ	1WM1	263	2,185	20	113	113
1 (Residue 232)	1WM1	1WM1	869	4,218	149	24	20
	2WUF	1WM1	2,651	1,255	22	3	23
	3ANS	1WM1	1,041	634	21	12	16
	3B12	1WM1	2,638	987	2	2	3
	3GZJ	1WM1	1,781	1,799	1	1	1
1 (Residue 204)	1WM1	1WM1	553	658	13	8	1
	2WUF	1WM1	3,204	1,149	113	4	13
	3ANS	1WM1	1,140	309	74	2	12
	3B12	1WM1	3,160	549	128	1	38
	3GZJ	1WM1	1,964	1,600	1	16	41
0	1WM1	1WM1	877	4,303	6	42	53
	2WUF	1WM1	2,818	1,050	17	2	1
	3ANS	1WM1	1,201	961	5	4	2
	3B12	1WM1	2,916	793	1	17	79
	3GZJ	1WM1	1,884	1,874	4	1	3

2.5.5 Supplying the fragment picker with more information on native loop residues increases the algorithm's propensity to sample near-native insert domain placements.

In order to measure the extent to which linker sequence and structural information can shift the distribution of insert domain placements, we also examined the effect of biasing the selection of linker peptides toward more native-like amino acid sequences and secondary structures through reduction of the 'molten' residue positions, defined as lacking sequence and secondary structure bias. As currently implemented, libraries of linkers are generated with two input parameters corresponding to the amino acid sequence and secondary structure sequence $I_1I_2XXXXP_1P_2$, where I_i denotes a known insert domain residue, X denotes a 'molten' residue position, and P_i denotes a known parent domain residue. This input will generate libraries containing eight-residue linker sequences biased to contain sequences with more native-like terminal residues. If instead the sequences are biased using a molten region of two residues ($I_1I_2L_1XXL_4P_1P_2$) or zero residues ($P_1P_2L_1L_2L_3L_4P_1P_2$), where L_i denotes a native linker residue, the sampling of low insert domain RMSD structures before loop closure remains the same or shifts towards more near-native insert domain conformations (**Table 2-1**). Furthermore, the same experiment with a larger set of insert-domain proteins (Gray set) shows similar results in which lower RMSD structures are more consistently sampled when the entire linker sequences are specified (**Figure 2-7**). Therefore, the LooDo algorithm may be limited by the composition of its fragment libraries, and more accurate modeling of conformational flexibility of the domain-domain linker segments is likely to improve its performance.

Table 2-7: Effect of decreasing the molten region on lowest insert domain RMSD in the benchmark set.

PDB Accession Code	Lowest Insert Domain RMSD (Before Loop Closure)		
	Molten Region = 4	Molten Region = 2	Molten Region = 0
1WM1	1.33	1.33	1.33
2WUF	3.08	3.04	4.25
3ANS	0.75	0.33	0.33
3B12	0.64	0.82	1.15
3GZJ	1.71	1.71	0.58

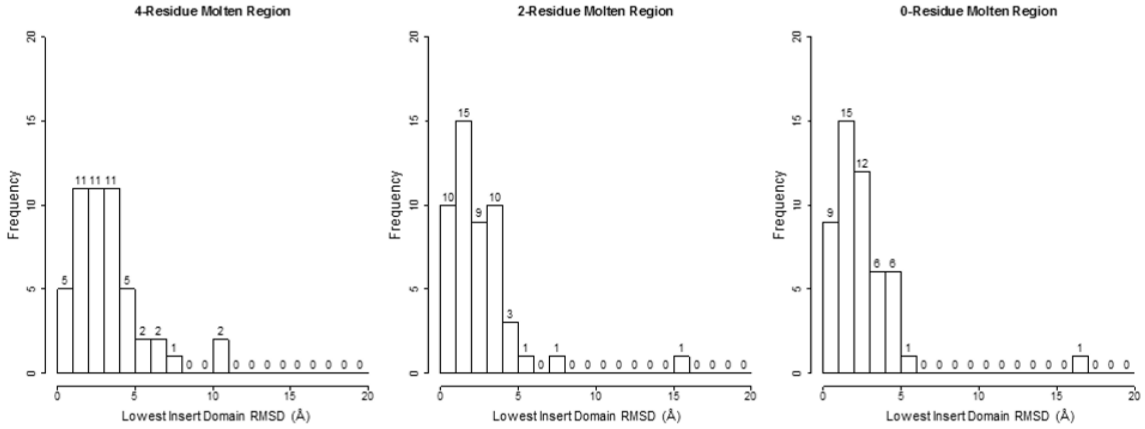


Figure 2-7: Decreasing the molten region during fragment picking shifts the distribution of the lowest insert domain RMSD toward more near-native conformations in the Gray set.RMSD distributions observed for 4-residue (left), 2-residue (middle), and 0-residue (right) molten regions are indicated.

2.5.6 Using pairs of low-energy interface residues at the parent-insert interface as geometrical constraints in non-ligand binding proteins does not result in energy funnels.

To test how the choice of residues used as geometric constraints affects the performance of the LooDo algorithm, we performed similar analyses as described above for 10 systems from the Gray benchmark set of domain-inserted proteins[18]. These proteins do not bind ligands at the domain-domain interface in their crystal structures, and we asked if constraining the LooDo sampling using two geometrical constraints between the parent and insert domains would lead to funneled landscapes as observed for our benchmark set of ligand-binding proteins (**Figure 2-3**). Specifically, we chose one residue

from the parent domain and one residue from the insert domain whose interaction energy was calculated to be the lowest of all such pairs of residues at the domain-domain interface (**Figure 2-8:A, Figure 2-9**). The distance between these pseudo-catalytic sidechains in this set (~6-12 Å; **Figure 2-8:A**, **Figure 2-9**), the lengths of the inserted domains (65-217) residues, and the number of designable (interface) residues (10-66) were similar to the enzyme benchmark set (**Table 2-1, Table 2-2, Table 2-8**). We find that these proteins, in which geometric constraints without any obvious functional role were used to guide LooDo, show dramatically different results in that non-native orientations of the two domains can be designed to be equally low in energy (**Figure 2-8:B,C**). These results indicate that funnels observed in the enzyme benchmark set (**Figure 2-3**) are not the result of simply geometrically constraining the system, and highlight the key role of catalytic residues at the interface.

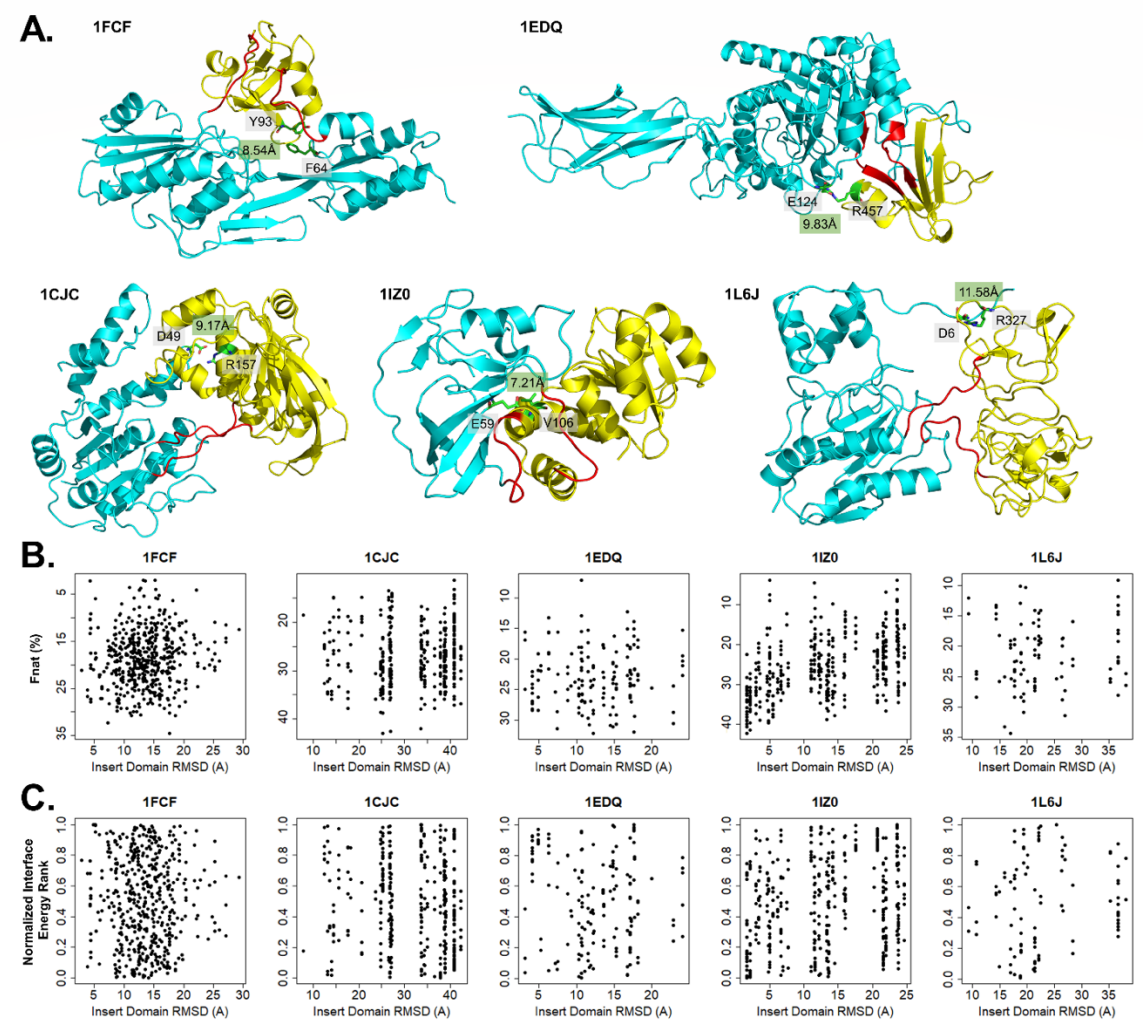


Figure 2-8: LooDo analyses for five PDBs from the Gray set where non-catalytic conformational constraints were used during the Matching stage show a lack of FNat or Interface Energy funnels.(A) Cartoon representations of five Gray Set proteins colored by parent domain (cyan), insert domain (yellow), and eight-residue linker regions (red). Constrained residues are shown as green sticks and labeled. The Cα-Cα distance between the constrained residues is depicted in the green boxes. (B) FNat versus Cα-RMSD. (C) Interface Energy rankings normalized to a range between 0 (lowest rank) and 1 (highest rank). Results for five other systems from this set are presented in **Figure 2-9**.

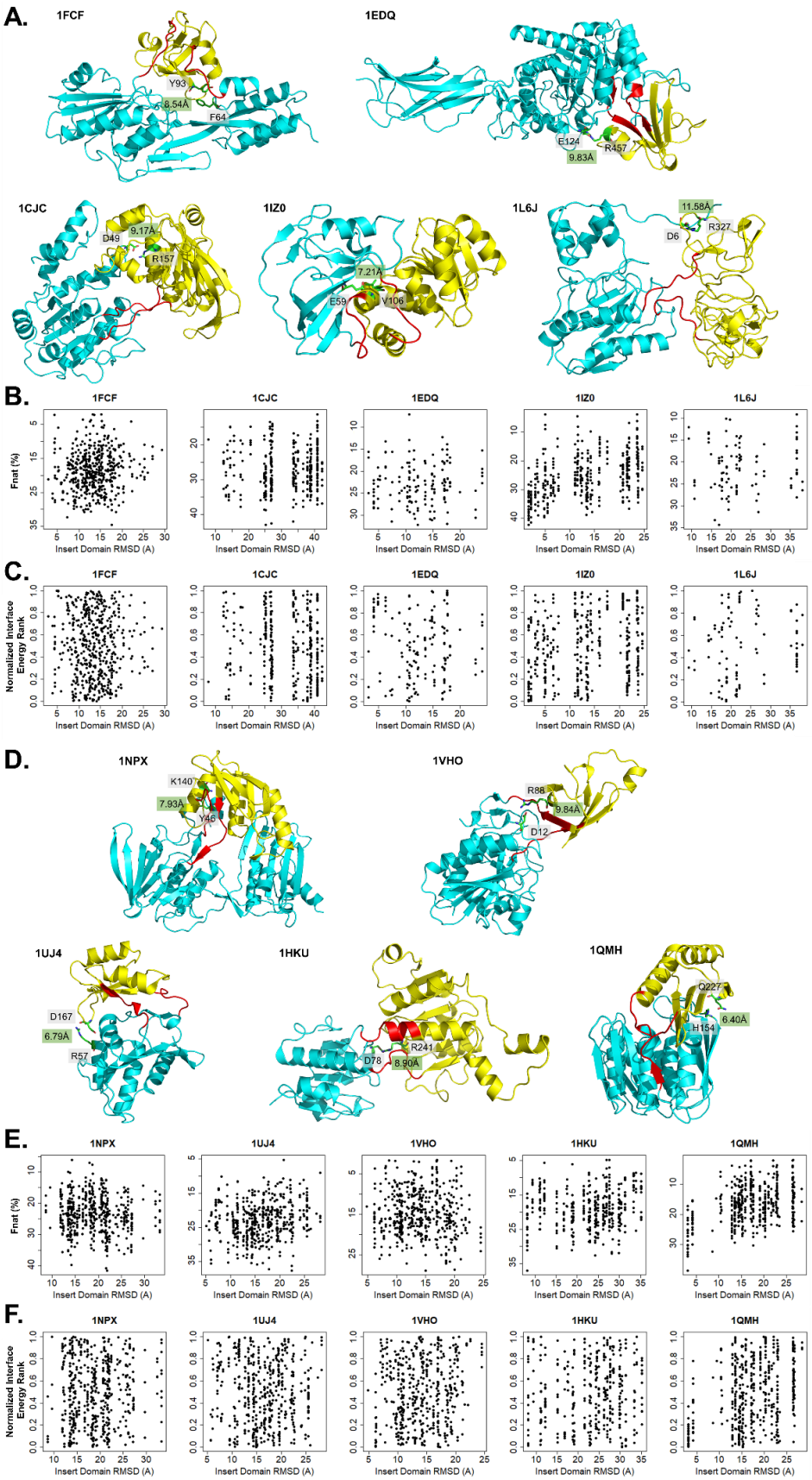

A.
1FCF
Y93
8.54Å
F64
1EDQ
E124
R457
9.83Å
1CJC
D49
9.17Å
R157
1IZ0
7.21Å
E59
V106
1L6J
11.58Å
D6
R327
B.
1FCF
1CJC
1EDQ
1IZ0
1L6J
Fnat (%)
Insert Domain RMSD (A)
C.
Normalized Interface Energy Rank
D.
1NPX
K140
7.93Å
Y46
1VHO
R88
9.84Å
D12
1UJ4
D167
6.79Å
R57
1HKU
R241
D78
8.90Å
1QMH
Q227
H154
6.40Å
E.
F.

Figure 2-9: LooDo analyses for ten PDBs from the Gray set where non-catalytic conformational constraints were used during the Matching stage show a lack of FNat or Interface Energy funnels.(A) Cartoon representations of five Gray Set proteins colored by parent domain (cyan), insert domain (yellow), and eight-residue linker regions (red). Constrained residues are shown as green sticks and labeled. The Cα-Cα distance between the constrained residues is depicted in the green boxes. (B) FNat versus Cα-RMSD. (C) Interface Energy rankings normalized to a range between 0 (lowest rank) and 1 (highest rank). Results for five other systems from this set are presented in Supplementary Figure 1. (D) Same representations as (A) but for the remaining five Gray Set proteins. (E) and (F) are similar to (C) and (D), but for the five proteins in (D).

Table 2-8: Subset of 10 PDBs from the Gray set.

PDB Accession Code	Total Number of Residues:		Number of Residues at Parent/Insert Interface
	Parent Domain	Insert Domain	
1FCF	287	84	28
1CJC	222	217	35
1EDQ	459	65	56
1IZ0	120	154	48
1L6J	222	167	18
1NPX	315	115	66
1UJ4	140	65	26
1VHO	231	66	14
1HKU	132	183	10
1QMH	235	87	34

2.6 Discussion

A major pathway by which protein evolution occurs involves the modular incorporation of multiple protein domains into single, connected complex, which has given rise to a vast diversity of complex functionalities. This complex structural coupling of domains provides a means for radically altering the topology of proteins and for generating new domain-domain interfaces amenable to functional modulation. Computationally designing domain insertion proteins would constitute a generalizable, rational approach for altering binding affinity and/or substrate selectivity of a given parent enzyme or protein domain. In this article, we have developed and benchmarked an algorithm to sample insert domain placements with respect to a parent domain while simultaneously designing structure and function at the domain-domain interface. We have shown that the LooDo algorithm can successfully recapitulate native domain-insertion protein structure and sequence, as well as discriminate between native and nonnative domain conformations

and combinations. One limitation of the algorithm lies in its reliance on the conformations of loops observed in protein structure databases, as linker regions wholly dictate initial cap placement and could produce non-optimal placements in certain cases. However, our results also suggest that optimal placement of domains can be recovered by using non-native loops upon post-placement conformational refinement: for example, the linkers in PDB 1WM1 can be found in the Vall database in their entirety whereas the linkers for the other four parent domains are not present in the database. Indeed, insert domain sampling for 1WM1 produced near-native results (**Table 2-3**, **Table 2-7**) as expected but, gratifyingly, the LooDo algorithm was also able to find sub-Angstrom insert domain orientations for 3ANS, 3B12, and 3GZJ.

The prevalence of multi-domain enzymes to acquire complex active sites at domain-domain interfaces alludes to the co-evolution between stability and function within protein structures[55]. The exploration of inter-domain sequences leading to novel activities has led to a vast functional repertoire in multi-domain enzymes, despite probable adverse effects on intramolecular domain stabilities. Therefore, modern insert-domain proteins with complex interface-based activities likely cannot be described solely by measurements of overall stability, but also by the stability and geometry of the interaction specifically present at the domain-domain interface. Consequently, our method is unique in that it also incorporates scoring terms that reflect this importance of the domain-domain interface residues and geometric relationship between catalytic residues (captured via CST energy).

With a framework in hand for the addition of insert domains given two inter-domain points of attachment, generalizing the current method for the addition of a linearly attached insert domain, a domain with three or more points of attachment, or any combination of insert domains will be a straightforward extension of the method.

Furthermore, if the sequence is known, the LooDo algorithm may also be used as a structure-prediction tool for domain-insertion proteins, which we have also tested with the benchmark set in this study (**Figure 2-10**). In these experiments, we used the native sequences of the connecting loops and no small molecule placement was performed. Strong funnel-shaped distributions in graphs of Insert-domain RMSD versus Total Energy are found to be present when decoy generation was altered by supplying the entire sequence of the native linker region fragment picking, indicating that the algorithm robustly samples and accurately scores near-native conformations.

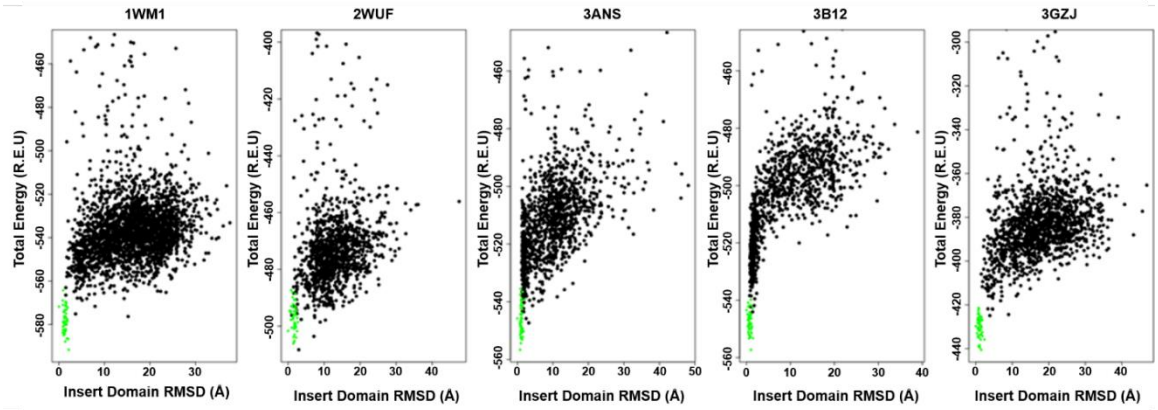


Figure 2-10: Total Energy versus Insert Domain RMSD for structure prediction benchmark.RMSD is calculated over all Cα atoms in the insert domain. Black points represent decoy structures; green points represent 50 native structures subjected to five iterations of FastRelax.

2.7 Conclusion

We describe a computational method to enable the structure prediction and design of novel nested proteins. The method, implemented in the context of the Rosetta software, is able to robustly recapitulate the structures and the sequences of these proteins, guided by Rosetta energy and catalytic/binding site geometry. However, two facets of each protein structure were known prior to the experiments: (1) the delineation between insert and parent domain and, (2) the geometric constraints dictating important relationships

between residues in the insert domain and a native ligand molecule. In order to expand the method to be applicable to a general set of domain insertion proteins, the outstanding questions of how to choose the insertion site and catalytic residues remain to be investigated. While global structure-based statistics on domain-insertion proteins have been gathered, showing that inserts are typically smaller and located in the latter two-thirds of the primary sequence[1], these studies are ongoing in our laboratory. Recent work by Dokholyan and co-workers[16] as well as King et. al[56] also provide encouraging leads. The ability to design new domain inserted proteins is expected to aid design efforts for generating novel protein functions at the resulting domain-domain interfaces.

2.8 References

1. Aroul-Selvam, R., Hubbard, T. & Sasidharan, R. Domain insertions in protein structures. *J. Mol. Biol.* **338,** 633–641 (2004).

2. Bashton, M. & Chothia, C. The Generation of New Protein Functions by the Combination of Domains. *Structure* **15,** 85–99 (2007).

3. Gilbert, W. Why genes in pieces? *Nature* **271,** 501 (1978).

4. Blake, C. C. F. Do genes-in-pieces imply proteins-in-pieces? *Nature* **273,** 267 (1978).

5. Dwyer, M. A. & Hellinga, H. W. Periplasmic binding proteins: A versatile superfamily for protein engineering. *Current Opinion in Structural Biology* **14,** 495–504 (2004).

6. Baker, P. J., Turnbull, A. P., Sedelnikova, S. E., Stillman, T. J. & Rice, D. W. A role for quaternary structure in the substrate specificity of leucine dehydrogenase. *Structure* **3,** 693–705 (1995).

7. McLachlan, A. D. Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* **128,** 49–79 (1979).

8. Burroughs, A. M., Allen, K. N., Dunaway-Mariano, D. & Aravind, L. Evolutionary Genomics of the HAD Superfamily: Understanding the Structural Adaptations and Catalytic Diversity in a Superfamily of Phosphoesterases and Allied Enzymes. *J. Mol. Biol.* **361,** 1003–1034 (2006).

9. Aravind, L., Mazumder, R., Vasudevan, S. & Koonin, E. V. Trends in protein evolution inferred from sequence and structure analysis. *Current Opinion in Structural Biology* **12,** 392–399 (2002).

10. Mascotti, M. L., Juri Ayub, M., Furnham, N., Thornton, J. M. & Laskowski, R. A. Chopping and Changing: the Evolution of the Flavin-dependent Monooxygenases. *J. Mol. Biol.* **428,** 3131–3146 (2016).

11. Doi, N. & Yanagawa, H. Insertional gene fusion technology. *FEBS Lett.* **457,** 1–4 (1999).

12. Hemanth Giri Rao, V. V. & Gosavi, S. In the Multi-domain Protein Adenylate Kinase, Domain Insertion Facilitates Cooperative Folding while Accommodating Function at Domain Interfaces. *PLoS Comput. Biol.* **10,** (2014).

13. Ehrmann, M., Boyd, D. & Beckwith, J. Genetic analysis of membrane protein topology by a sandwich gene fusion approach. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 7574–7578 (1990).

14. Ostermeier, M. Engineering allosteric protein switches by domain insertion. *Protein Engineering, Design and Selection* **18,** 359–364 (2005).

15. Yasui, N. *et al.* Directed network wiring identifies a key protein interaction in embryonic stem cell differentiation. *Mol. Cell* **54,** 1034–1041 (2014).

16. Dagliyan, O. *et al.* Engineering extrinsic disorder to control protein activity in living cells. *Science (80-.).* **354,** 1441–1444 (2016).

17. Dagliyan, O. *et al.* Rational design of a ligand-controlled protein conformational switch. *Proc. Natl. Acad. Sci.* **110,** 6800–6804 (2013).

18. Berrondo, M., Ostermeier, M. & Gray, J. J. Structure Prediction of Domain Insertion Proteins from Structures of Individual Domains. *Structure* **16,** 513–527 (2008).

19. Guntas, G. & Ostermeier, M. Creation of an Allosteric Enzyme by Domain Insertion. *J. Mol. Biol.* **336,** 263–273 (2004).

20. Guntas, G., Mansell, T. J., Kim, J. R. & Ostermeier, M. Directed evolution of protein switches and their application to the creation of ligand-binding proteins. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 11224–11229 (2005).

21. Edwards, W. R., Busse, K., Allemann, R. K. & Jones, D. D. Linking the functions of unrelated proteins using a novel directed evolution domain insertion method. *Nucleic Acids Res.* **36,** (2008).

22. Wright, C. M., Wright, R. C., Eshleman, J. R. & Ostermeier, M. A protein therapeutic modality founded on molecular regulation. *Proceedings of the National Academy of Sciences* **108,** 16206–16211 (2011).

23. Kanwar, M., Wright, R. C., Date, A., Tullman, J. & Ostermeier, M. Protein switch engineering by domain insertion. *Methods Enzymol.* **523,** 369–388 (2013).

24. Nadler, D. C., Morgan, S.-A., Flamholz, A., Kortright, K. E. & Savage, D. F. Rapid construction of metabolite biosensors using domain-insertion profiling. *Nat. Commun.* **7,** 12266 (2016).

25. Desjarlais, J. R. & Handel, T. M. Side-chain and backbone flexibility in protein core design. *J. Mol. Biol.* **290,** 305–318 (1999).

26. Kuhlman, B., O'Neill, J. W., Kim, D. E., Zhang, K. Y. J. & Baker, D. Accurate computer-based design of a new backbone conformation in the second turn of protein L. *J. Mol. Biol.* **315,** 471–477 (2002).

27. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science* **302,** 1364–1368 (2003).

28. Dahiyat, B. I., Sarisky, C. A. & Mayo, S. L. De novo protein design: towards fully automated sequence selection. *J. Mol. Biol.* **273,** 789–796 (1997).

29. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. High-resolution protein design with backbone freedom. *Science* **282,** 1462–1467 (1998).

30. Hua, S., Guo, T., Gough, J. & Sun, Z. Proteins with class α/β fold have high-level participation in fusion events. *J. Mol. Biol.* **320,** 713–719 (2002).

31. Nardini, M. & Dijkstra, B. W. α/β hydrolase fold enzymes: The family keeps growing. *Current Opinion in Structural Biology* **9,** 732–737 (1999).

32. Lenfant, N. *et al.* ESTHER, the database of the α/β-hydrolase fold superfamily of proteins: Tools to explore diversity of functions. *Nucleic Acids Res.* **41,** (2013).

33. Pavlova, M. *et al.* Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nat. Chem. Biol.* **5,** 727–733 (2009).

34. Jochens, H. *et al.* Protein Engineering of α/β-Hydrolase Fold Enzymes. *ChemBioChem* **12,** 1508–1517 (2011).

35. Hu, X., Wang, H., Ke, H. & Kuhlman, B. High-resolution design of a protein loop. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 17668–17673 (2007).

36. Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6,** 551–552 (2009).

37. Kortemme, T. & Baker, D. Computational design of protein-protein interactions. *Current Opinion in Chemical Biology* **8,** 91–97 (2004).

38. Fleishman, S. J. *et al.* Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite. *PLoS One* **6,** (2011).

39. Misura, K. M. S. & Baker, D. Progress and challenges in high-resolution refinement of protein structure models. *Proteins Struct. Funct. Genet.* **59,** 15–29 (2005).

40. Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. M. & Baker, D. Generalized fragment picking in rosetta: Design, protocols and applications. *PLoS One* **6,** (2011).

41. Zanghellini, A. *et al.* New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* **15,** 2785–2794 (2006).

42. Coutsias, E. A., Seok, C., Wester, M. J. & Dill, K. A. Resultants and loop closure. in *International Journal of Quantum Chemistry* **106,** 176–189 (2006).

43. Bhardwaj, G., Mulligan, V. K., Bahl, C. D. & Gilmore, J. M. Accurate de novo design of hyperstable constrained peptides. *Nature* **538,** 329–335 (2016).

44. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

45. Inoue, T. *et al.* Novel inhibitor for prolyl aminopeptidase from Serratia marcescens and studies on the mechanism of substrate recognition of the enzyme using the inhibitor. *Arch. Biochem. Biophys.* **416,** 147–154 (2003).

46. Nakayama, T. *et al.* Substrate specificity of fluoroacetate dehalogenase: An insight from crystallographic analysis, fluorescence spectroscopy, and theoretical computations. *Chem. - A Eur. J.* **18,** 8392–8402 (2012).

47. Lack, N. A. *et al.* Characterization of a carbon-carbon hydrolase from Mycobacterium tuberculosis involved in cholesterol metabolism. *J. Biol. Chem.* **285,** 434–443 (2010).

48. Tanaka, D. *et al.* A practical use of ligand efficiency indices out of the fragment-based approach: Ligand efficiency-guided lead identification of soluble epoxide hydrolase inhibitors. *J. Med. Chem.* **54,** 851–857 (2011).

49. Yang, L., Hill, M., Wang, M., Panjikar, S. & Stöckigt, J. Structural basis and enzymatic mechanism of the biosynthesis of C 9from C10-monoterpenoid indole alkaloids. *Angew. Chemie - Int. Ed.* **48,** 5211–5213 (2009).

50. Park, H. *et al.* Simultaneous optimization of biomolecular energy function on features from small molecules and macromolecules. *J. Chem. Theory Comput.* **12,** 6201–6212 (2016).

51. Leaver-Fay, A. *et al.* Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol.* **523,** 109–143 (2013).

52. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13,** 3031–3048 (2017).

53. Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S. & Baker, D. De novo enzyme design using Rosetta3. *PLoS One* **6,** (2011).

54. Beadle, B. M. & Shoichet, B. K. Structural bases of stability-function tradeoffs in enzymes. *J. Mol. Biol.* **321,** 285–296 (2002).

55. Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. A relationship between protein stability and protein function. *Proceedings of the National Academy of Sciences* **92,** 452–456 (1995).

56. King, I. C. *et al.* Precise assembly of complex beta sheet topologies from de novo designed building blocks. *Elife* **4,** 3–11 (2015).

Chapter 3: Computational Design of a Photo-controlled Cytosine Deaminase

3.1 Preface

A version of Chapter 3 has been published [Blacklock, K. M., Yachnin, B. J., Woolley, G. A. & Khare, S. D. Computational Design of a Photocontrolled Cytosine Deaminase. *J. Am. Chem. Soc.* **140,** 14–17 (2018)]. I was the lead investigator, responsible for all major areas of concept formation, data collection and analysis, as well as manuscript composition. Brahm Yachnin and G. Andrew Woolley contributed to early stages of concept formation, experimental guidance, and manuscript edits. G. Andrew Woolley also graciously provide reagents. Sagar Khare was the supervisory author on this project and was involved throughout the project in concept formation and manuscript composition.

3.2 Abstract

There is growing interest in designing spatiotemporal control over enzyme activities using noninvasive stimuli such as light. Here, we describe a structure-based, computation-guided predictive method for reversibly controlling enzyme activity using covalently attached photo-responsive azobenzene groups. Applying the method to the therapeutically useful enzyme yeast cytosine deaminase, we obtained a ~3-fold change in enzyme activity by the photo-controlled modulation of the enzyme's active site lid structure, while fully maintaining thermostability. Multiple cycles of switching, controllable in real time, are possible. The predictiveness of the method is demonstrated by the construction of a variant that does not photoswitch as expected from computational modeling. Our design approach opens new avenues for optically controlling enzyme function. The designed

photo-controlled cytosine deaminases may also aid in improving chemotherapy approaches that utilize this enzyme.

3.3 Introduction

Externally modulating the activity of an enzyme allows for the precise control over the location, timing, and amplitude of its activity[1]. Many successful advances in photo-control over enzymes have arisen through the use of photocaged unnatural amino acids[2], or fusing split enzymes to naturally occurring photoswitchable protein-protein interactions[3,4]. In many biological systems, however, it is beneficial to design regulatory mechanisms with reversible, photoswitching properties without splitting proteins, and to this end, labeling enzymes with azobenzene derivatives is an attractive option. Azobenzene is a photochromic molecule capable of undergoing *trans*-to-*cis* isomerization upon illumination with UV light and *cis*-to-*trans* isomerization via thermal relaxation or illumination with visible (VIS) light, which causes reversible end-to-end length changes[5,6]. Coupling these changes to the conformational modulation of the attached protein has enabled the design of a switchable cadherin[7], restriction enzyme[8], leucine zipper[9], and an allosterically modulated ionotropic glutamate receptor[10]. However, azobenzene attachment is generally performed using trial-and-error as a generalizable, predictive design protocol for the crosslinking of a protein structure with an azobenzene derivative is not available.

Directed enzyme prodrug therapy (DEPT) is a strategy that aims to reduce undesirable side effects of chemotherapy by selectively activating prodrugs at tumor sites while sparing normal tissue from damage[11]. The zinc-containing dimeric metalloenzyme cytosine deaminase from yeast (yCD) is a prototypical DEPT enzyme, that catalyzes the conversion of cytosine to uracil, and is used with the prodrug 5-fluorocytosine (5FC) to

generate the toxin 5-fluorouracil (5FU). A reversibly photoactivatable yCD variant would be desirable in DEPT to site-specifically induce toxicity with an external, non-invasive optical signal. Previous structural and mechanistic studies have shown that yCD activity involves a dynamic active site lid region comprising an amphipathic helical segment at its C-terminal[12–15]. The lid helix covers the active site, and is spatially proximal to the dimer interface of the enzyme.

3.4 Results and Discussion

We developed a generalizable computational protocol for the design of a photo-controlled yCD, a prototypical DEPT enzyme, *via* the crosslinking of two suitably placed cysteine residues with a chosen azobenzene derivative. In a first step, we used the RosettaMatch algorithm[16] to query all pairs of residue positions in a thermostabilized variant of yCD (called yCD-Triple[17]) crystal structure (PDB:1YSB) to find suitable cysteine substitutions that could sterically accommodate the attachment of an azobenzene derivative in the UV-light induced *cis*-conformation. We chose the bifunctional azobenzene derivative 4,4'-bis(maleimido)azobenzene (ABDM) (**Figure 3-1:A**) that can be attached with high efficiency to designed cysteines. The Rosetta Match algorithm requires the generation of a conformational ensemble of the azobenzene derivative and a set of defined parameters for the maleimide-cysteine attachment geometry, (**Figure 3-1:B**), which were derived from a search of the Cambridge Crystallographic Database (CCDC)[18]. Once geometrically compatible locations were obtained (**Figure 3-1:C**), further sequence optimizations were performed using RosettaDesign (**Figure 3-1:D**) and five designs were chosen for further characterization, based on (1) distance of the designed cysteine pairs from the active site, (2) Rosetta energies, and (3) the conformational strain at the azobenzene-cysteine connection (**Table 3-1**), and (4) the predicted ability of *cis-trans*

isomerization to induce distortions in the active site structures, as determined by Rosetta-based simulations of modified proteins. We illustrate the ABDM cross-linked models for two designs, called yCD-Opt1 and yCD-Opt2 (**Figure 3-1:E**).

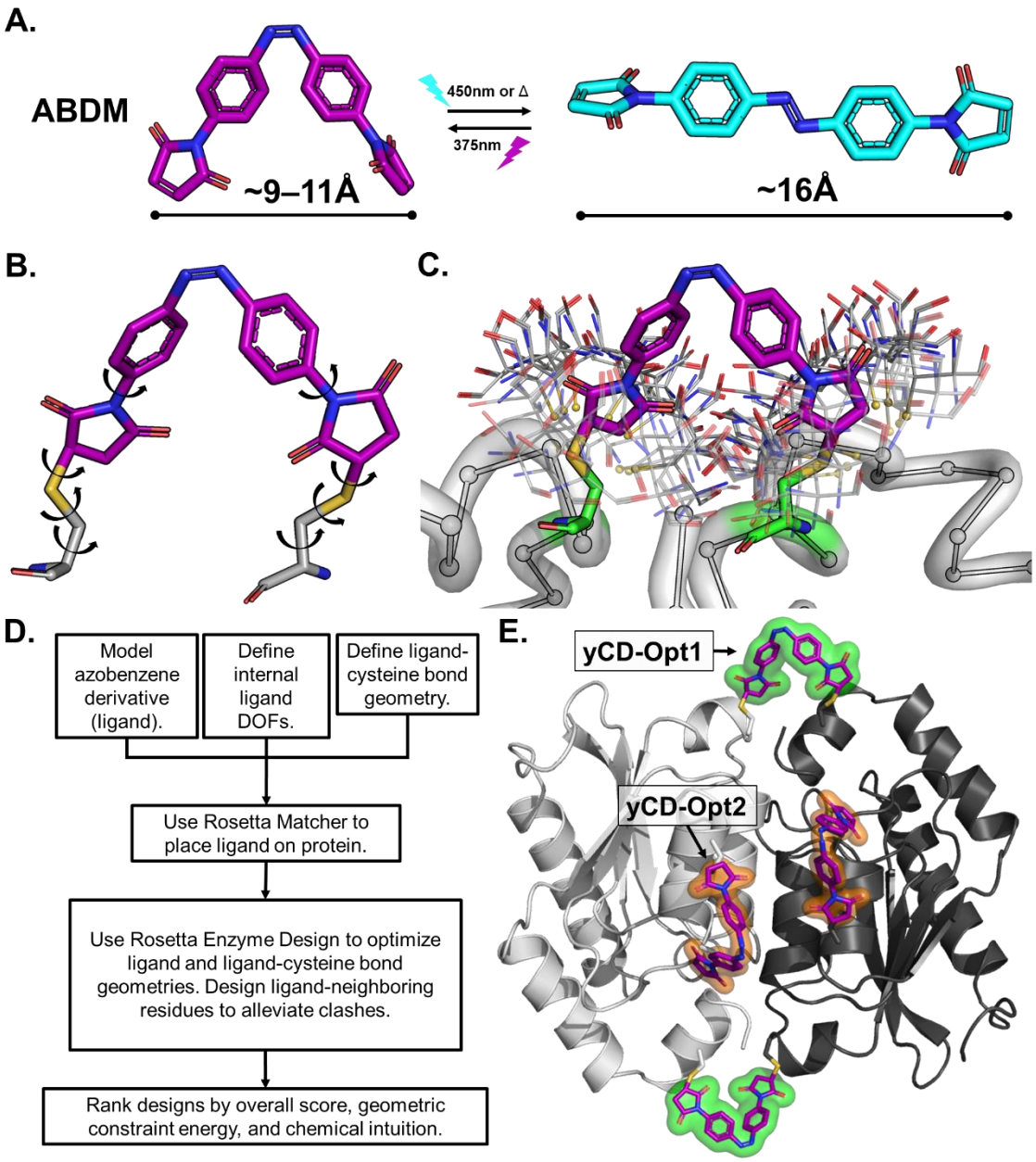


Figure 3-1: Computational design of ABDM-labeled yCD-dC3 variants.(A) Schematic of the *cis* (purple) and *trans* (cyan) states of ABDM. Isomerization from *cis* to *trans* is induced by UV light, while *trans* to *cis* is induced by blue light or thermal relaxation. The distance between maleimide groups in the *cis* state is shorter than in the *trans* state. (B) The rotatable degrees of freedom within the *cis*-ABDM ligand (purple) and between the

maleimide moieties and cysteine (grey) sidechains are shown with circular arrows. (C) The inverse rotamer ensemble (grey lines) generated by the specified degrees of freedom in (B) is shown surrounding *cis*-ABDM (purple). A pair of successful rotamers (green sticks) have been matched with a protein backbone (grey tube, Cα atoms shown as spheres). (D) Flowchart of the RosettaMatch-based design approach for *cis*-ABDM labeled proteins. (E) Designed yCD-Opt1 and yCD-Opt2 ABDM attachments to the yCD dimer.

Table 3-1: yCD Variants.

Design	Substitutions	Rosetta Score	Geometric Constraint Score	Distance from Active Site (Cys-Ca to ZN) (Å)	Experimentally Characterized?
yCD-Opt1	E75C, K77A, F153C	-509.385	0.451	22.7, 23.8	Yes.
yCD-Opt2	K117C, T124C	-517.987	2.727	11.7, 18.0	Yes.
yCD-Opt3	Q123C, V120C	-512.731	2.607	16.6, 16.8	No, did not express.
yCD-Opt4	K117S, K120C, T124C, R125C	-517.002	0.049	14.4, 16.7	No, did not express.
yCD-Opt5	K13C, S44C	-509.629	2.926	21.6, 23.5	No, inactive.

To computationally predict the impact of light-induced *cis-trans* isomerization on the active site structure, we simulated the *cis*-to-*trans* isomerization of the designed ABDM-crosslinked enzymes by "mutating" the azobenzene derivatives to their *trans* isomers and performing Rosetta FastRelax[19,20] simulations. We allowed for the regions flanking the secondary structure elements that contained the cysteine mutations to sample alternative backbone and sidechain movements in order to conservatively sample structural changes during the conformational switching (**Figure 3-2:A,B**). The degree of active site disruption and reconstitution was measured by Cα-RMSDs with respect to the native state.

The lowest scoring *trans* conformations of the yCD-Opt1 (**Figure 3-2:C**) and yCD-Opt2 (**Figure 3-2:D**) designs show widely different Cα-RMSDs in each of their respective subsets of moveable residues, indicating the yCD-Opt1 cross-linking position requires a

wider range of motion to accommodate the *trans* ABDM conformation as compared to yCD-Opt2. All of the movement exhibited by yCD-Opt1 in the *trans* state was shown to be in the swinging of active site lid helix α6 away from the active site pocket (**Figure 3-2:E**). In stark contrast, the design yCD-Opt2 shows little movement from the native conformation in the *trans*-ABDM state, demonstrating an ability to absorb the motion of the ABDM conformational switch using local flexibility in glycine residues adjacent to the helix (**Figure 3-2:F**).

For experimental testing, first, we reduced the number of non-designed crosslinkable residues by substituting three of the four non-catalytic native cysteine residues in yCD-Triple. The resulting variant, yCD-Triple-ΔC3, featured three mutations, C36A, C71S, and C137A, and this template was used as the backbone on which yCD-Opt1 (E75C/F153C) and yCD-Opt2 (K117C/T124C) substitutions were made. Both designed proteins were expressed in the *E. coli* strain BL21(DE3)pLysS using the pET15b expression vector, and purified with standard Ni^{2+}-affinity chromatography followed by thrombin cleavage to remove the His_6-tag (**Materials and Methods**). The cysteine pair in yCD-Opt1 is adjacent to one another on different monomers of the dimer and is expected to covalently link the dimer post-labeling (**Figure 3-1:E**). In yCD-Opt2, the pair of cysteines is located on the same interior alpha-helix, with the cysteines flanking a helical segment (**Figure 3-1:E**).

We first investigated the labeling efficiency of yCD-Opt1 and yCD-Opt2 with ABDM, as well as the ability of the post-labeled ABDM molecule to reversibly isomerize under UV and blue light signals. Denaturing and reducing SDS-PAGE results show an increased intensity of the dimer-sized molecular weight band in the labeled yCD-Opt1 sample versus the unlabeled yCD-Opt1 sample, indicating the existence of a majority

population of yCD-Opt1 species with at least one bridging ABDM molecule between the two monomers (**Figure 3-5:A**).

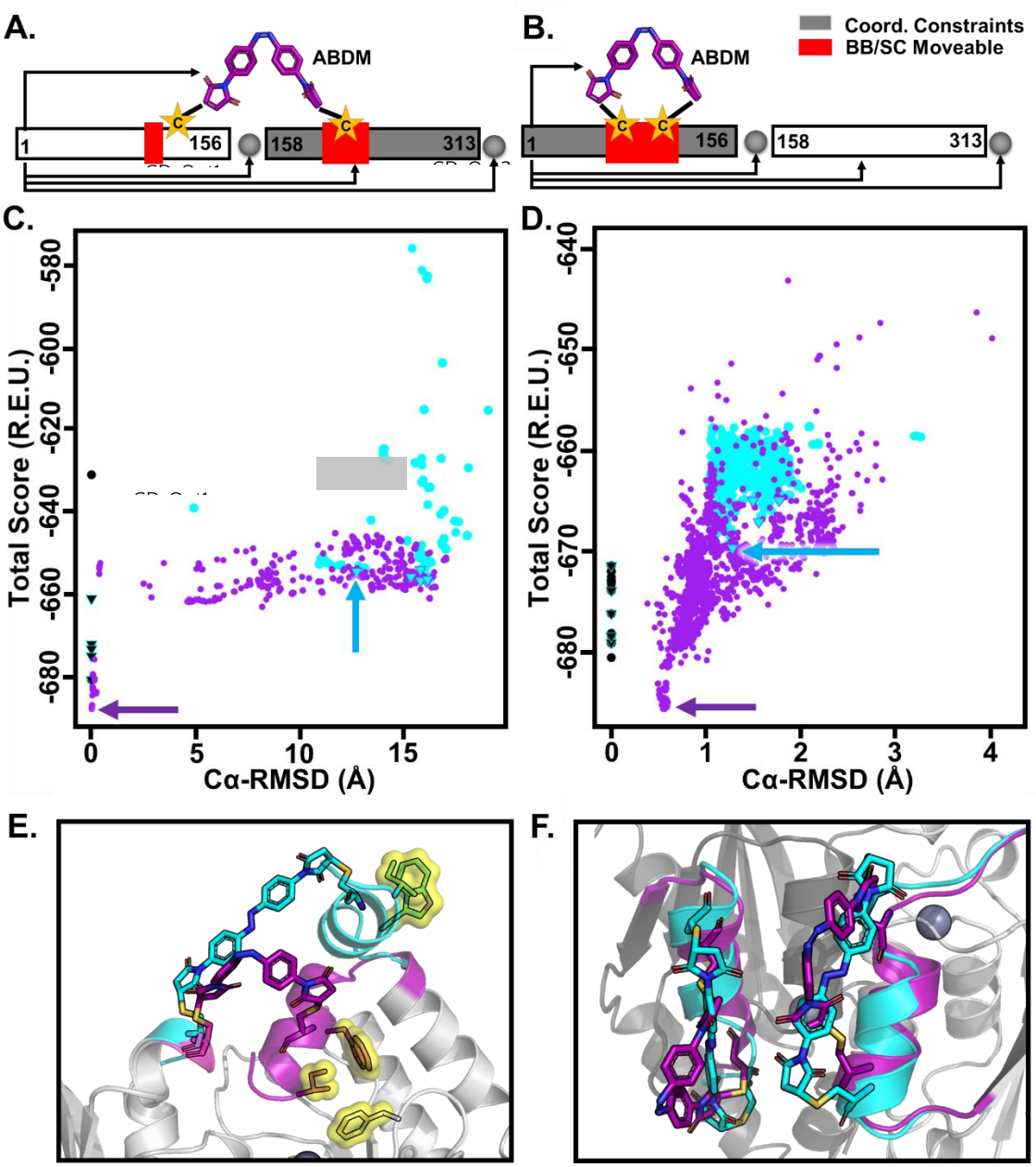


Figure 3-2: *Cis*-to-*trans* simulations of yCD-Opt1 and yCD-Opt2.Purple and cyan are used to denote the *cis* and *trans* states, respectively, throughout. The Rosetta FoldTree used during *cis*-*trans* isomerization simulations for (A) yCD-Opt1 and (B) yCD-Opt2. Cysteine substitutions are denoted with yellow stars. (C) Total Score (R.E.U.) vs Cα-RMSD of the terminal helix residues in the *cis*-Designed (black), *cis*-to-*trans*, and *trans*-to-*cis* states. The cyan points show that the *trans* state disrupts yCD-Opt1, while the purple points demonstrate that the *cis* state can retain the native structure. (D) Total Score (R.E.U.) vs Cα-RMSD of the residues within and flanking the helix in the *cis*-Designed (black), *cis*-to-*trans*, and *trans*-to-*cis* states of yCD-Opt2. Both cyan and

purple points have low Cα-RMSDs, indicating that neither conformation severely disrupts yCD-Opt2. (E, F) Lowest-energy decoys from the simulations in (C) and (D) (structures denoted with arrows). In (E), there is considerable displacement of two of three lid-region residues (yellow-bordered sticks), while (F) shows minor displacement of secondary structure near the active site (catalytic zinc is shown as grey sphere).

ESI-MS of yCD-Opt1 variants (**Figure 3-5:B**) revealed a mass change of 820 Da upon labeling, which was close to the expected change of 816 Da for two ABDM moieties per dimer, assuming the maleimide moieties were hydrolyzed during workup[8]. Electrophoretic patterns of yCD-Opt2, on the other hand, show a majority of the monomer species in both labeled and unlabeled samples although some labeling induced multimerization was observed indicating a more heterogeneous labeling pattern (**Figure 3-3:B**). We further tested the completeness of ABDM crosslinking by confirming the lack of reactivity of cysteine residues to a PEG-maleimide derivative post-crosslinking (**Figure 3-4**).

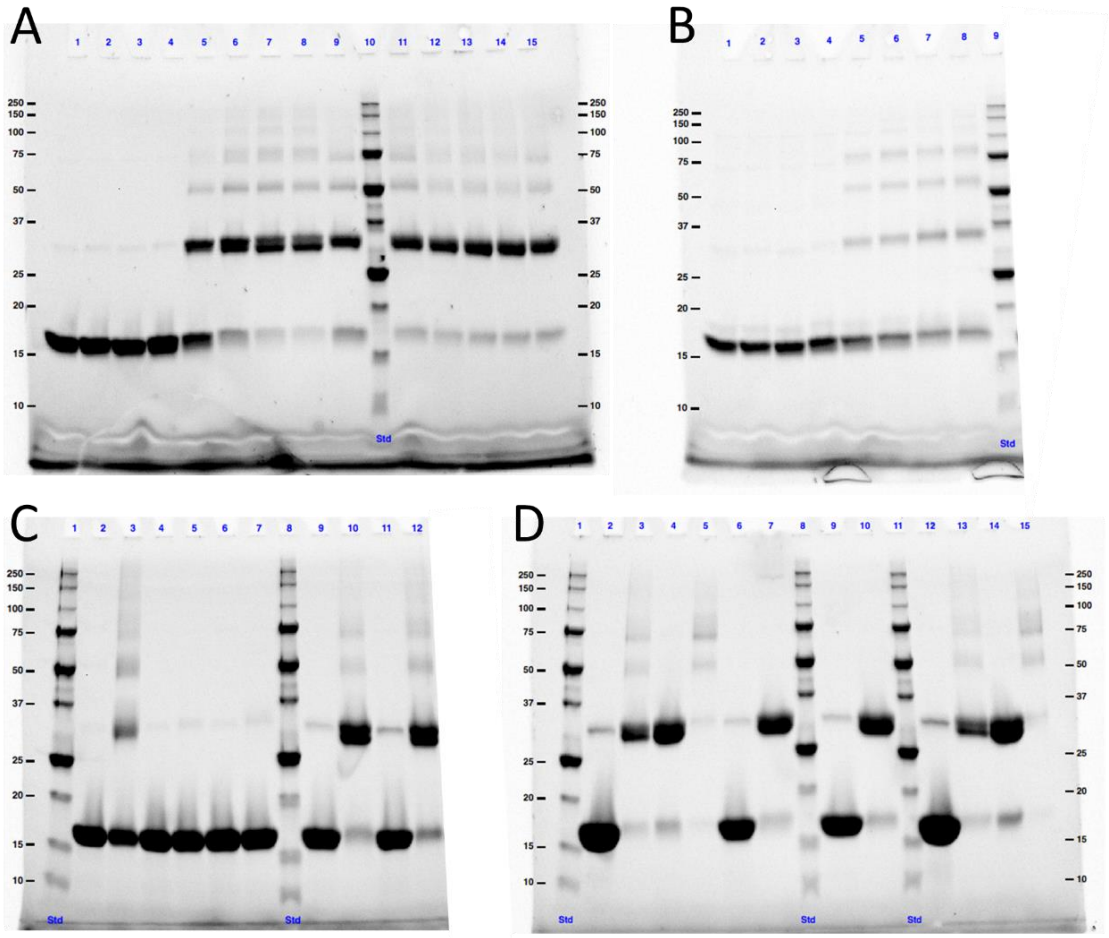


Figure 3-3: Denaturing, reducing SDS-PAGE Gels showing labeling patterns of azobenzene derivatives conjugated to yCD variants. (A) yCD-Opt1 unlabeled (lanes 1-4: -UV/-DMSO, -UV/+DMSO, +UV/-DMSO, +UV/+DMSO) compared to labeled yCD-Opt1 samples with an iodoacetamide azobenzene derivative (lanes 5-9: 1:1 enzyme:azobenzene molar ratio, 1:2 molar ratio, 1:3 molar ratio, 1:4 molar ratio, 1:5 molar ratio) or ABDM (lanes 11-15: 1:1 molar ratio, 1:2 molar ratio, 1:3 molar ratio, 1:4; molar ratio, 1:5 molar ratio). Lane 10 is a standard, with molecular weights for each band labeled on either side of the gel. (B) yCD-Opt2 unlabeled (lanes 1-4: -UV/-DMSO, -UV/+DMSO, +UV/-DMSO, +UV/+DMSO) compared to labeled yCD-Opt2 samples with an iodoacetamide azobenzene derivative (lanes 5-6: 1:1 enzyme:azobenzene molar ratio, 1:2 molar ratio) or ABDM (lanes 7-8: 1:1 molar ratio, 1:2 molar ratio). Lane 9 is the standard, with molecular weights labeled on the side of the gel. (C) Labeling of yCD variants with ABDM at a 1:4 molar ratio. Lanes 2-3: yCD-Triple Unlabeled/Labeled. Lanes 4-5: yCD-Triple-ΔC2 Unlabeled/Labeled. Lanes 6-7: yCD-Triple-ΔC3 Unlabeled/Labeled. Lanes 9-10: yCD-Opt1 mutations on the yCD-Triple-ΔC2 background Unlabeled/Labeled. Lanes 11-12: yCD-Op1 Unlabeled/Labeled. Lanes 1 and 8 are standards, and the molecular weights are labeled on the side of the gel. (D) Pre- and post-gel filtration of labeled yCD-Opt1 variants to remove aggregates. Lanes 2-7 are data for yCD-Opt1 on yCD-Triple-ΔC2 background (lane 2: concentrated unlabeled sample, lane 3: concentrated labeled sample, pre-gel filtration, lane 4: concentrated labeled sample, post-gel filtration and pooling of dimer species, lane 5: unlabeled sample at 50 μM, lane 6: labeled sample post-gel filtration at 50 μM). Lanes 9,10,12-15 are data for yCD-Opt1 (lane 9: unlabeled sample at 50 μM, lane 10: labeled sample post-gel filtration at 50 μM, lane 12: concentrated unlabeled sample, lane 13: concentrated labeled sample, pre-gel filtration, lane 4: concentrated labeled sample, post-gel filtration and pooling of dimer species, lane 15: concentration labeled sample, post-gel filtration and pooling of trimers and higher species).

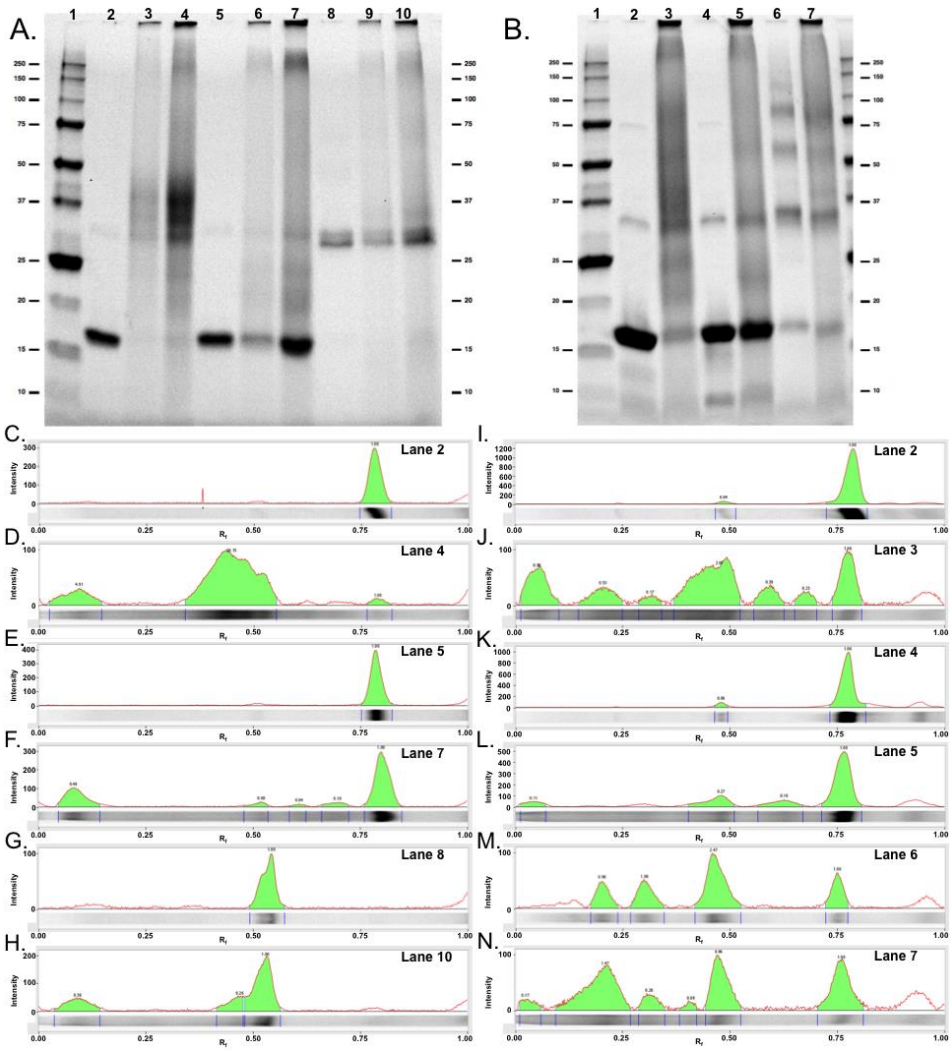


Figure 3-4: Post-crosslinking analysis via free-cysteine PEGylation.(A) Denaturing, reducing SDS-PAGE gel of yCD-Opt1 showing unlabeled, free-maleimide (FM) labeled, and ABDM-labeled samples, as well as their incubation with PEG-mal post-crosslinking. Lanes 2-4: Unlabeled sample, unlabeled sample incubated with PEG-mal, concentrated unlabeled sample incubated with PEG-mal. Lanes 5-7: FM-labeled sample, FM-labeled sample incubated with PEG-mal, concentrated FM-labeled sample incubated with PEG-mal. Lanes 8-10: ABDM-labeled sample, Lane 1 is a protein standard. (B) Denaturing, reducing SDS-PAGE gel of yCD-Opt2 showing unlabeled, free-maleimide (FM) labeled, and ABDM-labeled samples, as well as their incubation with PEG-mal post-crosslinking. Lanes 2-3: Unlabeled sample, unlabeled sample incubated with PEG-mal. Lanes 4-5: FM-labeled sample, FM-labeled sample incubated with PEG-mal. Lanes 6-7: ABDM-labeled sample, ABDM-labeled sample incubated with PEG-mal. (C-H) Intensity analysis of select lanes from yCD-Opt1 gel (A). (C-D) Calculated intensities of bands found in unlabeled and unlabeled+PEG-mal lanes, each relative to the monomer peak (or approximated monomer peak) for that lane. (E-F) Similar to C-D but for FM-labeled lanes. (G-H) Similar to C-D but for ABDM-labeled lanes. (I-N) Intensity analysis of lanes from yCD-Opt2 gel (B). (I-J) Calculated intensities of bands found in unlabeled and unlabeled+PEG-mal lanes, each relative to the monomer peak. (K-L) Similar to I-N but for FM-labeled lanes. (M-N) Similar to I-J but for ABDM-labeled lanes.

To investigate the conformational changes upon labeling in yCD-Opt1, we performed ^{19}F-NMR spectroscopy with protein samples containing genetically incorporated 5-fluorotryptophan (5-FW) (**Figure 3-5:C**). yCD-Triple and yCD-Opt1 contain two tryptophan residues: W10 (distal from active site) and W152 (in the active site lid). Comparisons of the spectra of yCD-Triple and yCD-Opt1 (before and after labeling) indicate that one tryptophan residue (likely W152) experiences an altered environment due to cysteine mutations and labeling by ABDM (detected in the *trans* state), in accord with our simulations (**Figure 3-2** and **Figure 3-5:C**). Intrinsic tryptophan fluorescence experiments also showed clear differences in UV-irradiated and blue-irradiated ABDM-labeled yCD-Opt1 spectra further supporting the change in the environment of W152 in the active site lid upon irradiation (**Figure 3-5:D**).

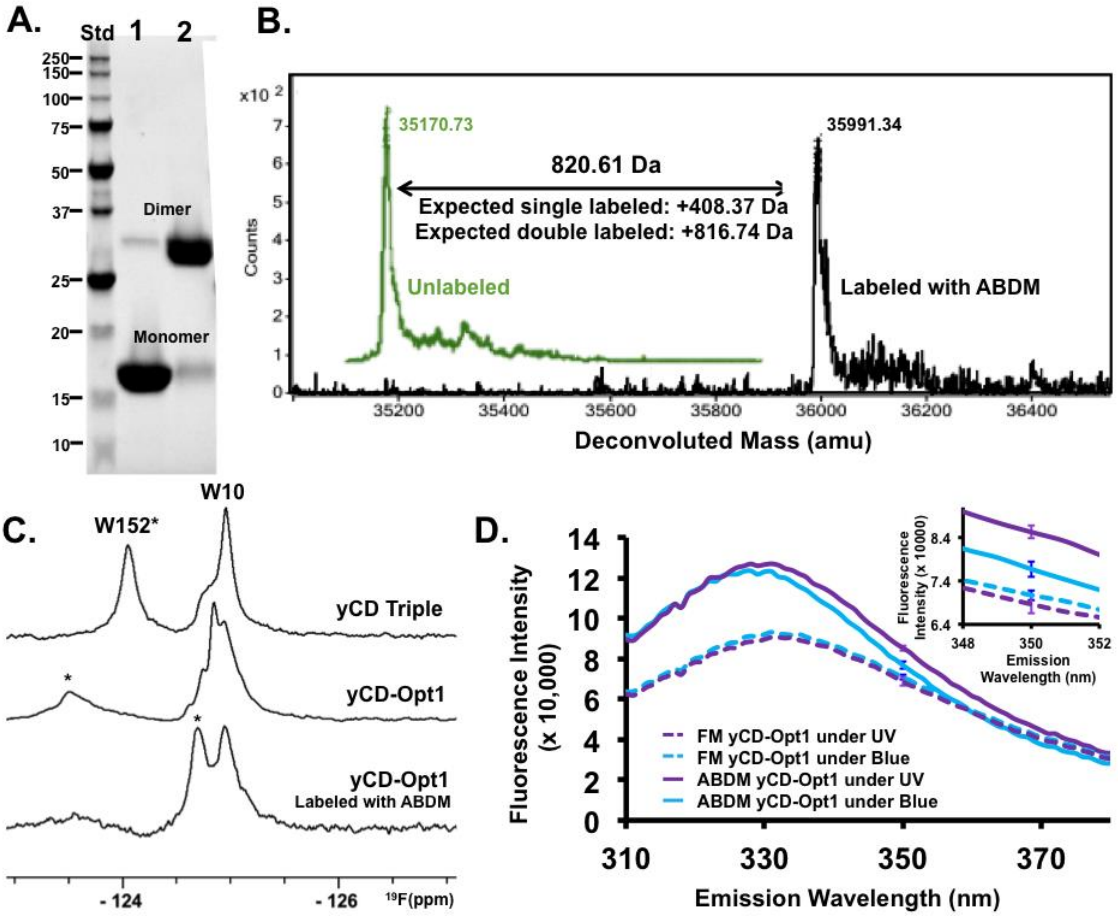


Figure 3-5: yCD-Opt1 Labeling with ABDM.(A) yCD-Opt1 SDS-PAGE Gel showing unlabeled (Lane 1) and labeled (Lane 2) yCD-Opt1. (B) ESI mass spectra of unlabeled (green) and cis-ABDM labeled (black) yCD-Opt1. (C) 19F NMR spectra of 5-F-Trp-labeled yCD Triple, yCD-Opt1, and Labeled yCD-Opt1. Resonances of W10 and W152 (denoted with an asterisk) have been inferred from the proximity of W152 to the yCD-Opt1 mutations and the ABDM molecule in the labeled sample. (D) Fluorescence emission spectra of yCD-Opt1 labeled with free-maleimide (FM) or ABDM under blue and UV light. The inset highlights the differences between the spectra at an emission wavelength of 350 nm, where error bars are +1 SD from three independent experiments.

Repeated irradiation cycles of UV and blue light demonstrated the reversible photoisomerization of ABDM post-crosslinking with no loss of intensity over 15 cycles for yCD-Opt1 (**Figure 3-6:A**). In addition, labeled yCD-Opt1 is able to produce similar and right-shifted characteristic absorbance spectra changes when illuminated with blue light (455 nm) followed by UV light (375 nm) compared to the uncrosslinked azobenzene derivative, in which the *trans* isomer has a local maximum absorbance at 363 nm that decreases upon illumination (**Figure 3-6:B**). The half-life for thermal relaxation of the

azobenzene moiety in the context of bound ABDM to yCD-Opt1 is approximately 90 minutes at 25°C (**Figure 3-6:B**). These experiments were not performed with yCD-Opt2 as it did not show activity changes upon irradiation (*vide infra*). The longer half-life of the ABDM upon attachment to the yCD-Opt1 protein compared to its reported half-life in the unattached state (~30 minutes[21]) indicates that the protein structure stabilizes the otherwise higher-energy *cis* state, and indicates that further modulation of the half-life of the more active state, an important parameter for chemotherapeutic utilization of the enzyme, may be feasible by altering the local stability of the protein environment.

Next, we tested the activity of the cross-linked yCD variants using a previously-described cytosine consumption assay[22]. The cross-linked yCD variant, yCD-Opt1, showed increased activity under UV light when the azobenzene adopted predominantly the *cis* conformation compared to the *trans* (**Figure 3-6:C, Figure 3-7:A**). The second yCD variant, yCD-Opt2, showed no difference in activity toward cytosine under both UV and blue light after labeling, and the activities of both were higher than the unlabeled yCD-Opt2. Gratifyingly, this behavior of yCD-Opt1 and yCD-Opt2 is in agreement with our computational prediction (**Figure 3-2:D**,**F**).

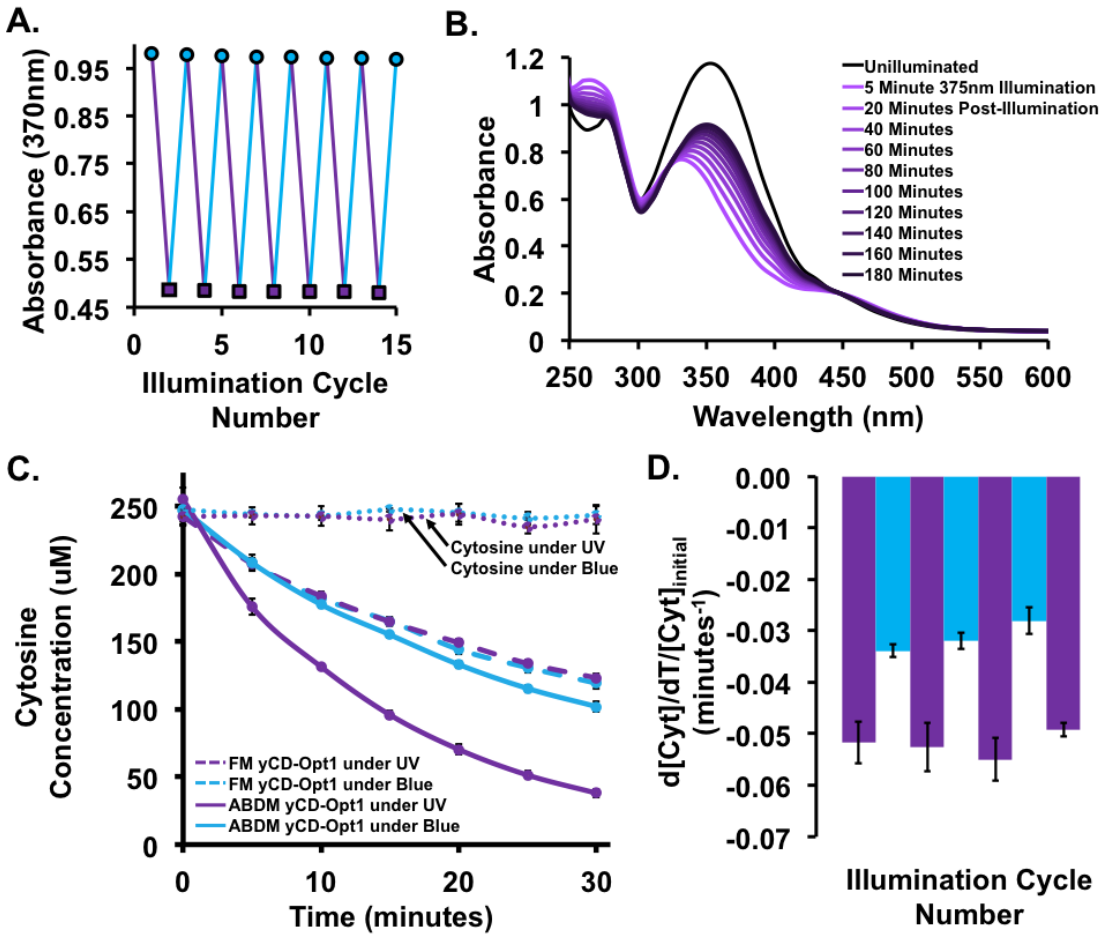


Figure 3-6: Bound ABDM characterization and optical control of cross-linked yCD-Opt1 cytosine deaminase activity.(A) The photoisomerizability of bound ABDM under UV (purple squares) and blue (cyan circles) light is reversible. (B) Thermal relaxation of bound ABDM from the *cis*-state (purple) toward unilluminated, *trans*-state (black) monitored for three hours at 25°C. (C) Cytosine concentration vs. time with FM-labeled (short-dashed lines) and cross-linked (solid lines) yCD-Opt1 under UV and blue light, and cytosine-only controls under UV and blue light (dotted lines). Error bars are +1 s.d. from four independent experiments. (D) Slopes of cytosine consumption assays with labeled yCD-Opt1 as a function of repeating five-minute intervals of UV light (purple bars) followed by blue light (cyan bars), normalized to the starting cytosine concentration for that illumination cycle.

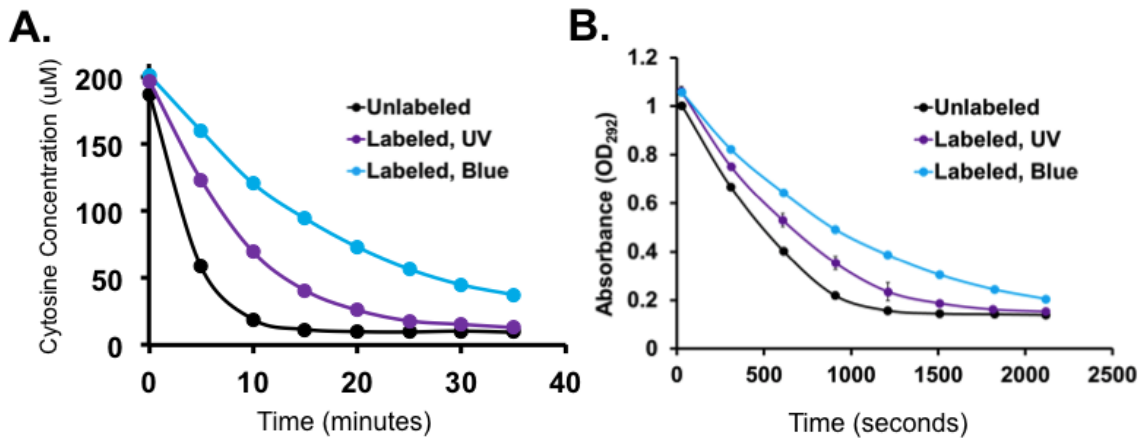


Figure 3-7: yCD-Opt1 activity assays.(A) yCD-Opt1 cytosine consumption assay. Cytosine concentration (μM) as a function of time with unlabeled yCD-Opt1 (black line), labeled yCD-Opt1 under UV light (purple line) and labeled yCD-Opt1 under blue light (blue line) Enzyme concentration = 150 nM, substrate concentration at time zero = 250 μM. (B) yCD-Opt1 5-fluorocytosine consumption assay. Absorbance at OD292 as a function of time with unlabeled yCD-Opt1 (black line), labeled yCD-Opt1 under UV light (purple line), and yCD-Opt1 under blue light (blue). Enzyme concentration = 333 nM, substrate concentration at time zero = 1 mM.

Further characterization of yCD-Opt1 specific activity and Michaelis Menten kinetics under UV and blue light indicated that UV-light irradiation leads to an approximately 2.5-fold increase in the specific activity of the enzyme compared to the dark state (**Figure 3-6:C**), and that the increase appears to primarily affect the k_{cat} value (2.2 ± 0.10 s^{-1} and 1.4 ± 0.07 s^{-1} under UV and blue light, respectively; **Table 3-2**). These results are consistent with previous mechanistic studies which have found that product release is rate-limiting in the overall mechanism of the enzyme[13,15], and Trp152 plays a key role in this process. Our NMR and fluorescence experiments suggest that the environment of Trp152 is altered in yCD-Opt1 by labeling and irradiation, potentially leading to the observed changes in k_{cat}.

Table 3-2: Michaelis-Menten kinetics for yCD-Opt1 under UV or blue light condtions.

	K_m (mM substrate)	kcat (M Prod / sec M E)	k_{cat}/K_m (1 / sec M E)
UV Light (*cis*)	0.532 ± 0.059	2.20 ± 0.10	4131 ± 192
Blue Light (*trans*)	0.473 ± 0.056	1.40 ± 0.07	2946 ± 141

These values of the switching ratio for both specific activity and Michaelis-Menten parameters are likely underestimated because not all of the UV-irradiated protein contains the ABDM moiety in the *cis* state.

Optical control over the activity of yCD-Opt1 could be observed in real time by switching the external signal from UV light to blue light over several five-minute iterations during the course of the assay (**Figure 3-6:D**). The more negative slopes of each UV-light interval compared to each blue-light interval after normalizing the slopes to the starting concentrations of cytosine at the beginning of each interval indicates that the photomodulation of yCD-Opt1 is nearly fully reversible.

We next tested the photosensitized activity of yCD-Opt1 against its DEPT prodrug, 5FC, and found a smaller (~1.5-fold) difference in activity between UV and blue-irradiated samples compared to the cytosine assays (**Figure 3-7:B**). The *cis*-conformation azobenzene derivative has a higher activity toward 5FC than the *trans* version, but both are closer in activity to the unlabeled yCD-Opt1 than in the cytosine case.

Having demonstrated the photosensitization of yCD-Opt1 (**Figure 3-6**), we investigated whether the light-induced change in activity was due to global destabilization of the enzyme. We used CD thermal melt experiments to probe protein stability under increasing temperatures, and found similar melting curves and temperatures for the design starting point variant (yCD-Triple) and yCD-Opt1 labeled with ABDM in the *trans* conformation (**Figure 3-8**), which further suggest that a local change in structure as detected by ^{19}F-NMR and intrinsic fluorescence measurements is likely the cause of the photosensitization rather than an overall global instability imparted to the enzyme by the ABDM cross-link in the *trans* state.

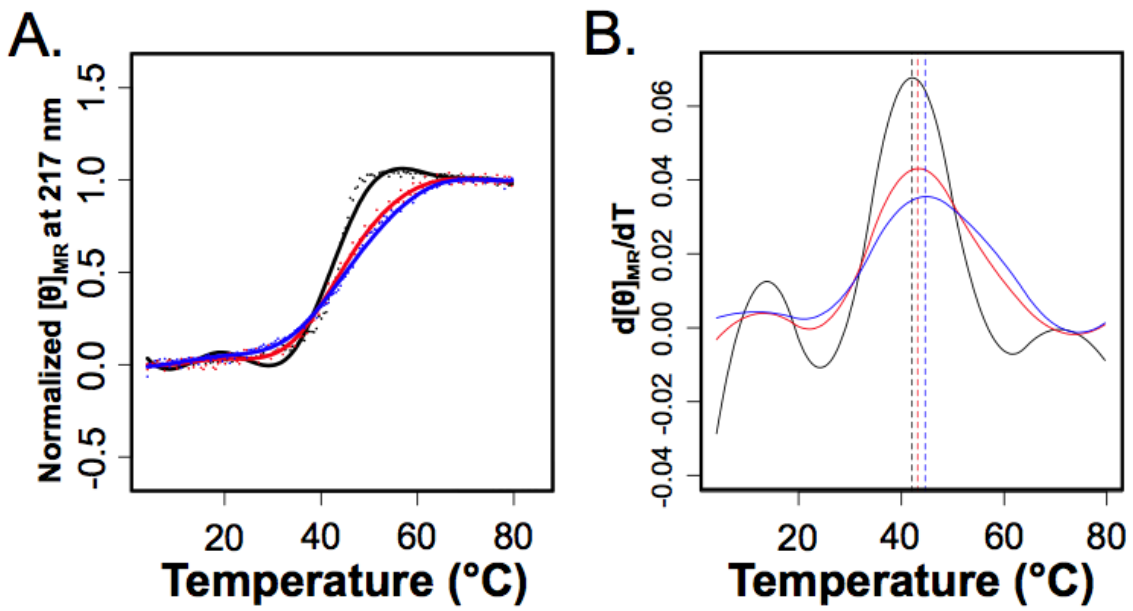


Figure 3-8: CD Thermal Melt Spectra.(A) CD spectra as a function of temperature comparing the melting spectrum of yCD-Triple (black line) to yCD-Opt1 labeled with ABDM (red line) or an iodoacetamide azobenzene derivative (blue line). (B) Analysis of the midpoint of the unfolding transitions from (A) show that the melting temperatures of labeled yCD-Opt1 are higher than yCD-Triple.

3.5 Conclusion

In summary, we have developed a predictive computational approach for designing photo-control over enzyme activity, and validated our method using the therapeutically relevant enzyme yCD. Designs that were predicted to show (yCD-Opt1) and to not show (yCD-Opt2) photoswitchability behaved in accord with the computational models. Our results provide some insights into design rules enzyme photoswitchability using azo-benzene crosslinking: labeling sites must be placed close enough to the active site to affect substrate binding or catalytic turnover, but not so close as to disrupt catalytic activity in a non-switchable manner. In comparing the two yCD variants, it appears that placing the azobenzene cross-linker in a position that reversibly disrupts local dynamics that are required for catalysis (lid-opening and closing) is likely to be a good strategy for selection of azobenzene attachment sites. It is also likely that greater conformational effects

due to photoswitching are inducible at protein termini as opposed to internal positions. While further improvements in switching efficacy, especially with 5FC, are being pursued in our laboratory, we expect that the described computational method should assist with the predictive design of azobenzene attachment-mediated photo-control over a variety of other enzymes.

3.6 Methods:

3.6.1 Computational Methods

Crystal structures of azobenzene in the *cis*-conformation (AZBENC01) and *trans*-conformation (AZOBEN12) were downloaded from the Cambridge Crystallographic Data Centre (CCDC). The average C-N=N-C dihedral angle was investigated using the CCDC Mogul software[23] in order to derive parameters for the torsion value necessary to generate an azobenzene-derivative conformational library. The *cis*-conformation, which was found as a substructure in 56 crystal structures, was subsequently sampled by the *gen_ligens* executable in Rosetta at 0°+- 10° in increments of 10° and the *trans*-conformation, which was found as a substructure in 3117 of the structures, was fixed at 180 degrees (**Figure 3-9:A**).

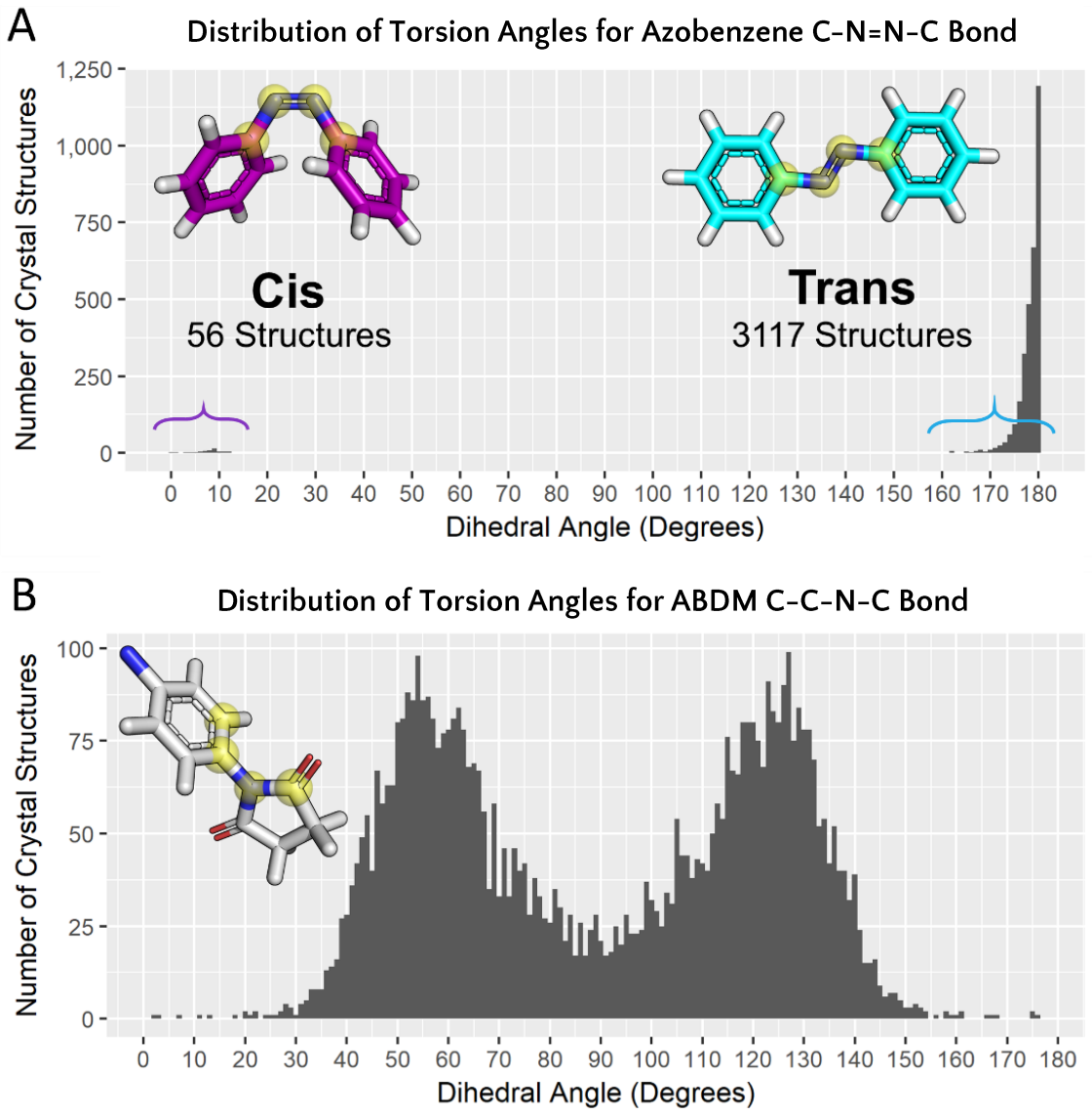


Figure 3-9: ABDM internal degrees of freedom.(A) Histogram of the dihedral angle from the four interior atoms (highlighted with yellow spheres) of all structures in the CCDC containing azobenzene as a substructure. Cis-conformations show a dihedral range between 0° and 12°, while the more numerous trans-conformations show a range between 160° and 180°. (B) Histogram of the dihedral angle formed by two carbon atoms on the benzene ring and a nitrogen and carbon atom on the maleimide substituent of ABDM.

The maleimide substituents of ABDM were built in Avogadro[24] at the para-position to the NN bond on the both benzene rings to give an initial structure of ABDM. Molecular geometry was optimized using default optimization protocol. The dihedral angle between the maleimide substituent and benzene ring was investigated for torsional value

distribution in all structures in the CCDC (**Figure 3-9:B**) and was allowed to sample 40° to 140° in increments of 10° during conformer library generation.

Geometric restraints describing the cysteine-maleimide bond interaction were determined by probing the CCDC database for similar interactions found in substructures of crystal structures. The C-C-S angle from two ABDM carbons to the cysteine sulfur atom were determined by finding all structures in the CCDC database containing the fragment shown in **Figure 3-10:A**, and so was set to 112° +- 6° degrees (**Figure 3-10:A**). The C-S-C angle from one ABDM carbon to the S-Cβ of the cysteine sidechain was determined by substructure searches of the fragment shown in Supplementary Figure S6B, and subsequently set to 101° +- 3° degrees. The torsion involving atoms N-C-C-S, where the N-C-C bonded atoms are from the maleimide fragment of ABDM and the sulfur atom is from the cysteine sidechain was set to 120° +- 20 degrees, 240° +- 20° degrees, or 0° +- 20° degrees, where structures containing 0° dihedrals were removed from the set of decoy structures post-design. The torsion C-C-S-C, where the first C-C bond was from the maleimide substituent of ABDM and the S-C bond from the cysteine sidechain was left free, while the torsion C-S-C-C, in which the first carbon came from ABDM and the S-C-C consisted of the sulfur-Cβ-Cα atoms from cysteine was set to 60° +- 20°, 180° +- 20°, or 300° +- 20°, as found in the Dunbrack backbone independent rotamer library.

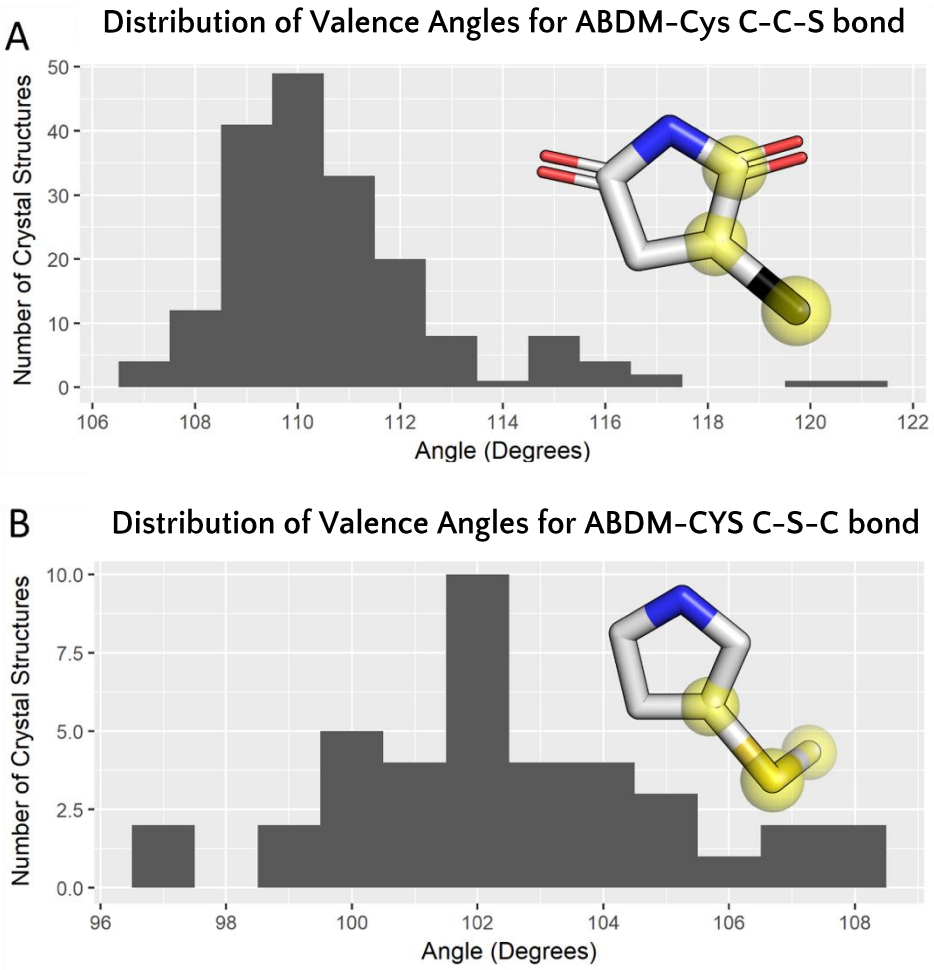


Figure 3-10: ABDM-cysteine geometric degrees of freedom.(A) Histogram of the angle between two atoms on a maleimide fragment (highlighted with yellow spheres) to any atom (shown in black, highlight with a yellow sphere). (B) Histogram of the angle between one carbon on a pyrrolidine fragment and the sulfur-Cβ bond of a cysteine sidechain.

The crystal structure of yCD-Triple (1YSB) was downloaded from the RCSB PDB and the following substitutions were made in PyMOL to generate the starting structure of yCD-Triple-ΔC3: C36A, C71S, C137A. This structure was subjected to 50 independent simulations of RosettaFastRelax[19,20] to relieve internal clashes and optimize sidechain rotamers.

The initial structure of ABDM as a Rosetta parameters file, the conformer library of ABDM, the relaxed structure of yCD-Triple-ΔC3, a list of all possible residue positions in

yCD-Triple-ΔC3 for azobenzene-derivative conjugation, and the desired restraints of the external geometry involved between the protein sidechains and the azobenzene derivative, were used as inputs to the RosettaMatch application[16,25]. Briefly, the Rosetta matcher finds positions on a protein backbone scaffold capable of accommodating the binding site for a small molecule. The output structures underwent iterative rounds of constraint optimization and neighboring residue identity and rotamer minimization by the RosettaEnzDes[25] algorithm.

Cis-to-*trans* and *trans*-to-*cis* isomerizations of the azobenzene moiety were simulated by "mutating" ABDM to the desired conformation in Rosetta and allowing Rosetta FastRelax to adjust certain residue backbone and sidechain geometries to re-bridge the cysteine-ABDM-cysteine interaction. Residues that were allowed to move during the simulation of yCD-Opt1 included the entire loop region flanking one of the two cysteine substitutions and the adjacent loop region N-terminal to the second cysteine. Moveable residues in the yCD-Opt2 case included the loop residues adjacent to each cysteine and the helical region interior to both cysteines.

3.6.2 Experimental Methods

3.6.2.1 Cloning

The yCD-Triple[17] gene cloned into pET15b with a thrombin cleavable 6-His tag (gift from Dr. Barry Stoddard) was used as the template for site-directed mutagenesis to generate all yCD variants (**Table 3-1**). PCR amplified plasmids were transformed into XL10-Gold cells (Stratagene). Plasmid DNA was isolated from cultures of individual colonies using the GeneJET Plasmid Miniprep Kit (Thermo Fisher Scientific) and the sequence was

confirmed by DNA sequencing (GenScript). Positive sequences were transformed into BL21(DE3)pLysS chemically competent cells (Novagen) for expression.

3.6.2.2 Protein Expression and Purification

Individual colonies were used to inoculate a 10 ml culture of Luria-Bertani (LB) containing 100 µg/ml ampicillin and 30 µg/ml chloramphenicol. Cultures were grown at 37°C overnight, and 10 ml of the cultures were subcultured into 500 ml of autoinduction media[26] supplemented with 250 µM ZnSO4. Cultures were grown at 37°C for 3 hours, then 18°C overnight.

Cells were harvested by centrifugation at 4,000xg for 25 minutes, and the pellets were resuspended at 40mL/L of culture in lysis buffer (300 mM NaCl, 20 mM Tris-HCl, pH 7.5, 50 mM imidazole), and supplemented with 1 mM phenylmethylsulfonyl fluoride (PMSF), DNase I, and lysozyme. Cells were lysed using sonication, and cell lysates were cleared by centrifugation at 21,000xg for 45 minutes at 4°C.

A packed bed of Ni-NTA resin (Qiagen) at 4 ml/L of culture was equilibrated with lysis buffer in a 20 mL gravity-flow chromatography column, and the lysis supernatant was applied to the resin twice. The resin was subsequently washed with 20 ml of lysis buffer, and the protein was eluted from the column using 15 ml of 300 mM NaCl, 20 mM Tris-HCl, pH 7.5, 300 mM Imidazole. Dithiothreitol (DTT) was added at a 2 mM final concentration to keep the proteins from precipitating.

Protein samples were desalted into activity buffer (300 mM NaCl, 20 mM Tris-HCl, pH 7.5) using Sephadex G-25 PD-10 desalting columns (GE #1708510) and supplemented with 2 mm DTT. Thrombin was added at 6U/mg protein and samples were nutated at room temperature for one hour. 6XHis tags, contaminating proteins, and uncleaved protein

were removed using a second round of Ni-NTA purification by equilibrating a 2 ml packed bed of resin with activity buffer and applying the Thrombin-cleaved supernatants to the resin. Resin was washed with 15 ml activity buffer and combined with the supernatant flow-through. Protein purity was verified with SDS-PAGE.

For 5-fluoro-tryptophan incorporation, cultures of pLysS cells expressing the yCD-Triple and yCD-Opt1 variants were grown at 37°C in MJ9 minimal medium[9] containing ^{12}C-glucose and $(^{14}NH_4)_2SO_4$ as the sole carbon and nitrogen sources and supplemented with chloramphenicol and ampicillin. Incorporation of 5-F-Trp into yCD-Triple and yCD-Opt1 was performed as previously described[10]. Protein purification was carried out in the same manner as described above.

3.6.2.3 Protein Labeling with ABDM

4,4'-bis(maleimido)azobenzene (ABDM) (gift from A. Woolley) was freshly reconstituted in DMSO before use and the concentration was determined using the absorbance value at 370 nm determined by a NanoDrop Microvolume Spectrophotometer. The extinction coefficient used was ε_{370}=29000 M^{-1} cm^{-1}.

Cleaved and purified protein samples were incubated at room temperature for 2.5 hours with 2 mM final concentration of freshly reconstituted DTT to ensure adequate reduction of cysteine sidechains. Samples were desalted into 300 mM NaCl, 20 mM Tris-HCl, pH 8.5 using PD-10 columns and diluted to 10 µM. ABDM was added at a 4-fold molar excess over enzyme concentration to the desalted and diluted protein samples under constant UV light (Thor Labs, M375L3) to keep the ABDM in the *cis*-conformation. Specifically, one-third of the total volume of ABDM to be added to the protein samples was added every 10 minutes for 30 minutes at room temperature. During each of the 10-

minute intervals, the stock solution of ABDM was irradiated for 5 minutes before the addition to the protein samples but kept in the dark otherwise. After the last addition of *cis*-ABDM, excess ABDM and protein aggregates were removed *via* desalting using a HiLoad 26/600 Superdex 200 pg column (GE #28-9893-36) attached to an FPLC system and equilibrated with activity buffer. Dimer fractions were pooled. "Unlabeled" samples were made by adding an equivalent amount of DMSO to duplicate desalted and diluted protein samples, irradiating with UV light for 30 minutes, and desalting into activity buffer using PD-10 columns. "Free-maleimide" labeled samples were made by adding an equivalent amount of DMSO with a large excess of free-maleimide to duplicate desalted and diluted protein samples such that the concentration of free-maleimide was in 50-molar excess of protein. These were also irradiated with UV light for 30 minutes and desalted into activity buffer using PD-10 columns.

Aliquots of post-crosslinking samples were removed after the last addition of *cis*-ABDM, DMSO, or DMSO+maleimide, and incubated at room temperature with a 50-molar excess of poly(ethylene glycol) methyl ether maleimide, average M_n 2,000. (Sigma-Aldrich). These samples were concentrated and run on denaturing, reducing SDS-PAGE gels for analysis.

3.6.2.4 Protein Concentration Determination

Pre-labeled protein concentrations were determined using OD_{280} in conjunction with molecular weights and extinction coefficients predicted by the ExPASy Server's ProtParam tool[11] for determining various physicochemical parameters, which are yCD-Triple (His tag cleaved) = 17794 Da, yCD-Opt1 (cleaved) = 17587 Da, yCD-Opt2 (cleaved) = 17691 Da. The extinction coefficient used for all yCD variants was 19940 M^{-1} cm^{-1}. Post-

labeling, protein sample concentrations for labeled and unlabeled samples were determined using a Bradford Assay (Sigma Aldrich).

3.6.2.5 Isomerization Relaxation Measurements

An absorbance scan from 320 to 600 nm of 100 µL of 66 µM thermally-relaxed labeled yCD-Opt1 in activity buffer a 96-well UV-transparent was measured using a Tecan Infinite M200 Pro plate reader. The sample was irradiated with a UV LED light for 5 minutes and another absorbance scan was taken immediately, followed by absorbance scans taken every 20 minutes for 3 hours.

3.6.2.6 Isomerization Reversibility

Labeled yCD-Opt1 at 66 µM in activity buffer was illuminated with UV light for 5 minutes and the absorbance at 370 nm was measured. The protein was illuminated with blue light (Thor Labs, M455L3) for 5 minutes and the absorbance at 370 nm was again measured. This process was repeated for 15 cycles.

3.6.2.7 Circular Dichroism (CD)

CD data were collected on an Aviv model 420SF CD spectrometer with a five cuvette rotor. Temperature melts were run with 5 µM protein in activity buffer using 1 mm pathlength cuvettes and a wavelength of 217 nm. A temperature range of 5°C to 80°C in 0.5°C increments was used, and temperature was regulated with a Peltier device. Samples were allowed to equilibrate for 2 minutes before measurements were collected, and measurements were averaged over 10 seconds. Identical measurements of buffer-only samples were taken for each cuvette, and the buffer spectra were subtracted from the sample spectra from the same cuvette.

3.6.2.8 Mass Spectrometry

Electrospray Ionization (ESI) mass spectrometry was carried out on samples desalted into activity buffer and concentrated at the Advanced Instrumentation for Molecular Structure (AIMS) laboratory at the University of Toronto.

3.6.2.9 ^{19}F NMR

1D ^{19}F NMR spectroscopy was performed locked and at 25°C on a Bruker 600MHz NMR spectrometer equipped with a 5mm Bruker TCI cryoprobe. A 1.7mm tube was used in the 5mm probe using 45µL of 400 µM sample. A sweep width of 8.416 ppm, an acquisition time of 1.726 s, a 1.0s delay time, and a 10 µs 90° pulse length were used. 1D ^{19}F NMR spectra were processed with 10 Hz line broadening and displayed using Bruker TopSpin 3.5. The buffer used for ^{19}F NMR was the same as for the cytosine consumption assay but diluted to 10% ^{2}H2O. The number of transients for yCD-Triple was 1024 and was 4000 for both yCD-Opt1 (unlabeled) and yCD-Opt1 labeled with ABDM.

3.6.2.10 Intrinsic Tryptophan Fluoresence

300 µL of 40 µM protein samples were placed into Greiner 96 flat black well plates (VWR, 655900) and fluorescence emission spectra were recorded on a Tecan Infinite M200 Pro plate reader. Using an excitation wavelength of 280 nm, spectra were recorded from 310 to 380 nm with a step size of 1 nm and an integration time of 20 µs with 25 flashes.

3.6.2.11 SDS-PAGE Band Intensity Analysis

SDS-PAGE gels were visualized using a Bio-Rad Gel Doc EZ Imager and analyzed with the Image Lab 6.0 software (Bio-Rad). Lanes and bands were picked manually, and

reference bands were chosen as the monomer bands for all samples except yCD-Opt1 labeled with ABDM, in which the dimer band was used as the reference band. Relative intensities of other bands in the same lane as the reference band were obtained using the Image Lab Quantity Tools.

3.6.2.12 Quenched-Reaction Cytosine and 5-Fluorocytosine (5FC) Assays

Enzyme assays with cytosine or 5FC as the substrate were performed in 96-well plates using a protocol similar to Mahan, et. al[12]. A stock of 1.5 mM cytosine and 5.8 mM 5FC were made in activity buffer. Two identical 175 µL reactions containing the desired concentration of cytosine or 5FC, enzyme, and activity buffer were placed under a UV LED or Blue light LED at room temperature. All yCD-Opt1 substrate consumption assays consisted of 250 µM cytosine and 250 nM enzyme or 1 mM 5FC and 333 nM enzyme. Before the addition of the substrate to the reaction wells, the enzyme and activity buffer were incubated under the light sources for 5 minutes to equilibrate the enzyme in the *cis* or *trans* states. Aliquots (20 µL) of the reaction were taken every 5 minutes over a 30-minute time period and immediately quenched in 0.1N HCl. Readings of cytosine were taken at OD_{286} and 5FC at OD_{292} using a plate reader. Measurements of activity were taken in triplicate and averaged to reduce error.

3.6.2.13 Kinetics

Kinetic parameters were determined for yCD-Opt1 under UV light and Blue light using the assay described above over a range of cytosine concentrations between 0.1 – 1.2 mM. Enzyme concentration remained the same in all assays and all time points were gathered in the linear region of enzymatic activity. All kinetics and substrate consumption assays were carried out with fresh, unfrozen protein stored at 4°C for less than two weeks

and performed in triplicate. K_m, V_{max}, and k_{cat} were determined by fitting the data to the Michaelis-Menten equation using nonlinear regression.

3.7 References

1. Beierle, J. M., Kistemaker, H. A. V, Velema, W. A. & Feringa, B. L. Reversible Photocontrol of Biological Systems by the Incorporation of Molecular Photoswitches. *Chem. Rev.* **113,** 6114–6178 (2013).

2. Riggsbee, C. W. & Deiters, A. Recent advances in the photochemical control of protein function. *Trends Biotechnol.* **28,** 468–475 (2010).

3. Brechun, K. E., Arndt, K. M. & Woolley, G. A. Strategies for the photo-control of endogenous protein activity. *Curr. Opin. Struct. Biol.* **45,** 53–58 (2017).

4. Guntas, G. *et al.* Engineering an improved light-induced dimer (iLID) for controlling the localization and activity of signaling proteins. *Proc. Natl. Acad. Sci.* **112,** 112–117 (2015).

5. Beharry, A. A. & Woolley, G. A. Azobenzene photoswitches for biomolecules. *Chem. Soc. Rev.* **40,** 4422–4437 (2011).

6. Pozhidaeva, N., Cormier, M. E., Chaudhari, A. & Woolley, G. A. Reversible photocontrol of peptide helix content: Adjusting thermal stability of the cis state. *Bioconjug. Chem.* **15,** 1297–1303 (2004).

7. Ritterson, R. S., Kuchenbecker, K. M., Michalik, M. & Kortemme, T. Design of a photoswitchable cadherin. *J. Am. Chem. Soc.* **135,** 12516–12519 (2013).

8. Schierling, B. *et al.* Controlling the enzymatic activity of a restriction enzyme by light. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 1361–1366 (2010).

9. Kumita, J. R., Flint, D. G., Woolley, G. A. & Smart, O. S. Achieving photo-control of protein conformation and activity: producing a photo-controlled leucine zipper. *Faraday Discuss.* **122,** 89-103; discussion 171-190 (2003).

10. Volgraf, M. *et al.* Allosteric control of an ionotropic glutamate receptor with an optical switch. *Nat. Chem. Biol.* **2,** 47–52 (2006).

11. Karjoo, Z., Chen, X. & Hatefi, A. Progress and problems with the use of suicide genes for targeted cancer therapy. *Adv. Drug Deliv. Rev.* **99,** 113–128 (2016).

12. Ko, T.-P. *et al.* Crystal structure of yeast cytosine deaminase. Insights into enzyme mechanism and evolution. *J. Biol. Chem.* **278,** 19111–7 (2003).

13. Zhao, Y., She, N., Zhang, X., Wang, C. & Mo, Y. Product release mechanism and the complete enzyme catalysis cycle in yeast cytosine deaminase (yCD): A computational

study. *Biochim. Biophys. Acta - Proteins Proteomics* **1865,** 1020–1029 (2017).

14. Ireton, G. C., Black, M. E. & Stoddard, B. L. The 1.14 Å crystal structure of yeast cytosine deaminase: Evolution of nucleotide salvage enzymes and implications for genetic chemotherapy. *Structure* **11,** 961–972 (2003).

15. Yao, L., Li, Y., Wu, Y., Liu, A. & Yan, H. Product Release Is Rate-Limiting in the Activation of the Prodrug 5-Fluorocytosine by Yeast Cytosine Deaminase. *Biochemistry* **44,** 5940–5947 (2005).

16. Zanghellini, A. *et al.* New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* **15,** 2785–2794 (2006).

17. Korkegian, A., Black, M. E., Baker, D. & Stoddard, B. L. Computational thermostabilization of an enzyme. *Science* **308,** 857–860 (2005).

18. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge structural database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **72,** 171–179 (2016).

19. Fleishman, S. J. *et al.* Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite. *PLoS One* **6,** (2011).

20. Misura, K. M. S. & Baker, D. Progress and challenges in high-resolution refinement of protein structure models. *Proteins Struct. Funct. Genet.* **59,** 15–29 (2005).

21. Wu, D., Dong, M., Collins, C. V., Babalhavaeji, A. & Woolley, G. A. A Red-Light Azobenzene Di-Maleimide Photoswitch: Pros and Cons. *Adv. Opt. Mater.* **4,** 1402–1409 (2016).

22. Mahan, S. D., Ireton, G. C., Knoeber, C., Stoddard, B. L. & Black, M. E. Random mutagenesis and selection of Escherichia coli cytosine deaminase for cancer gene therapy. *Protein Eng. Des. Sel.* **17,** 625–633 (2004).

23. Bruno, I. J. *et al.* Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.* **44,** 2133–2144 (2004).

24. Hanwell, M. D. *et al.* Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform.* **4,** 1–17 (2012).

25. Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S. & Baker, D. De novo enzyme design using Rosetta3. *PLoS One* **6,** (2011).

26. Studier, F. W. Protein production by auto-induction in high-density shaking cultures. *Protein Expr. Purif.* **41,** 207–234 (2005).

Chapter 4: Systematic comparison of Amber and Rosetta energy functions for protein structure evaluation

4.1 Preface

A version of Chapter 4 has been published [Rubenstein, A. B., Blacklock, K. M., Nguyen, H., Case, D. A., & Khare, S. D. Systematic Comparison of Amber and Rosetta Energy Functions for Protein Structure Evaluation. *J. Chem. Theory Comput.* XX**,** XX–XX (2018). doi: 10.1021/acs.jctc.8b00303], of which I am an author. I was responsible for contributing to concept formation, data collection and analysis, and manuscript composition regarding the loop modeling section. I was also responsible for data collection for the Amber minimizations in the Decoy Discrimination benchmark set and made minor manuscript edits. Hai Nguyen and David Case contributed to concept formation, experimental design, and manuscript edits. Sagar Khare was the supervisory author on this project and was involved throughout the project in concept formation and manuscript composition. Aliza Rubenstein was the lead investigator, responsible for all major areas of concept formation, data collection and analysis, as well as manuscript composition.

4.2 Abstract

An accurate energy function is an essential component of biomolecular structural modeling and design. The comparison of differently derived energy functions enables analysis of the strengths and weaknesses of each energy function, and provides independent benchmarks for evaluating improvements within a given energy function. We compared the molecular mechanics Amber empirical energy function to two versions of the Rosetta energy function (talaris2014 and REF2015) in decoy discrimination and loop

modeling tests. In decoy discrimination tests, both Rosetta and Amber (ff14SBonlySC) energy functions performed well in scoring the native state as the lowest energy conformation in many cases, but several false minima were found in with both talaris2014 and Amber ff14SBonlySC scoring functions The current default version of the Rosetta energy function, REF2015, which has more physically-derived terms than talaris2014, performs significantly better than talaris2014, highlighting the improvements made to the Rosetta scoring approach. There are no cases in Rosetta REF2015, and 8/140 cases in Amber, where a false minimum is found that is absent in the alternative landscape. In loop modeling tests Amber and REF2015 perform equivalently, although false minima are detected in several cases for both. The balance between dihedral, electrostatic, solvation and hydrogen bonding scores contribute to the existence of false minima. To take advantage of the semi-orthogonal nature of the Rosetta and Amber energy functions, we developed a technique that combines Amber and Rosetta conformational rankings to predict the most near-native model for a given protein. This algorithm improves upon predictions from either energy function in isolation, and should aid in model selection for structure evaluation and loop modeling tasks.

4.3 Introduction

Computational protein structure prediction is dependent on an accurate energy function. The native state of a protein is expected to be found uniquely at the minimum of the energy function[1]; therefore, the energy function must robustly discriminate between native and non-native conformations. A variety of energy functions to predict protein structure have been implemented over the past forty years[2–8]. These potentials largely fall into one of two categories: molecular mechanics force fields that rely on the combination of various empirical potentials such as Lennard-Jones, torsional energies, Coulombic

interactions, and desolvation penalties[3,4,7] and statistical or knowledge-based potentials that depend on characteristics of known protein structures[2,5,6]. While molecular mechanics force-fields are generally parameterized on small molecule properties[7,9–11], statistical potential parameter optimization is often guided by known biomolecular structures[12–14]. Each approach has its own drawback: since parameters in physically derived force-fields are fit based on small molecule properties, they may not be suited to macromolecules[15,16]: for example, force-fields will often display biases towards secondary structure propensities[15,17]. On the other hand, statistical potentials are trained on specific datasets of large biomolecules, and data sparseness may lead to overfitting[18].

The Rosetta macromolecular modeling program energy functions combine elements of both categories; they contains physical force-field terms (Lennard-Jones interactions, electrostatic interactions, desolvation penalties, etc.) and statistical potentials (probability of amino acid identity given backbone angles, probability of backbone angles given amino acid identity, probability of backbone-dependent rotamer, etc.)[19]. The most recent Rosetta energy function (REF2015) is parameterized on both small molecule properties and large sets of biomolecular structures[18], although previous energy functions were generally parameterized on known biomolecular structures alone[13]. While efforts have been made to compare the performance of various empirical force-fields[17,20,21], little attention has been focused on the comparison between the Rosetta energy function and empirical force-fields.

The Amber ff14SBonlySC force field[10] uses a standard fixed-charge molecular mechanics potential, with torsion potentials based entirely on fits to quantum chemistry data. It is very similar to the more commonly-used ff14SB protein force field, but does not include the empirical modifications to backbone torsion potentials that are present in ff14SB, and which provide an improved balance of secondary structure in explicit solvent

simulations. Hence, ff14SBonlySC is more "physics-based" than is ff14SB, and it arguably better suited for the implicit solvent energy evaluations used here, since the empirical backbone torsional potentials in ff14SB might be specific to its use of explicit solvent simulations. The ff14SBonlySC force field, in combination with a generalized Born implicit solvent model[22], has been shown to fold a variety of single-domain proteins using unrestrained molecular dynamics simulations[23].

Comparing the Amber force-field and Rosetta energy function performance at structure evaluation elucidates the strengths and areas of improvements for each energy function. As Rosetta energy functions have been developed based on improving performance for certain modeling datasets, testing their performance on the same macromolecular datasets may result in overfitting of the Rosetta energy function, while comparing their performance to that of a physics-based Amber energy function is a relatively unbiased comparison for evaluating performance improvements. Therefore, we chose two types of benchmark sets for energy function evaluation: first, decoy discrimination benchmark sets that were used for the development of the Rosetta REF2015 energy function[18], and second, a loop modeling benchmark set from the Kortemme group that were generated using Rosetta sampling but not utilized for REF2015 development. Comparison of two Rosetta energy functions with the same Amber energy function for these two benchmark sets provides insights into the improvements made in the energy functions and highlights areas for further improvement. Finally, selecting a correct near-native model for a given sequence is an elementary challenge; the combination of these two semi-orthogonal energy functions provides a method for model selection that is able to select more accurate models.

4.4 Methods

4.4.1 Benchmark Sets

To evaluate and compare the performance of Rosetta and Amber energy functions, we used two benchmark sets, a structure evaluation (decoy discrimination) set and a loop modeling set. The decoy discrimination benchmark set includes a total of 150 proteins, a combination of two independent decoy sets used in previous studies[18,24]. The proteins in the set are monomeric and have crystallographic native structures available in the RCSB PDB[25] with resolution < 2.0 Å. The protein lengths range from 50 to 200 residues and have a diverse range of topologies. The decoy sets were originally generated using biased and unbiased ab-initio sampling runs[26] followed by parallel loophash sampling (PLS)[27]. This produced 40,000-200,000 decoys per protein, ~1000 representative low-energy structures of which were chosen for each protein to cover the range of possible C-α RMSD values.

The loop modeling benchmark set consisted of the 45-PDB dataset for 12-residue loops in the monomeric protein loops training set of the 2016 Collaborative Assessment and Development of Rosetta Energetics and Sampling (CADRES) initiative in the Rosetta community. This loop modeling benchmark set was obtained from Shane O'Connor and Tanja Kortemme (personal communication).

4.4.2 Structure Preparation

4.4.2.1 Rosetta

In the decoy discrimination benchmark, the native crystallographic structures for each protein set were downloaded from the RSCB PDB and residues were trimmed from

the structure to match the sequence of the crystal with the decoy structure in the benchmark sets. Native structures were used to evaluate RMSD from native for decoy conformations. Native structures were relaxed using FastRelax[26] with the talaris2014[13] energy function to relieve any clashes. One hundred relaxation trajectories were simulated to generate one hundred relaxed native-like decoys. These native-like decoys were used for false minima analysis. Then, these one hundred native-like decoys, along with the ~1000 pre-sampled decoys, were subjected to backbone and sidechain minimization using talaris2014 and the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) minimizer implementation with inexact line search conditions (lbfgs_armijo_nonmonotone) over a maximum of 2000 iterations for convergence. C-α atom RMSD was calculated for all decoys.

The REF2015 decoy dataset was obtained from F. DiMaio and H. Park[18]. For this dataset, each decoy was relaxed with 3 cycles torsion-space minimization and 2 cycles Cartesian mode[24] using the REF2015 energy function[19]. Only 140 out of 150 protein systems were included in this set due to the lower quality of experimentally determined structures for 10 systems (H. Park and F. DiMaio, personal communication). Those 10 systems are ignored when comparing REF2015 to Amber.

In the loop modeling benchmark, the native crystal structures for each protein set were downloaded from the RCSB PDB and trimmed of residues that were not present in the decoy PDB structures. The backbone and sidechain geometries for residues in the loop region of each decoy structure were minimized in Rosetta using the talaris2014 or REF2015 energy function and the lbfgs_armijo_nonmonotone over a maximum of 2000 iterations for convergence. C-α RMSDs were calculated with respect to the crystal structure over loop residues only without fitting; since the protein scaffold was fixed during optimization,

this statistic describes the extent of loop deviation. Loop residues are defined in supplementary file LoopDefs.xlsx.

4.4.2.2 Amber

Hydrogens were removed from the crystal structures and decoy PDBs, and initial structures were built using the tLEaP module of AmberTools[28] with the ff14SBonlySC[10] forcefield parameters. Minimizations were carried out for a maximum of 1000 steps under the LBFGS quasi-Newton algorithm[29] with a convergence criterion of 0.01 kcal/mol-A. In the loop modeling benchmark, positional restraints were added to all non-loop-residue atoms except for hydrogens with a force constant of 10.0 kcal mol^{-1} A^{-2}. Solvent effects were treated with a generalized Born implicit solvent model (GB-Neck2[22]) implemented in the Amber16[28] package with mbondi3 radii and a cutoff value of 999A for nonbonded interactions. Total potential energies of minimized structures and C-α RMSDs with respect to the crystal structure were obtained using the pytraj 2.0.0 interactive molecular dynamics simulation data analysis Python package[30], which is a Python interface for cpptraj in AmberTools16[28]. In the loop modeling benchmark, C-α RMSDs were calculated over the loop residues only. Six sets of decoy structures for the loop modeling benchmark were unable to be minimized in Amber due to missing residues, and those sets were not considered in subsequent analyses (PDB codes 1cb0, 1dts, 1m3s, 1ms9, 1t1d, and 2pia).

4.4.2.3 Energy Landscape Generation

Energy landscapes (RMSD vs. normalized energy scatterplots) were generated for all proteins for both Rosetta and Amber. The ideal shape of an energy landscape is that of a funnel (i.e. **Figure 4-1:A**) where the lowest-scoring decoy conformations are of near-

native RMSD. We use the binned Boltzmann metric (see below) to evaluate the funnel shape of each energy landscape.

4.4.2.4 Energy Normalization

For each set of energies per energy function per protein, energies are normalized so that the gap between the 5th percentile of and the 95th percentile is equal to 1. This enables the comparison of energies between different structures and between different energy functions. This is accomplished via the following equation:

$$E_{i(norm)} = (E_i - E_{min})/E_{95th} - E_{5th}$$

E_i refers to the raw energy of decoy i. E_{min} is the minimum energy value, E_{95th} is the 95th percentile energy, and E_{5th} is the 5th percentile energy

4.4.2.5 Funnel Evaluation Metric

We use the binned Boltzmann metric, B, for energy landscape evaluation, as described previously by Park et al.[18]. This metric finds the Boltzmann probability of selecting native-like decoys over high-RMSD decoys based on their energy values. As in previous work[24], the metric is averaged over multiple thresholds for determining native-like status for each decoy.

$$B = \frac{\sum_j (\sum_i d_{ij} P_i / \sum_i P_i)}{N_j}$$

$$P_i = e^{-\beta E_{i(norm)}}$$

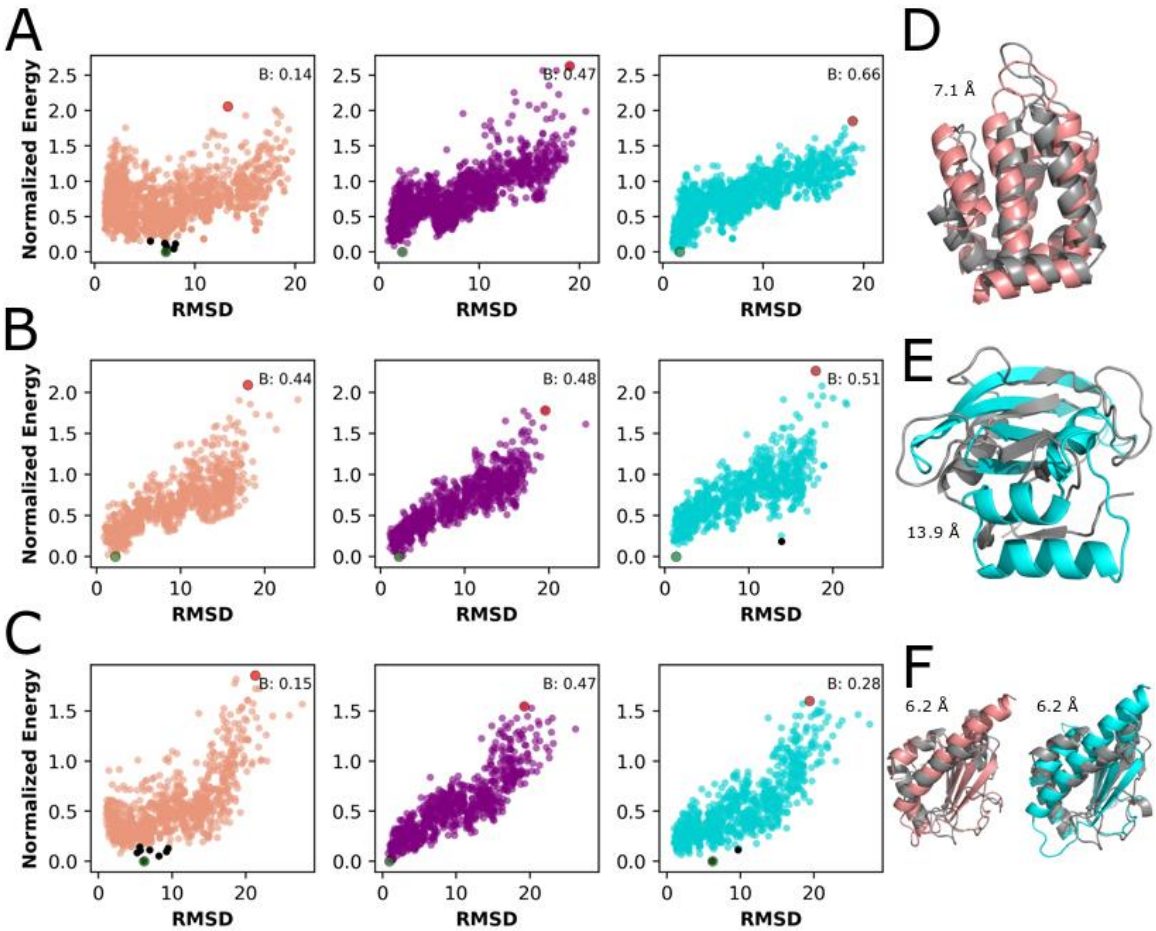


Figure 4-1: (A-C) Energy landscapes for 2QY7, 1T2I, and 1SEN respectively using Rosetta talaris2014 (salmon; left), Rosetta REF2015 (purple; middle) and Amber ff14SBonlySC (turquoise; right) energy functions. Each dot on the plot represents one decoy conformation. The x-axis is RMSD from native crystal structure and the y-axis is normalized energy. False minima (defined as decoys within top 10 energies but with RMSD > 5.0 Å) are depicted in black. The B metric, which represents the efficacy of the energy function at differentiating between native and non-native decoys, is shown at the top right corner of each plot. (D-F) Superimposed native (gray) and Rosetta talaris2014 lowest-ranking false minimum decoy (salmon) and/or superimposed native (gray) and Amber lowest-ranking false minimum decoy (turquoise) for 2QY7, 1T2I, and 1SEN respectively.

The conformation index is i and j is the native threshold definition index. Cutoffs are 0.5, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 3.5, 4.0, 5.0, 6.0 Å and N_j is thus 14. $E_{i(norm)}$ is the score of decoy i as determined in Rosetta or Amber and normalized as described above. The factor β refers to the Boltzmann factor and has a value of 0.1, as this was the value of β used by Park et al. previously[18]. d_{ij} determines whether decoy i is considered native at threshold j; it is set to 1 if it is native and 0 if it is not. As the sum of the probabilities of the non-native-like conformations approaches 0, the numerator $(\sum_i d_{ij} P_i)$

approaches the value of the denominator ($\sum_i P_i$), so that the value of B approaches to 1. As mentioned in Park et al.[18], the B metric is better than the previously used S metric[24] due to a larger increase in the metric for a poor energy landscape vs. a good energy landscape than the increase from an already good energy landscape to a steeper energy landscape. Additionally, it is a smoother metric that is less affected by single-decoy outliers.

4.4.2.6 Model Selection

For each protein, a model was selected by finding the decoy that had the lowest sum of Amber and Rosetta ranks; this decoy also satisfies the criteria for Pareto-optimality. First, the Amber scores and Rosetta scores were converted into ranks so that the rank of decoy a was less than the rank of decoy b if the energy of decoy a was less than the energy of decoy b. Second, the Pareto-optimal solutions are found as follows. Decoy a is defined as dominating decoy b if both ranks (Rosetta and Amber) of decoy a are <= both ranks of decoy b. Pareto-optimal decoys are decoys that dominate at least one other decoy and are not dominated by any decoys. From among the set of Pareto-optimal decoys, the decoy that has the lowest sum of ranks is chosen as the solution. In the rare case that more than one decoy has a minimum sum of ranks, a decoy is arbitrarily chosen from the minimum-sum-ranks decoys. Technically, the minimum-sum decoy must be found within the Pareto set of solutions; however, the use of the Pareto-minimization allows for easier visualization and interpretation of the minimum-sum solution. Additionally, the Pareto-optimal set of decoys may be useful for selecting a set of top-scoring n (n>1) decoys when performing consensus scoring according to the two energy functions.

4.5 Results

4.5.1 Performance of Amber and Rosetta energy functions in discriminating between native and non-native structures

Protein free energy landscapes involve folding funnels[31–33] which enable the folding chain to efficiently find the native state, and their existence implies that the higher energy of non-native (decoy) structures compared to the native (e.g., crystallographically-determined) structure drives protein folding. Therefore, a common test used for evaluating[23,26] and improving[18] energy functions is the decoy discrimination test, in which the evaluated scores of decoy structures are compared to that of near-native structures. High-RMSD decoys which have comparable energies to near-native structures are classified as "false minima", and are indicative of inaccuracies in the energy function. The B metric[18], ranging from 0 to 1, quantifies the existence of false minima in a set of structures upon evaluation with a given energy function, with values close to 1 indicating a smooth folding funnel with no false minima. Conversely, a lower B value indicates that one or more false minima exist.

We compared the performance of the Amber energy function and both of the Rosetta energy functions at ranking native state structures lower than decoy conformations for a set of 150 proteins (talaris2014 and Amber) or a set of 140 proteins (REF2015 and Amber). Amber ff14SBonlySC generally performed better than Rosetta talaris2014, scoring significantly higher B metrics for many systems (**Figure 4-2:A**). We also compared Amber to the default Rosetta energy function, REF2015[19], and found that while Amber did have a higher B metric for several systems, several other systems had a higher B metric when scored by REF2015 (**Figure 4-2:B**), thus showing the improvement of REF2015 over talaris2014 (**Figure 4-2:C**) when compared to Amber as an unbiased benchmark. The

REF2015 energy function was developed to maximize the discrimination between native and decoy structures using the same benchmark datasets (among other tests), so it is not surprising that improvements over talaris2014 are observed. While many terms in the Amber ff14SBonlySC energy function were developed using small molecule data, solvation parameters were fit to reproduce more accurate Poisson-Boltzmann calculations for a range of peptides and proteins[22]. Thus, this energy function, similar to REF2015, has been trained to discriminate decoy structures from natives. While there is no overlap between the decoy discrimination training sets used for ff14SBonlySC and REF2015, the performance of the two energy functions in decoy discrimination tests is similar (Table 1).

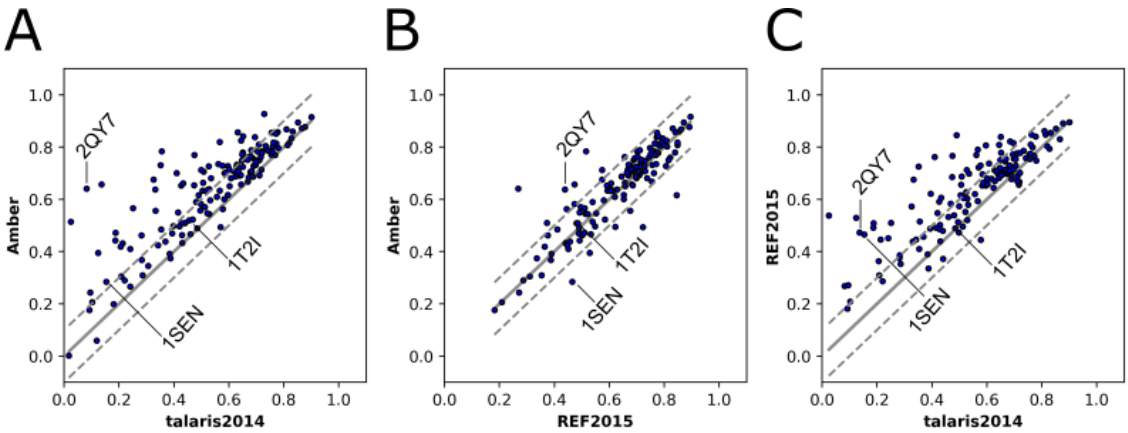


Figure 4-2: Scatterplots to depict general performance of the scoring functions in decoy discrimination.(A)Rosetta talaris2014 scoring function vs. Amber scoring function, (B) Rosetta REF2015 scoring function vs. Amber scoring function, and (C) Rosetta REF2015 scoring function vs. talaris2014 scoring function over the entire decoy discrimination set. Each dot represents the B metric for one system. The black line is x=y and the dashed lines show the B-values +/- 0.1 from the x=y line. Decoy sets discussed in more depth are annotated with the PDB ID of that set.

To investigate the sources of misprediction and improvements in energy functions, we examined cases in which either Amber, Rosetta, or both were unable to correctly rank high-RMSD decoy conformations, scoring them as low-scoring instead of high-scoring. A false minimum is defined as a decoy within the top-10 ranked decoys that has a C-α RMSD from native of greater than 5 Å. Three of these cases are shown in **Figure 4-1**. 2QY7 (**Figure 4-1:A,D**) has several false minima for Rosetta talaris2014 but none for Amber. Generally,

Rosetta talaris2014 alone had at least one false minimum in 16.0% of structures, and Rosetta REF2015 alone had no false minima that were absent in talaris2014 or Amber. 1T2I (**Figure 4-1:B,E**) has a false minimum for Amber but none for Rosetta talaris2014;1.3% and 5.7% of systems have at least one false minimum for Amber alone when compared to talaris2014 and REF2015, respectively. 1SEN (**Figure 4-1:C,F**) has false minima for both Amber and Rosetta talaris2014, as do 14% of overall structures in the Rosetta talaris2014/Amber comparison, and 7.1% in the Rosetta REF2015/Amber comparison (**Table 4-1**).

Table 4-1: *B* metric, false minima, and model selection summary comparisons for Amber ff14SBonlySC, Rosetta talaris2014, and Rosetta REF2015 energy functions.

	No. Cases/Total No. Proteins
Decoy Discrimination	
ff14SBonlySC B > talaris2014 B by 0.1	54/150
talaris2014 B > ff14SBonlySC B by 0.1	0/150
ff14SBonlySC B > REF2015 B by 0.1	6/140
REF2015 B > ff14SBonlySC B by 0.1	9/140

False minima in ff14SBonlySC only (not talaris2014)	2/150
False minima in talaris2014 only (not ff14SBonlySC)	24/150
False minima in ff14SBonlySC and talaris2014	21/150
False minima in ff14SBonlySC only (not REF2015)	8/140
False minima in REF2015 only (not ff14SBonlySC)	0/140
False minima in ff14SBonlySC and REF2015	10/140
Minimum-sum RMSD < ff14SBonlySC selected RMSD by 1.0 Å	14/140
Minimum-sum RMSD < REF2015 selected RMSD by 1.0 Å	13/140
Loop Modeling	
ff14SBonlySC B > talaris2014 B	15/39

talaris2014 B > ff14SBonlySC B	7/39

We next investigated if any structural patterns and features can be discerned in the landscapes where false minima are observed selectively for one energy function. As REF2015 landscapes had no false minima that were not present in Amber landscapes, we restricted our analyses to talaris2014 landscapes for these analyses. Superimpositions of false minima decoys with native decoys show their distinct non-native conformations involving both misprediction of secondary structure elements as well as their incorrect relative placement in tertiary structures. In the case of 2QY7, a Rosetta talaris2014 false minimum, the four helical bundle found in the structure is significantly perturbed in the false minimum as the order of the first two helices is reversed; thus, while they do not contact the other two helices as tightly as that of the native structure (**Figure 4-3:E, I-J**) their energies of interactions are found to be favorable. The difference between the native structure of 1T2I and its Amber false minimum is subtler. While the contact maps for the native and false minimum conformations are similar, except for a small contact region in the native structure between residues 40 and 59 that does not appear in the false minimum (**Figure 4-3:K-L**), the false minimum is slightly more compact and has a more ordered secondary structure. Two beta sheet regions in the false minimum are beta strands/unordered in the native structure and two alpha helices in the false minimum are beta strands in the native structure (**Figure 4-3:F-G**).

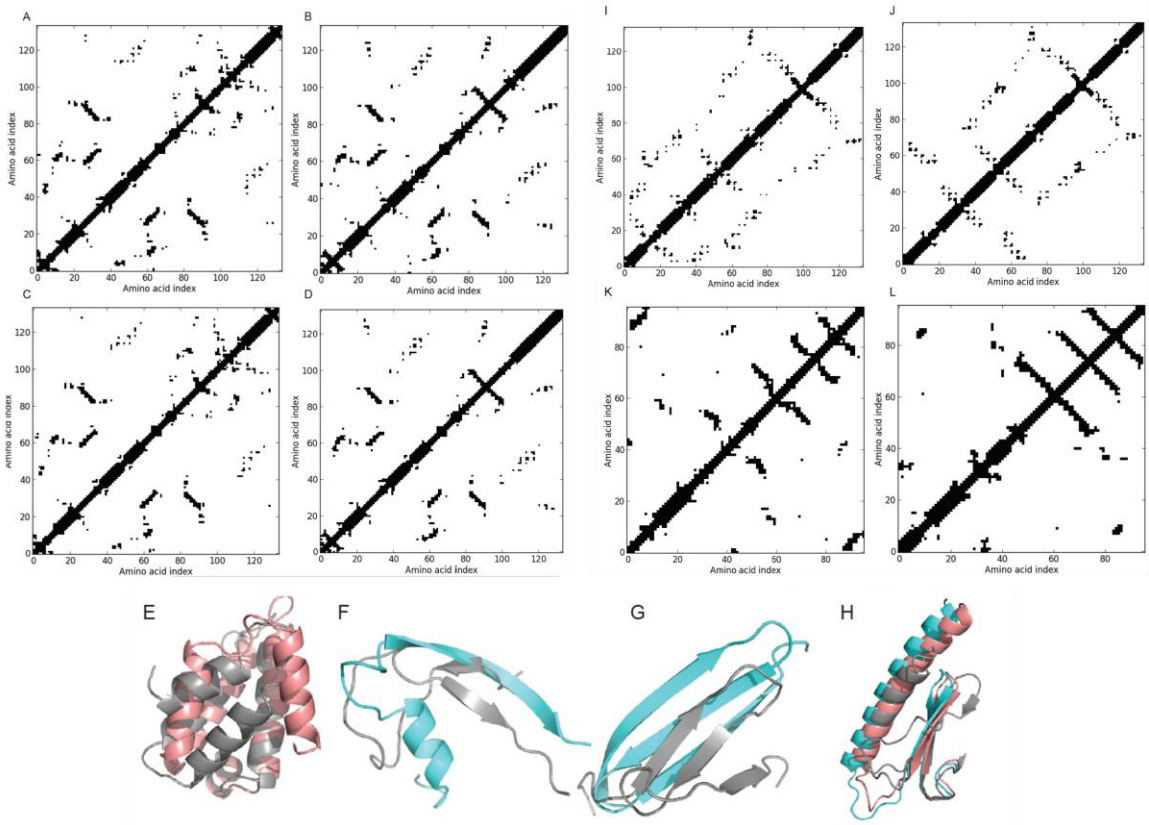


Figure 4-3: Contact maps and representative structure comparisons for 1SEN, 2QY7, and 1T2I. (A) Contact map for 1SEN lowest-Rosetta scored native structure. (B) Contact map for 1SEN Rosetta false minimum. (C) Contact map for 1SEN lowest-Amber scored native structure. (D) Contact map for 1SEN Amber false minimum. (E) Residues 1-81 of 2QY7. Gray is native, salmon is Rosetta false minimum. (F-G) Residues 31-59 and 61-96 of 1T2I respectively. Gray is native, turquoise is Amber false minimum. (H) Residues 75-135 of 1SEN. Gray is native, salmon is Rosetta false minimum, and turquoise is Amber false minimum. (I) Contact map for 2QY7 lowest-Rosetta scored native structure. (J) Contact map for 2QY7 Rosetta false minimum. (K) Contact map for 1T2I lowest-Amber scored native structure. (L) Contact map for 1T2I Amber false minimum. Contact maps were generated using Benjamin Rafferty; Zachary Carl Flohr; Ashlie Martini (2014), "Protein Contact Maps," https://nanohub.org/resources/contactmaps. (DOI: 10.4231/D35M62761).

The case of 1SEN, which has false minima for both Rosetta and Amber, is similar to 1T2I in that the false minima are more ordered than the native structure, although the native structure forms more contacts between remote regions than do the false minima (**Figure 4-3:A-D**). Residues 85-96 form a tight beta hairpin in the false minima, whereas the native residues 85-96 is a longer loop between the beta strands, resulting in a shorter, less tight, beta hairpin. Additionally, the stretch of residues 94-109 in the native does not have any secondary structure, while that of the false minimum begins as a beta strand and ends in an alpha helix (**Figure 4-3:H**). Decoys that are predicted as false minima often have the

same overall structure and contact maps as native structures, yet secondary structure differences may result in large structural deviation. In some cases, false minima contain more ordered secondary structures yet fewer contacts than native conformations; the propensity away from disordered loops may result in lower energies for these false minima.

To investigate whether certain residues, structural elements, or energy terms contribute more to false minima conformations, we analyzed the per-residue score decomposition for Rosetta talaris2014 scores for the three systems outlined above (2QY7, 1T2I, and 1SEN). We were unable to perform the same decomposition for Amber as the GB solvation term is not pairwise-decomposable[28]. We calculated the Z-scores for each residue over the lowest-scoring native and false minimum conformations. We identified residues as possibly implicated in false minima if the false minimum residue Z-score was lower than the native residue Z-score by at least one (i.e. the distance between the two was greater than one standard deviation). We have highlighted these residues (**Figure 4-4:A-C**). False minima contributing residues were distributed over the conformations and did not cluster to any particular region. Moreover, false minima contributing residues were found in various types of secondary structure: alpha helices, beta strands, and loops. It is therefore currently not possible to attribute Rosetta false minima to any single per-residue propensity, but as expected, several small errors in energy estimation may lead to the observed incorrect scoring.

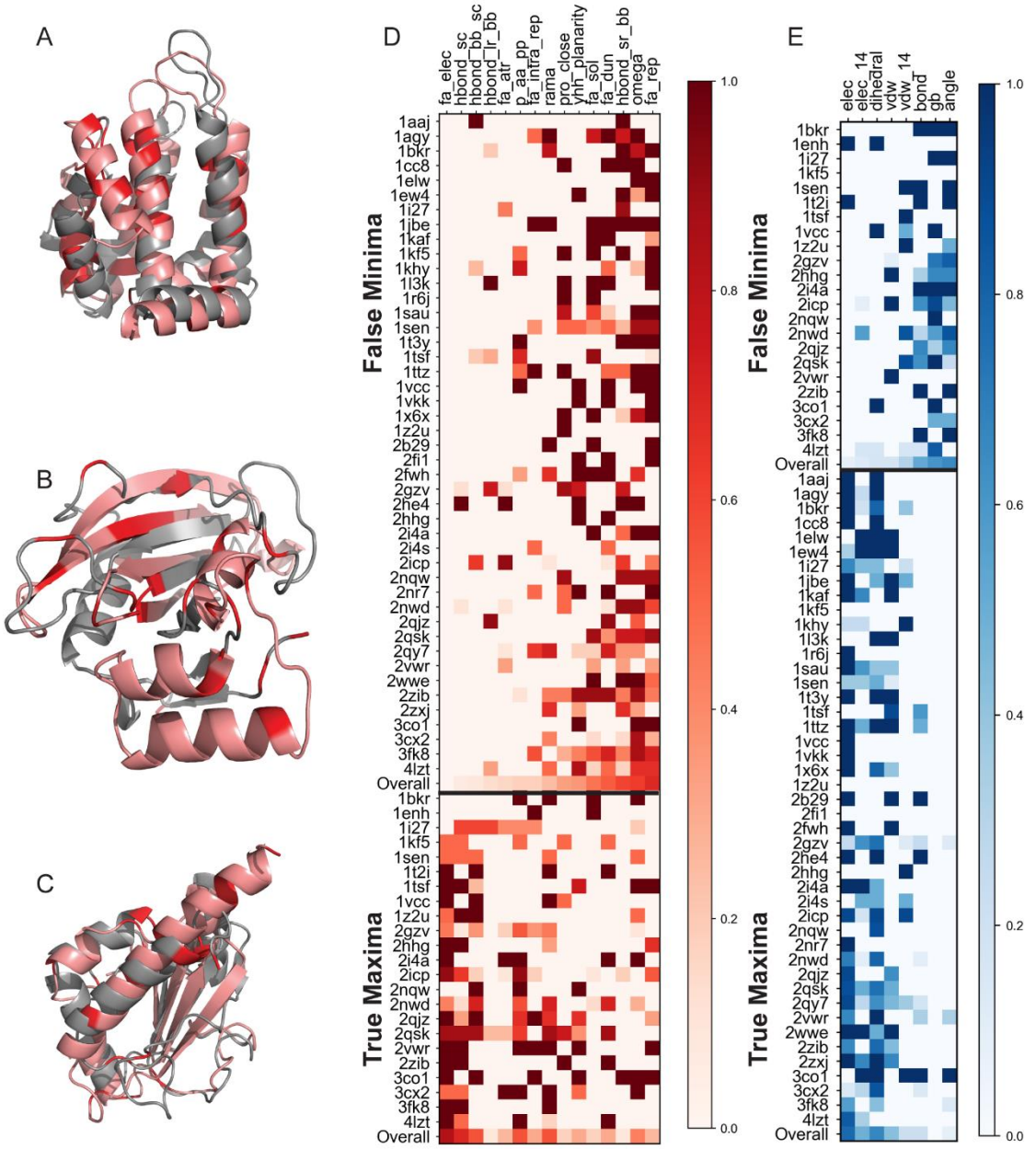


Figure 4-4: Per-residue and per-score-term propensity of energy functions toward false minima. (A-C) Native (gray) and Rosetta-talaris2014-minimized (salmon) structures of 2QY7, 1T2I, and 1SEN respectively. Rosetta-talaris2014-minimized residues that are scored by Rosetta talaris2014 as greater than 1 standard deviation away from the corresponding native residue are highlighted in red. Heatmaps of per-structure, score-term contribution to Rosetta-determined (D) and Amber-determined (E) false minima and true maxima. The row marked Overall shows the percentage of structures that indicate some degree of implication for that score-term.

4.5.2 Per score-term contributions in Amber and Rosetta energy functions

We reasoned that insight about the performance and pathologies of each energy function could be gained by identifying the energy terms that are responsible for correct and incorrect evaluations within the same energy function. For example, we asked which terms in the Amber energy function help it avoid mis-scoring a decoy (called Amber true maximum) that is identified as a false minimum in the Rosetta talaris2014 landscape (called Rosetta false minimum), and vice versa.

We identified terms that contribute to false minima and true maxima by calculating the Z-scores per decoy set and native set for each protein. If the lowest native score-term Z-score is greater than the false minimum score-term by at least one, that term is implicated in that false minimum. The reverse (i.e. true maximum score-term Z-score is greater than the lowest native score-term Z-score by at least one) is true for identifying true maximum contributing score-terms. The heatmap in **Figure 4-4:D** depicts the fraction of Rosetta talaris2014 false minima decoys (top) and true maxima decoys (bottom) that show some degree of implication for each score-term. This is calculated both on a per-protein basis and over the entire false minima/true maxima sets. Several score-terms, including hbond_sr_bb, fa_dun, fa_rep and omega, are implicated in a majority of false minima in the Rosetta talaris2014 energy function. A set of other score-terms contribute to a majority of Rosetta true maxima (or Amber false minima), including rama, hbond_bb_sc, hbond_sc, p_aa_pp, and fa_elec. These are score-terms that are not usually implicated in Rosetta false minima, thus demonstrating that the score-terms that contribute to the two trends (towards false minima and true maxima) are mutually exclusive. Except fa_elec, the other terms identified as helping "rescue" Amber false minima are all PDB-statistics

derived, and it is not surprising that they are implicated in correcting the errors of the more physics-based Amber energy function.

We next performed a similar analysis on Amber score-terms for both Amber false minima and Amber true maxima (**Figure 4-4:E**). We found that bond, angle, and gb are responsible for more than 50% of Amber false minima and that dihedral and elec are implicated in rescuing Rosetta false minima (Amber true maxima). We found that score-terms that are responsible for false minima are not implicated in true maxima and vice versa. Similar to the identification of statistically-derived terms in Rosetta as being responsible for correctly scoring Amber false minima, we find that physics-based terms, i.e., elec (which is counterbalanced by gb) and dihedral potentials, that are orthogonal to the talaris2014 Rosetta energy function, are implicated in the rescue of Rosetta talaris2014 false minima by Amber.

4.5.3 Combining rankings to select decoys improves decoy selection

Based on the results above indicating that use of one energy function can rescue the false minima in the landscape generated by the other energy function due to additional terms or different parameterization of terms, we sought to develop an approach to productively combine the two landscapes for model selection. In model selection (for example in protein structure prediction) the challenge is to select a near-native conformation from a set of decoy conformations based on one or more energy values or other features. Typically, an energy value obtained from a single energy function is used. In the current benchmark set, if model selection is performed by Rosetta REF2015 and Amber ff14SBonlySC energy functions individually, the Rosetta lowest-scored decoy has an RMSD of > 5.0 Å for three out of 140 systems, while the lowest-scored Amber decoy has an RMSD of > 5.0 Å for four other systems. We designed a minimum-sum based algorithm

(see 4.4.2.6) to select a decoy conformation based on both sets of ranks to improve the chances of selecting a near-native decoy.

We found that our minimum-sum algorithm generally improved model selection for both Rosetta and Amber rankings (**Figure 4-5:D-E**). Each minimum-sum selected decoy had an RMSD of < 5.0 Å; thus, the minimum-sum selected decoy rescued all seven systems for which either the Rosetta or the Amber lowest-scored decoy had an RMSD of > 5.0 Å. The minimum-sum selected decoy had an RMSD of at least 1 Å lower than that of the lowest-scoring Rosetta decoy for 13 out of 140 cases and an RMSD of at least 1 Å lower than that of the lowest-scoring Amber decoy for 14 out of 140 cases. More generally, the minimum-sum selected decoy had a lower RMSD than that of the Amber lowest-ranked decoy for 66/140 systems and a lower RMSD than that of the Rosetta lowest-ranked decoy for an overlapping but different set of 66/140 systems.

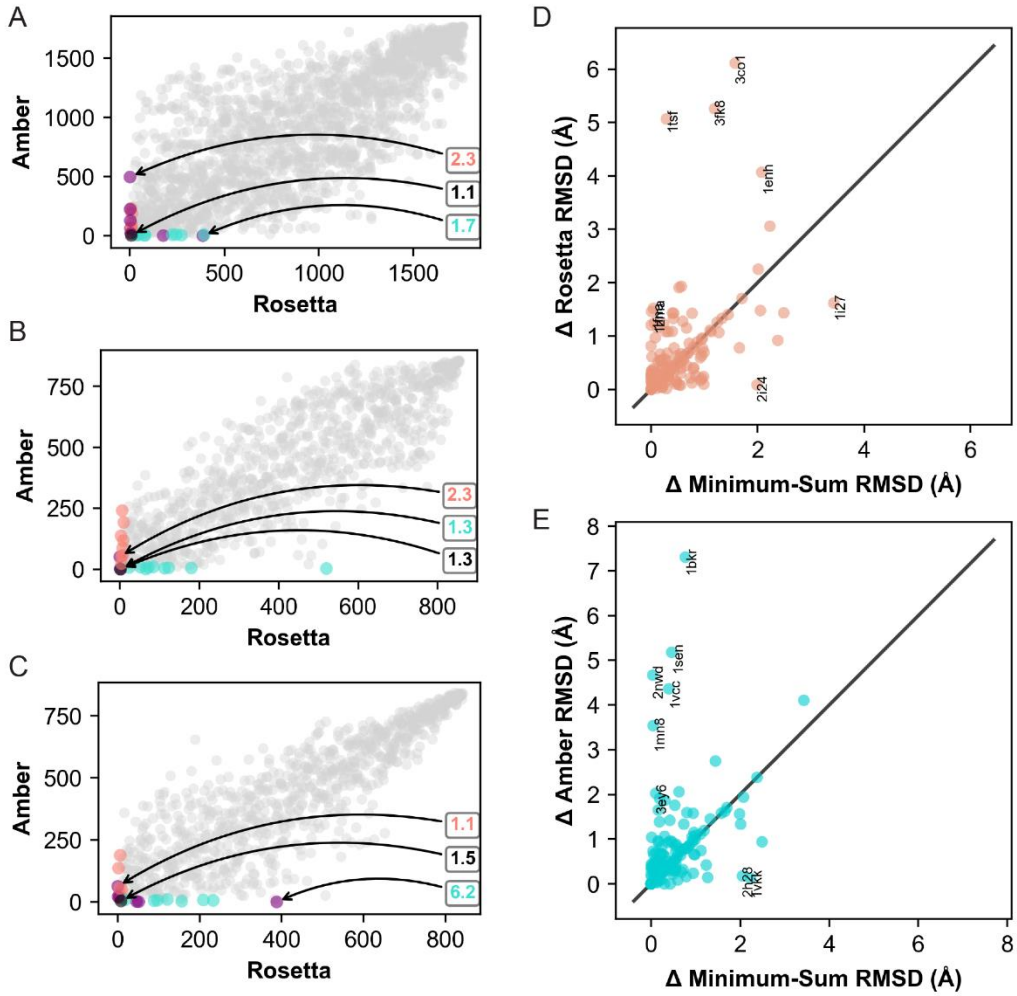


Figure 4-5: Minimum-sum model selection. (A-C) Scatterplots of Rosetta-rank (REF2015) vs. Amber-rank (ff14SBonlySC) for all decoys of 2QY7, 1T2I, and 1SEN respectively. Each point represents one decoy conformation. The set of Pareto solutions is purple, the top-10 ranked Amber decoys are turquoise, the top-10 ranked Rosetta decoys are salmon, and the minimum-sum solution is black. Annotations represent the RMSD in Å from native for the top-ranked Amber decoy (turquoise), top-ranked Rosetta decoy (salmon), and minimum-sum solution (black). Scatterplots that show the efficacy of the minimum-sum solution at minimizing the distance from native relative to the top-ranked Rosetta decoy (D) and top-ranked Amber decoy (E). Each point represents one system. The x-axis is the difference between the minimum-sum solution RMSD from native and the RMSD of the minimum-RMSD decoy conformation, while the y-axis is the difference between the Rosetta lowest-ranked conformation (D) or Amber lowest-ranked conformation (E) RMSD from native and the RMSD of the minimum-RMSD decoy conformation. Points that fall outside the 95% prediction intervals are annotated.

We examined the false minima cases described above (2QY7, 1T2I, and 1SEN) and found that the minimum-sum decoy generally had lower or equal RMSD as compared to those of Rosetta- or Amber-selected decoys (**Figure 4-5:A-C**). However, for 1SEN, which contains false minima for Rosetta talaris2014 and Amber ff14SBonlySC but not for Rosetta

REF2015, the Rosetta REF2015-selected decoy had a slightly lower RMSD than that of the minimum-sum decoy (1.1 Å vs. 1.5 Å, respectively). Nevertheless, the minimum-sum selected decoy RMSD is significantly lower than that of the Amber-selected decoy (1.5 Å vs. 6.2 Å). Thus, a minimum-sum framework allows combining the two energy functions productively to select a near-native model.

4.5.4 Loop Modeling

The conformational variability of loops plays a multi-functional role in protein structure and function. They are implicated in stability and folding pathways[34], binding and active sites[35,36], and binding other proteins[37,38]. Efficient sampling algorithms have been developed[36–39], but loop structure prediction efforts can be limited by energy functions, as the energy gaps between loops are smaller and minima are narrow[40]. Therefore, we tested both Rosetta energy functions and the Amber energy function on a loop modeling benchmark (see Methods). In this benchmark, most of the structure remains the same over the set of decoys; the difference lies in a small loop region, which can vary highly in RMSD. The energy gaps between structures are therefore smaller; thus, loop modeling provides a stringent test to distinguish between energy functions.

We found that Amber ranked loop conformations considerably more accurately than Rosetta talaris2014, as several systems had higher B values with Amber (**Figure 4-6:A, left**). REF2015 performance was more even: in 13 cases Amber B-values are significantly (B>0.1) greater than REF2015, whereas the converse is true in 7 cases (**Figure 4-6:A, middle**). The Rosetta REF2015 energy function also ranks loops more accurately than the Rosetta talaris2014 energy function, although the improvement is less dramatic (**Figure 4-6:A, right**) compared to that observed for whole structure decoy discrimination (**Figure 4-2:C**). **Figure 4-6:B** depicts the energy landscapes for a particular case where the Amber

B-value was significantly higher than the Rosetta B-values (PDB 1CYO). The Amber funnel is steeper than either of the Rosetta funnels, which is reflected in its higher B (0.86 vs. talaris2014 - 0.37, REF2015 – 0.19). We further investigated the energetic differences between the lowest-RMSD and lowest-energy decoys in the REF2015 landscape to elucidate which energy terms lead to this false minimum. We examined two residues of interest, His15 and Asn17 (**Figure 4-6:C**), which have extensive hydrogen bonding networks in the native and near-native structures compared to the low-energy decoy (**Figure 4-7:A,B**). REF2015 correctly identifies the His15 conformation in the low-RMSD decoy to be of lower energy than the false minimum decoy conformation, with more favorable hydrogen bonding scores (hbond_sc, hbond_lr_bb, hbond_sr_bb) as well as electrostatic and solvation terms (fa_elec, fa_sol), as the histidine sidechain is pointed toward the solvent in low-RMSD decoys and into the protein framework in the false minimum decoy. The energetic preference for this residue in its native-like conformation, however, is reduced by a penalization in the backbone dihedral (called rama_prepro in REF2015) term (**Figure 4-7:C**). For Asn17, the native-like decoy has a slightly energetically unfavorable conformation according to REF2015 that arises from a higher solvation (fa_sol) and rama_prepro score, despite the presence favorable hydrogen bonding and electrostatic terms (hbond_bb_sc, hbond_sr_bb, fa_elec, fa_atr) (**Figure 4-7:D**). Furthermore, the native-like loop for 1CYO scores more poorly compared to the false-minimum decoy across all 12 residues in the rama_prepro term, the amino acid probability (p_aa_pp), and sidechain energy (fa_dun) terms (**Figure 4-7:E**). Several pre-loop residues (five residues on the N-terminal side of the loop region) also score poorly in the low-RMSD structure, mostly due to large differences in the electrostatic (fa_elec) and hydrogen bonding terms (**Figure 4-7:F**). These results suggest that improper balancing of the backbone and sidechain dihedral, solvation and electrostatic/hydrogen bonding terms is responsible for the observed false energy minimum in the Rosetta landscape for 1CYO.

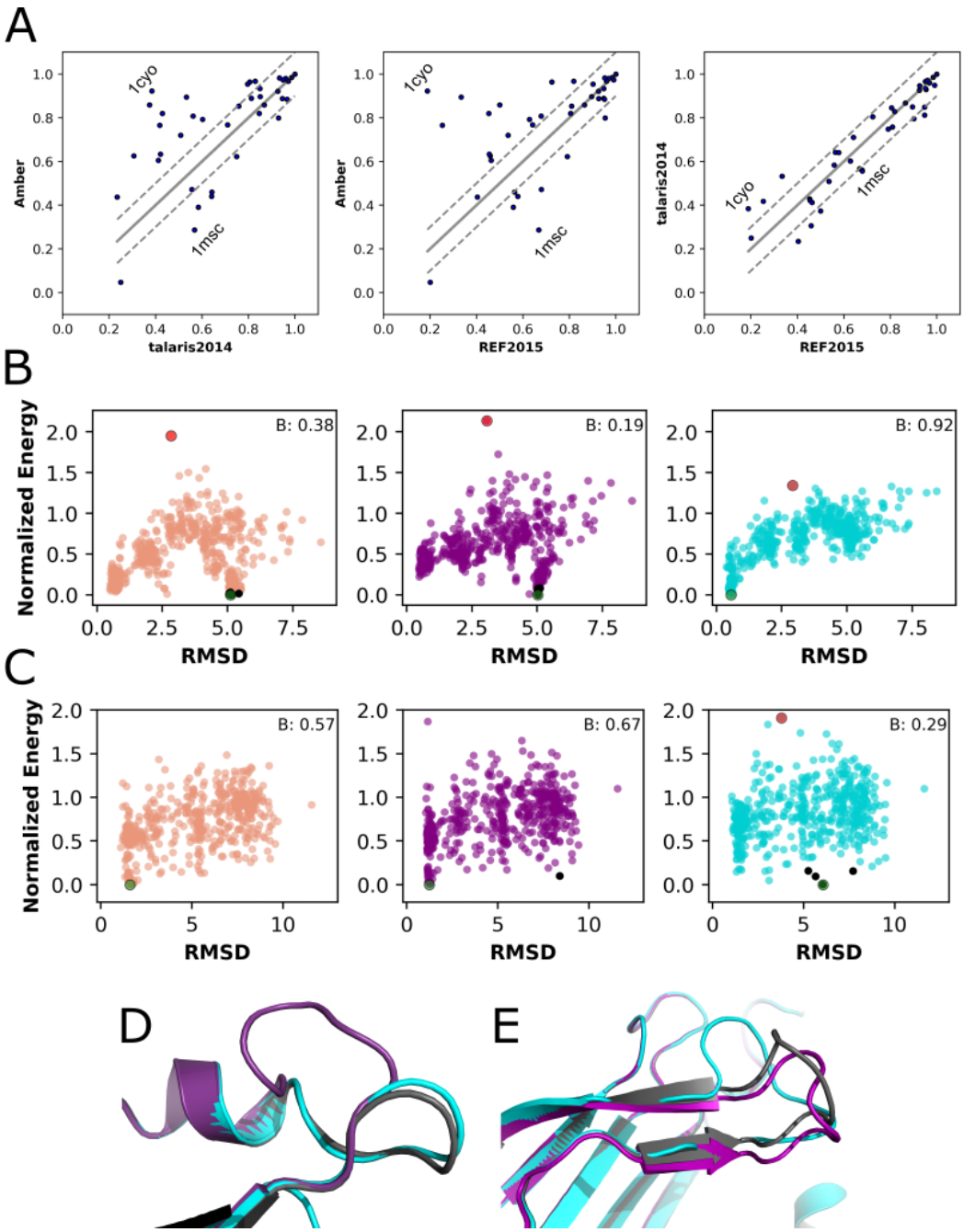


Figure 4-6: Loop modeling benchmark.General performance of Rosetta talaris2014 scoring function vs. Amber scoring function (left), Rosetta REF2015 scoring function vs. Amber scoring function (middle), and Rosetta talaris2014 scoring function vs. Rosetta REF2015 scoring function (right) over the entire loop modeling set. Each dot represents the B metric for one system. The black line is x=y and the dashed line represents B-values that are +/- 0.1 from the x=y line. Decoy sets discussed in more depth are annotated with the PDB ID of that set. (B) Energy landscapes for 1CYO. Each dot on the plot represents one decoy conformation. The x-axis is RMSD from native and the y-axis is normalized energy. The B metric, which represents the efficacy of the energy function at differentiating between native and non-native decoys, is shown at the top right corner of each plot. Rosetta talaris2014 and REF2015 plots are on the left (salmon) and middle (purple), respectively, and the Amber plot is on the right (turquoise). The lowest-energy decoy conformation in each plot is shown in green and the

highest-energy decoy conformation is shown in red. (C) Similar plot as (B) for 1MSC. The y-axis has been truncated to show normalized energies less than 2.0. (D) Loop conformations for 1CYO in the native structure (gray) and the lowest-energy decoys from REF2015 (purple) and Amber (turquoise). (E) Loop conformations for 1MSC in the native structure (gray) and lowest-energy decoys from REF2015 (purple) and Amber (turquoise).

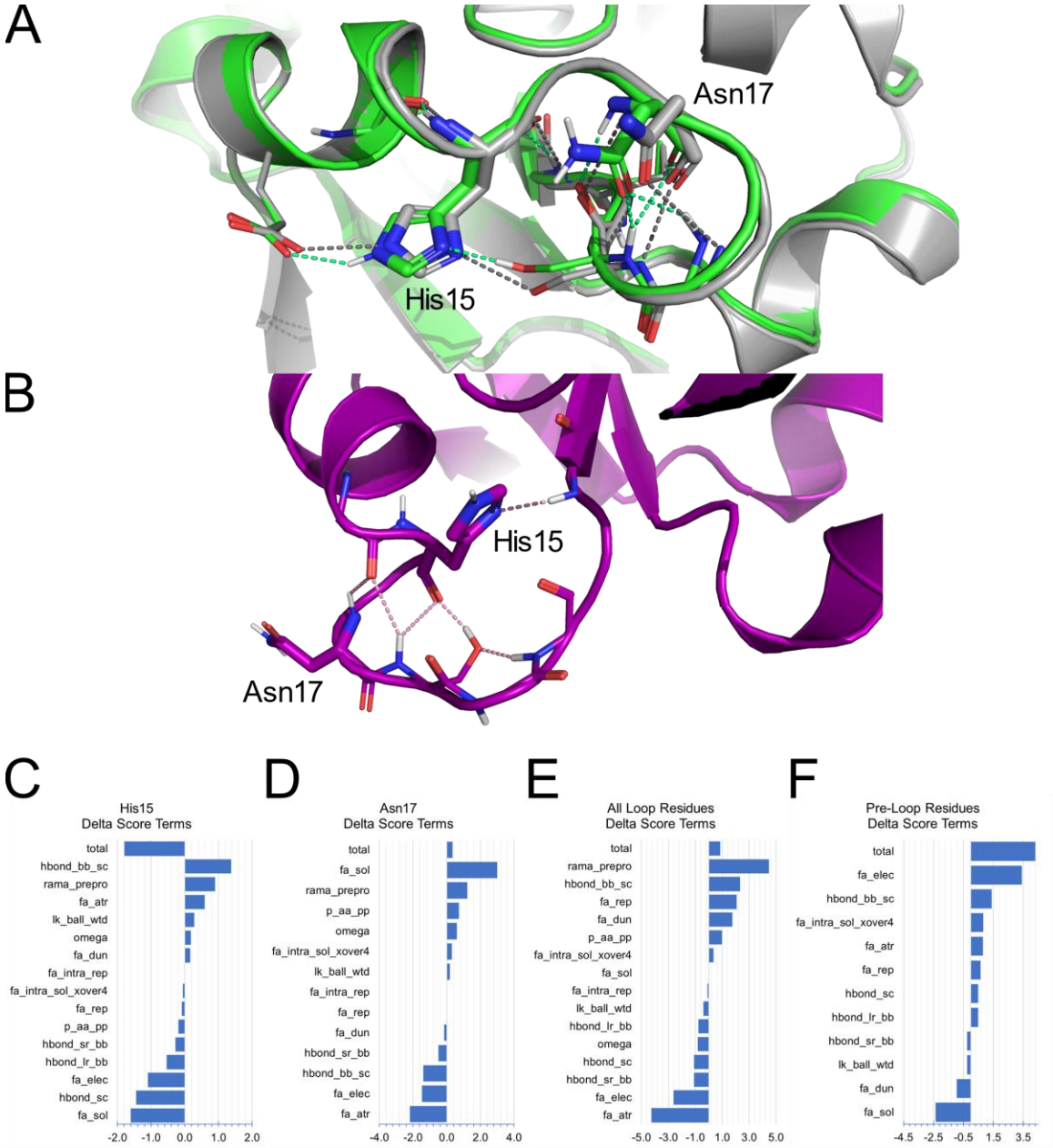


Figure 4-7: Comparison between the lowest energy decoy and lowest RMSD decoy for 1CYO. (A) The lowest RMSD decoy (green) has extensive hydrogen bond networks involving the backbones and chains of His15 and Asn17 (sticks), and is similar to the native 1CYO structure (grey). (B) The lowest energy decoy does not find native-like hydrogen bond networks for His15 and Asn17. (C) Clustered bar graph showing the delta between individual REF2015 score terms in the lowest RMSD decoy versus the lowest energy decoy for His15. More positive scores indicate terms where the lowest RMSD decoy scored worse. We only show the non-zero terms for clarity. (D) Similar graph to (C), but for Asn17. (E) Similar graph to (C), but for the sum over all residues in the

loop. (F) Similar graph to (C), but for the sum over five pre-loop residues.We similarly examined the lowest-RMSD and lowest-energy decoys in the Amber landscape to determine which ff14SBonlySC terms lead to the false minimum in the case of 1MSC, a case where Rosetta correctly discriminates against non-native conformations but Amber does not. Comparing the differences between the energy terms calculated over all residues for each structure, we found significantly unfavorable changes in the electrostatic (elec) and van der Waals (vdw) terms in the low-RMSD decoy (**Figure 4-8:A**) in Amber scoring. To specifically probe the contributions of the loop region and its effect on these overall energy differences, we compared the per-residue ff14SBonlySC energy decompositions for the low-RMSD decoy to those of the false-minimum. Although the GB term is not pair-wise decomposable over individual residues, comparing the differences for each of these terms can allow for a qualitative comparison of the two loop regions. We find that, although the majority of the near-native loop residues (eight out of 12) exhibit favorable total energies in the low-RMSD decoy, two residues have significantly unfavorable overall energies: Thr15 and Thr19. These residues and the C-terminal lobe of the loop in the false-minimum decoy are positioned such that an alternative hydrogen bond network is found between the loop and other portions of the protein (**Figure 4-8:B**), which weakens the energetic advantage of the near-native loop conformation. Furthermore, the false-minimum conformation allows for Thr15 to participate in more hydrogen bonds than the near-native conformation, where it has no hydrogen bond partners. This gives the false-minimum a considerably more favorable electrostatics score for this residue (**Figure 4-8C**). Similarly, Thr19 in the false-minimum structure participates in more hydrogen bonds than in the low-RMSD decoy, therefore allowing the false-minimum decoy to again have a more favorable electrostatics score for this residue (**Figure 4-8D**). Together, these results suggest an imbalance between the solvation terms (gb term when comparing whole decoys and polar solvation term in the

per-residue decomposition) and electrostatic terms are responsible for misprediction in this case.

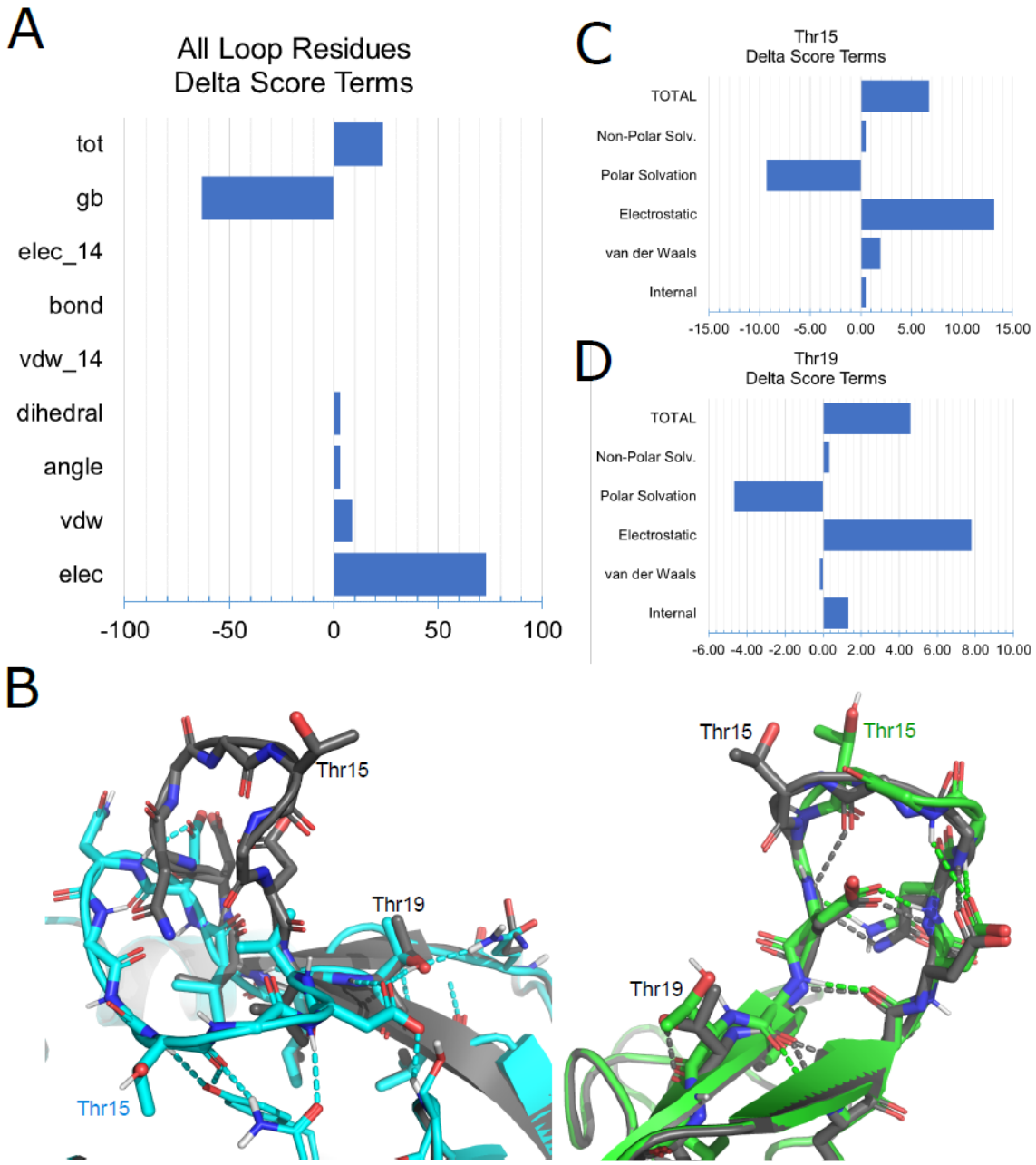


Figure 4-8: Comparison between the lowest-energy decoy and lowest-RMSD decoy for 1MSC.(A) Clustered bar graph showing the delta between individual ff14SBonlySC score terms in the lowest RMSD decoy versus the lowest energy decoy over all residues in the decoy structures. More positive scores indicate terms where the lowest RMSD decoy scored worse. (B) The lowest-energy decoy (turquoise) has extensive hydrogen bond networks involving the backbones and sidechains the C-terminal lobe of the loop to non-loop regions of the

protein. The lowest-RMSD decoy utilizes a similar hydrogen bond network to the native 1MSC structure (grey) where most of the hydrogen bonds are between backbone and sidechain groups in the loop. (C) Similar graph to (A), but for the decomposed scoreterms for residue Thr15. (D) Similar graph to (A), but for the decomposed scoreterms for residue Thr19.

4.6 Discussion

Systematic comparison of Amber ff14SBonlySC (a physically-derived energy function) and Rosetta talaris2014 and Rosetta REF2015 (both physical and statistical based) reveals the strengths and weaknesses of each energy function. Generally, Amber ff14SBonlySC performs better than Rosetta talaris2014 at both decoy discrimination and loop modeling. The comparison of Amber ff14SBonlySC to Rosetta REF2015 (the current default Rosetta energy function) reveals that REF2015, which has more physically-derived terms than talaris2014, performs comparably well or better in certain cases when compared to Amber ff14SBonlySC. Examination of Rosetta energy function score-terms that rescue Amber ff14SBonlySC false minima and Amber ff14SBonlySC score-terms that correct Rosetta false minima reveals two possible sources for the performance improvement of REF2015: improved dihedral potential and electrostatics model. While two of the Rosetta score-terms and two of the Amber score-terms that contribute to the correction of false minima are counterparts of each other (Amber dihedral and Rosetta rama, and Amber elec and Rosetta fa_elec), differences in their derivation and parameterization appear to influence the propensity of each energy function toward false minima. Although rama and dihedral both score the propensity of the backbone dihedral angles, rama does so to fit dihedral distributions in native protein structures while dihedral is based on fits to quantum chemistry data on small molecules. Similarly, while both elec and fa_elec are derived from a Coulombic model, yet they are differently parameterized; the Amber elec is parameterized via small-molecule properties, whereas fa_elec is optimized on larger biomolecular structures. The improvement of Rosetta REF2015 over Rosetta talaris2014 may be caused by its greater inclusion of physical-derived terms (bond,

angle, etc.) and/or its extensive parameterization on both small-molecule properties (more typical of so-called "physics-based" force fields) and larger biomolecular structures. Conversely, the training of ff14SBonlySC solvation parameters[22] using the computed energy landscapes of a diverse set of proteins, which is more typical of so-called "knowledge-based" force field development, may contribute to its success observed here. Thus, a promising direction for force field development involves combining both small molecule and macromolecular benchmarks, using fitting to diverse experimental and simulation data.

We note that the sampling method used for decoy generation and choice of benchmark sets also contribute to the apparent performance of a given energy function. In our tests, all decoys were generated by extensive sampling using Rosetta talaris2014, so the pathologies of this energy function are more likely to be highlighted. REF2015 was developed to overcome these pathologies using the same decoy sets (used either for training or evaluation), some pathologies remain as evidenced by loop modeling tests (**Figure 4-6**) which were not considered by Park et al.[18] It will be interesting to perform the converse experiment: extensive sampling with Amber ff14SBonlySC followed by Rosetta scoring may reveal the pathologies of the former and aid in its further development. Thus, iterative sampling and cross-scoring with differently derived energy functions of sufficiently high quality may be a generally valuable approach for force-field improvement.

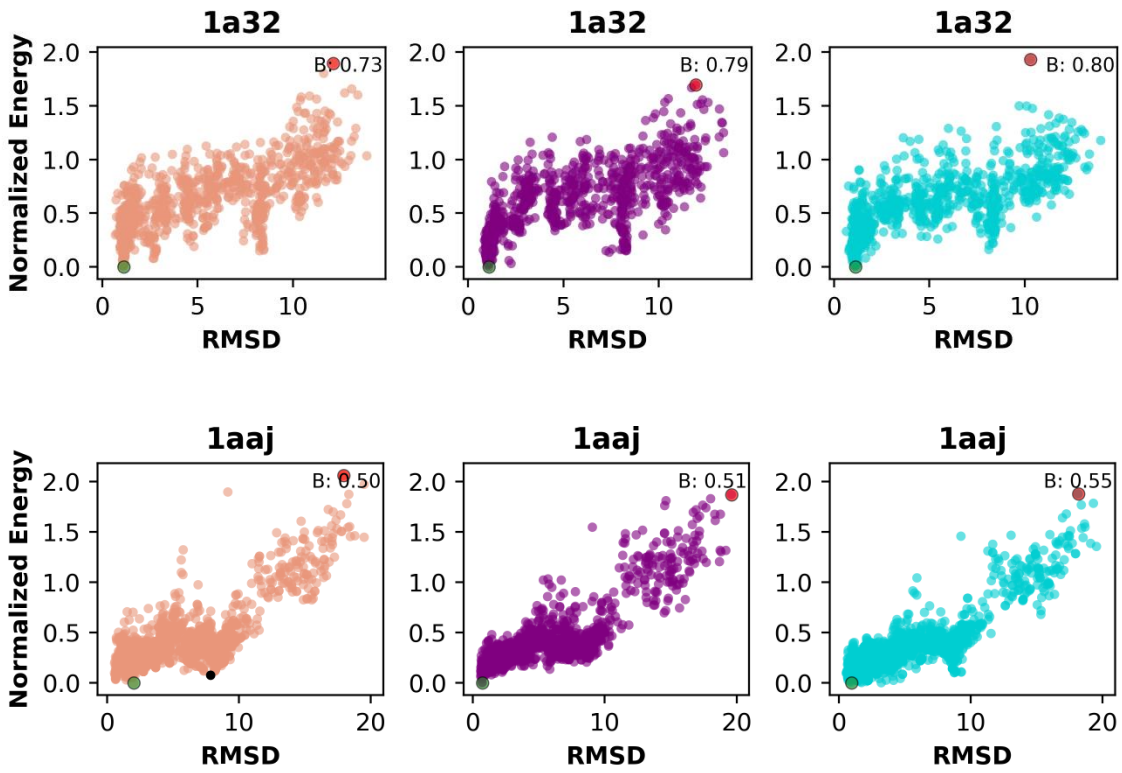
Model selection, or the ability to select a near-native decoy from a set of decoy conformations is a general problem in protein structure prediction. If low-energy decoys exist in false minima in the energy landscape, it is difficult to identify conformations that are near-native. Since Amber and Rosetta provide different, semi-orthogonal information, a framework to combine the two rankings enables more efficient identification of near-native decoys. The minimum-sum based algorithm that we have implemented improves

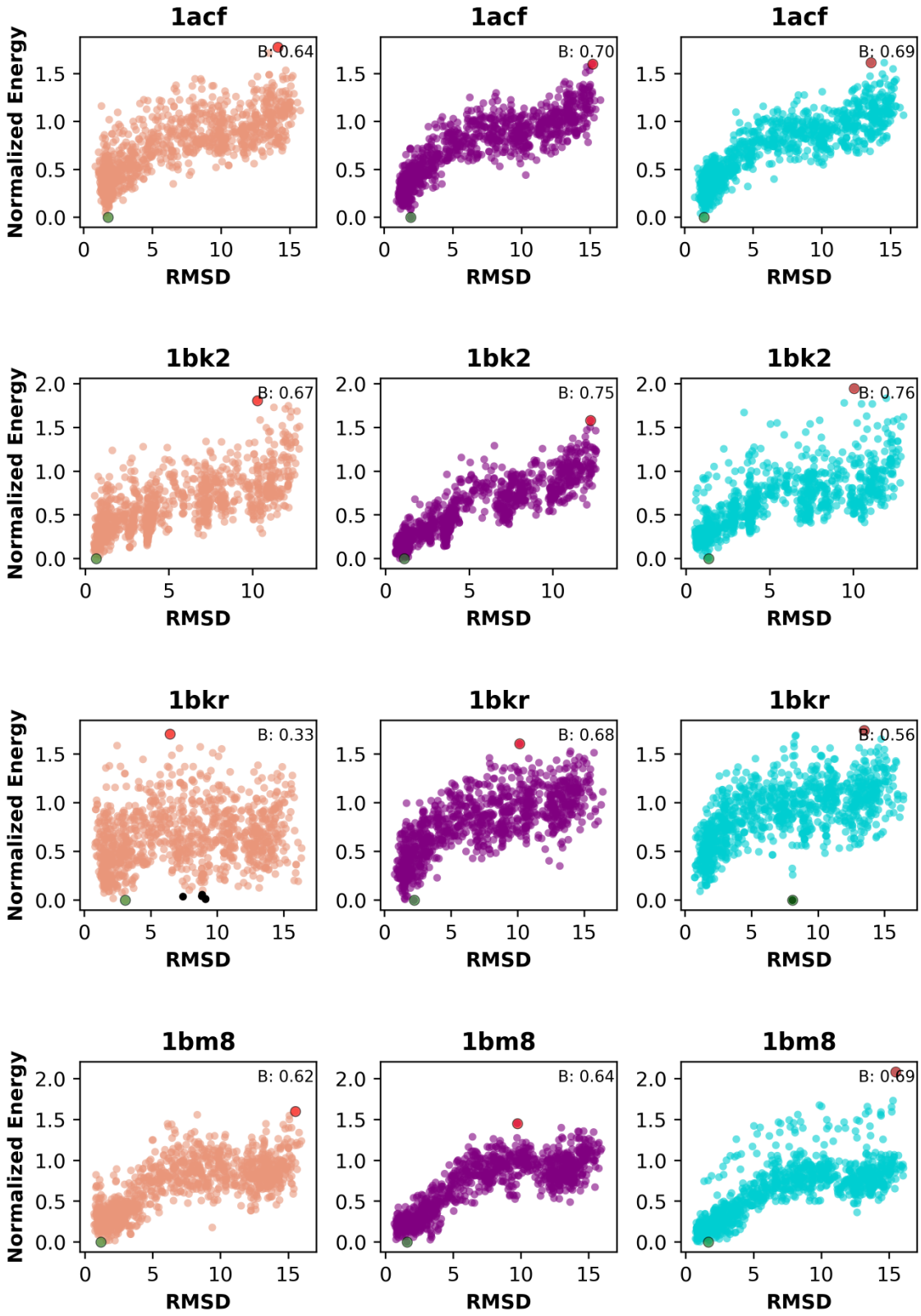
model selection for 9% of structures over Rosetta REF2015 model selection and 10% of structures over Amber ff14SBonlySC model selection (**Figure 4-5**, **Table 4-1**). The model selection algorithm is extensible to any two sets of energy functions or model ranks for one set of models and can thus be used to combine any two sources of information to produce meaningful improvements in near-native decoy selection. Similarly, scores from three or more energy functions can be combined using a Pareto optimality framework.

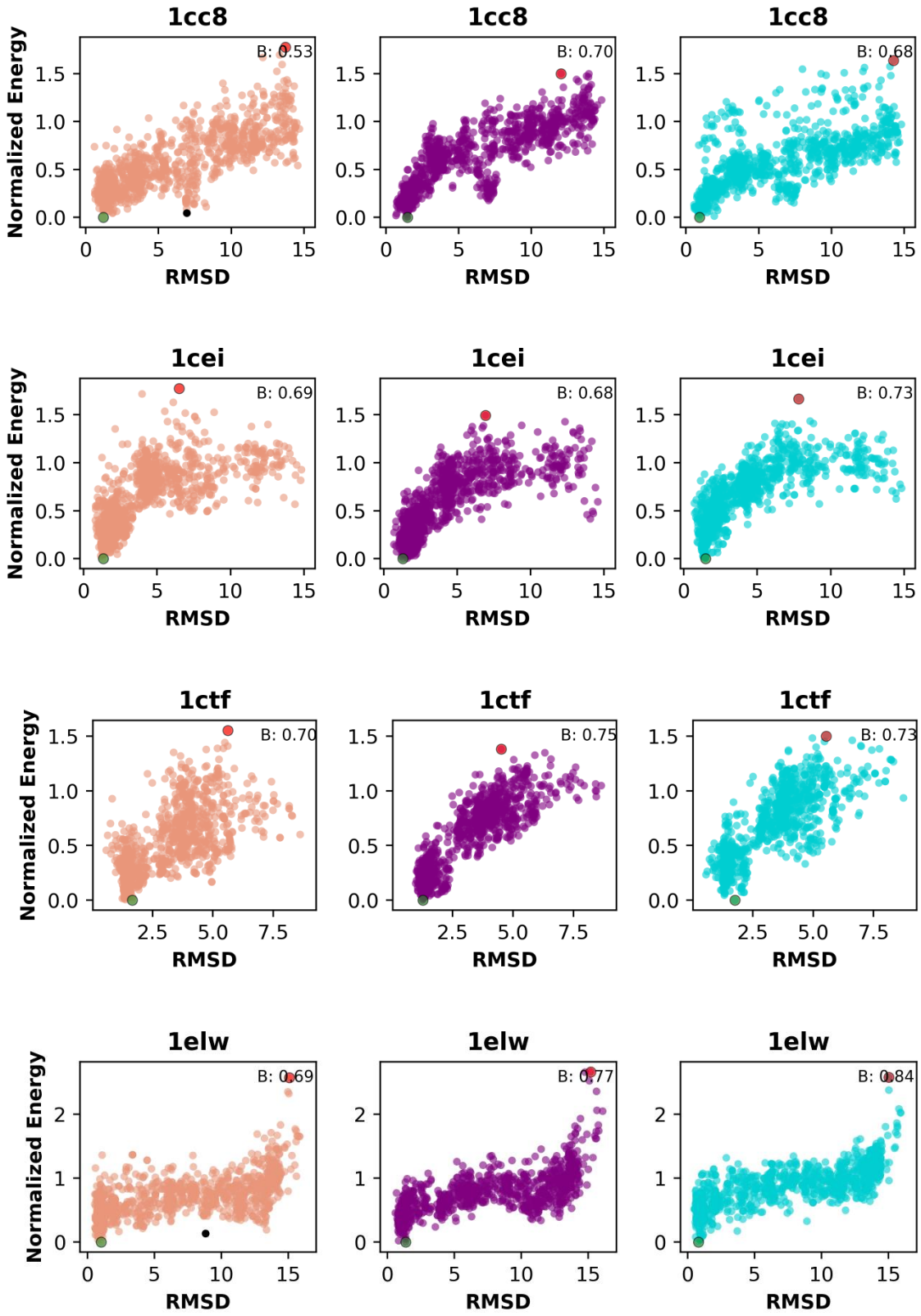
The approach described here should enable comparative analysis and combination of future versions of both Amber and Rosetta scoring functions and enable a variety of biomolecular modeling tasks.

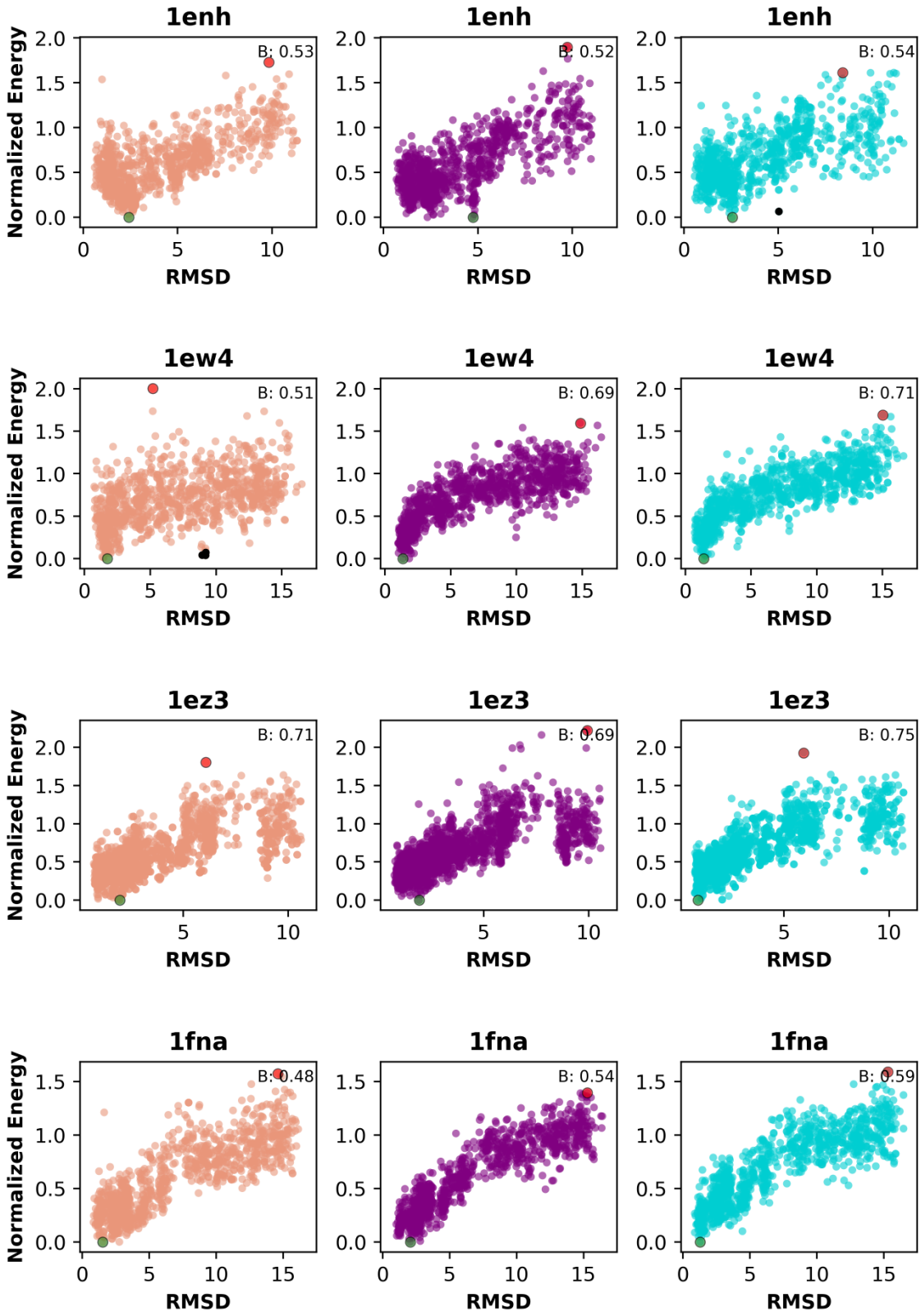
4.7 Supplementary Figures

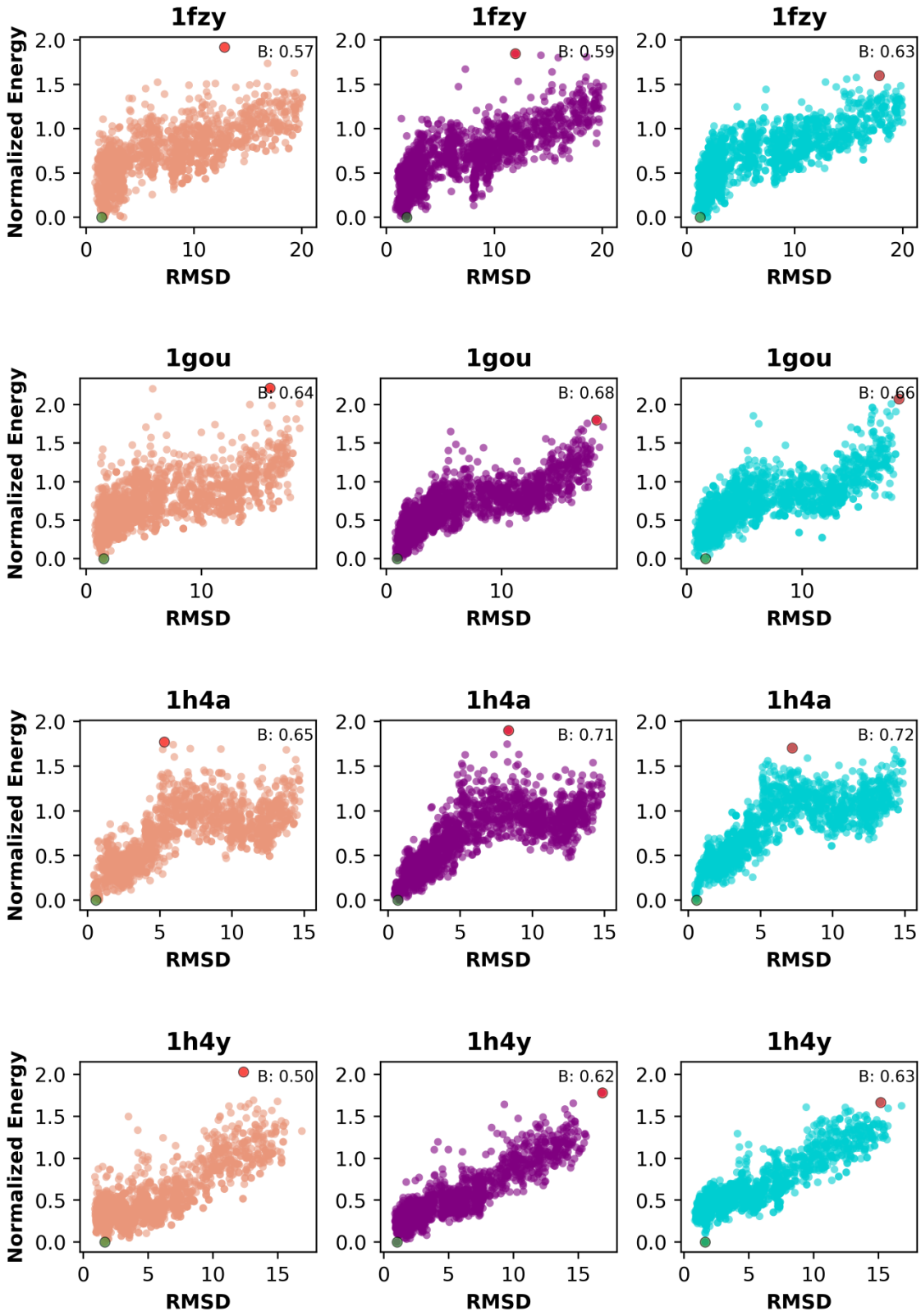
Figure 4-9: Energy landscapes for all decoy discrimination proteins.(Talaris2014, Left/Salmon; REF2015, Middle/Purple; ff14sbonlysc Right/Turquoise)

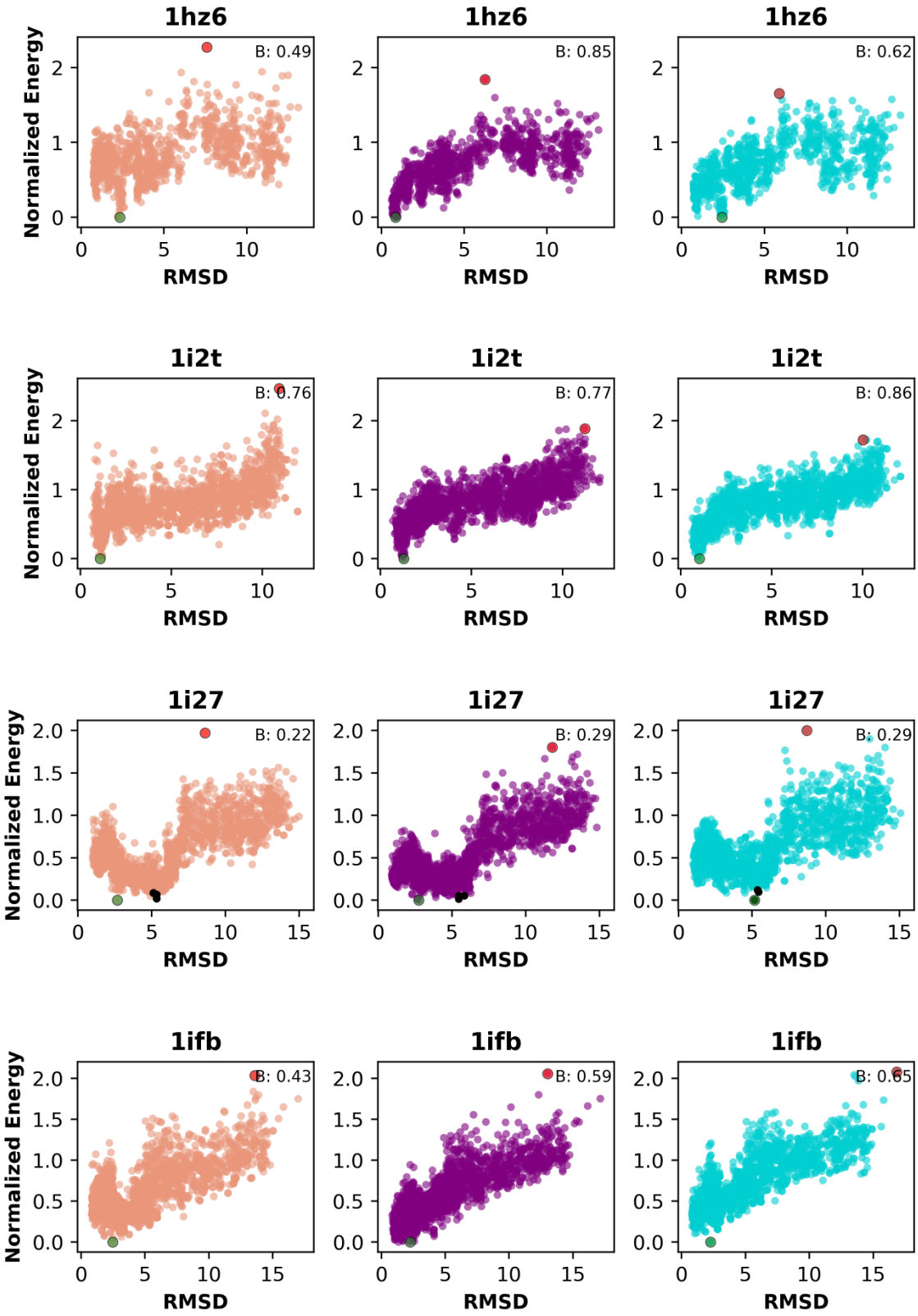


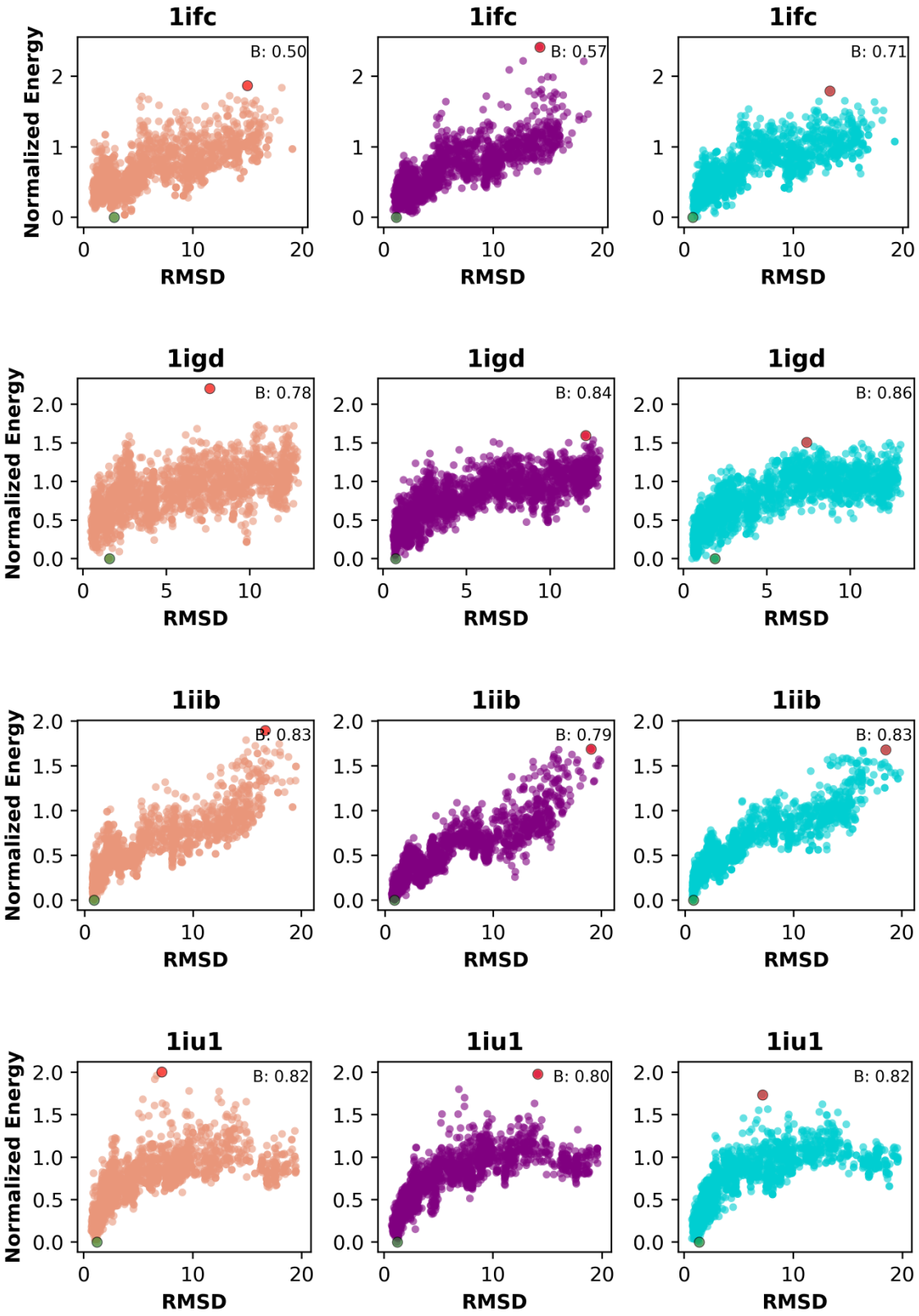

1acf
B: 0.64
1acf
B: 0.70
1acf
B: 0.69
1bk2
B: 0.67
1bk2
B: 0.75
1bk2
B: 0.76
1bkr
B: 0.33
1bkr
B: 0.68
1bkr
B: 0.56
1bm8
B: 0.62
1bm8
B: 0.64
1bm8
B: 0.69
Normalized Energy
RMSD

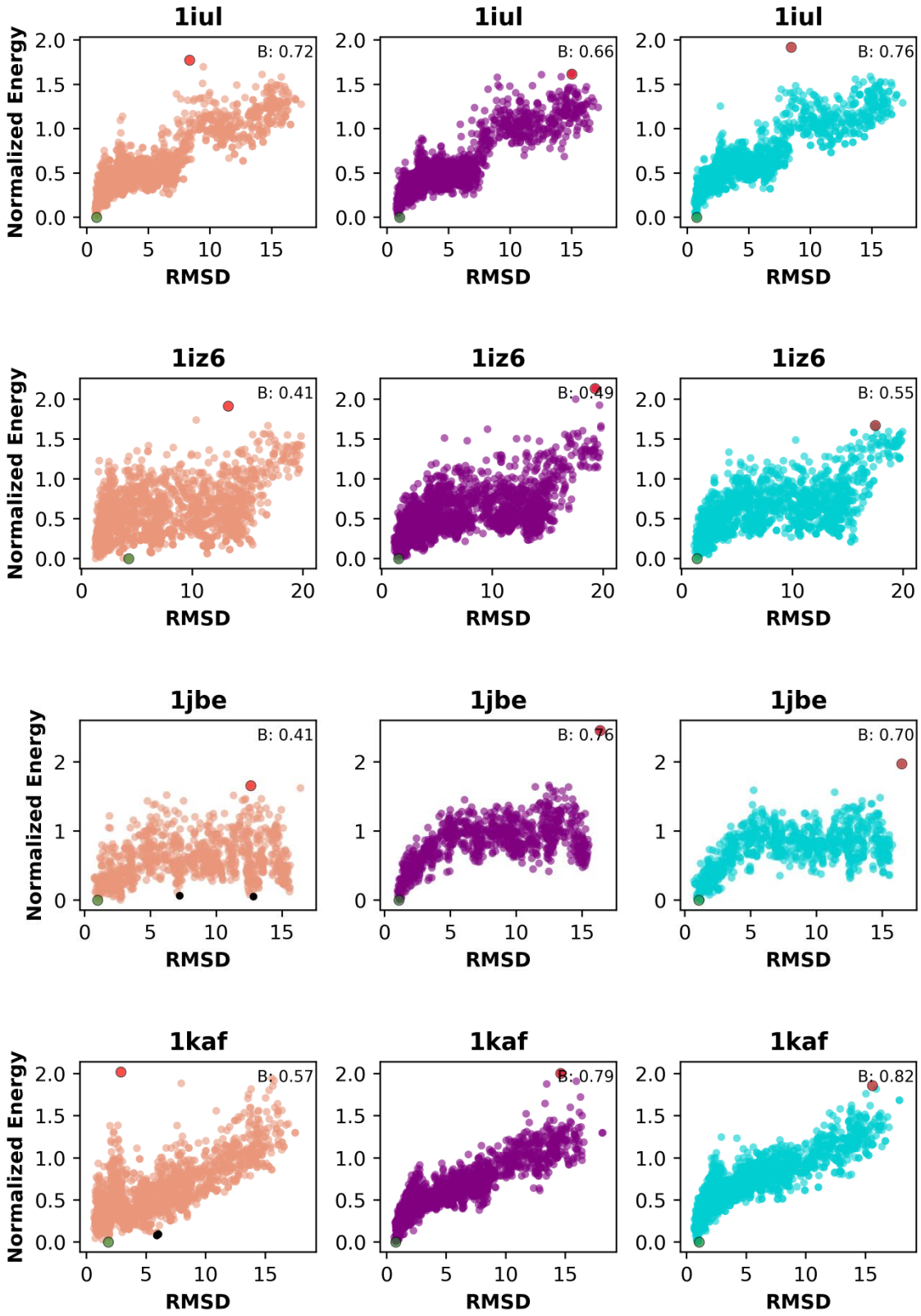

1cc8
B: 0.53
1cc8
B: 0.70
1cc8
B: 0.68
1cei
B: 0.69
1cei
B: 0.68
1cei
B: 0.73
1ctf
B: 0.70
1ctf
B: 0.75
1ctf
B: 0.73
1elw
B: 0.69
1elw
B: 0.77
1elw
B: 0.84
Normalized Energy
RMSD

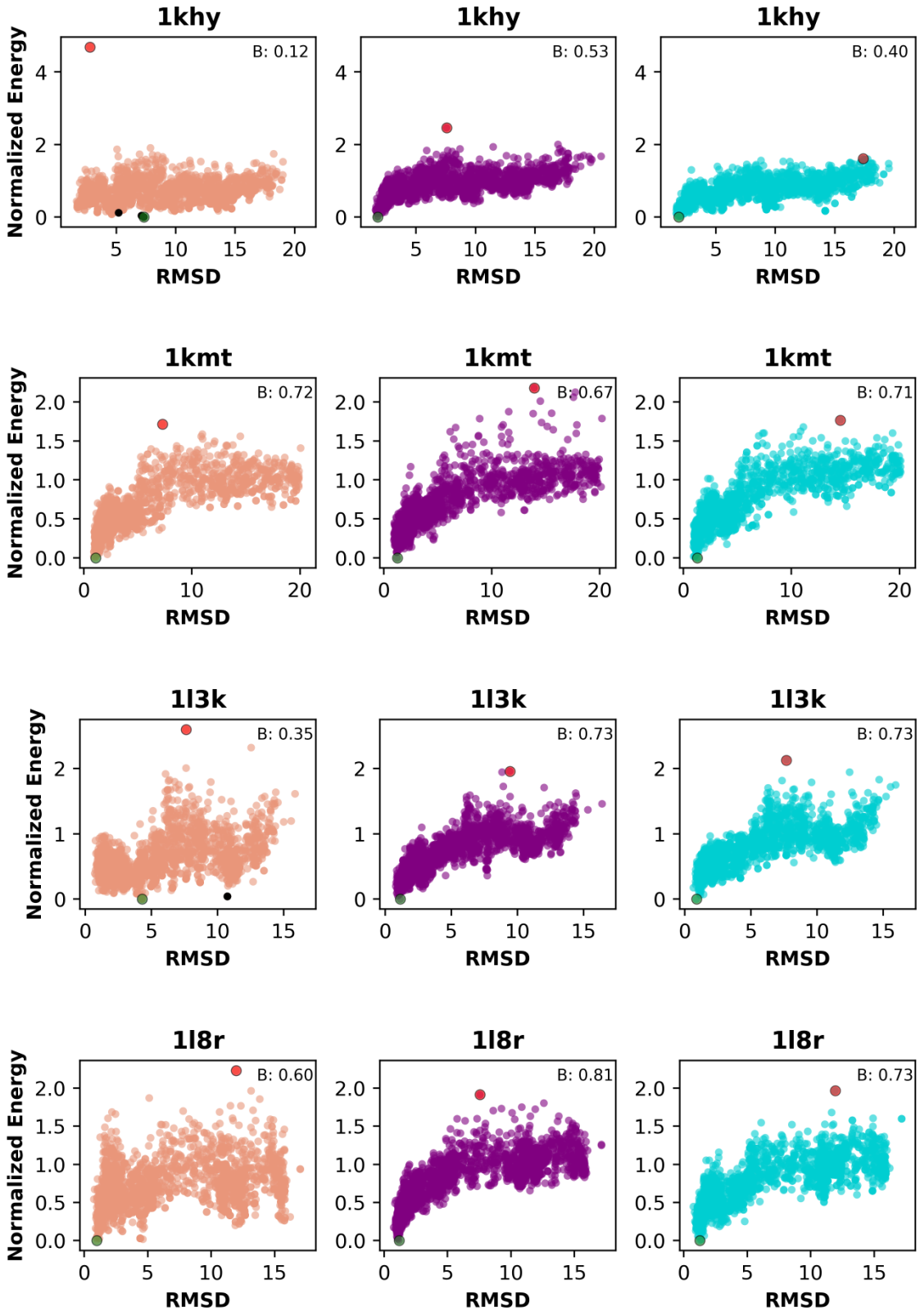

1enh
B: 0.53
1enh
B: 0.52
1enh
B: 0.54
1ew4
B: 0.51
1ew4
B: 0.69
1ew4
B: 0.71
1ez3
B: 0.71
1ez3
B: 0.69
1ez3
B: 0.75
1fna
B: 0.48
1fna
B: 0.54
1fna
B: 0.59
Normalized Energy
RMSD

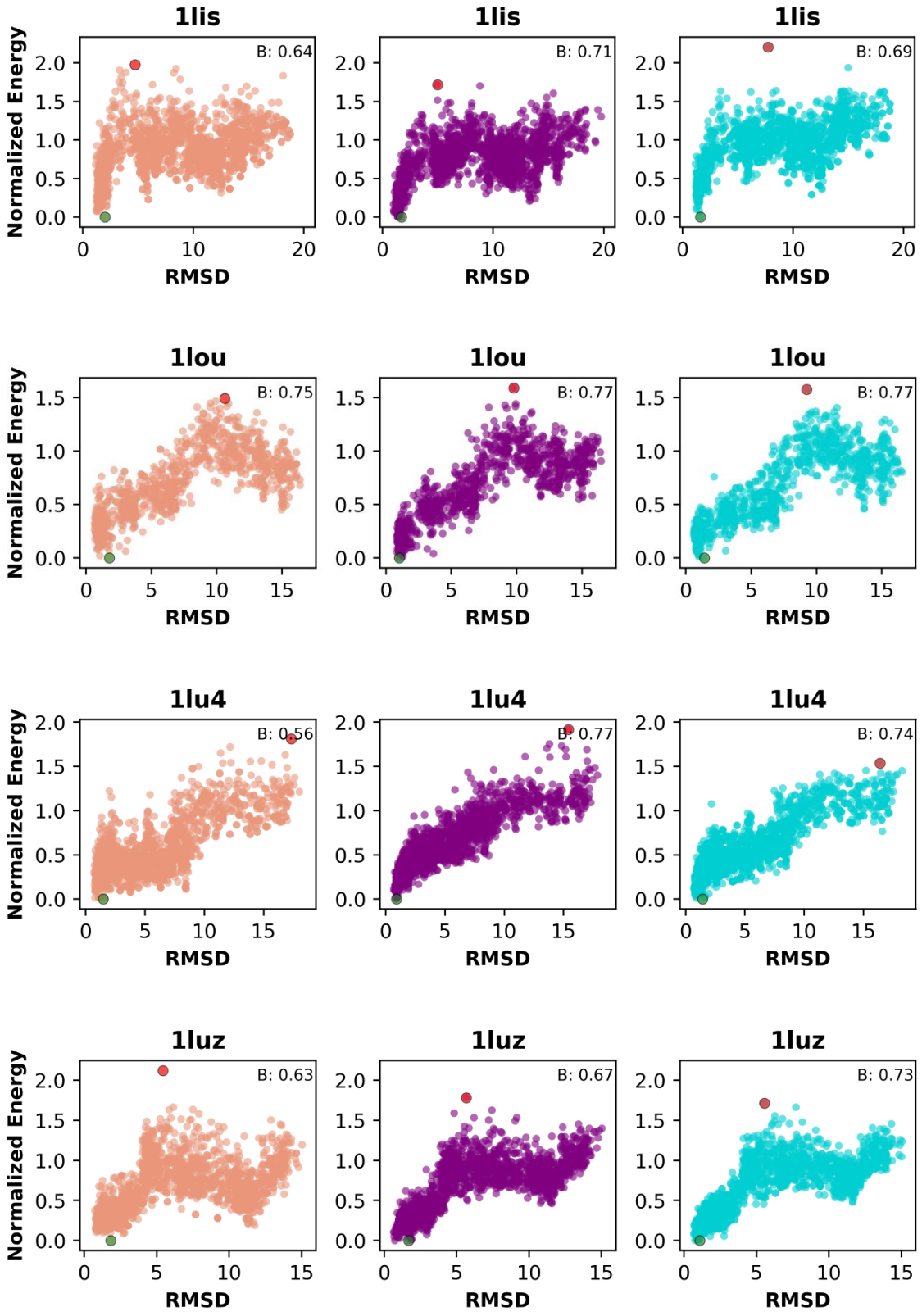

1fzy
B: 0.57
1fzy
B: 0.59
1fzy
B: 0.63
1gou
B: 0.64
1gou
B: 0.68
1gou
B: 0.66
1h4a
B: 0.65
1h4a
B: 0.71
1h4a
B: 0.72
1h4y
B: 0.50
1h4y
B: 0.62
1h4y
B: 0.63
Normalized Energy
RMSD

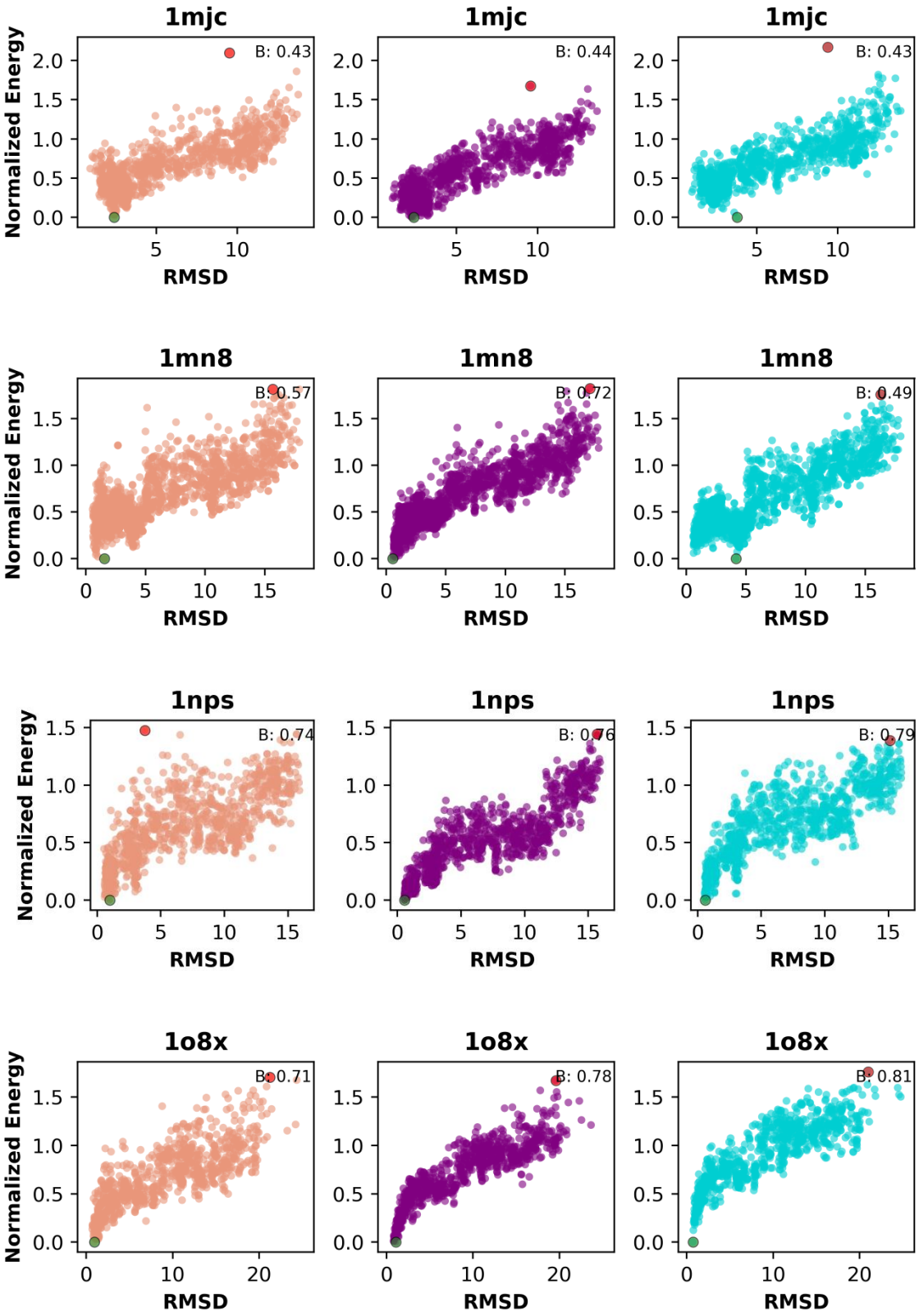

1hz6
B: 0.49
1hz6
B: 0.85
1hz6
B: 0.62
1i2t
B: 0.76
1i2t
B: 0.77
1i2t
B: 0.86
1i27
B: 0.22
1i27
B: 0.29
1i27
B: 0.29
1ifb
B: 0.43
1ifb
B: 0.59
1ifb
B: 0.65
Normalized Energy
RMSD

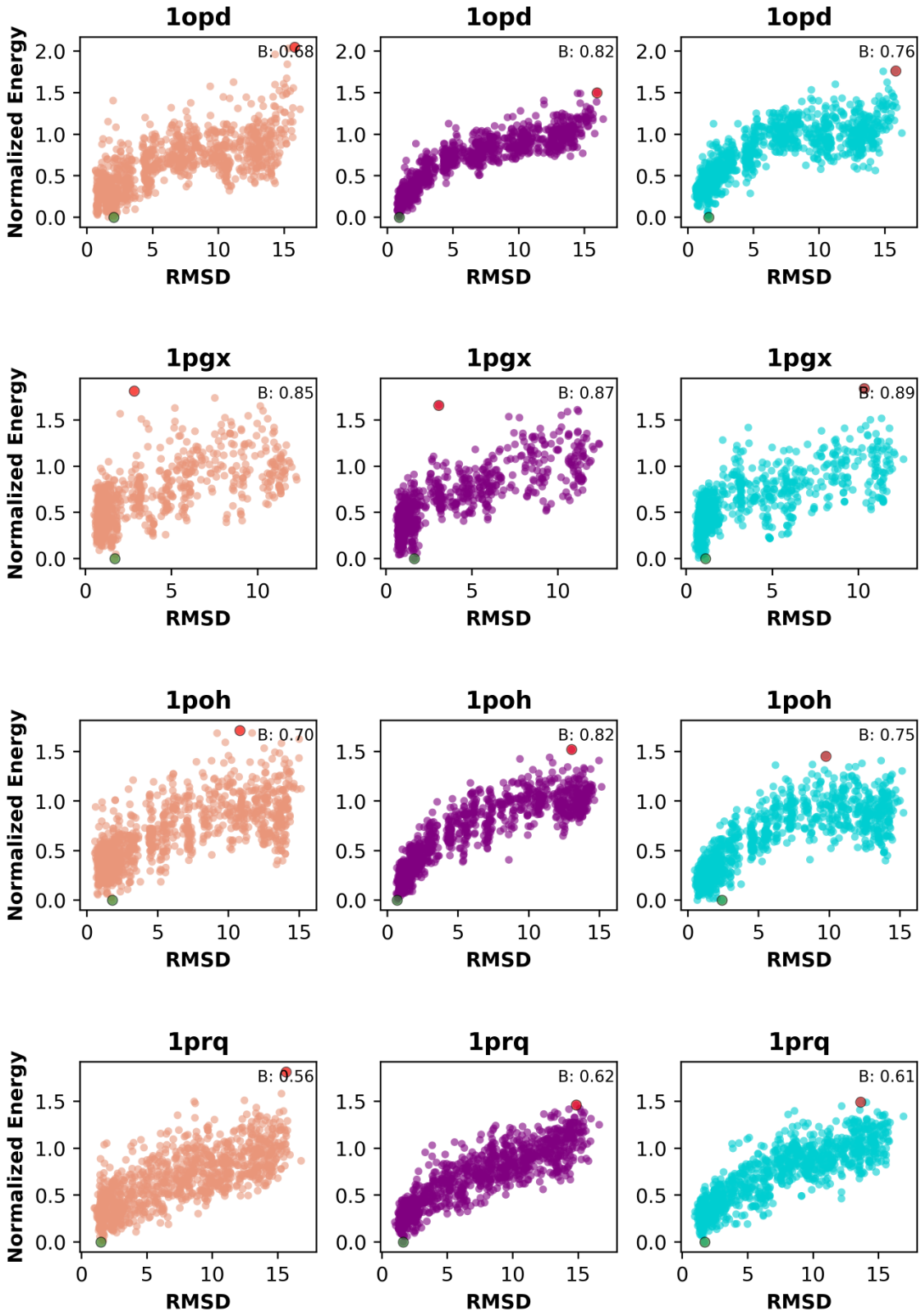

1ifc
B: 0.50
1ifc
B: 0.57
1ifc
B: 0.71
1igd
B: 0.78
1igd
B: 0.84
1igd
B: 0.86
1iib
B: 0.83
1iib
B: 0.79
1iib
B: 0.83
1iu1
B: 0.82
1iu1
B: 0.80
1iu1
B: 0.82
Normalized Energy
RMSD

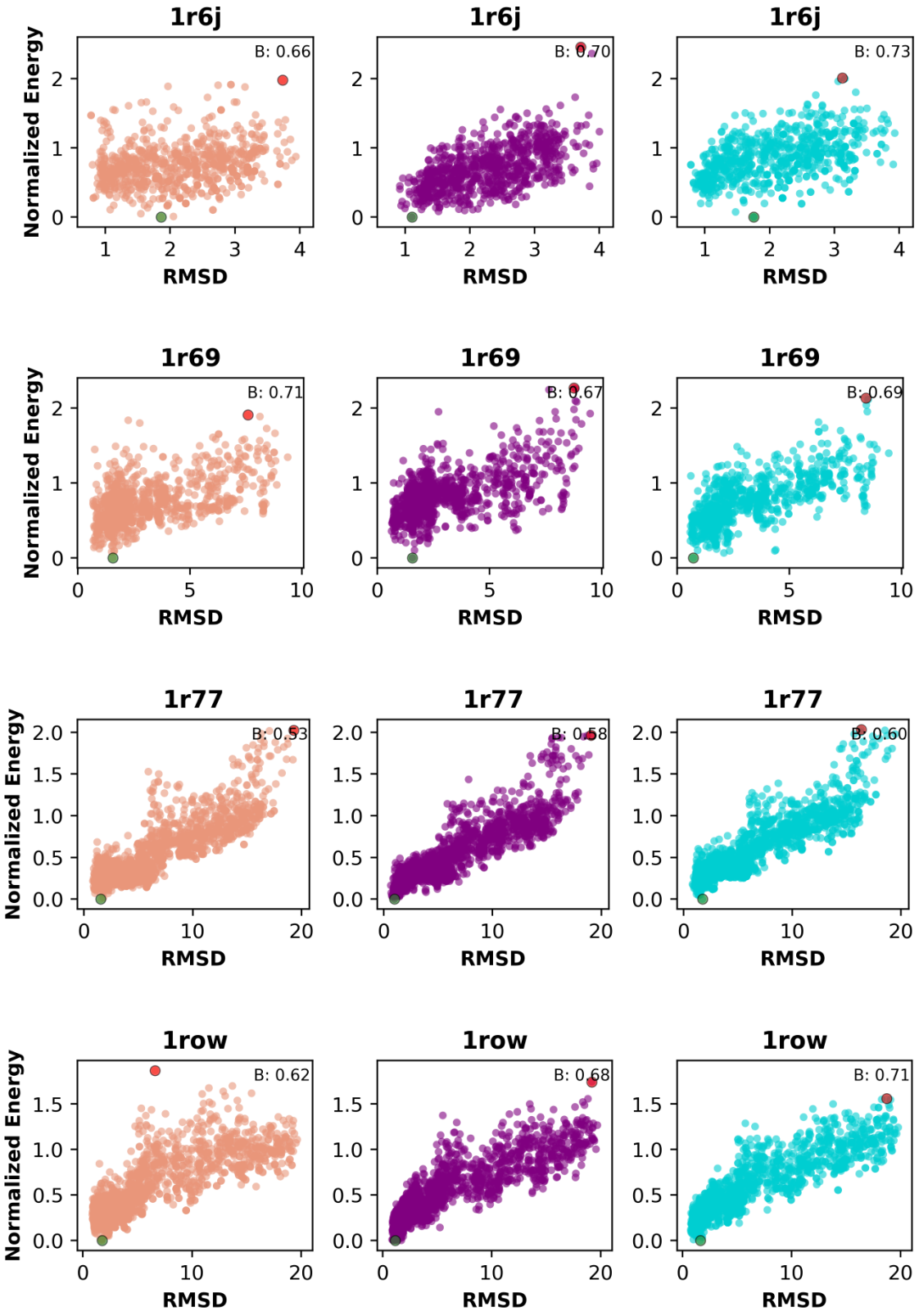

1iul
B: 0.72
1iul
B: 0.66
1iul
B: 0.76
1iz6
B: 0.41
1iz6
B: 0.49
1iz6
B: 0.55
1jbe
B: 0.41
1jbe
B: 0.76
1jbe
B: 0.70
1kaf
B: 0.57
1kaf
B: 0.79
1kaf
B: 0.82
Normalized Energy
RMSD

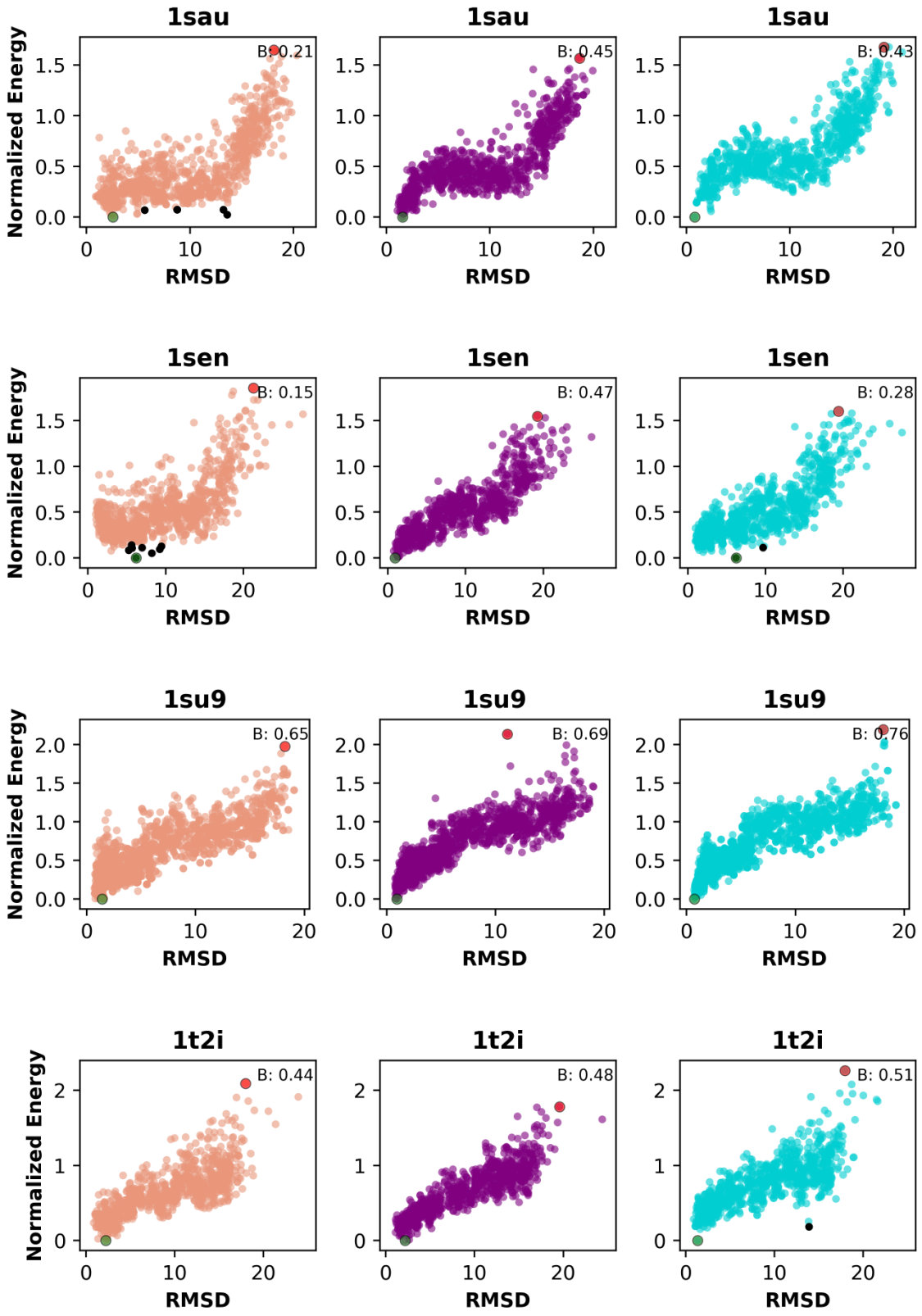

1khy
B: 0.12
1khy
B: 0.53
1khy
B: 0.40
1kmt
B: 0.72
1kmt
B: 0.67
1kmt
B: 0.71
1l3k
B: 0.35
1l3k
B: 0.73
1l3k
B: 0.73
1l8r
B: 0.60
1l8r
B: 0.81
1l8r
B: 0.73
Normalized Energy
RMSD

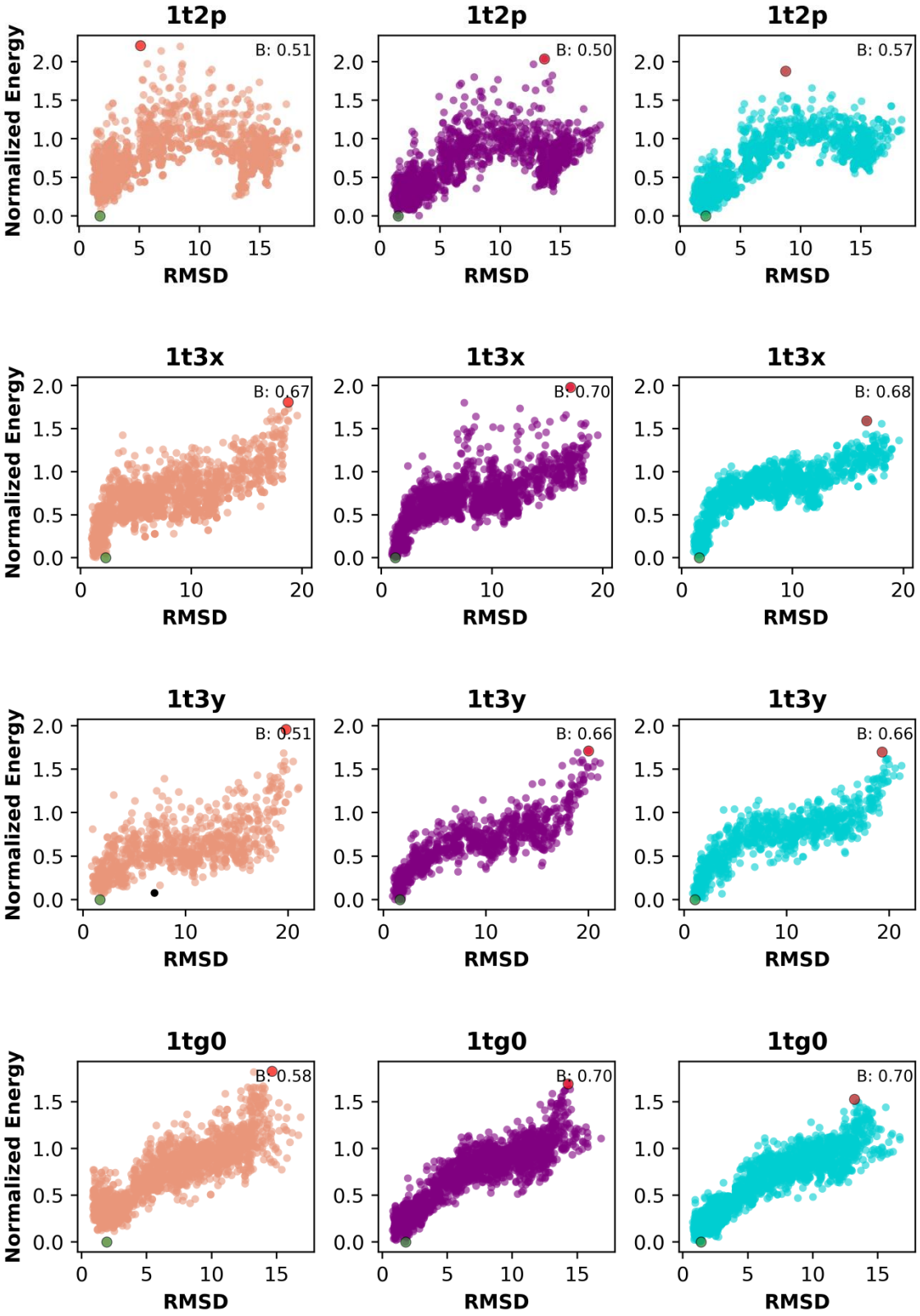

1lis
B: 0.64
1lis
B: 0.71
1lis
B: 0.69
1lou
B: 0.75
1lou
B: 0.77
1lou
B: 0.77
1lu4
B: 0.56
1lu4
B: 0.77
1lu4
B: 0.74
1luz
B: 0.63
1luz
B: 0.67
1luz
B: 0.73
Normalized Energy
RMSD

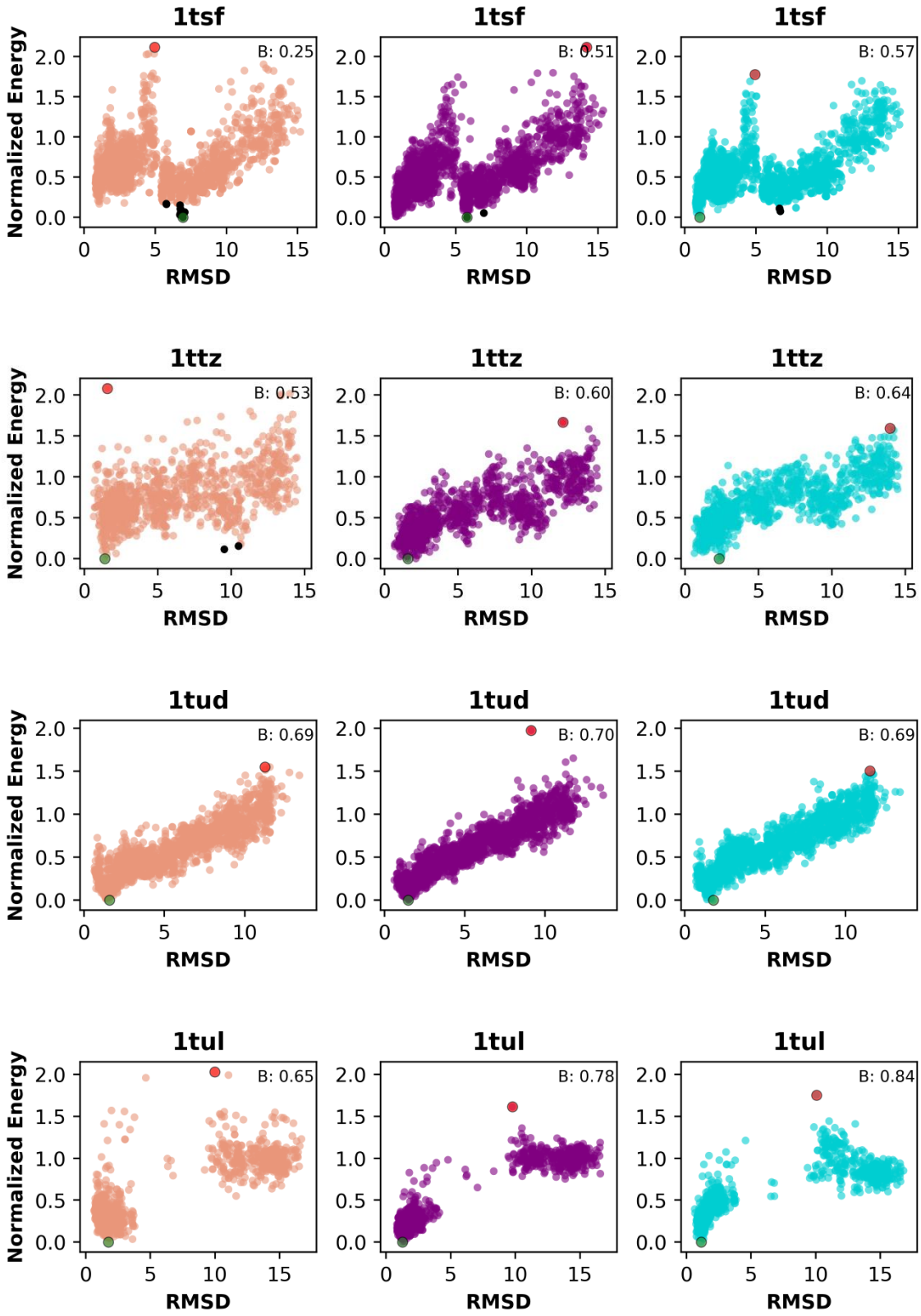

1mjc
B: 0.43
1mjc
B: 0.44
1mjc
B: 0.43
1mn8
B: 0.57
1mn8
B: 0.72
1mn8
B: 0.49
1nps
B: 0.74
1nps
B: 0.76
1nps
B: 0.79
1o8x
B: 0.71
1o8x
B: 0.78
1o8x
B: 0.81
Normalized Energy
RMSD

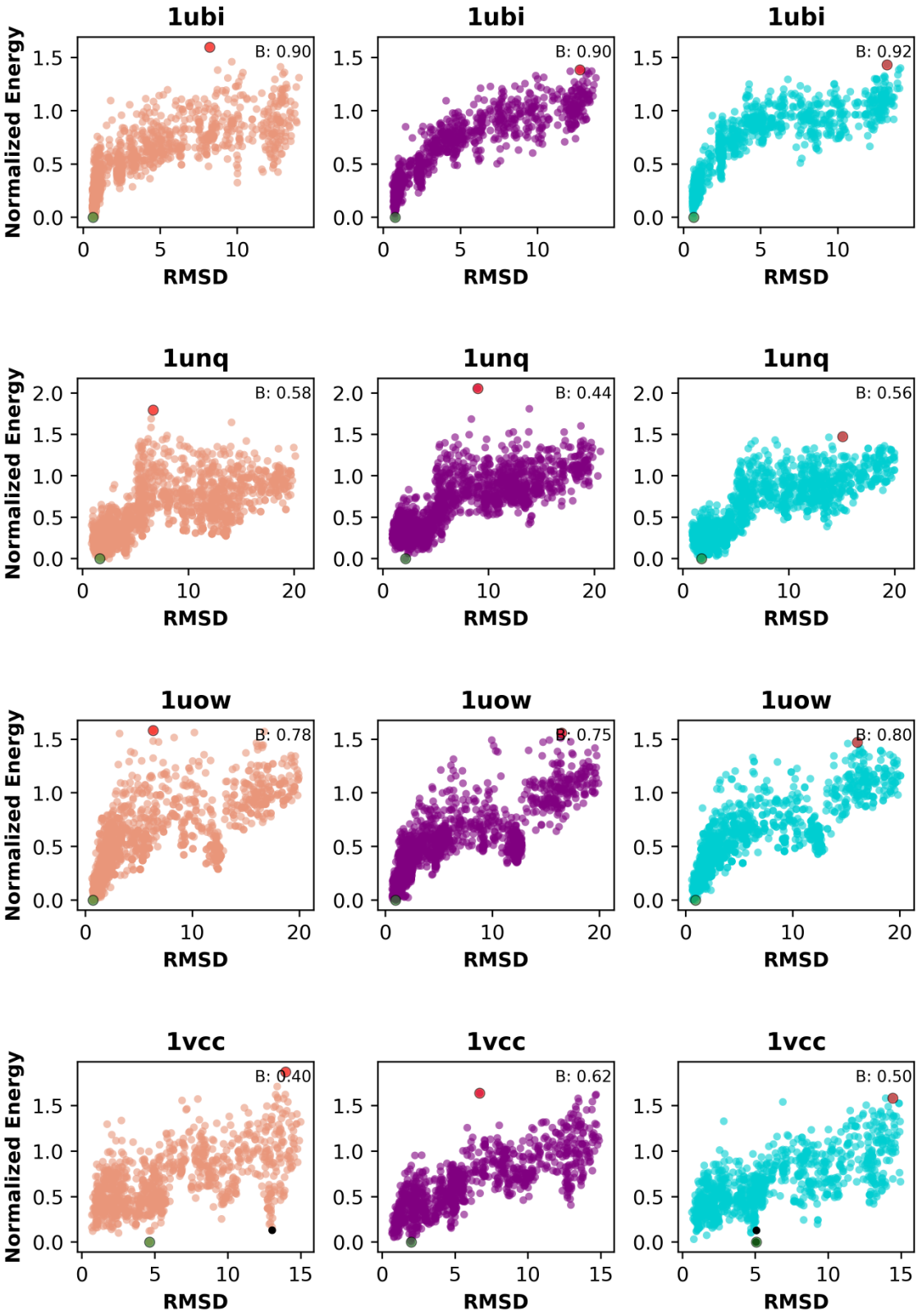

1opd
B: 0.68
1opd
B: 0.82
1opd
B: 0.76
1pgx
B: 0.85
1pgx
B: 0.87
1pgx
B: 0.89
1poh
B: 0.70
1poh
B: 0.82
1poh
B: 0.75
1prq
B: 0.56
1prq
B: 0.62
1prq
B: 0.61
Normalized Energy
RMSD

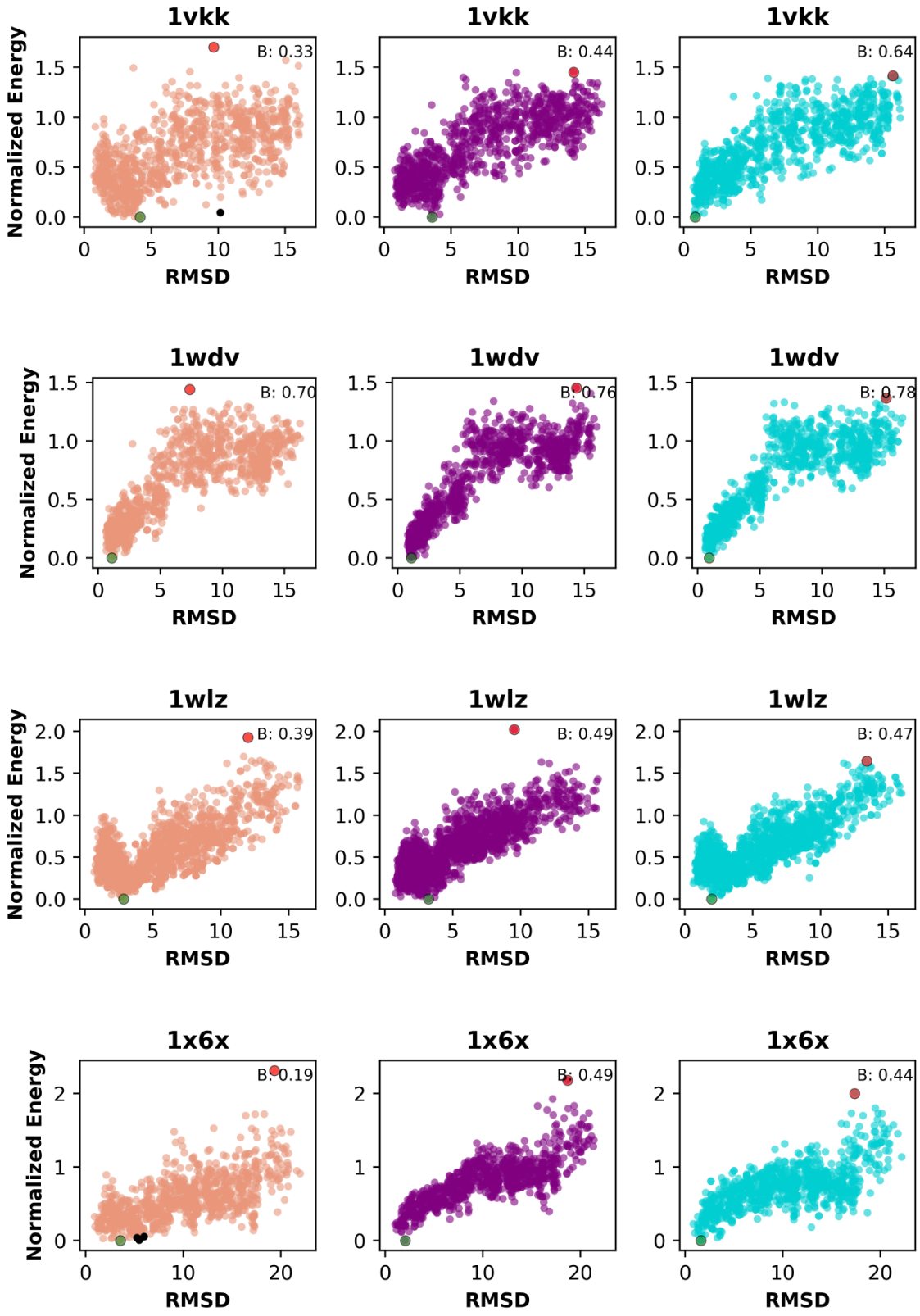

1r6j
B: 0.66
1r6j
B: 0.70
1r6j
B: 0.73
1r69
B: 0.71
1r69
B: 0.67
1r69
B: 0.69
1r77
B: 0.53
1r77
B: 0.58
1r77
B: 0.60
1row
B: 0.62
1row
B: 0.68
1row
B: 0.71
Normalized Energy
RMSD

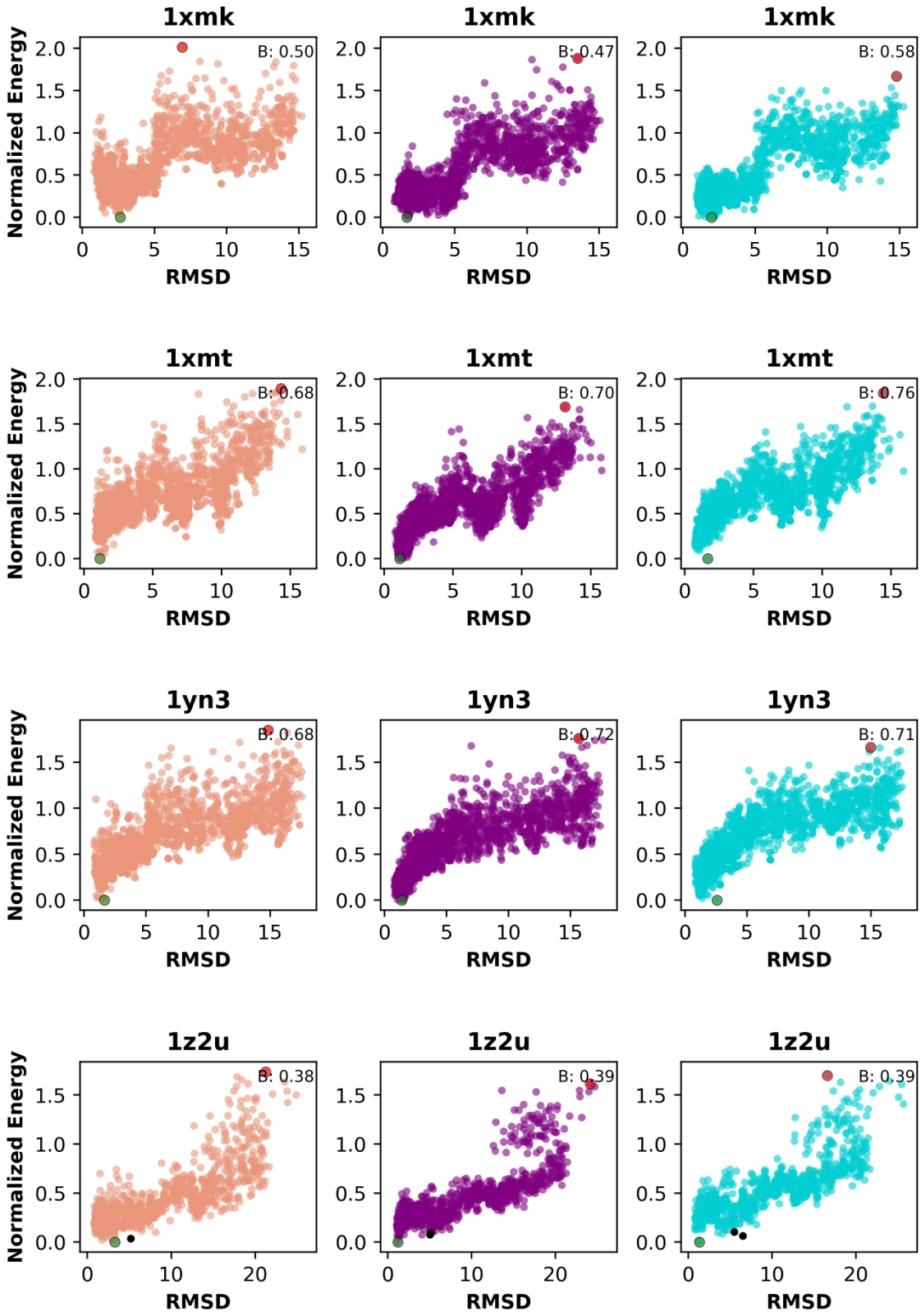

1sau
B: 0.21
1sau
B: 0.45
1sau
B: 0.43
1sen
B: 0.15
1sen
B: 0.47
1sen
B: 0.28
1su9
B: 0.65
1su9
B: 0.69
1su9
B: 0.76
1t2i
B: 0.44
1t2i
B: 0.48
1t2i
B: 0.51
Normalized Energy
RMSD

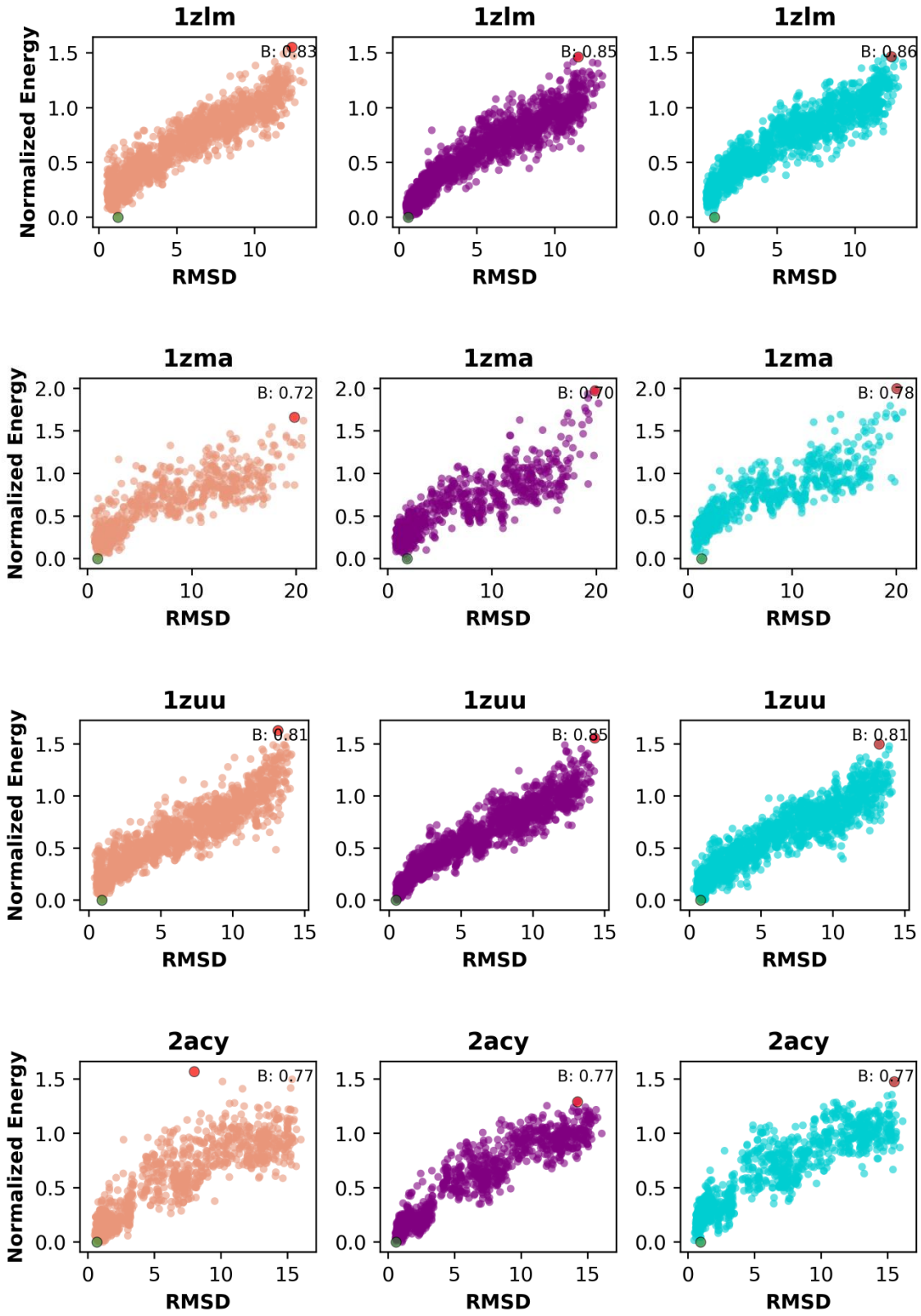

1t2p
B: 0.51
1t2p
B: 0.50
1t2p
B: 0.57
1t3x
B: 0.67
1t3x
B: 0.70
1t3x
B: 0.68
1t3y
B: 0.51
1t3y
B: 0.66
1t3y
B: 0.66
1tg0
B: 0.58
1tg0
B: 0.70
1tg0
B: 0.70
Normalized Energy
RMSD

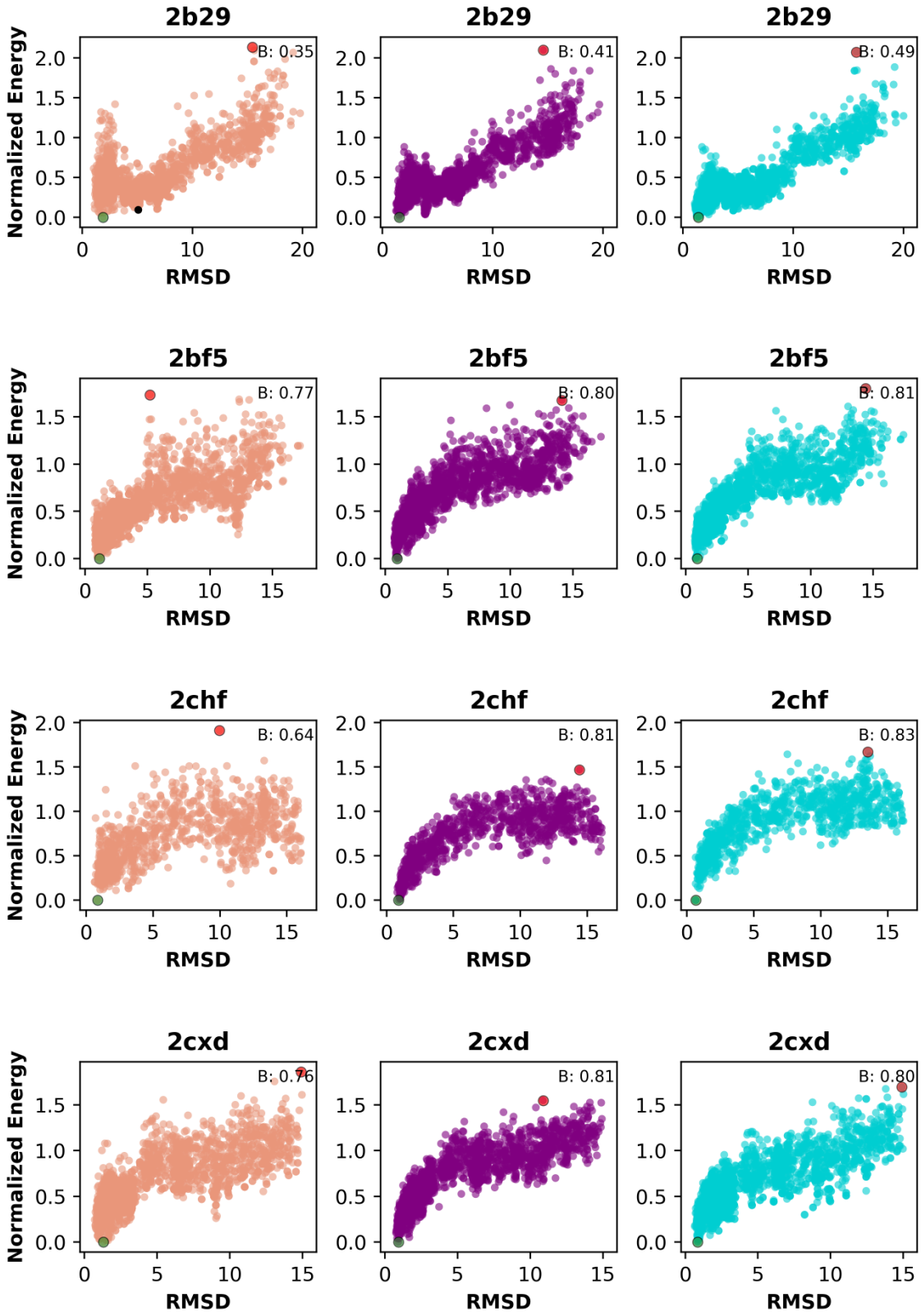

1tsf
B: 0.25
1tsf
B: 0.51
1tsf
B: 0.57
1ttz
B: 0.53
1ttz
B: 0.60
1ttz
B: 0.64
1tud
B: 0.69
1tud
B: 0.70
1tud
B: 0.69
1tul
B: 0.65
1tul
B: 0.78
1tul
B: 0.84
Normalized Energy
RMSD

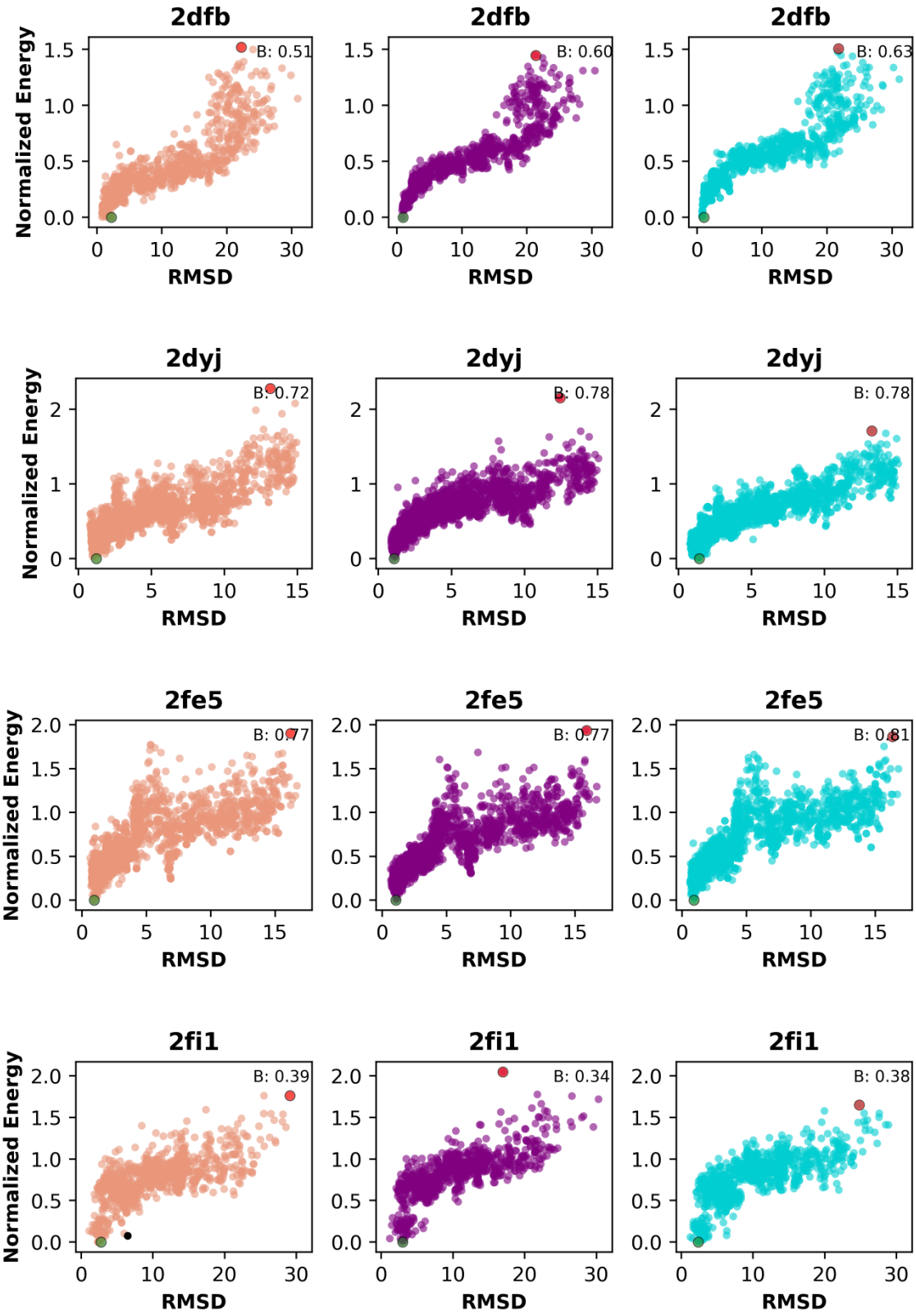

1ubi
B: 0.90
1ubi
B: 0.90
1ubi
B: 0.92
1unq
B: 0.58
1unq
B: 0.44
1unq
B: 0.56
1uow
B: 0.78
1uow
B: 0.75
1uow
B: 0.80
1vcc
B: 0.40
1vcc
B: 0.62
1vcc
B: 0.50
Normalized Energy
RMSD

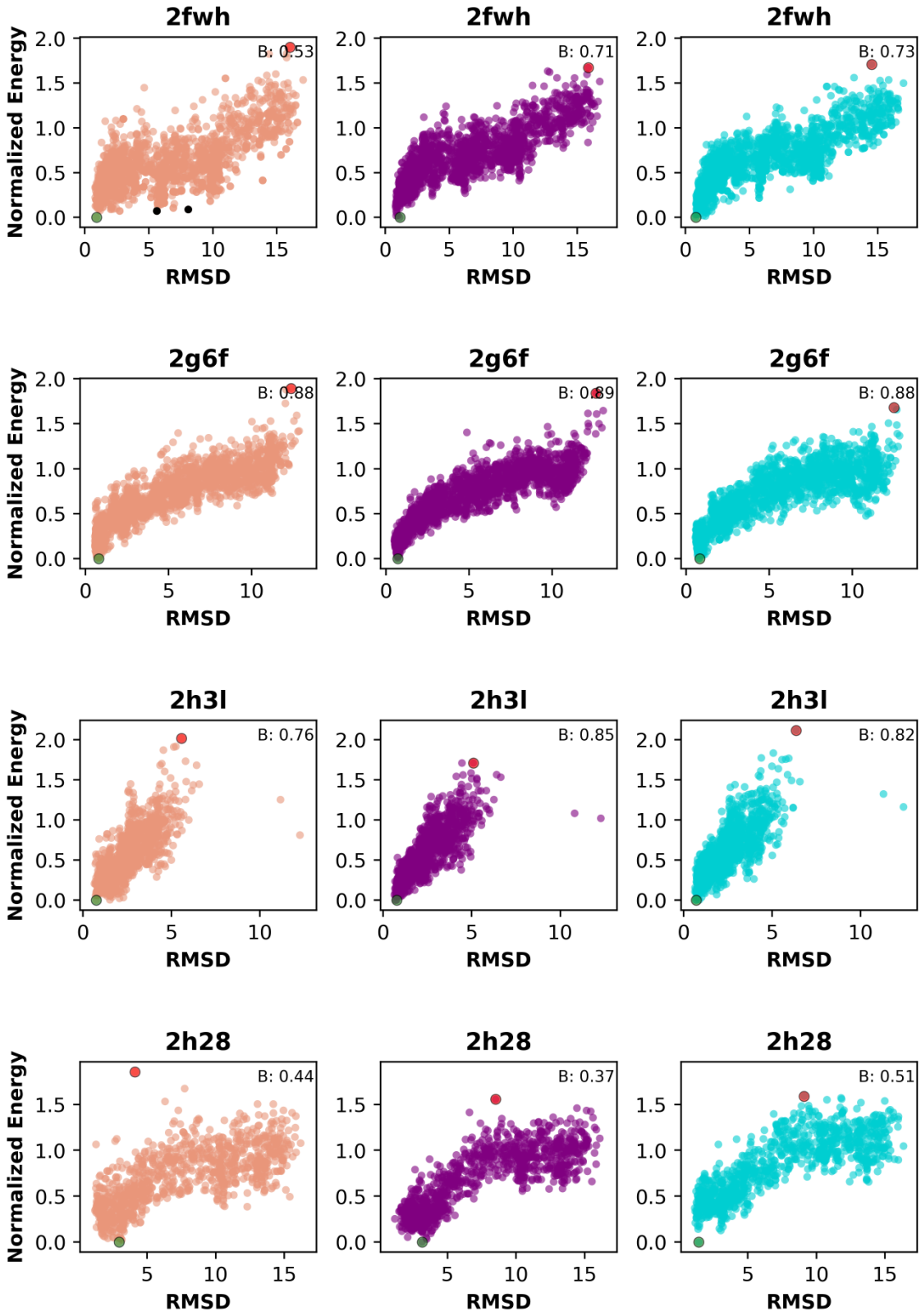

1vkk
B: 0.33
1vkk
B: 0.44
1vkk
B: 0.64
1wdv
B: 0.70
1wdv
B: 0.76
1wdv
B: 0.78
1wlz
B: 0.39
1wlz
B: 0.49
1wlz
B: 0.47
1x6x
B: 0.19
1x6x
B: 0.49
1x6x
B: 0.44
Normalized Energy
RMSD

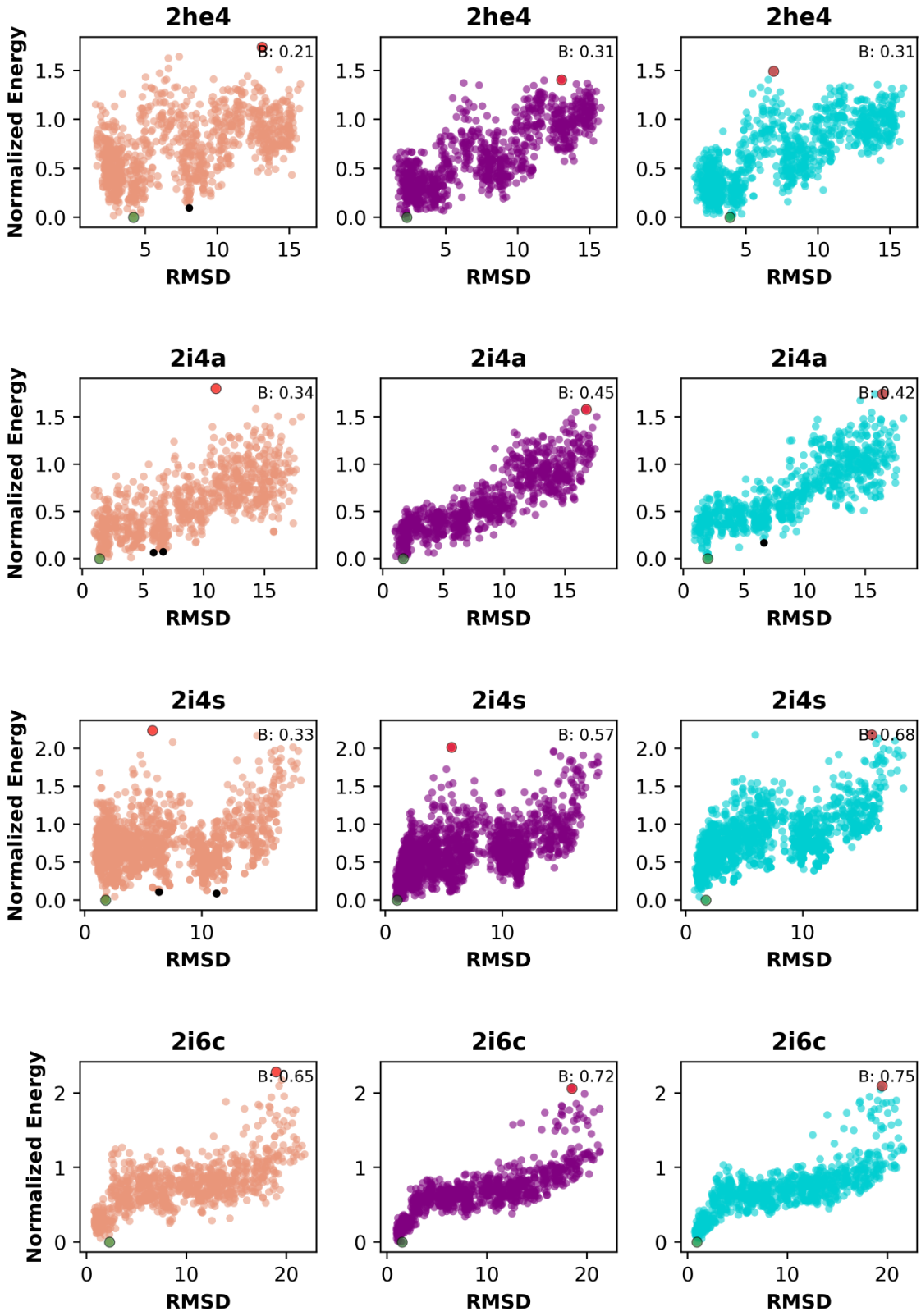

1xmk
B: 0.50
1xmk
B: 0.47
1xmk
B: 0.58
1xmt
B: 0.68
1xmt
B: 0.70
1xmt
B: 0.76
1yn3
B: 0.68
1yn3
B: 0.72
1yn3
B: 0.71
1z2u
B: 0.38
1z2u
B: 0.39
1z2u
B: 0.39
Normalized Energy
RMSD

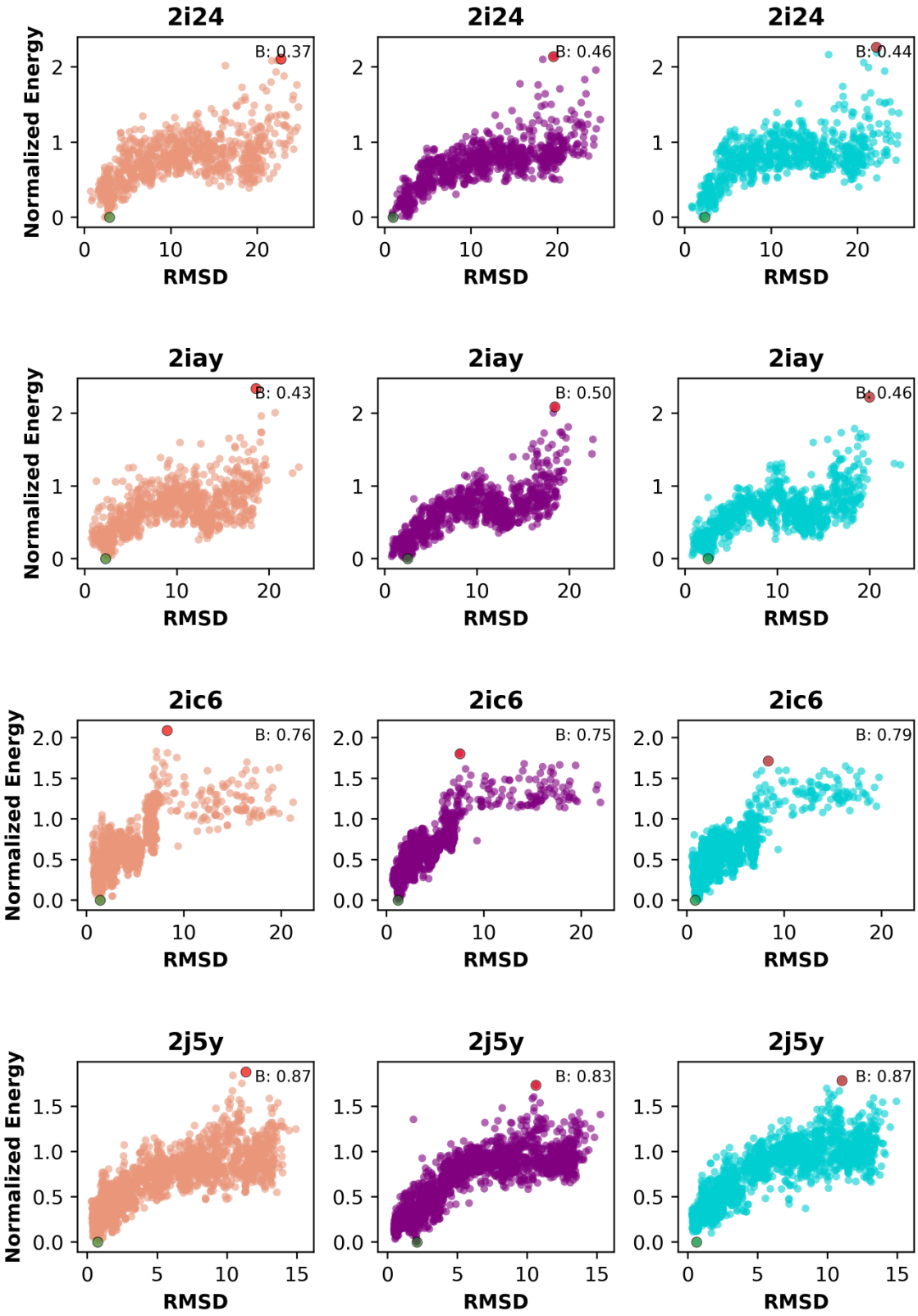

1zlm
B: 0.83
1zlm
B: 0.85
1zlm
B: 0.86
1zma
B: 0.72
1zma
B: 0.70
1zma
B: 0.78
1zuu
B: 0.81
1zuu
B: 0.85
1zuu
B: 0.81
2acy
B: 0.77
2acy
B: 0.77
2acy
B: 0.77
Normalized Energy
RMSD

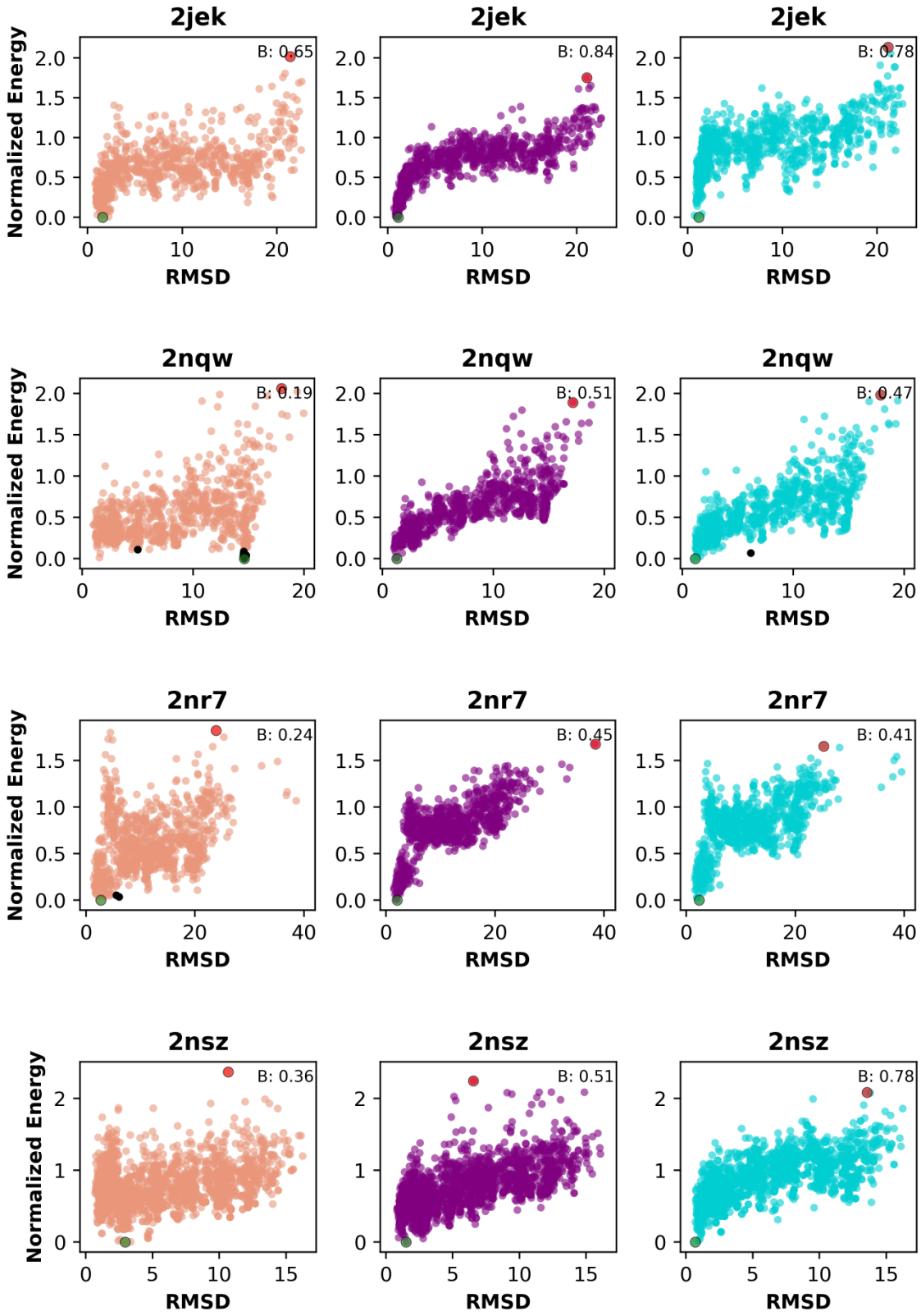

2b29
B: 0.35
2b29
B: 0.41
2b29
B: 0.49
2bf5
B: 0.77
2bf5
B: 0.80
2bf5
B: 0.81
2chf
B: 0.64
2chf
B: 0.81
2chf
B: 0.83
2cxd
B: 0.76
2cxd
B: 0.81
2cxd
B: 0.80
Normalized Energy
RMSD

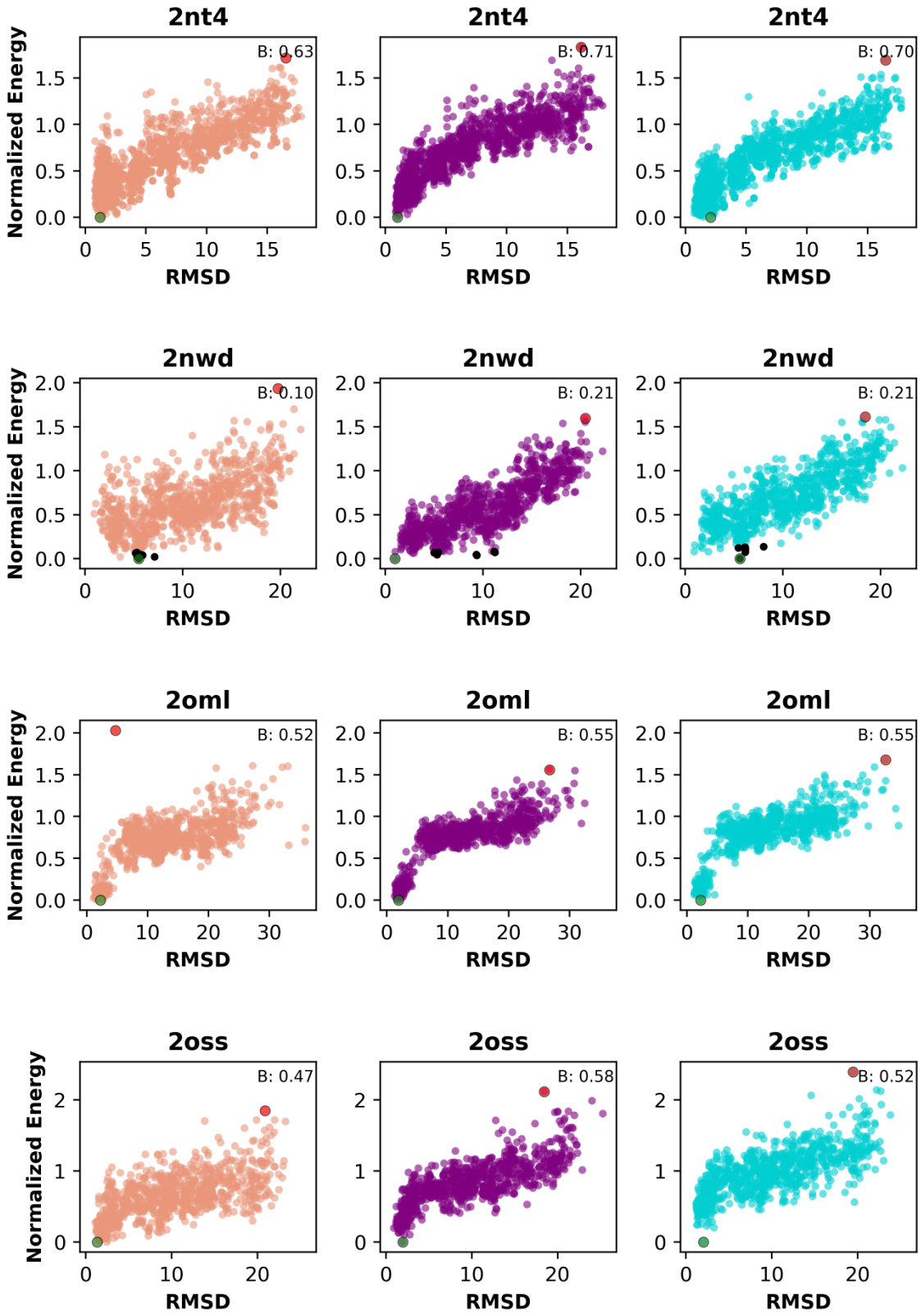

2dfb
B: 0.51
2dfb
B: 0.60
2dfb
B: 0.63
2dyj
B: 0.72
2dyj
B: 0.78
2dyj
B: 0.78
2fe5
B: 0.77
2fe5
B: 0.77
2fe5
B: 0.81
2fi1
B: 0.39
2fi1
B: 0.34
2fi1
B: 0.38
Normalized Energy
RMSD

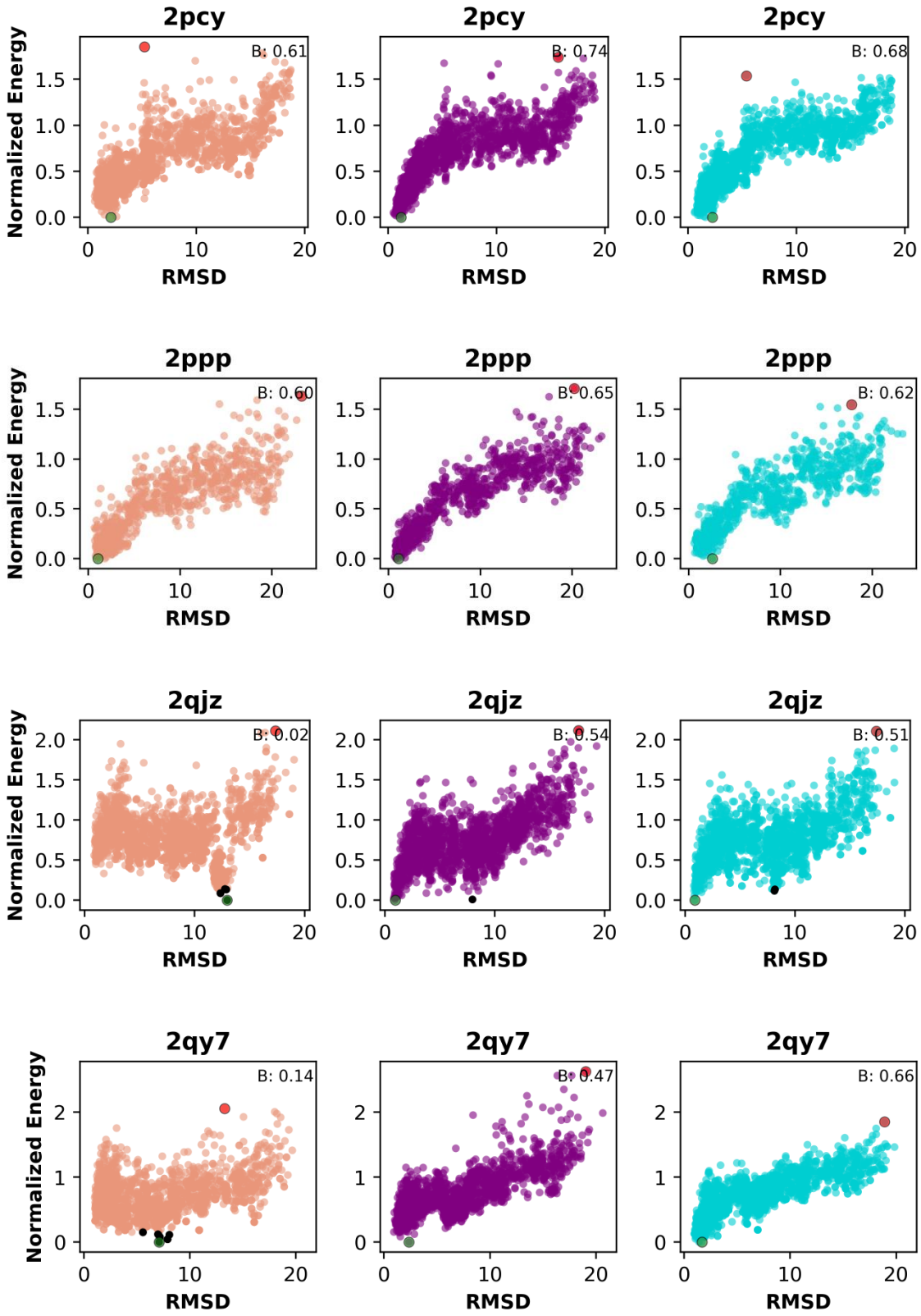

2fwh
B: 0.53
2fwh
B: 0.71
2fwh
B: 0.73
2g6f
B: 0.88
2g6f
B: 0.89
2g6f
B: 0.88
2h3l
B: 0.76
2h3l
B: 0.85
2h3l
B: 0.82
2h28
B: 0.44
2h28
B: 0.37
2h28
B: 0.51
Normalized Energy
RMSD

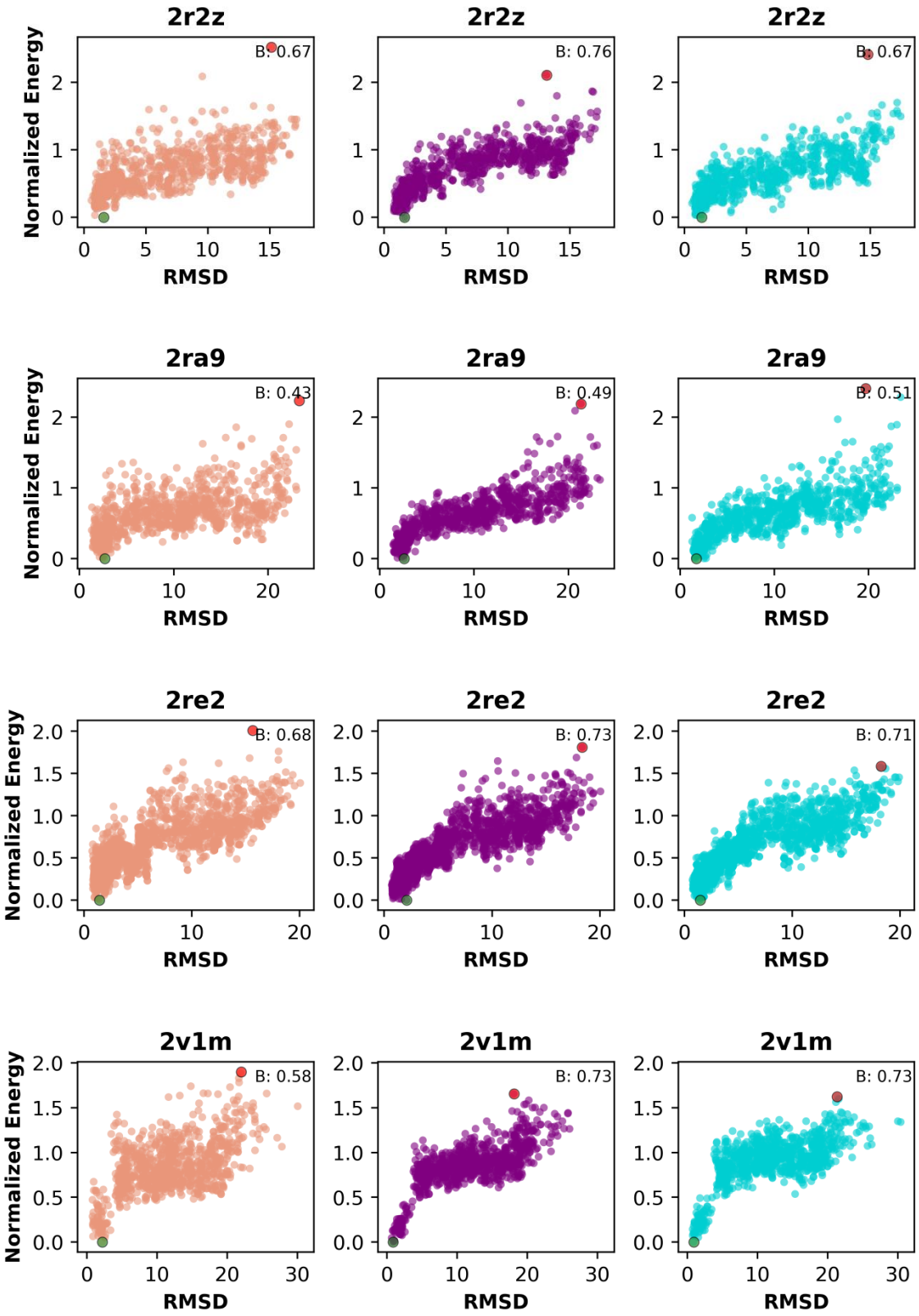

2he4
B: 0.21
2he4
B: 0.31
2he4
B: 0.31
2i4a
B: 0.34
2i4a
B: 0.45
2i4a
B: 0.42
2i4s
B: 0.33
2i4s
B: 0.57
2i4s
B: 0.68
2i6c
B: 0.65
2i6c
B: 0.72
2i6c
B: 0.75
Normalized Energy
RMSD

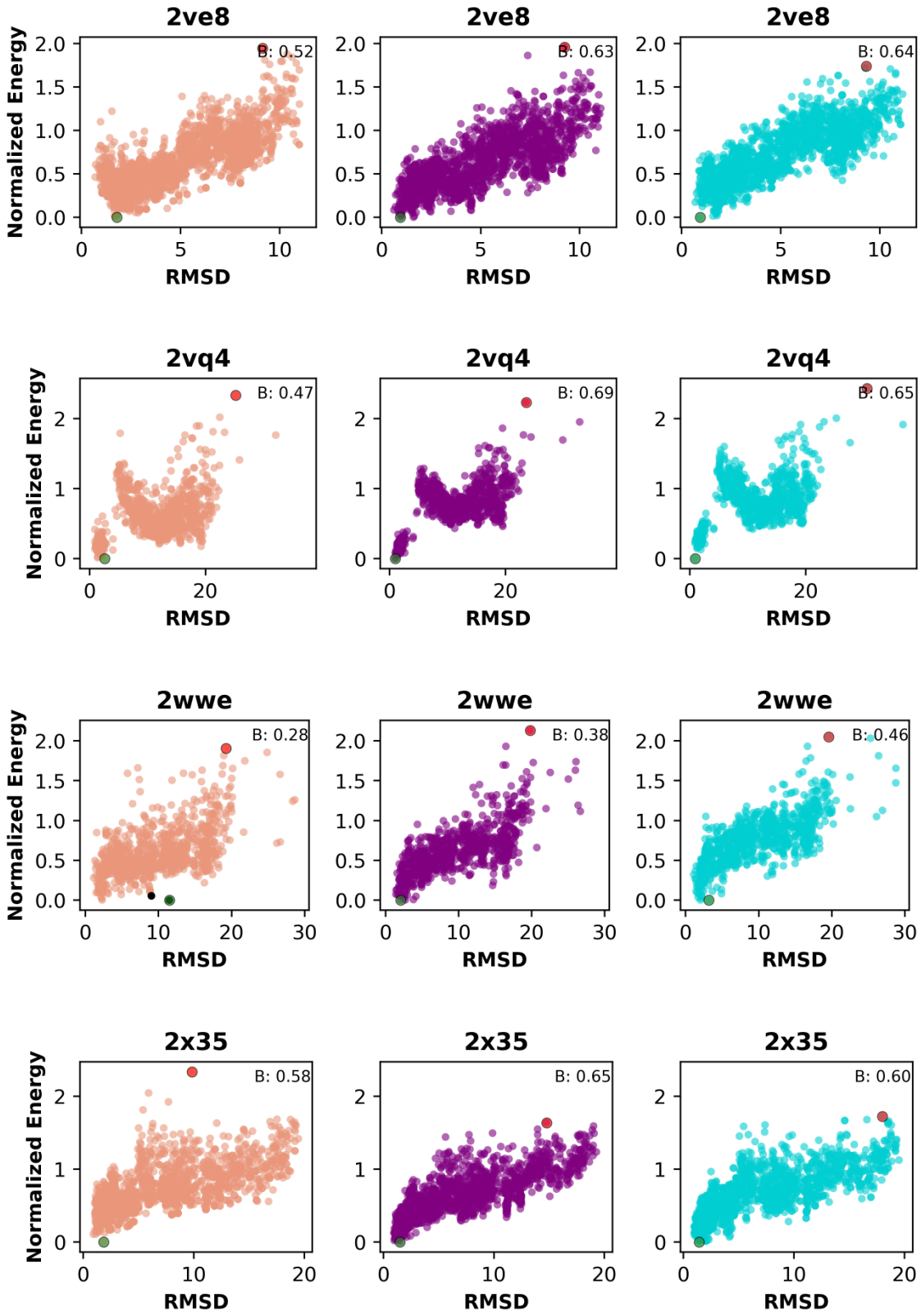

2i24
B: 0.37
2i24
B: 0.46
2i24
B: 0.44
2iay
B: 0.43
2iay
B: 0.50
2iay
B: 0.46
2ic6
B: 0.76
2ic6
B: 0.75
2ic6
B: 0.79
2j5y
B: 0.87
2j5y
B: 0.83
2j5y
B: 0.87
Normalized Energy
RMSD

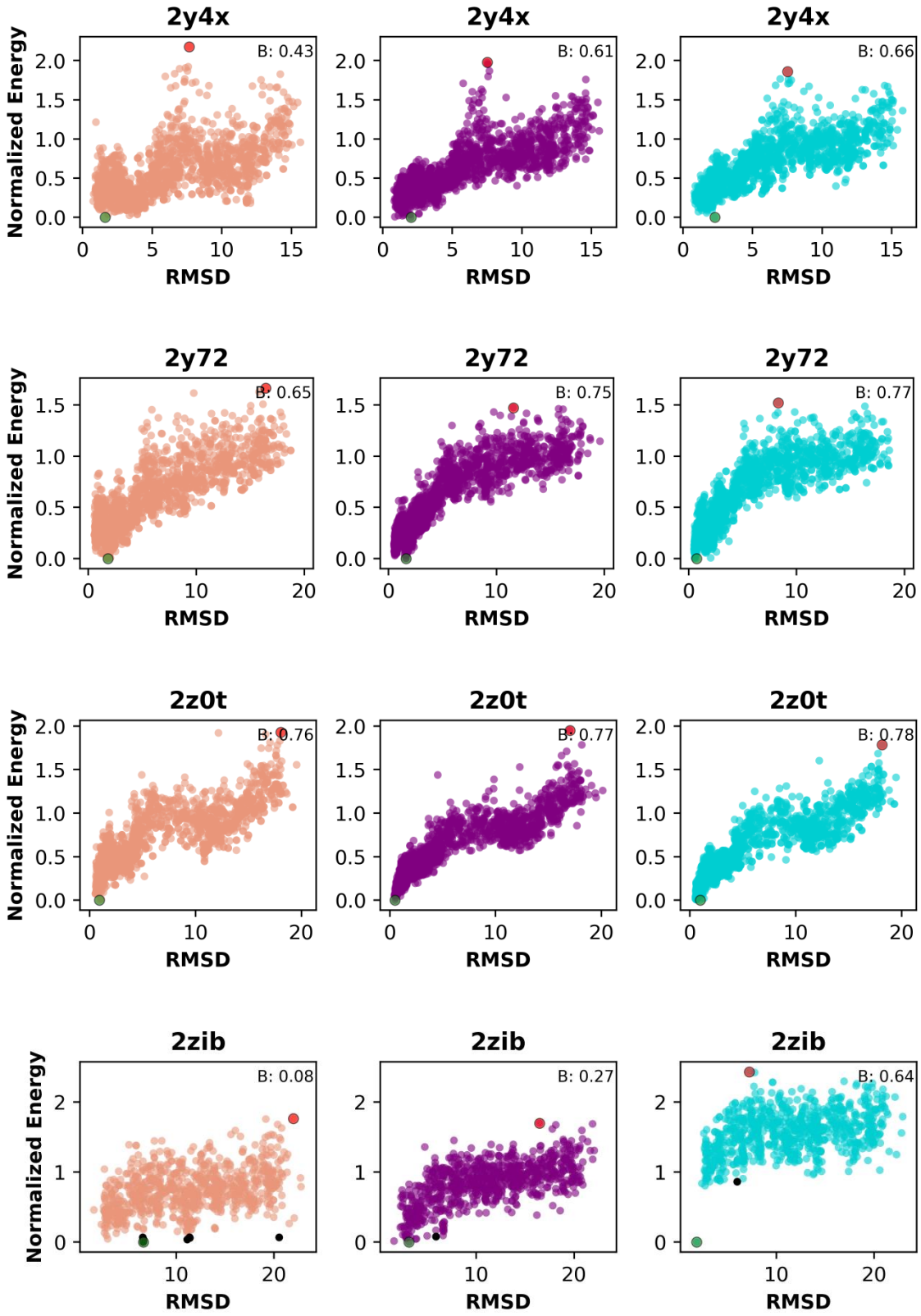

2jek
B: 0.65
2jek
B: 0.84
2jek
B: 0.78
2nqw
B: 0.19
2nqw
B: 0.51
2nqw
B: 0.47
2nr7
B: 0.24
2nr7
B: 0.45
2nr7
B: 0.41
2nsz
B: 0.36
2nsz
B: 0.51
2nsz
B: 0.78
Normalized Energy
RMSD

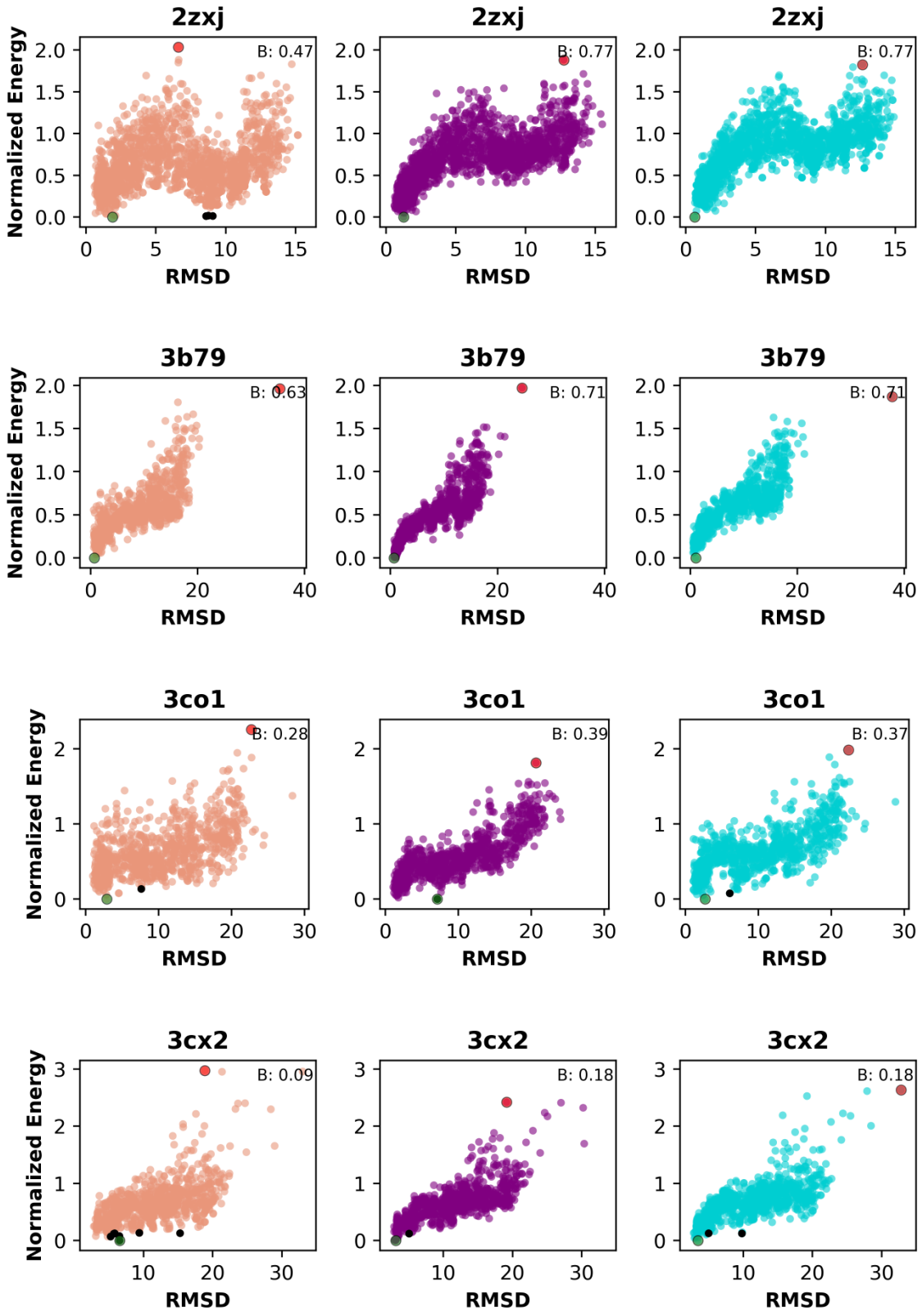

2nt4
B: 0.63
2nt4
B: 0.71
2nt4
B: 0.70
2nwd
B: 0.10
2nwd
B: 0.21
2nwd
B: 0.21
2oml
B: 0.52
2oml
B: 0.55
2oml
B: 0.55
2oss
B: 0.47
2oss
B: 0.58
2oss
B: 0.52
Normalized Energy
RMSD

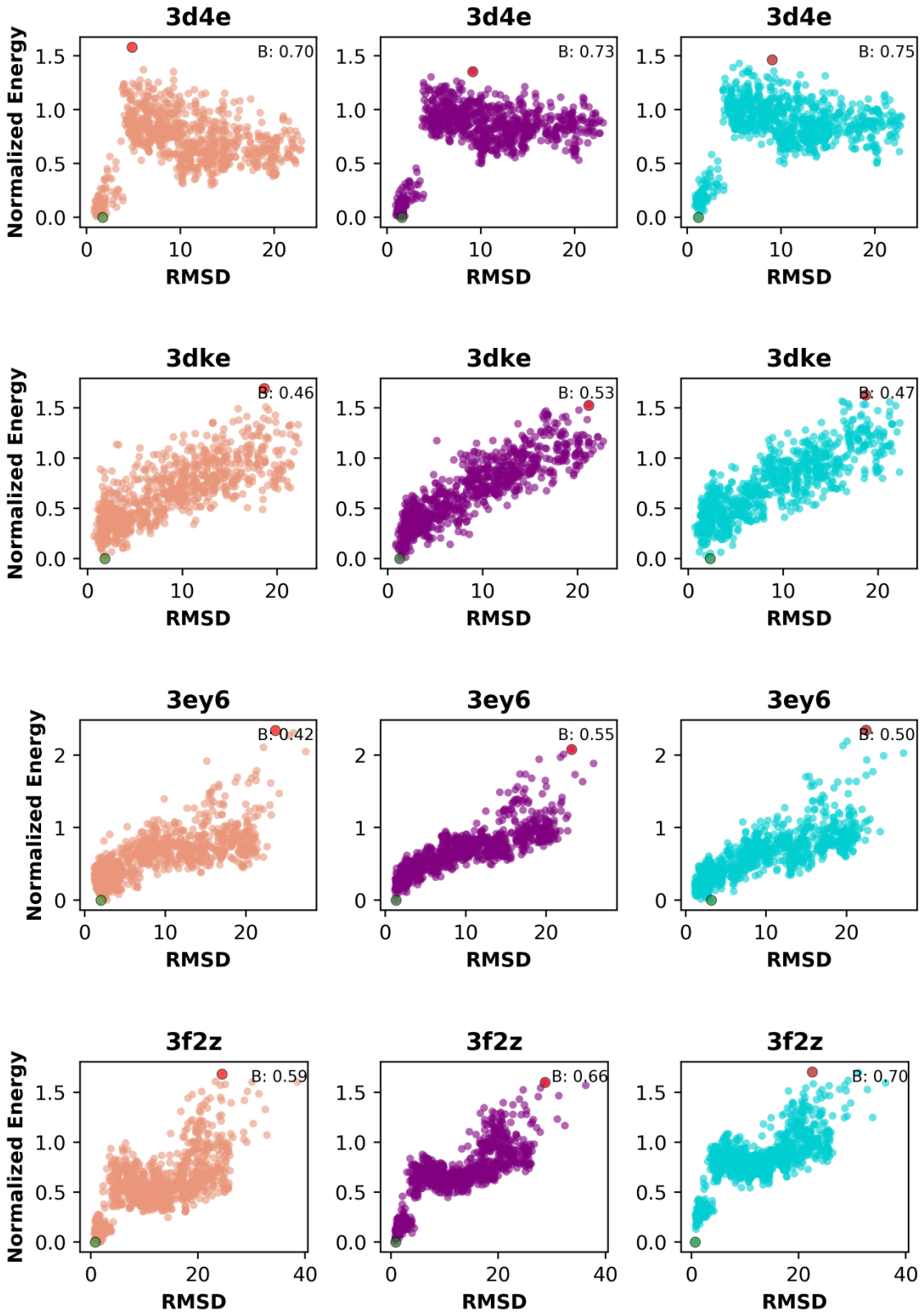

2pcy
B: 0.61
2pcy
B: 0.74
2pcy
B: 0.68
2ppp
B: 0.60
2ppp
B: 0.65
2ppp
B: 0.62
2qjz
B: 0.02
2qjz
B: 0.54
2qjz
B: 0.51
2qy7
B: 0.14
2qy7
B: 0.47
2qy7
B: 0.66
Normalized Energy
RMSD

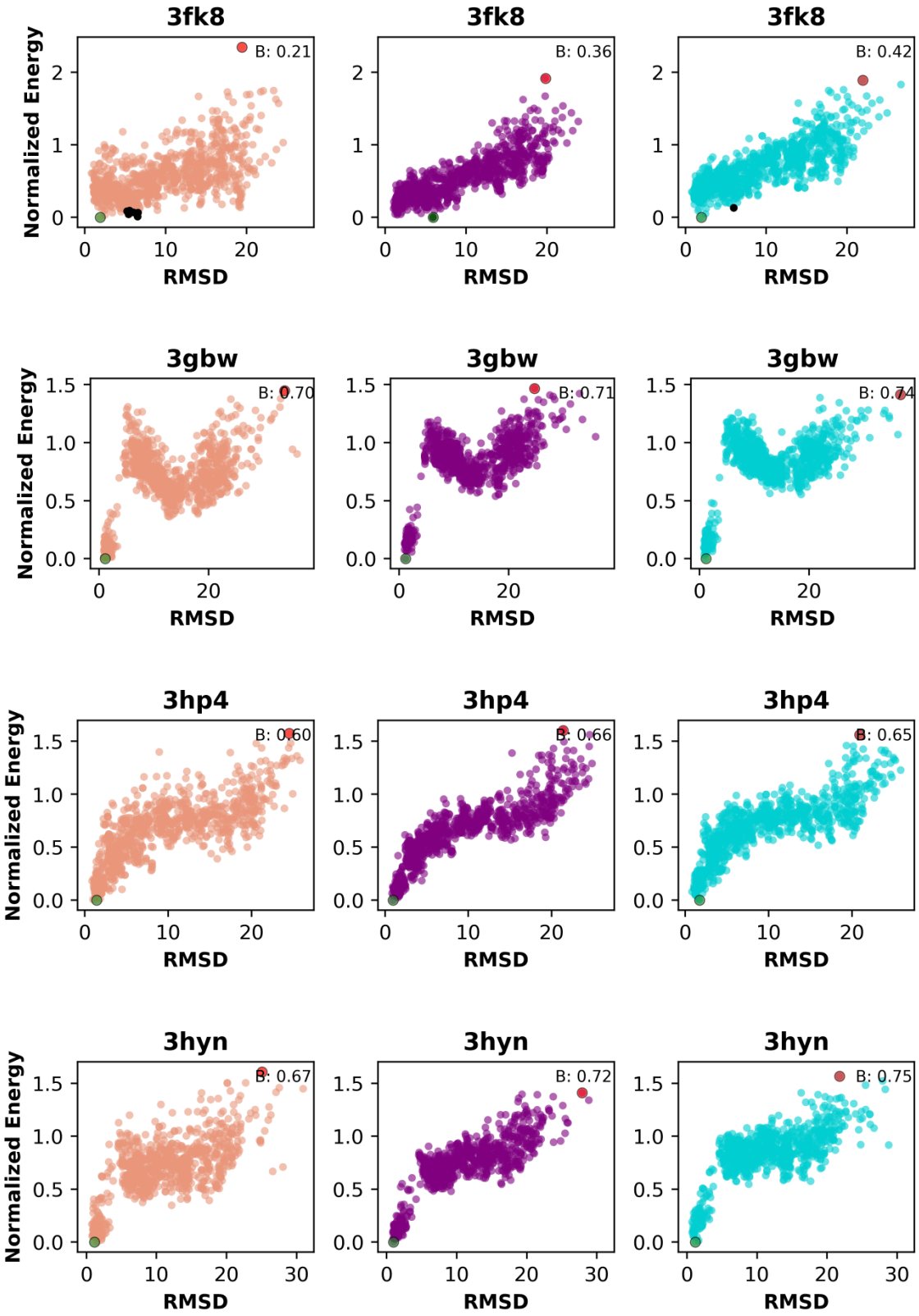

2r2z
B: 0.67
2r2z
B: 0.76
2r2z
B: 0.67
2ra9
B: 0.43
2ra9
B: 0.49
2ra9
B: 0.51
2re2
B: 0.68
2re2
B: 0.73
2re2
B: 0.71
2v1m
B: 0.58
2v1m
B: 0.73
2v1m
B: 0.73
Normalized Energy
RMSD

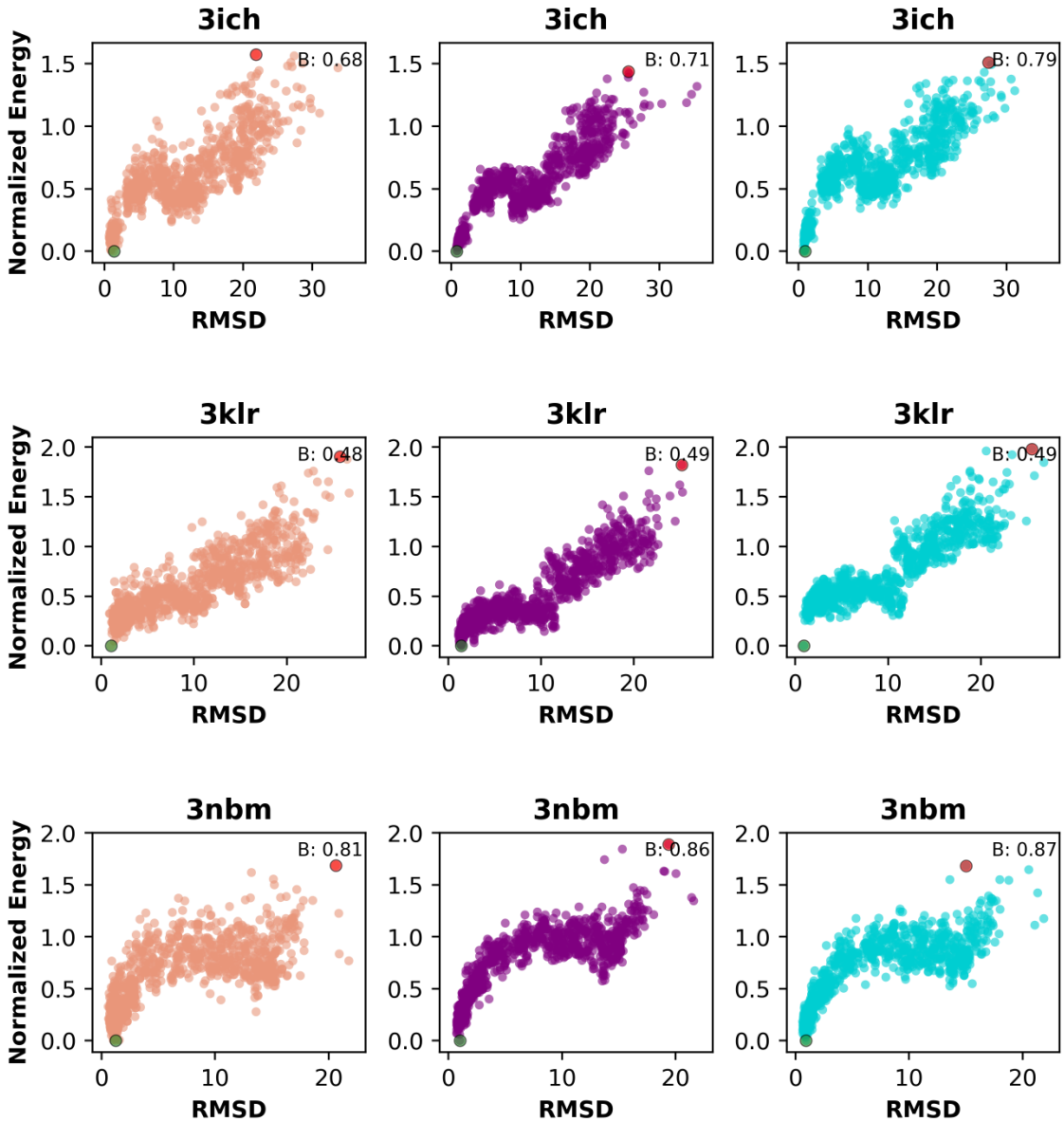

2ve8
B: 0.52
2ve8
B: 0.63
2ve8
B: 0.64
2vq4
B: 0.47
2vq4
B: 0.69
2vq4
B: 0.65
2wwe
B: 0.28
2wwe
B: 0.38
2wwe
B: 0.46
2x35
B: 0.58
2x35
B: 0.65
2x35
B: 0.60
Normalized Energy
RMSD

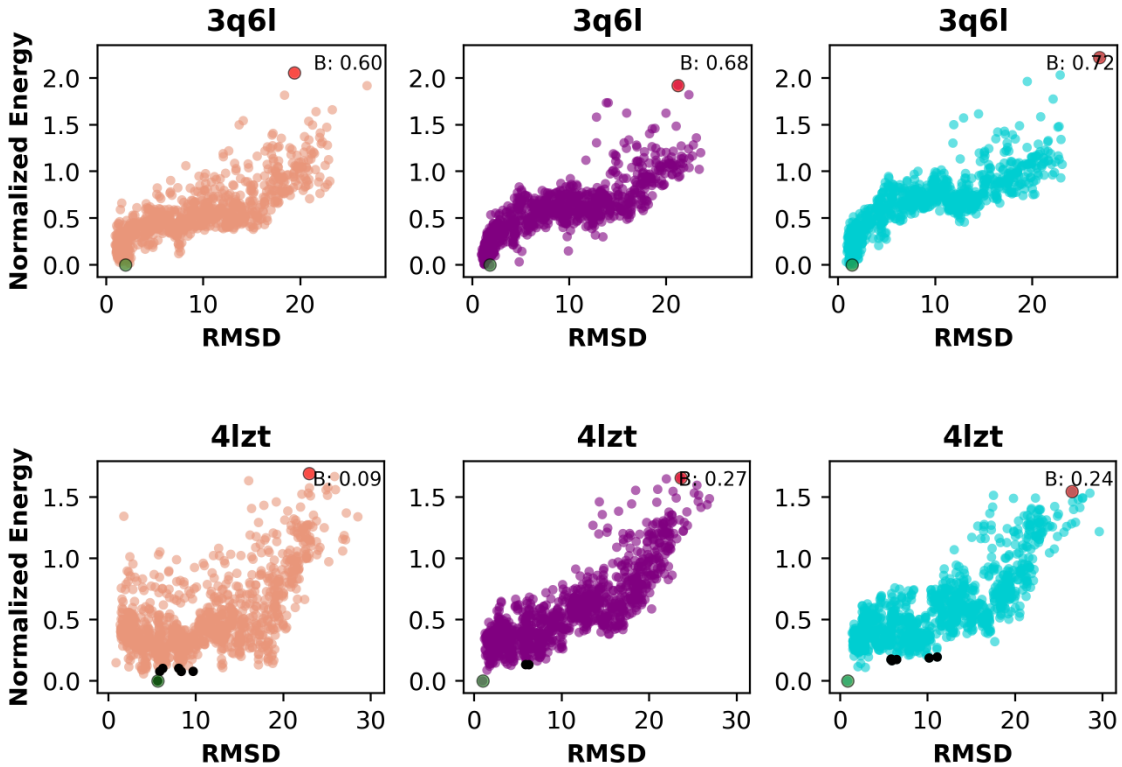

2y4x
B: 0.43
2y4x
B: 0.61
2y4x
B: 0.66
2y72
B: 0.65
2y72
B: 0.75
2y72
B: 0.77
2z0t
B: 0.76
2z0t
B: 0.77
2z0t
B: 0.78
2zib
B: 0.08
2zib
B: 0.27
2zib
B: 0.64
Normalized Energy
RMSD


2zxj
B: 0.47
2zxj
B: 0.77
2zxj
B: 0.77
3b79
B: 0.63
3b79
B: 0.71
3b79
B: 0.71
3co1
B: 0.28
3co1
B: 0.39
3co1
B: 0.37
3cx2
B: 0.09
3cx2
B: 0.18
3cx2
B: 0.18
Normalized Energy
RMSD

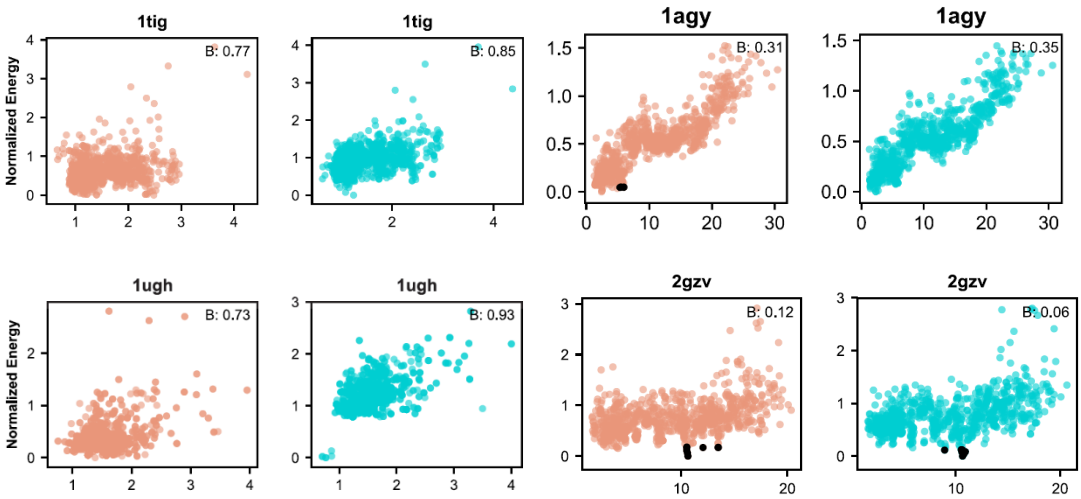

3d4e
B: 0.70
3d4e
B: 0.73
3d4e
B: 0.75
3dke
B: 0.46
3dke
B: 0.53
3dke
B: 0.47
3ey6
B: 0.42
3ey6
B: 0.55
3ey6
B: 0.50
3f2z
B: 0.59
3f2z
B: 0.66
3f2z
B: 0.70
Normalized Energy
RMSD


3fk8
B: 0.21
3fk8
B: 0.36
3fk8
B: 0.42
3gbw
B: 0.70
3gbw
B: 0.71
3gbw
B: 0.74
3hp4
B: 0.60
3hp4
B: 0.66
3hp4
B: 0.65
3hyn
B: 0.67
3hyn
B: 0.72
3hyn
B: 0.75
Normalized Energy
RMSD


3ich
B: 0.68
3ich
B: 0.71
3ich
B: 0.79
3klr
B: 0.48
3klr
B: 0.49
3klr
B: 0.49
3nbm
B: 0.81
3nbm
B: 0.86
3nbm
B: 0.87
Normalized Energy
RMSD



These PDB sets were not included in the REF2015 benchmark set.



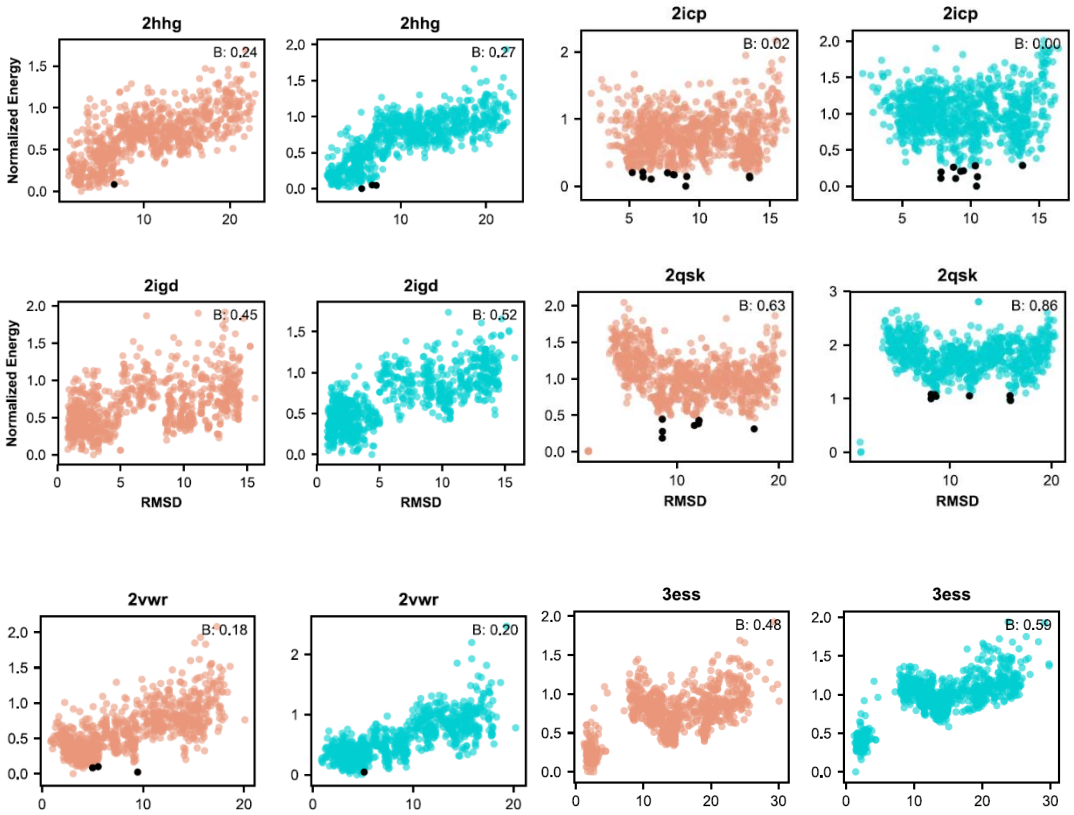


Figure 4-10: Plot of RMSD for the decoy selected by various criteria: minimum RMSD overall decoys (All), Minimum-sum-selected decoy, Rosetta-selected decoy, and Amber-selected decoy. Each tick on the x-axis represents one protein system.

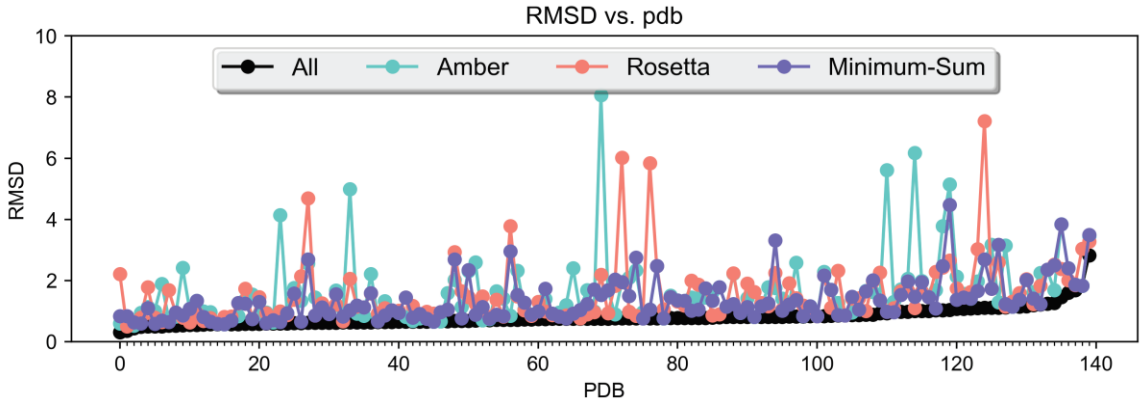
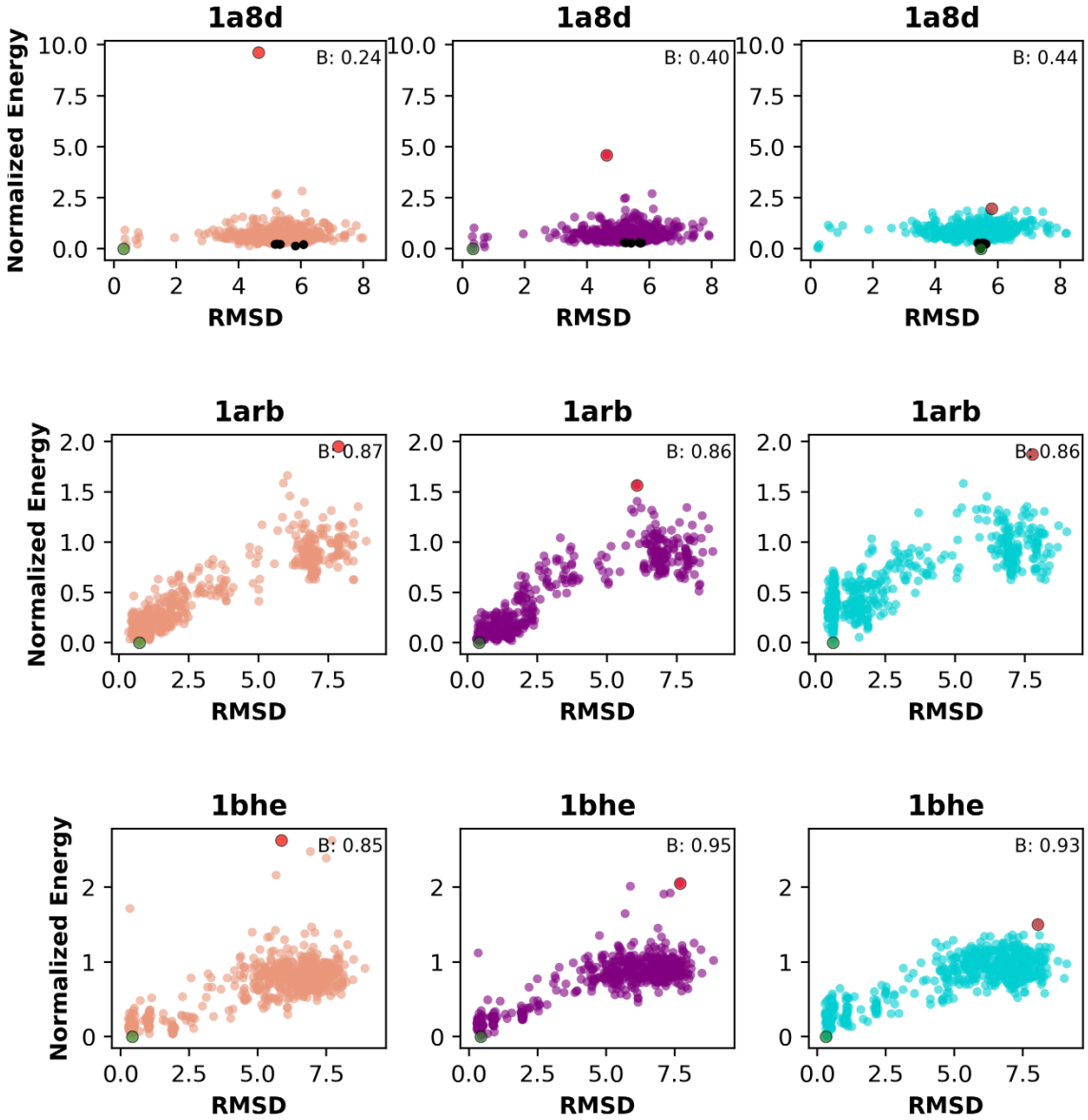
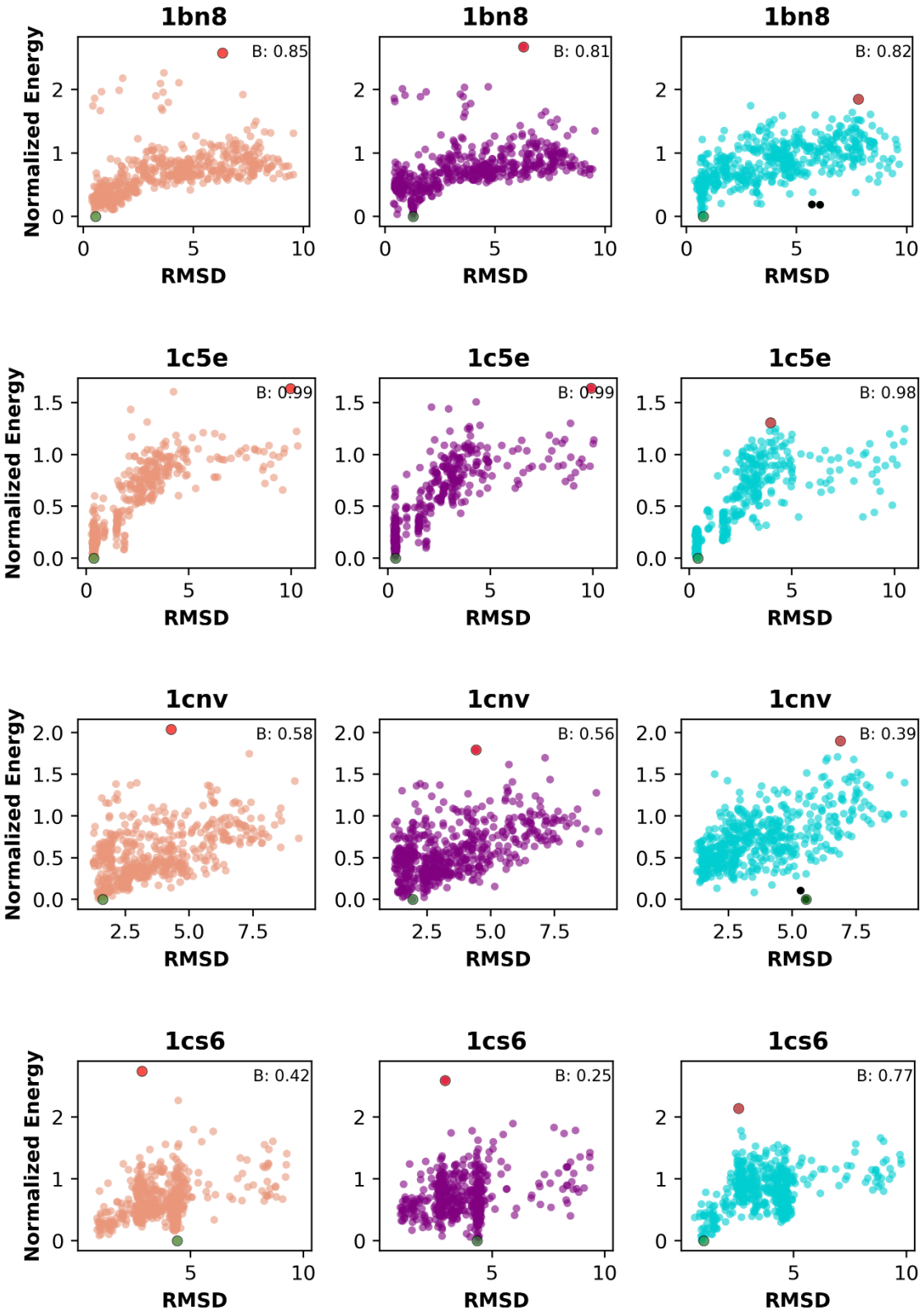
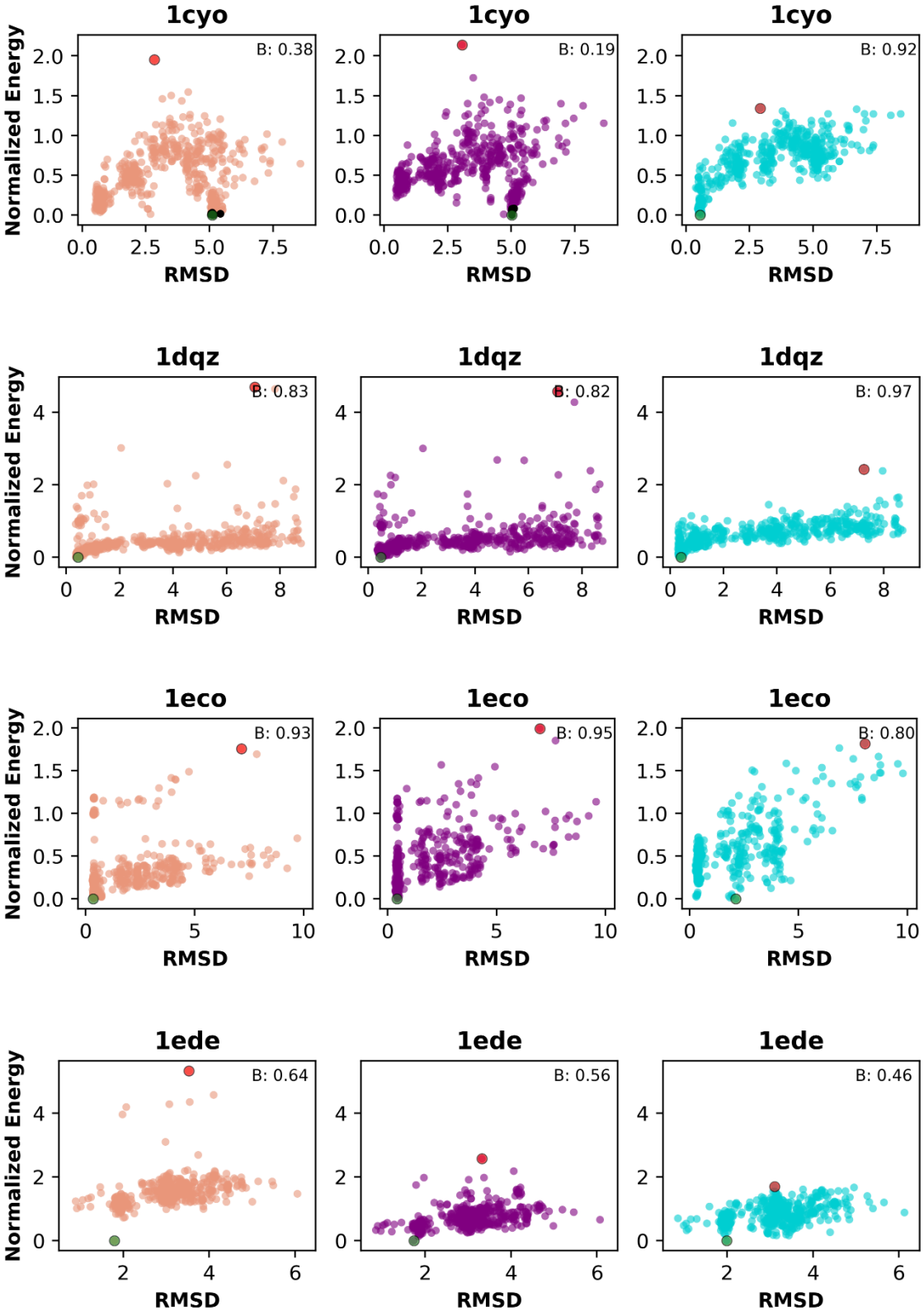
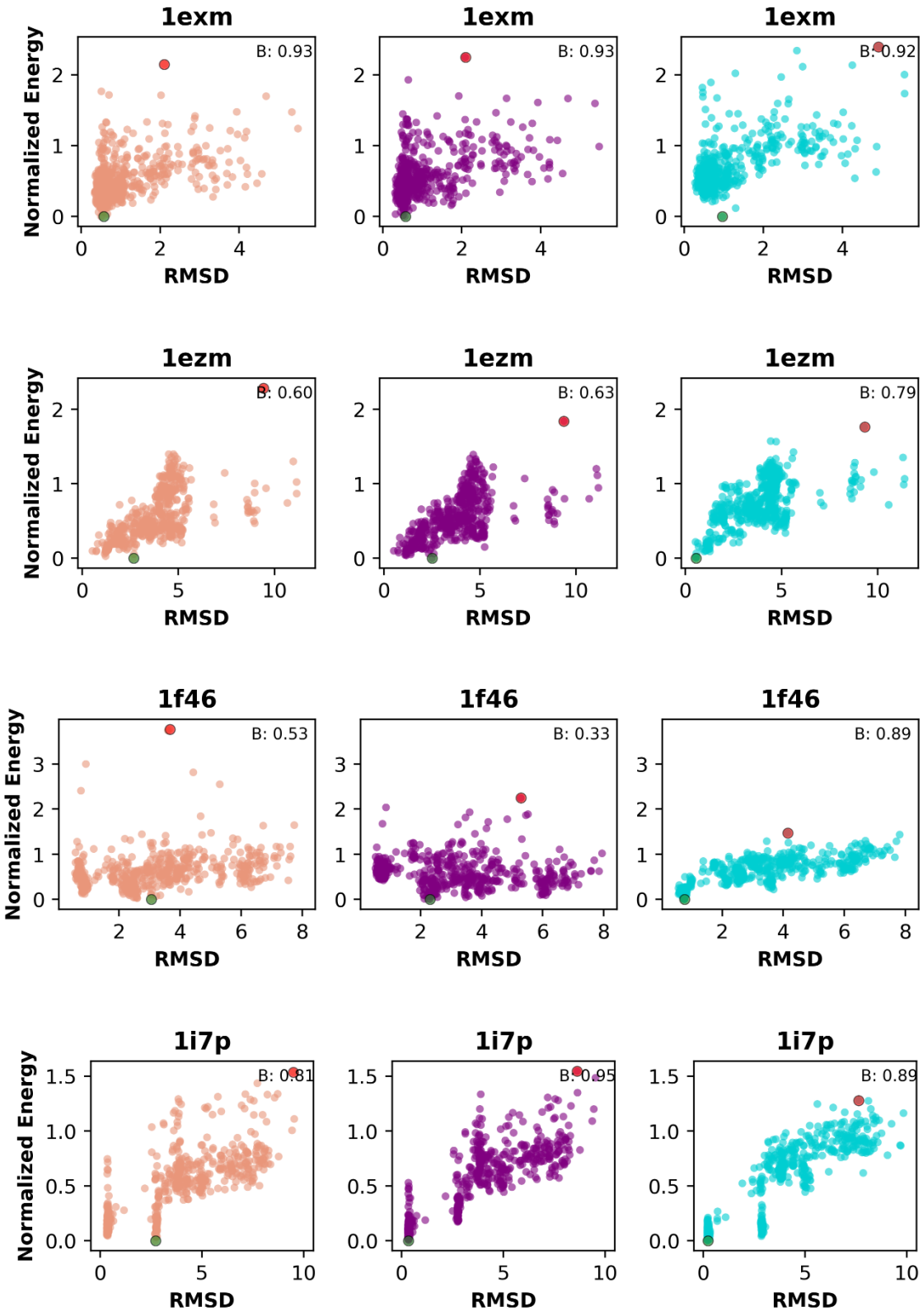


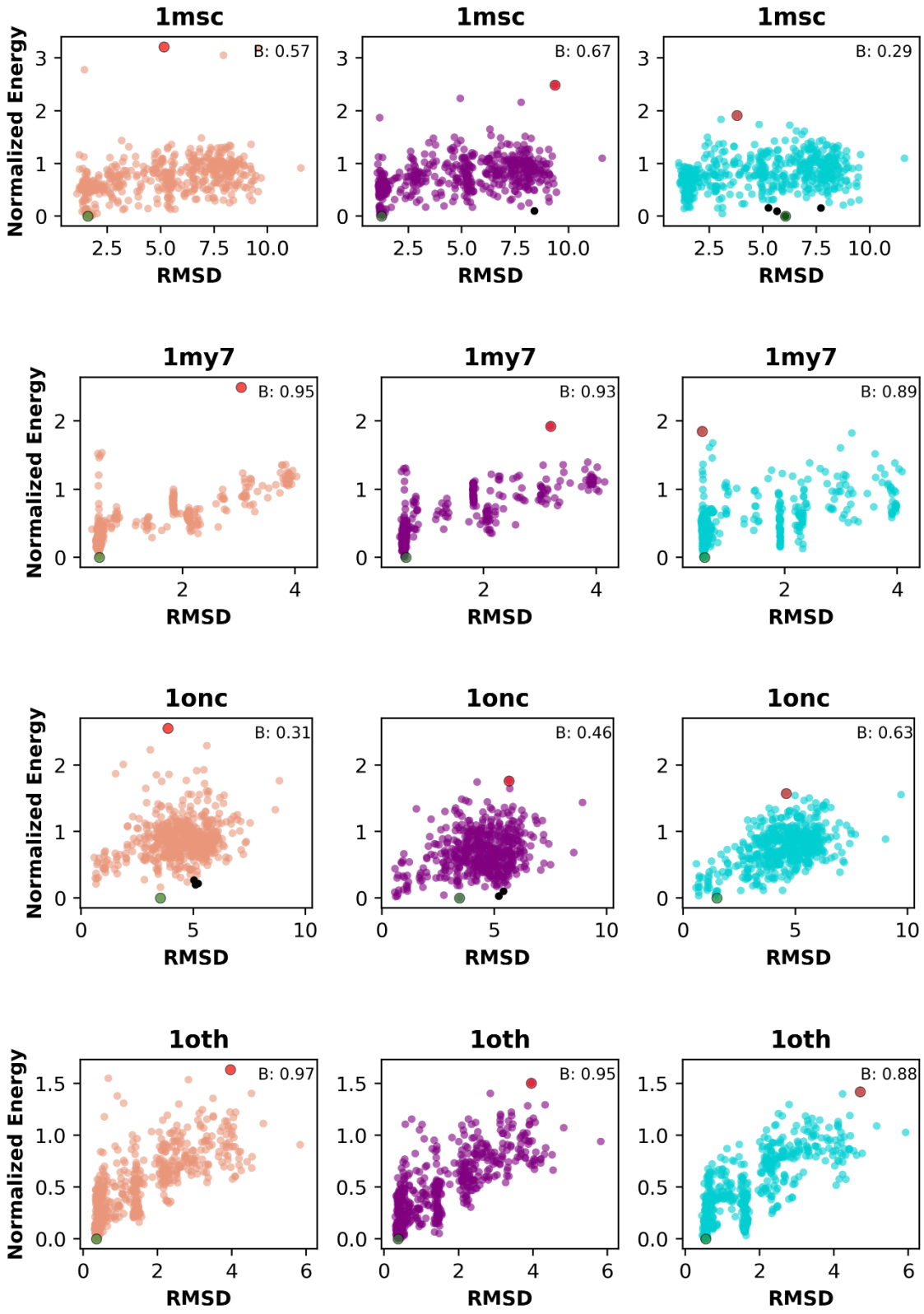
Figure 4-11: Energy landscapes for all loop modeling proteins.(Talaris2014, Left/Salmon ; REF2015, Middle/Purple; ff14sbonlysc Right/Turquoise)

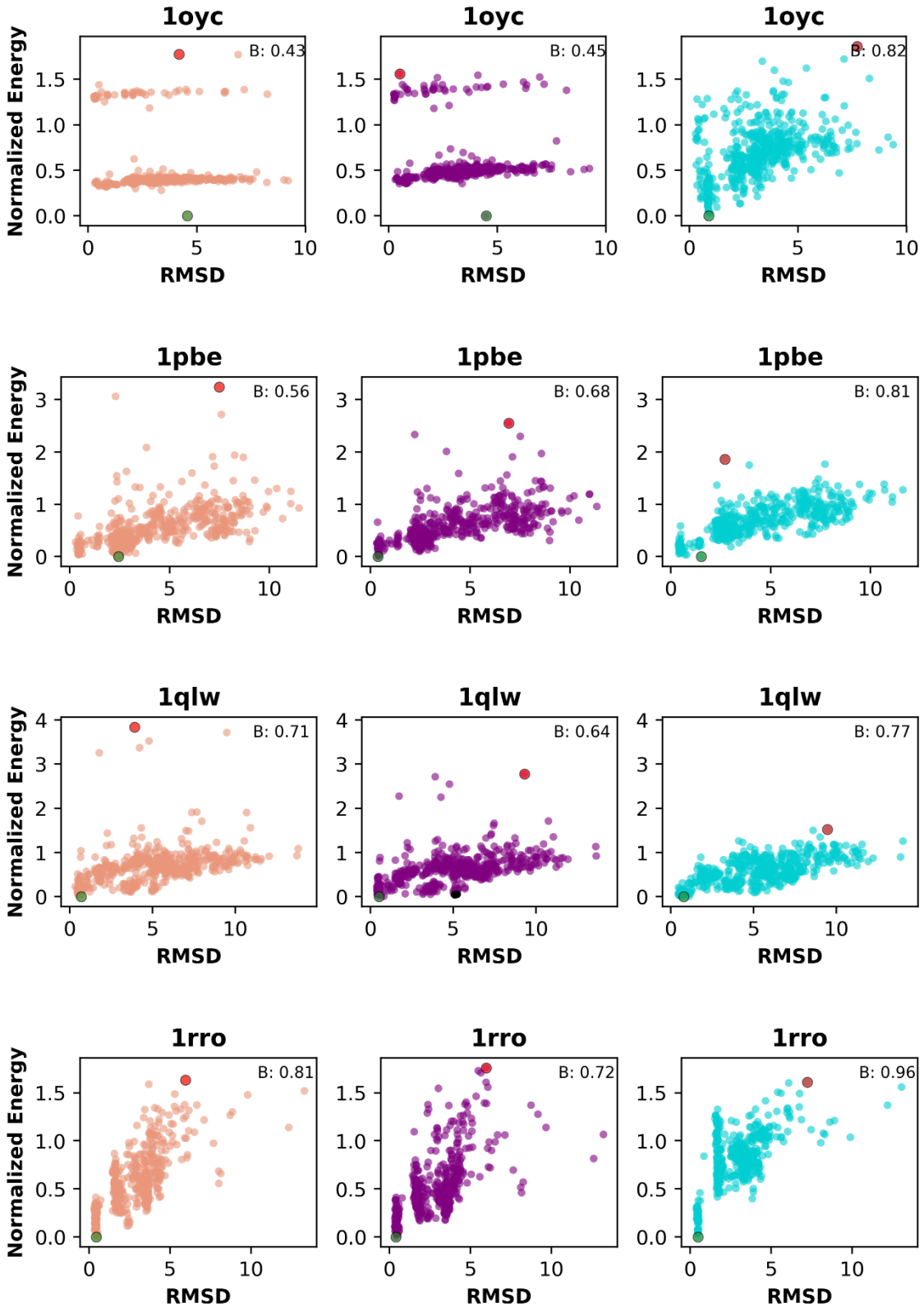


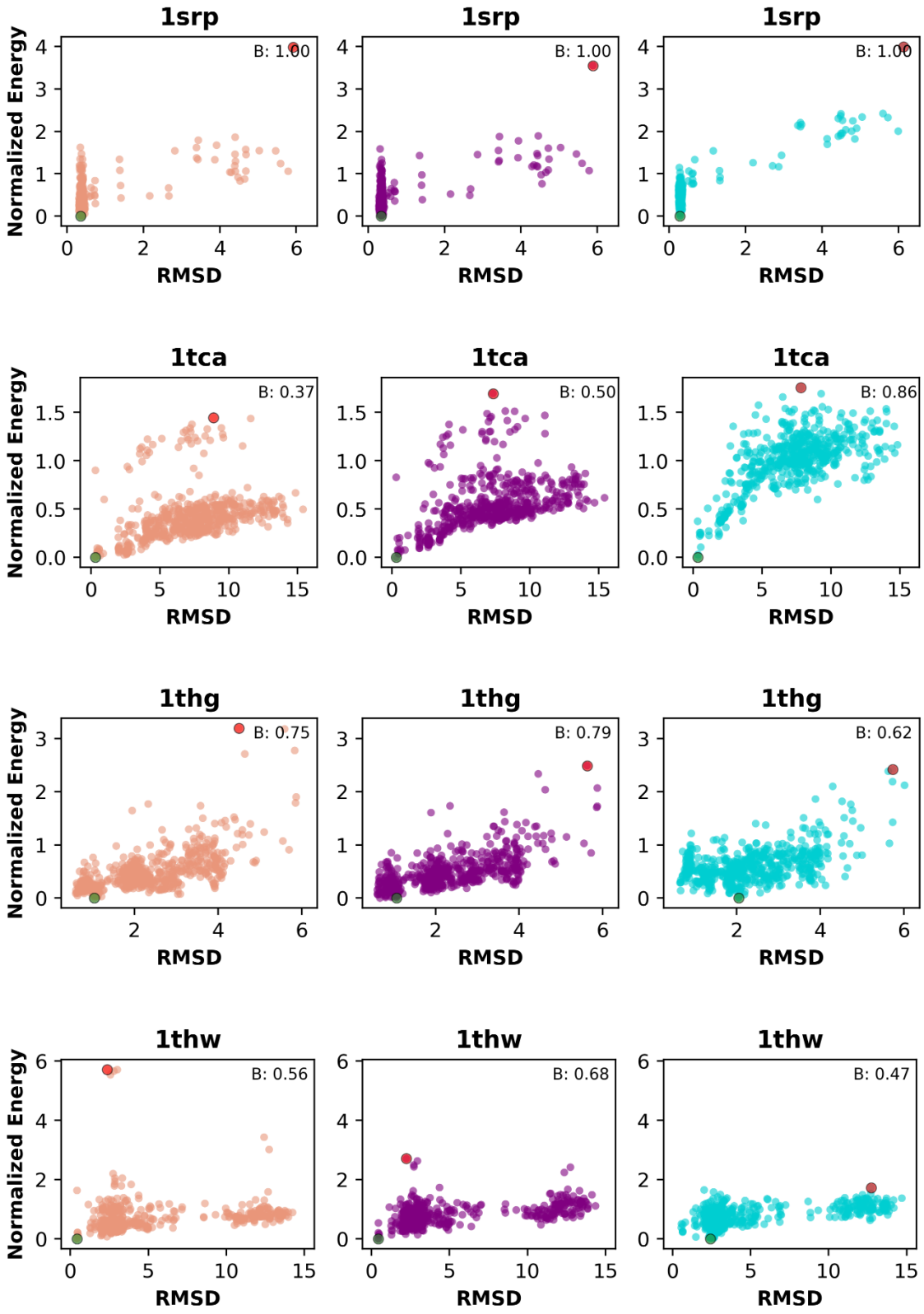

1bn8
B: 0.85
1bn8
B: 0.81
1bn8
B: 0.82
1c5e
B: 0.99
1c5e
B: 0.99
1c5e
B: 0.98
1cnv
B: 0.58
1cnv
B: 0.56
1cnv
B: 0.39
1cs6
B: 0.42
1cs6
B: 0.25
1cs6
B: 0.77
Normalized Energy
RMSD

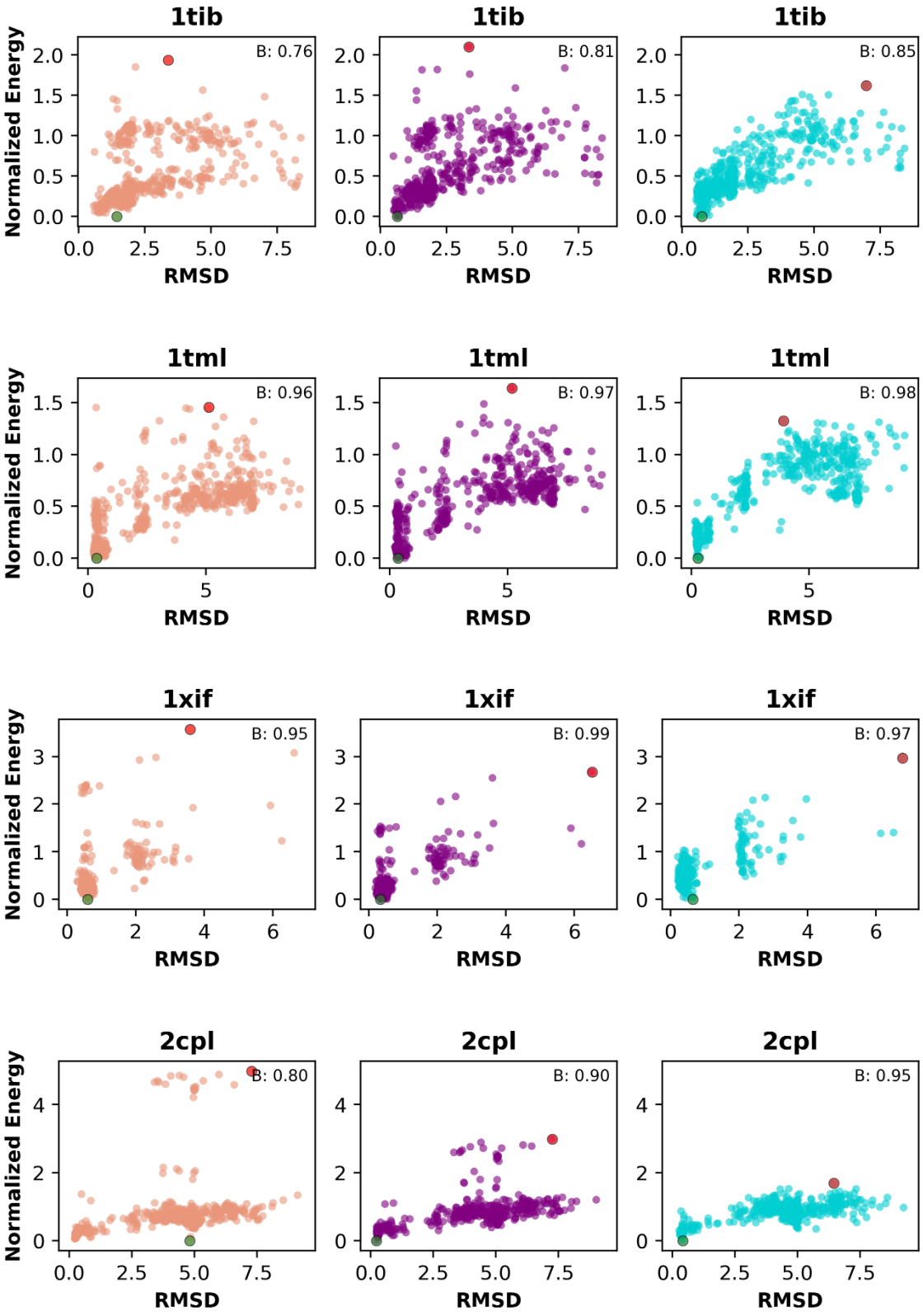

1cyo
B: 0.38
1cyo
B: 0.19
1cyo
B: 0.92
1dqz
B: 0.83
1dqz
B: 0.82
1dqz
B: 0.97
1eco
B: 0.93
1eco
B: 0.95
1eco
B: 0.80
1ede
B: 0.64
1ede
B: 0.56
1ede
B: 0.46
Normalized Energy
RMSD

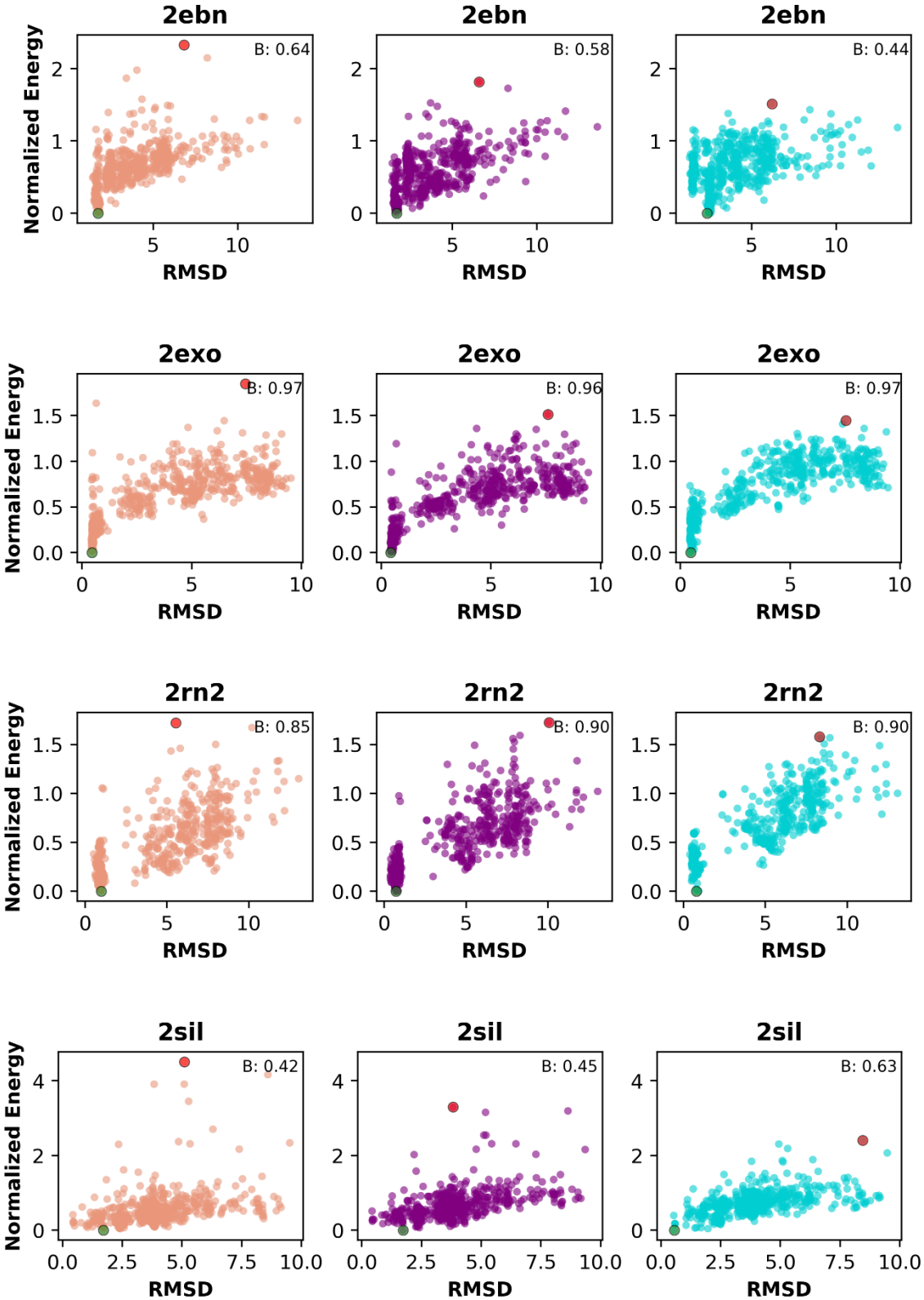

1exm
B: 0.93
1exm
B: 0.93
1exm
B: 0.92
1ezm
B: 0.60
1ezm
B: 0.63
1ezm
B: 0.79
1f46
B: 0.53
1f46
B: 0.33
1f46
B: 0.89
1i7p
B: 0.81
1i7p
B: 0.95
1i7p
B: 0.89
Normalized Energy
RMSD

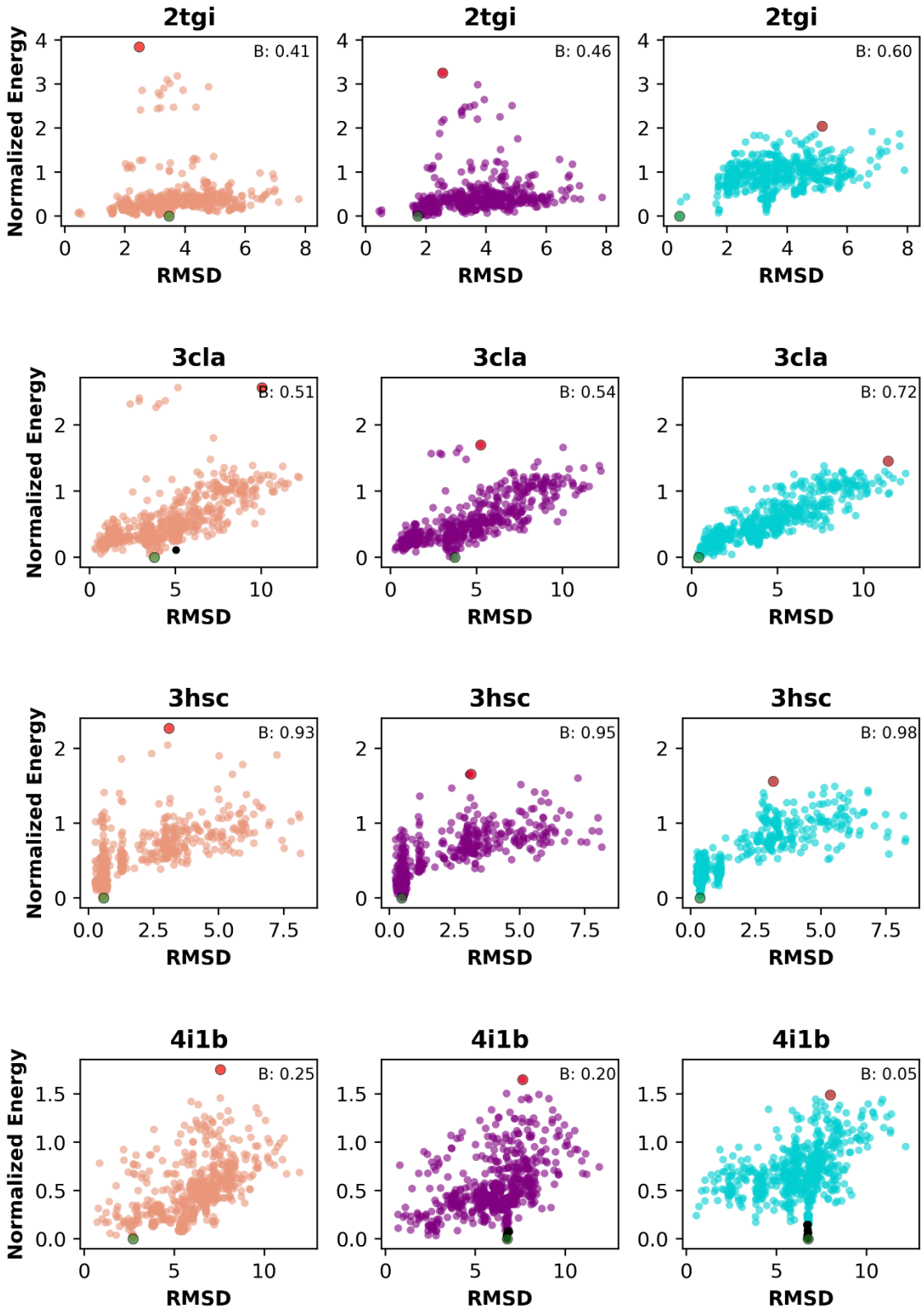

1msc
B: 0.57
1msc
B: 0.67
1msc
B: 0.29
1my7
B: 0.95
1my7
B: 0.93
1my7
B: 0.89
1onc
B: 0.31
1onc
B: 0.46
1onc
B: 0.63
1oth
B: 0.97
1oth
B: 0.95
1oth
B: 0.88
Normalized Energy
RMSD


1oyc
B: 0.43
1oyc
B: 0.45
1oyc
B: 0.82
1pbe
B: 0.56
1pbe
B: 0.68
1pbe
B: 0.81
1qlw
B: 0.71
1qlw
B: 0.64
1qlw
B: 0.77
1rro
B: 0.81
1rro
B: 0.72
1rro
B: 0.96
Normalized Energy
RMSD


1srp
B: 1.00
1srp
B: 1.00
1srp
B: 1.00
1tca
B: 0.37
1tca
B: 0.50
1tca
B: 0.86
1thg
B: 0.75
1thg
B: 0.79
1thg
B: 0.62
1thw
B: 0.56
1thw
B: 0.68
1thw
B: 0.47
Normalized Energy
RMSD


1tib
B: 0.76
1tib
B: 0.81
1tib
B: 0.85
1tml
B: 0.96
1tml
B: 0.97
1tml
B: 0.98
1xif
B: 0.95
1xif
B: 0.99
1xif
B: 0.97
2cpl
B: 0.80
2cpl
B: 0.90
2cpl
B: 0.95
Normalized Energy
RMSD


2ebn
B: 0.64
2ebn
B: 0.58
2ebn
B: 0.44
2exo
B: 0.97
2exo
B: 0.96
2exo
B: 0.97
2rn2
B: 0.85
2rn2
B: 0.90
2rn2
B: 0.90
2sil
B: 0.42
2sil
B: 0.45
2sil
B: 0.63
Normalized Energy
RMSD


2tgi
B: 0.41
2tgi
B: 0.46
2tgi
B: 0.60
3cla
B: 0.51
3cla
B: 0.54
3cla
B: 0.72
3hsc
B: 0.93
3hsc
B: 0.95
3hsc
B: 0.98
4i1b
B: 0.25
4i1b
B: 0.20
4i1b
B: 0.05
Normalized Energy
RMSD

4.8 Supplementary Tables

Table 4-2: *B*-metric values for Rosetta talaris2014, Rosetta REF2015, and Amber ff14sbonlysc for the Decoy Discrimination benchmark set.

	Rosetta		Amber
PDB	**Talaris2014**	**REF2015**	**ff14SBonlySC**
1a32	0.73	0.79	0.80
1aaj	0.50	0.49	0.56
1acf	0.64	0.70	0.69
1agy	0.31		0.35
1bk2	0.67	0.75	0.76
1bkr	0.33	0.68	0.56
1bm8	0.62	0.64	0.69
1cc8	0.53	0.70	0.68
1cei	0.69	0.68	0.73
1ctf	0.70	0.75	0.73
1elw	0.69	0.77	0.84
1enh	0.53	0.52	0.54
1ew4	0.51	0.69	0.71
1ez3	0.71	0.68	0.75
1fna	0.48	0.54	0.59
1fzy	0.57	0.58	0.63
1gou	0.64	0.68	0.66

1h4a	0.65	0.70	0.72
1h4y	0.50	0.62	0.63
1hz6	0.49	0.85	0.62
1i27	0.22	0.29	0.29
1i2t	0.76	0.77	0.86
1ifb	0.43	0.59	0.65
1ifc	0.50	0.56	0.71
1igd	0.78	0.84	0.86
1iib	0.83	0.79	0.83
1iu1	0.82	0.80	0.82
1iul	0.72	0.65	0.76
1iz6	0.41	0.48	0.55
1jbe	0.41	0.76	0.70
1kaf	0.57	0.79	0.82
1kf5	0.29	0.35	0.31
1khy	0.12	0.52	0.40
1kmt	0.72	0.65	0.71
1l3k	0.35	0.72	0.73
1l8r	0.60	0.80	0.73
1lis	0.64	0.69	0.69
1lou	0.75	0.77	0.77
1lu4	0.56	0.76	0.74

1luz	0.63	0.67	0.73
1mjc	0.43	0.44	0.43
1mn8	0.57	0.71	0.49
1nps	0.74	0.76	0.79
1o8x	0.71	0.78	0.81
1opd	0.68	0.82	0.76
1pgx	0.85	0.87	0.89
1poh	0.70	0.82	0.75
1prq	0.56	0.62	0.61
1r69	0.71	0.67	0.69
1r6j	0.63	0.68	0.73
1r77	0.53	0.58	0.60
1row	0.62	0.67	0.71
1sau	0.21	0.45	0.43
1sen	0.15	0.47	0.28
1su9	0.65	0.69	0.76
1t2i	0.44	0.48	0.51
1t2p	0.51	0.49	0.57
1t3x	0.67	0.67	0.68
1t3y	0.51	0.66	0.66
1tg0	0.58	0.69	0.70
1tig	0.77		0.85

1tsf	0.25	0.50	0.57
1ttz	0.53	0.60	0.64
1tud	0.69	0.68	0.69
1tul	0.65	0.78	0.84
1ubi	0.90	0.90	0.92
1ugh	0.73		0.93
1unq	0.58	0.44	0.56
1uow	0.78	0.73	0.80
1vcc	0.40	0.62	0.50
1vkk	0.33	0.44	0.64
1wdv	0.70	0.76	0.78
1wlz	0.39	0.49	0.47
1x6x	0.19	0.49	0.44
1xmk	0.50	0.47	0.58
1xmt	0.68	0.70	0.76
1yn3	0.68	0.71	0.71
1z2u	0.38	0.39	0.39
1zlm	0.83	0.85	0.86
1zma	0.72	0.70	0.78
1zuu	0.81	0.84	0.81
2acy	0.77	0.77	0.77
2b29	0.35	0.41	0.49

2bf5	0.77	0.79	0.81
2chf	0.64	0.81	0.83
2cxd	0.76	0.81	0.80
2dfb	0.51	0.60	0.63
2dyj	0.72	0.77	0.78
2fe5	0.77	0.76	0.81
2fi1	0.39	0.34	0.38
2fwh	0.53	0.71	0.73
2g6f	0.88	0.89	0.88
2gzv	0.12		0.06
2h28	0.44	0.37	0.51
2h3l	0.76	0.84	0.82
2he4	0.21	0.31	0.31
2hhg	0.24		0.27
2i24	0.37	0.46	0.44
2i4a	0.34	0.45	0.42
2i4s	0.33	0.55	0.68
2i6c	0.65	0.72	0.75
2iay	0.43	0.50	0.46
2ic6	0.76	0.75	0.79
2icp	0.02		0.00
2igd	0.45		0.52

2j5y	0.87	0.83	0.87
2jek	0.65	0.84	0.78
2nqw	0.19	0.51	0.47
2nr7	0.24	0.45	0.41
2nsz	0.36	0.50	0.78
2nt4	0.63	0.71	0.70
2nwd	0.10	0.21	0.21
2oml	0.52	0.55	0.55
2oss	0.47	0.58	0.52
2pcy	0.61	0.73	0.68
2ppp	0.60	0.65	0.62
2qjz	0.02	0.52	0.51
2qsk	0.63		0.86
2qy7	0.14	0.45	0.66
2r2z	0.67	0.76	0.67
2ra9	0.43	0.49	0.51
2re2	0.68	0.72	0.71
2v1m	0.58	0.73	0.73
2ve8	0.52	0.62	0.64
2vq4	0.47	0.69	0.65
2vwr	0.18		0.20
2wwe	0.28	0.38	0.46

2x35	0.58	0.64	0.60
2y4x	0.43	0.60	0.66
2y72	0.65	0.75	0.77
2z0t	0.76	0.77	0.78
2zib	0.08	0.27	0.64
2zxj	0.47	0.76	0.77
3b79	0.63	0.71	0.71
3co1	0.28	0.39	0.37
3cx2	0.09	0.18	0.18
3d4e	0.70	0.73	0.75
3dke	0.46	0.53	0.47
3ess	0.48		0.59
3ey6	0.42	0.55	0.50
3f2z	0.59	0.66	0.70
3fk8	0.21	0.36	0.42
3gbw	0.70	0.71	0.74
3hp4	0.60	0.66	0.65
3hyn	0.67	0.72	0.75
3ich	0.68	0.71	0.79
3klr	0.48	0.49	0.49
3nbm	0.81	0.86	0.87
3q6l	0.60	0.68	0.72

4lzt	0.09	0.27	0.24

Table 4-3: Values of RMSD for Min-sum-selected, Rosetta-selected, and Amber-selected models, and minimum RMSD model in the set (All).

PDB	All	Amber	Minimum-Sum	Rosetta
1a32	0.621	1.121	1.124	1.257
1aaj	0.556	0.993	0.817	0.684
1acf	0.781	1.441	1.745	1.745
1bk2	0.549	1.342	1.342	0.936
1bkr	0.758	8.065	1.529	2.185
1bm8	0.715	1.647	0.884	1.377
1cc8	0.589	0.908	1.307	1.453
1cei	0.773	1.504	1.453	1.453
1ctf	0.612	1.762	1.453	1.379
1elw	0.527	0.829	0.957	0.957
1enh	0.619	2.558	2.696	4.686
1ew4	0.745	1.396	1.731	0.988
1ez3	0.76	0.902	2.034	1.827
1fna	0.862	1.268	0.873	2.323
1fzy	0.821	1.22	1.22	1.908
1gou	0.672	1.596	1.057	1.057
1h4a	0.441	0.521	0.637	0.637

1h4y	0.883	1.665	1.652	1.007
1hz6	0.753	2.405	0.917	0.916
1i27	1.038	5.143	4.469	2.658
1i2t	0.649	1.032	1.032	1.032
1ifb	0.809	2.229	1.233	2.238
1ifc	0.77	0.77	0.77	1.063
1igd	0.506	1.894	0.698	0.643
1iib	0.662	0.805	0.758	0.979
1iu1	0.647	1.324	0.863	1.094
1iul	0.642	0.818	0.67	0.919
1iz6	1.148	1.375	1.375	1.591
1jbe	0.727	1.068	1.281	1.031
1kaf	0.628	1.057	0.828	0.671
1kf5	1.095	1.977	1.664	3.027
1khy	1.683	1.871	1.871	1.853
1kmt	0.963	1.3	0.967	1.099
1l3k	0.633	0.847	1.125	1.125
1l8r	0.779	1.241	1.34	1.181
1lis	1.196	1.627	1.405	1.237
1lou	0.62	1.449	0.861	0.861
1lu4	0.715	1.415	0.833	0.829
1luz	0.683	1.13	0.763	1.654

1mjc	1.03	3.777	2.478	2.433
1mn8	0.599	4.138	0.648	0.987
1nps	0.575	0.587	0.586	0.6
1o8x	0.667	0.667	0.972	0.972
1opd	0.584	1.546	0.746	0.747
1pgx	0.5	1.14	1.095	1.777
1poh	0.543	2.415	0.858	0.788
1prq	0.977	1.709	1.535	1.645
1r69	0.657	0.817	1.45	1.45
1r6j	0.795	1.785	1.785	0.894
1r77	0.817	1.777	1.254	1.162
1row	0.784	1.632	1.339	0.863
1sau	0.815	0.815	0.815	1.629
1sen	0.999	6.178	1.465	1.101
1su9	0.71	0.74	0.75	0.812
1t2i	0.932	1.349	1.349	2.261
1t2p	1.07	2.134	1.376	1.749
1t3x	1.075	1.568	1.446	1.446
1t3y	0.874	1.052	1.052	1.423
1tg0	0.78	1.44	1.052	1.864
1tsf	0.768	1.048	1.055	5.834
1ttz	0.722	2.316	1.529	1.478

1tud	0.626	1.679	1.551	1.222
1tul	0.624	1.021	0.914	0.938
1ubi	0.596	0.637	0.709	0.709
1unq	0.817	1.757	3.31	2.252
1uow	0.603	0.93	0.93	0.873
1vcc	0.628	4.991	1.03	2.06
1vkk	0.719	0.846	2.951	3.777
1wdv	0.63	0.926	1.181	1.142
1wlz	0.681	2.016	2.694	2.933
1x6x	0.887	1.631	2.013	1.97
1xmk	1.002	1.99	1.953	1.638
1xmt	0.758	1.691	1.691	0.93
1yn3	0.828	2.587	1.363	1.419
1z2u	0.777	1.351	1.351	1.318
1zlm	0.503	1.007	0.587	0.784
1zma	0.614	1.344	0.66	2.136
1zuu	0.36	0.748	0.834	0.507
2acy	0.487	0.942	0.595	0.785
2b29	1.093	1.41	1.414	1.622
2bf5	0.751	0.96	0.823	0.823
2chf	0.662	0.662	0.662	0.916
2cxd	0.748	0.912	0.916	0.87

2dfb	0.753	1.044	1.044	0.77
2dyj	0.817	1.413	1.012	1.054
2fe5	0.656	0.94	0.945	1.011
2fi1	1.244	2.354	2.354	2.354
2fwh	0.828	0.837	0.837	1.182
2g6f	0.577	0.81	0.696	0.822
2h28	1.115	1.292	3.169	2.594
2h3l	0.661	0.765	0.907	0.907
2he4	1.461	3.842	3.842	2.381
2i24	0.765	2.329	2.755	0.852
2i4a	0.979	2.061	1.975	1.676
2i4s	0.755	1.689	1.234	0.894
2i6c	0.735	0.992	0.992	1.307
2iay	0.769	2.475	2.475	2.475
2ic6	0.577	0.841	1.266	1.251
2j5y	0.306	0.601	0.832	2.214
2jek	0.808	1.147	1.147	1.147
2nqw	1.147	1.147	1.147	1.315
2nr7	1.601	2.401	2.401	2.011
2nsz	0.704	0.704	1.137	1.481
2nt4	0.762	2.097	1.5	0.977
2nwd	0.938	5.607	0.979	0.979

2oml	1.215	2.254	1.233	1.828
2oss	1.159	2.027	2.027	2.049
2pcy	0.636	2.214	1.585	1.576
2ppp	0.69	2.599	0.885	1.239
2qjz	0.814	0.948	0.948	0.958
2qy7	1.026	1.685	1.091	2.279
2r2z	0.583	1.37	1.249	1.735
2ra9	1.264	1.685	2.505	2.522
2re2	0.779	1.445	1.015	2.002
2v1m	0.734	0.879	0.879	0.851
2ve8	0.756	0.994	1.706	0.964
2vq4	0.864	0.864	0.864	0.864
2wwe	1.109	3.167	1.739	1.739
2x35	1.005	1.459	1.459	1.457
2y4x	0.835	2.286	2.167	2.167
2y72	0.525	0.78	0.644	1.682
2z0t	0.568	0.973	0.656	0.776
2zib	1.835	1.835	1.835	3.041
2zxj	0.657	0.683	0.792	1.173
3b79	0.544	1.071	1.071	0.633
3co1	1.107	2.687	2.687	7.218
3cx2	2.824	3.462	3.49	3.272

3d4e	0.816	1.165	1.165	1.268
3dke	0.685	2.307	2.345	1.464
3ey6	1.125	3.149	1.233	1.148
3f2z	0.576	0.576	0.576	0.817
3fk8	0.76	1.952	1.952	6.017
3gbw	0.833	1.161	1.161	1.027
3hp4	0.854	1.696	1.696	1.111
3hyn	0.752	1.195	0.766	0.856
3ich	0.766	0.964	0.766	0.775
3klr	0.87	0.957	1.474	1.415
3nbm	0.594	0.961	0.611	0.938
3q6l	0.814	1.428	1.135	1.896
4lzt	0.834	0.834	0.834	0.834

Table 4-4: *β*-metric values for Rosetta talaris2014, Rosetta REF2015, and Amber ff14SBonlySC for the Loop Modeling benchmark set.

	Rosetta		**Amber**
PDB	**Talaris2014**	**REF2015**	**ff14SBonlySC**
1a8d	0.24	0.40	0.44
1arb	0.87	0.86	0.86
1bhe	0.85	0.95	0.93
1bn8	0.85	0.81	0.82

1c5e	0.99	0.99	0.98
1cnv	0.58	0.56	0.39
1cs6	0.42	0.25	0.77
1cyo	0.38	0.19	0.92
1dqz	0.83	0.82	0.97
1eco	0.93	0.95	0.80
1ede	0.64	0.56	0.46
1exm	0.93	0.93	0.92
1ezm	0.60	0.63	0.79
1f46	0.53	0.33	0.89
1i7p	0.81	0.95	0.89
1msc	0.57	0.67	0.29
1my7	0.95	0.93	0.89
1onc	0.31	0.46	0.62
1oth	0.97	0.95	0.88
1oyc	0.43	0.45	0.82
1pbe	0.56	0.68	0.81
1qlw	0.71	0.64	0.77
1rro	0.81	0.72	0.96
1srp	1.00	1.00	1.00
1tca	0.37	0.50	0.86
1thg	0.75	0.79	0.62

1thw	0.56	0.68	0.47
1tib	0.76	0.81	0.85
1tml	0.96	0.97	0.98
1xif	0.95	0.99	0.97
2cpl	0.80	0.90	0.95
2ebn	0.64	0.58	0.44
2exo	0.97	0.96	0.97
2rn2	0.85	0.90	0.90
2sil	0.42	0.45	0.63
2tgi	0.41	0.46	0.60
3cla	0.51	0.54	0.72
3hsc	0.93	0.95	0.98
4i1b	0.25	0.20	0.05

4.9 Supplementary Software

4.9.1 Native Relax

4.9.1.1 Inputs

1. **<PATH_TO_XTAL_PDB>** Crystallographic pdb of protein-peptide complex

   Retrieve from pdb

2. **<XML_FILE>**

   Sample xml:

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
      <ScoreFunction name="myscore" weights="talaris2014.wts"/>
    </SCOREFXNS>
    <TASKOPERATIONS>
      <InitializeFromCommandline name="init"/>
    </TASKOPERATIONS>
    <FILTERS/>
    <MOVERS>
        <FastRelax name="fastrelax" repeats="8" scorefxn="myscore" task_operations="init"/>
    </MOVERS>
    <APPLY_TO_POSE/>
    <PROTOCOLS>
      <Add mover_name="fastrelax"/>
    </PROTOCOLS>
</ROSETTASCRIPTS>
```

4.9.1.2 Process

Run on initial crystallographic pdb 100 times to get rid of internal clashes.

4.9.1.3 Command Line:

<ROSETTA_BIN>rosetta_scripts.static.linuxgccrelease -database <ROSETTA_DB> -ex1 -ex2 -extrachi_cutoff 1 -use_input_sc -s <PATH_TO_XTAL_PDB> -parser:protocol <XML_FILE> -score:weights talaris2014

4.9.1.4 Outputs

100 relaxed pdb files.

4.9.2 Rosetta Minimize

4.9.2.1 Inputs

1. **<PATH_TO_DECOY>** path to decoy conformation or native-like conformation to minimize
2. **<PATH_TO_XTAL_PDB>** path to crystallographic native structure from which to calculate RMSD
3. **<XML_FILE>**

Sample xml:

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
        <ScoreFunction name="myscore" weights="talaris2014" />
    </SCOREFXNS>
    <TASKOPERATIONS>
    </TASKOPERATIONS>
    <FILTERS>
    <Rmsd name="rmsd" threshold="100" superimpose="1" >
    </Rmsd>
    </FILTERS>
    <MOVERS>
    <MinMover name="min_sc_bb" scorefxn="myscore" chi="1" bb="1" tolerance="0.0001"
max_iter="2000" type="lbfgs_armijo_nonmonotone" >
    </MinMover>
    </MOVERS>
<APPLY_TO_POSE>
</APPLY_TO_POSE>

    <PROTOCOLS>
    <Add mover_name="min_sc_bb"/>
    <Add filter_name="rmsd"/>
</PROTOCOLS>
</ROSETTASCRIPTS>
```

4.9.2.2 Process

Run on all decoy conformations and relaxed native conformations for each protein.

4.9.2.3 Command Line:

<ROSETTA_BIN>rosetta_scripts.static.linuxgccrelease -database <ROSETTA_DB> -ex1 -ex2 -extrachi_cutoff 1 -use_input_sc -s <PATH_TO_DECOY> -parser:protocol

<XML_FILE> -score:weights talaris2014 -in:file:native <PATH_TO_XTAL_PDB> -nblist_autoupdate

4.9.2.4 Outputs

1 minimized pdb file and one line in a score file, including total_score and rmsd values for each minimized conformation.

4.9.3 REF2015 Minimize

4.9.3.1 Inputs

Rosetta Git SHA1: 7deda350387909882721e82407e30054a16def51

1. **<PATH_TO_DECOY_LIST_FILE>** path to a file that contains a list of decoy conformations or native-like conformations to minimize
2. **<PATH_TO_XTAL_PDB>** path to crystallographic native structure from which to calculate RMSD
3. **<XML_FILE>**

```
<ROSETTASCRIPTS>

  <SCOREFXNS

    <ScoreFunction name="myscore" weights="ref2015.wts" />

  </SCOREFXNS>

  <RESIDUE_SELECTORS>

    <Index name="loop" resnums="%%loop_start%%-%%loop_end%%" />

    <Not name="not_loop" selector="loop" />

  </RESIDUE_SELECTORS>

  <TASKOPERATIONS>

    Turn off rep/des all residues that are not in the loop region. Leave loop designable.

    <OperateOnResidueSubset name="loop_region_designable" selector="not_loop" >

      <PreventRepackingRLT/>

    </OperateOnResidueSubset>

  </TASKOPERATIONS>



  <FILTERS>

    CA RMSD of loop when surrounding is superimposed (shape and position of loop)

    <Rmsd name="rmsd" threshold="100" superimpose="0" >
```

```
            <span begin_res_num="%%loop_start%%" end_res_num="%%loop_end%%" />
          </Rmsd>
          CA RMSD of loop when loop is superimposed (shape of loop)
          <Rmsd name="rmsd_suploop" threshold="100" superimpose="1" >
            <span begin_res_num="%%loop_start%%" end_res_num="%%loop_end%%" />
          </Rmsd>
          Energy for set of residues.
          <TaskAwareScoreType name="loop_energy" task_operations="loop_region_designable"
scorefxn="myscore" score_type="total_score" threshold="100000" mode="total" />
        </FILTERS>
        <MOVERS>
          Superimpose onto Native by 1:(loop start -1) residues.
          <Superimpose name="super_notloop" ref_start="1" ref_end="%%loop_start_pre%%"
target_start="1" target_end="%%loop_start_pre%%" CA_only="1" />
          Minimize the loop region only.
          <TaskAwareMinMover name="min_sc_bb_looponly" scorefxn="myscore" chi="1" bb="1"
tolerance="0.0001" max_iter="2000" type="lbfgs_armijo_nonmonotone"
task_operations="loop_region_designable" />
        </MOVERS>
        <PROTOCOLS>
          <Add mover="min_sc_bb_looponly"/>
          <Add mover="super_notloop"/>
                <Add filter="rmsd"/>
          <Add filter="rmsd_suploop"/>
        </PROTOCOLS>
      </ROSETTASCRIPTS>
```

4.9.3.2 Process

Run on all decoy conformations and relaxed native conformations for each protein.

4.9.3.3 Command Line:

rosetta_scripts.static.linuxgccrelease -database <ROSETTA_DB> -ex1 -ex2 -extrachi_cutoff 1 -use_input_sc -s -parser:protocol <XML_FILE> -score:weights talaris2014 -in:file:native <PATH_TO_XTAL_PDB> -nblist_autoupdate

<ROSETTA_BIN>rosetta_scripts.default.linuxgccrelease -ex1 -ex2 -extrachi_cutoff 1 -use_input_sc -nblist_autoupdate –l <PATH_TO_DECOY_LIST_FILE> -parser:protocol <PATH_TO_XML> -score:weights ref2015 -in:file:native <PATH_TO_XTAL_PDB> -parser:script_vars loop_start_pre=202 loop_start=203 loop_end=214

4.9.3.4 Outputs

Minimized pdb files and a score file containing lines for each pdb with metrics, including total_score and rmsd values for each minimized conformation.

4.9.4 Rosetta Relax (REF2015)

4.9.4.1 Inputs

1. **<PATH_TO_DECOY>** path to decoy conformation
2. **<XML_FILE>**

Sample xml:

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
      <ScoreFunction name="myscore" weights="REF2015.wts"/>
    </SCOREFXNS>
    <TASKOPERATIONS>
      <InitializeFromCommandline name="init"/>
    </TASKOPERATIONS>
   <FILTERS>
    <Rmsd name="rmsd" threshold="100" superimpose="1" >
    </Rmsd>
   </FILTERS>
    <MOVERS>
        <FastRelax name="fastrelax" repeats="8" scorefxn="myscore" task_operations="init"/>
    </MOVERS>
    <APPLY_TO_POSE/>
    <PROTOCOLS>
      <Add mover_name="fastrelax"/>
    <Add filter_name="rmsd"/>
    </PROTOCOLS>
```

```
</ROSETTASCRIPTS>
```

3. <RELAX_SCRIPT>

```
switch:torsion
repeat 3

ramp_repack_min 0.02  0.01    1.0  50
ramp_repack_min 0.250 0.01    0.5  50
ramp_repack_min 0.550 0.01    0.0 100
ramp_repack_min 1     0.00001  0.0 200
accept_to_best
endrepeat

switch:cartesian
repeat 2
ramp_repack_min 0.02  0.01    1.0  50
ramp_repack_min 0.250 0.01    0.5  50
ramp_repack_min 0.550 0.01    0.0 100
ramp_repack_min 1     0.00001  0.0 200
accept_to_best
endrepeat
```

4.9.4.2 Process

Run on all decoy conformations for each protein.

4.9.4.3 Command Line:

<ROSETTA_BIN>rosetta_scripts.static.linuxgccrelease –database <ROSETTA_DB> –s <PATH_TO_DECOY> –parser:protocol <XML_FILE> –score:weights REF2015 –relax:script <RELAX_SCRIPT>

4.9.4.4 Outputs

1 relaxed pdb file and one line in a score file, including total_score and rmsd values for each minimized conformation.

4.9.5 Pareto Solution

4.9.5.1 Inputs

1. **<Amber scores>** score-file that contains Amber energies for each decoy for a given protein
2. **<Rosetta scores>** score-file that contains Rosetta energies for each decoy for a given protein

4.9.5.2 Process

Python script that includes the following four methods (dominates and cull taken with slight modifications from Yakym Pirozhenko under Creative Commons license 3.0 and gen_ranks taken from mjolka under Creative Commons license 3.0):

```
def dominates(row, rowCandidate):
    return all(r <= rc for r, rc in zip(row, rowCandidate))


def cull(pts, dominates):
    dominated = []
    cleared = []
    remaining = pts
    while remaining:
        candidate = remaining[0]
        new_remaining = []
        for other in remaining[1:]:
            [new_remaining, dominated][dominates(candidate, other)].append(other)
        if not any(dominates(other, candidate) for other in new_remaining):
            cleared.append(candidate)
        else:
            dominated.append(candidate)
```

```
        remaining = new_remaining
    return cleared, dominated

def find_lowest_point( list_pts ):
    first_rank_list = [ p[0] for p in list_pts ]
    second_rank_list = [ p[1] for p in list_pts ]
    min_rank = min(first_rank_list + second_rank_list)
    min_point = [ (e1, e2, r) for e1, e2, r in list_pts if min_rank == e1 or min_rank == e2
][0]
    return min_point

def gen_ranks(list_energies):
    indices = list(range(len(list_energies)))
    indices.sort(key=lambda x: list_energies[x])
    output = [0] * len(indices)
    for i, x in enumerate(indices):
        output[x] = i
    return output

def main(amber_scorefile, rosetta_scorefile):

    #read in amber_scores and rosetta_scores using external library methods READ
    amber_scores = READ(amber_scorefile)
    rosetta_scores = READ(rosetta_scorefile)

    #generate ranks for scores
```

```
amber_ranks = gen_ranks(amber_scores)

rosetta_ranks = gen_ranks(rosetta_scores)


pts = map(list, zip(d1e_ranks, d2e_ranks))


#find the set of Pareto-optimal solutions

cleared, dominated = cull(pts, dominates)


#find the set of solutions that have a minimal sum of ranks

pareto_equal_min = min([ e1+e2 for e1,e2 in cleared.items() ])

list_pts =  [ (rosetta,amber) for rosetta,amber in cleared if amber+rosetta ==
pareto_equal_min ]


#find the subset of solutions that have the lowest rank of all the solutions with a
minimal sum of ranks

pareto_solution = find_lowest_point( list_pts )
```

4.9.6 Outputs

This gives the ranks of the Pareto-optimal solution, which can be used to find the decoy name or rmsd or any other information about the decoy.

4.9.7 Amber – Generate Initital Structures with tLEaP

4.9.7.1 Inputs

1. <PDB_ID>

2. <TLEAP_INFILE>

```
source leaprc.protein.ff14SBonlysc

m = loadpdb NoH_<PDB_ID>.pdb

set default pbradii mbondi3

saveamberparm m <PDB_ID>.parm7 NoH_<PDB_ID>.rst7

quit
```

4.9.7.2 Process

Replaces hydrogen atoms and converts PDB files of crystal structures and decoys to one topology file (parm7) and individual coordinate files (rst7).

4.9.7.3 Command Line:

tleap –f <TLEAP_INFILE>

4.9.7.4 Outputs

One topology file <PDB_ID>.parm7 and all decoy structures as coordinates files <PDB_ID>.rst7

4.9.8 Amber – Minimization

4.9.8.1 Inputs

1. <PDB_ID>

2a. . <MIN_SCRIPT> (without restraints)

```
$cntrl
  imin = 1, maxcyc=1000,
  ntx = 1,
  ntxo = 2,
  ntwr = 100, ntpr = 100,
  cut = 999.0,
  ntb = 0, igb = 8,
  ntr = 1,
  ntmin=3, drms=0.01,
/
```

OR 2b. <MIN_SCRIPT> (with restraints)

```
$cntrl
  imin = 1, maxcyc=1000,
  ntx = 1,
  ntxo = 2,
```

```
    ntwr = 100, ntpr = 100,
    cut = 999.0,
    ntb = 0, igb = 8,
    ntr = 1,
    restraint_wt = 10,
    restraintmask = "!:<LOOP_RESIDUES> & !@H=",
    ntmin=3, drms=0.01,
  /
```

4.9.8.2 Process

Run Amber minimization on all decoy conformations and native crystal conformations for each protein.

Command Line:

sander -O -i <MIN_SCRIPT> -p <PDB_ID>.parm7 -c <PDB_ID>.rst7 -r <PDB_ID>.min.rst7 -ref <PDB_ID>.rst7 -o <PDB_ID>.min.out

4.9.9 Amber – Scoring

4.9.9.1 Inputs

1. <PDB_ID>
2. <OUTPUT_FILE>

4.9.9.2 Process

Command Line: python GetEnergies.py –i <PDB_ID> -o <OUTPUT_FILE>

Sample python script for gathering scores from loop modeling benchmark set:

```
import pytraj as pt
import sander
import os, sys, getopt
from glob import glob
from mpi4py import MPI
import pandas as pd
```

```
loopdef = pd.read_json('loopdefs.json')

# create mpi handler to get cpu rank
comm = MPI.COMM_WORLD

PATH_TO_NATIVES = '<PATH/TO/NATIVES>'
PATH_TO_DECOYDISC = "<PATH/TO/DECOYS/>"

def chunks(l,n):
    n = max(1,n)
    return [l[i:i+n] for i in range(0, len(l), n)]

def get_ca_rmsd( traj, mask, metric ):

    ca_rmsd = pt.pairwise_rmsd(traj, mask=mask, metric=metric)
    return ca_rmsd;

def get_energies( traj, igb_value ):



    energy_data = pt.pmap_mpi(pt.energy_decomposition, traj, igb=igb_value )
    return energy_data;



def get_energy_term( traj, term ):
```

```
    energy_data = pt.pmap_mpi(pt.energy_decomposition, traj, igb=8)
    return energy_data[term];



def main(argv):
    args = sys.argv

    input_pdb = ''
    outfile = ''

    try:
        opts, args=getopt.getopt(sys.argv[1:], "ho:i:o:", ["in:file:i=", "out:file:scorefile="])
    except getopt.GetoptError:
        print('Unknown flag given.\nKnown flags:\n\t-h\n\t-n <native>')
        sys.exit()

    for opt, arg in opts:
        if opt == '-h':
            print('GetEnergies.py --in:file:s <input_pdb_id> --out:file:scorefile
<output_filename>')
            sys.exit()
        elif opt in ("-i", "--in:file:i"):
            input_pdb = arg
        elif opt in ("-o", "--out:file:scorefile"):
            outfile = arg
```

```
    if input_pdb == '':
        print('ERROR: No PDB ID supplied.')
        sys.exit()
    elif len(input_pdb) != 4:
        print("ERROR: Input PDB should be 4 letter PDB code.")
        sys.exit()
    input_pdb = input_pdb.lower()

    if outfile == '':
        outfile = 'Scores.sc'
    ##############################################################
######
##############################################################
#
    print( "=== Native RST7: {PDB} ===".format(PDB=input_pdb) )
    #Native is crystal structure.
    native_rst7 = PATH_TO_NATIVES+"/{PDB}.rst7".format( PDB=input_pdb )
    native_parm = PATH_TO_NATIVES+"/{PDB}.parm7".format( PDB=input_pdb )

    print("Native RST7\t{natrst}".format(natrst=native_rst7))
    print("Native PARM7\t{natparm}".format(natparm=native_parm))

    os.chdir(PATH_TO_DECOYDISC+"{PDB}/".format( PDB=input_pdb ) )
    print(PATH_TO_DECOYDISC+"{PDB}/".format( PDB=input_pdb ) )
    min_decoys = glob('min*.rst7')  ## List of rst7s ["<rst7_1>", "<rst7_2>", ...]

    print("== Analyzing %i mols==" % len(min_decoys))
```

```
        min_decoys.insert(0, native_rst7)


        print("\t=== Loading Trajectories ===")

        traj = pt.iterload(min_decoys, native_parm)

        print(traj)


        print("\t=== Getting RMSDs ===")

        ca_mask = '@CA & :{loop_start}-
{loop_end}'.format(loop_start=int(loopdef[input_pdb]['StartResidueID'].strip()),
loop_end=int(loopdef[input_pdb]['EndResidueID'].strip() ))

        print ca_mask


        ca_rmsd_nofit = get_ca_rmsd( traj, ca_mask, 'nofit' )

        ca_rmsd_fit = get_ca_rmsd( traj, ca_mask, 'rms' )


        print("\t=== Getting Energy Decomposition ===")

        energy_data = get_energies( traj, 8 )

        print("\t\tFinished")


        if energy_data:

            energy_data['rmsd'] = ca_rmsd_nofit[0]

            energy_data['rmsd_suploop'] = ca_rmsd_fit[0]


            ekeys = energy_data.keys()

            ekeys.sort()
```

```
        print("Outfile: " + outfile)
        with open(outfile,'w') as scorefile:
            header = 'description\t'
            for key in ekeys:
                header += key + '\t'
            scorefile.write(header+"\n")

            for pdb_index in range(len(min_decoys)):
                scoreline = min_decoys[pdb_index]+'\t'
                for key in ekeys:
                    scoreline += '%s\t' % str(energy_data[key][pdb_index])
                scorefile.write(scoreline+"\n")

    if __name__ == "__main__":
        main(sys.argv[1:])
```

4.9.9.3 Outputs

One output file with Amber scores and RMSDs for native crystal structure and decoys.

4.10 References

1. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science (80-.).* **181,** 223–230 (1973).

2. Lu, H. & Skolnick, J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins Struct. Funct. Genet.* **44,** 223–232 (2001).

3. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30,** 1545–1614 (2009).

4. Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117,** 5179–5197 (1995).

5. Jernigan, R. L. & Bahar, I. Structure-derived potentials and protein simulations. *Current Opinion in Structural Biology* **6,** 195–209 (1996).

6. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15,** 2507–24 (2006).

7. Ponder, J. W. & Case, D. A. Force fields for protein simulations. *Advances in Protein Chemistry* **66,** 27–85 (2003).

8. 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.* **268,** 209–25 (1997).

9. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118,** 11225–11236 (1996).

10. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11,** 3696–3713 (2015).

11. MacKerell, A. D. *et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. *J. Phys. Chem. B* **102,** 3586–3616 (1998).

12. Xu, D. & Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins Struct. Funct. Bioinforma.* **80,** 1715–1735 (2012).

13. O'Meara, M. J. *et al.* Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theory Comput.* **11,** 609–622 (2015).

14. Shapovalov, M. V. & Dunbrack, R. L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19,** 844–858 (2011).

15. Best, R. B. *et al.* Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ??, ?? and side-chain ?? 1 and ?? 2 Dihedral Angles. *J. Chem. Theory Comput.* **8,** 3257–3273 (2012).

16. Raval, A., Piana, S., Eastwood, M. P., Dror, R. O. & Shaw, D. E. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins Struct. Funct. Bioinforma.* **80,** 2071–2079 (2012).

17. Lindorff-Larsen, K. *et al.* Systematic validation of protein force fields against experimental data. *PLoS One* **7,** (2012).

18. Park, H. *et al.* Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **12,** 6201–6212 (2016).

19. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular

Modeling and Design. *J. Chem. Theory Comput.* **13,** 3031–3048 (2017).

20. Beauchamp, K. A., Lin, Y. S., Das, R. & Pande, V. S. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J. Chem. Theory Comput.* **8,** 1409–1414 (2012).

21. Cino, E. A., Choy, W. Y. & Karttunen, M. Comparison of secondary structure formation using 10 different force fields in microsecond molecular dynamics simulations. *J. Chem. Theory Comput.* **8,** 2725–2740 (2012).

22. Nguyen, H., Roe, D. R. & Simmerling, C. Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* **9,** 2020–2034 (2013).

23. Nguyen, H., Maier, J., Huang, H., Perrone, V. & Simmerling, C. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J. Am. Chem. Soc.* **136,** 13959–13962 (2014).

24. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23,** 47–55 (2014).

25. Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

26. Tyka, M. D. *et al.* Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405,** 607–618 (2011).

27. Tyka, M. D., Jung, K. & Baker, D. Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers. *J. Comput. Chem.* **33,** 2483–2491 (2012).

28. Case, D. A. *et al.* AMBER 2016. (2016).

29. Liu, D. C. & Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **45,** 503–528 (1989).

30. Nguyen, H., Roe, D. R., Swails, J. M. & Case, D. A. PYTRAJ: Interactive data analysis for molecular dynamics simulations. *Manuscr. Prep.* (2017).

31. Dill, K. a & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **338,** 1042–1046 (2012).

32. Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. Navigating the folding routes. *Science* **267,** 1619–1620 (1995).

33. Shakhnovich, E. Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet. *Chemical Reviews* **106,** 1559–1588 (2006).

34. Xiang, Z., Soto, C. S. & Honig, B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 7432–7437 (2002).

35. Fiser, A., Kinh Gian Do, R. & Sali. Modeling Loops in Protein Structures. *Protein Sci.* **9,** 1753–1773 (2000).

36. Murphy, P. M., Bolduc, J. M., Gallaher, J. L., Stoddard, B. L. & Baker, D. Alteration of enzyme specificity by computational loop remodeling and design. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 9215–20 (2009).

37. Wang, C., Bradley, P. & Baker, D. Protein-Protein Docking with Backbone Flexibility. *J. Mol. Biol.* **373,** 503–519 (2007).

38. Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6,** 551–552 (2009).

39. Ollikainen, N., Smith, C. A., Fraser, J. S. & Kortemme, T. Flexible backbone sampling methods to model and design protein alternative conformations. *Methods Enzymol.* **18,** 1199–1216 (2013).

40. Stein, A. & Kortemme, T. Improvements to Robotics-Inspired Conformational Sampling in Rosetta. *PLoS One* **8,** (2013).

Chapter 5: Computational assessment of thioether isosteres

5.1 Preface

A version of Chapter 5 has been published [Barrows, R. D., Blacklock, K. M., Rablen, P. R., Khare, S. D. & Knapp, S. Computational assessment of thioether isosteres. *J. Mol. Graph. Model.* (2018). doi:10.1016/j.jmgm.2018.01.018], of which I am an author. I was responsible for contributing to concept formation, data collection and analysis, and manuscript composition regarding the Rosetta and Amber analysis of prospective isosteres. Sagar Khare performed the RosettaDock protocol and contributed to manuscript composition. Paul Rablen contributed to concept formation, experimental design, and manuscript composition regarding the QM calculations. Spencer Knapp was the supervisory author on this project and was involved throughout the project in concept formation and manuscript composition. Robert Barrows was the lead investigator, responsible for all major areas of concept formation, data collection and analysis, as well as manuscript composition.

5.2 Abstract

Replacement of the sulfur atom in biologically active diaryl and heteroaryl thioethers (Ar–S–Ar', HAr–S–Ar, and HAr–S–HAr') with any of several one-atom or two-atom linkers can be expected to reduce the susceptibility of the analogue to metabolic oxidation, a well-documented problem for thioethers intended for medicinal chemistry applications. Ab initio calculations indicate how well various proposed thioether isosteric groups, including some new and unusual ones, may perform structurally and electronically in replacing the bridging sulfur atom. Four of these are calculationally evaluated as proposed substructures in Axitinib analogues. The predicted binding behavior of the latter

within two different previously crystallographically characterized protein–Axitinib binding sites (VEGFR2 kinase and ABL1 T315I gatekeeper mutant kinase), and an assessment of their suitability and anticipated shortcomings, are presented.

5.3 Introduction

An important tactic in medicinal chemistry is the replacement of one synthesized substructure with a "bioisostere"[1,2] that matches the original in most respects, but offers improved performance in others. As one example, tetrazol-5-yl has been used as a carboxyl (—CO_2H) equivalent to take advantage of its similar size and hydrogen bonding properties, but its resistance to some, if not all, unwanted metabolic conversions[3–5]. Similarly, suggestions for bioisosteric groups have been compiled for replacing various other commonly occurring linkers, functionalities, and substituents in medicinal chemistry, such as phenyl, carboxamide, ester, ether, hydroxyl, methoxyl, alkyl, and halogen.

Thioethers, especially Ar–S–Ar', HAr–S–Ar, and HAr–S–HAr' (Ar = aryl; HAr = heteroaryl) systems, occur commonly among bioactive compounds, "hit" and "lead" nominees, and drug candidates. **Figure 5-1** displays a representative variety of thioethers **A**–**Q** gathered from the medicinal chemistry literature and the Protein Data Bank (PDB)[6–29] A search of the latter turned up 148 examples of structures in which a bound thioether of the form Ar–S–Ar', HAr–S–Ar, or HAr–S–HAr' is found. Additional examples of alkyl thioethers (e. g., Ar–S-alkyl) are also available, and some of these will display similar properties, but for the present purpose only the aromatic thioethers will be considered. To summarize the 148 protein structure examples: the sulfur atom in bound aryl/heteroaryl thioethers does not play a role as a hydrogen bond acceptor, but rather as a hydrophobic structural element. According to PDB Poseview[30,31] representations, often the sulfur atom is buried deep within a hydrophobic pocket lined by protein side chains

such as those from phenylalanine and leucine. Therefore, the important aspects for a potential thioether bioisostere to mimic include the unique structural features (**Figure 5-1**), as well as the relatively small size and lipophilicity of the sulfur atom itself. An additional aspect of covalent sulfur, the "sulfur hole effect," is discussed in Section 5.4.3.

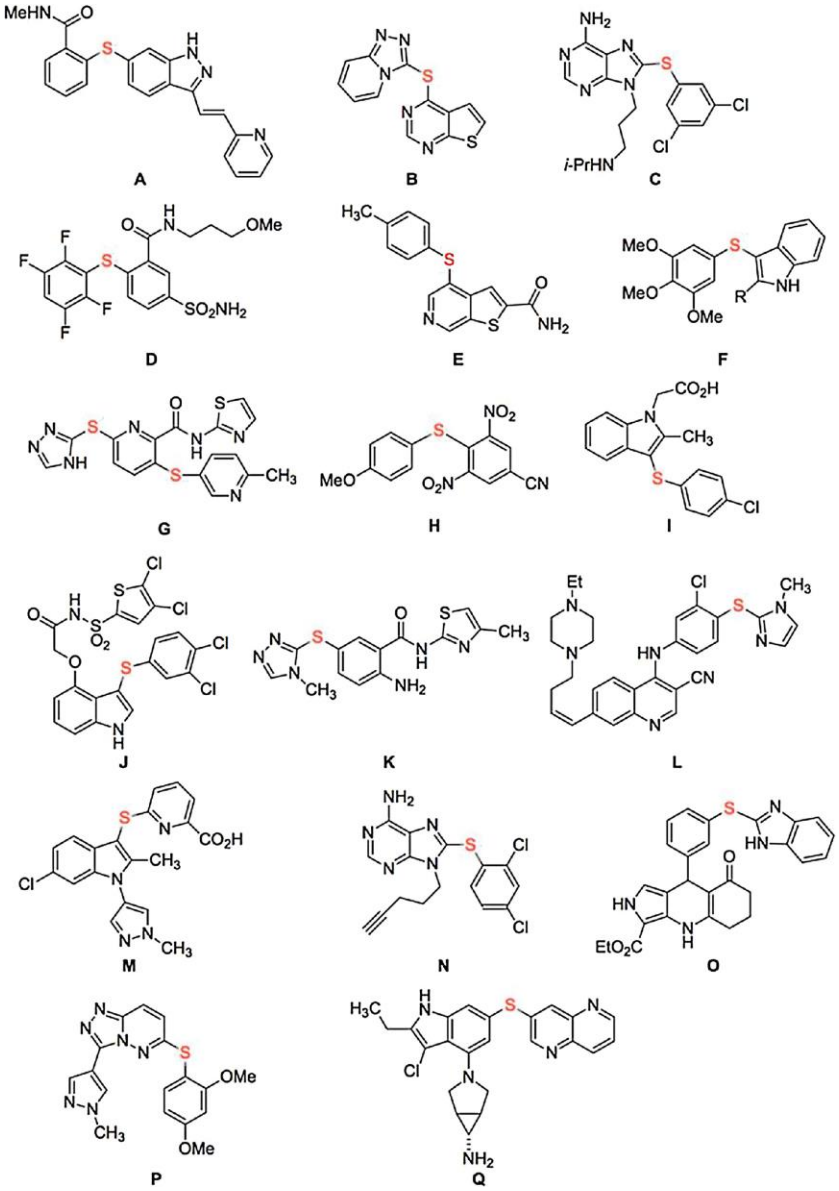
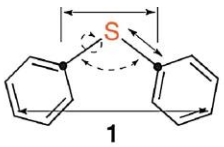
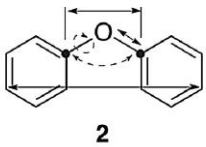


Figure 5-1: Representative aryl and heteroaryl thioethers from the medicinal chemistry literature.6 The bridging sulfur atoms are shown in red.

Oxidation at the sulfur atom, often under the action of cytochrome P450 isoforms and, to a lesser extent, flavin containing monoxygenases, is a commonly observed metabolic process that can detract from the biological effectiveness of the intended thioether[32,33]. Among the **Figure 5-1** thioethers, several (**A**, **E**, **G**, **I**, and **O**) are explicitly described as undergoing metabolic oxidation to the corresponding sulfoxide [—S(O)—] or sulfone [—S(O_2)—]; in other cases (**B**, **C**, **F**, **J**, **K**, **L**, and **P**), unspecified rapid metabolism or clearance of the thioether is reported. For some examples, the corresponding sulfoxide or sulfone is described as being less active than the sulfide (**A**, **C**, and **D**), although for three (**I**, **J**, and **N**) the sulfoxide or sulfone does possess activity. For several (**A**, **C**, **D**, **H**, **M**, and **P**), attempted isosteric replacement of the sulfur atom by —O—, —CH_2—, or other group is described, but the resulting analogue is less active. For **P** specifically, replacement of —S— by, variously, — O—, —NH—, —CH_2—, —CF_2—, —NMe—, or —CHOH— gave in each case a less active analogue. In some examples (**F**, **I**, **K**, **N**, **O**, and **Q**) the corresponding ether or other one atom bridge is only slightly worse or actually an improvement over the thioether.

The common occurrence of thioethers in medicinal chemistry is due in part to their ease of synthesis, as the S–Ar/HAr bond can be assembled in a variety of useful and efficient ways[34–43]. Additionally, sulfur as a bridging atom offers unique structural characteristics, as displayed in **Figure 5-2**. Calculated values for bond lengths, angles, and interatomic distances are shown for Ph—S—Ph (**1**) and Ph—O—Ph (**2**). C(•) represents the ipso carbon. Thus, the divalent bridging sulfur atom, relative to —O—, features much longer bonds, a narrower central bond angle, and aromatic rings that are less co-planar. Interestingly, the C(para)— C(para) distances taken alone are a good match. Nevertheless, the positioning of the attached aromatic rings in **1** differs considerably from that in **2** and other commonly used one-atom bridged structures, including those with —NH—, or —CH_2—.





1	2
103.6° C(•)–S–C(•): bond angle	120.9° C(•)–O–C(•): bond angle
1.79 Å C(•)–S: bond length	1.38 Å C(•)–O: bond length
46.5° CC(•)SC(•): torsional angle	39.9° CC(•)OC(•): torsional angle
2.82 Å C(•)–C(•): ipso–ipso distance	2.40 Å C(•)–C(•): ipso–ipso distance
7.44 Å C(para)–C(para): distance	7.46 Å C(para)–C(para): distance

Figure 5-2: Comparison of the calculated structural parameters of diphenyl thioether (1) and diphenyl ether (2).

In this paper, we evaluate by *ab initio* calculation the structural characteristics of a variety of thioether isosteres, including several that feature two-atom bridges and one with a three-atom bridge, in comparison with diphenyl sulfide. We also examine the electrostatic charge characteristics of several bis(pyrid-4-yl) isosteres as stand-ins for bis(heteroaryl) sulfide systems. Finally, we simulate the incorporation of four isosteres of Axitinib into two crystallographically characterized protein binding sites, and comment on which aspects of the isosteres lead to favorable and unfavorable interactions.

5.4 Results and discussion

5.4.1 Calculated parameters for isosteres of diphenyl thioether and bis(pyridin-4-yl) thioether

A selection of 22 proposed thioether isosteres for diphenyl thioether (**3a = 1**) is shown in **Figure 5-3**. Calculated geometrical parameters for the most stable conformation of these compounds (**4a**–**25a**), optimized by using the B3LYP/6-31G(d) procedure, are displayed in **Table 5-1**. The isosteres are arranged according to the increasing theoretical distortion energy required to geometrically match each with **3a**. By this measurement, the dimethylsilylene isostere (**4a**, 4.2 kcal/mol) is a relatively close geometrical match, whereas the simple ether Ph—O—Ph (**22a**, 30.1 kcal/mol) is poor. Depending on the proposed

application, however, other factors may be important. For example, the ketimine **9a** best matches **3a** in terms of C(ipso)—C(ipso) distance, but otherwise shows poor overlap in the relative positioning of the respective phenyl rings. The 1,1-cyclopropylidene isostere, **7a**, resembles **3a** well in terms of C(para)—C(para) distance (7.35 vs 7.44 Å, respectively), but not the central C(·)—X—C(·) angle (116.6° vs 103.6°). Dimethylsilylene (**4a**) and difluorosilylene (**5a**) as linkers provide the best geometrical correspondence with **3a**, but the methyls of **4a** are considerably larger than the sulfur lone pairs of **3a**, and the Si-F bonds of **5a** may prove unacceptable because of hydrolytic lability[44,45]. Tetravalent silicon (SiR_4) has been discussed as an isostere for tetravalent carbon (CR_4)[46,47]. Fluoroethene-1,2-diyl (compare **19a**) has been investigated in the alternative Z geometry as an isostere for peptide carboxamide [—NHC(O)—][48–52]. 1,1-Cyclopropylidene (see **7a**) has been incorporated into bio-active compounds, although not explicitly as a thioether mimic[53–55]. Trans-1,2-cyclopropylidene (see **25a**) has been suggested as a replacement for —CH_2SCH_2—[56].

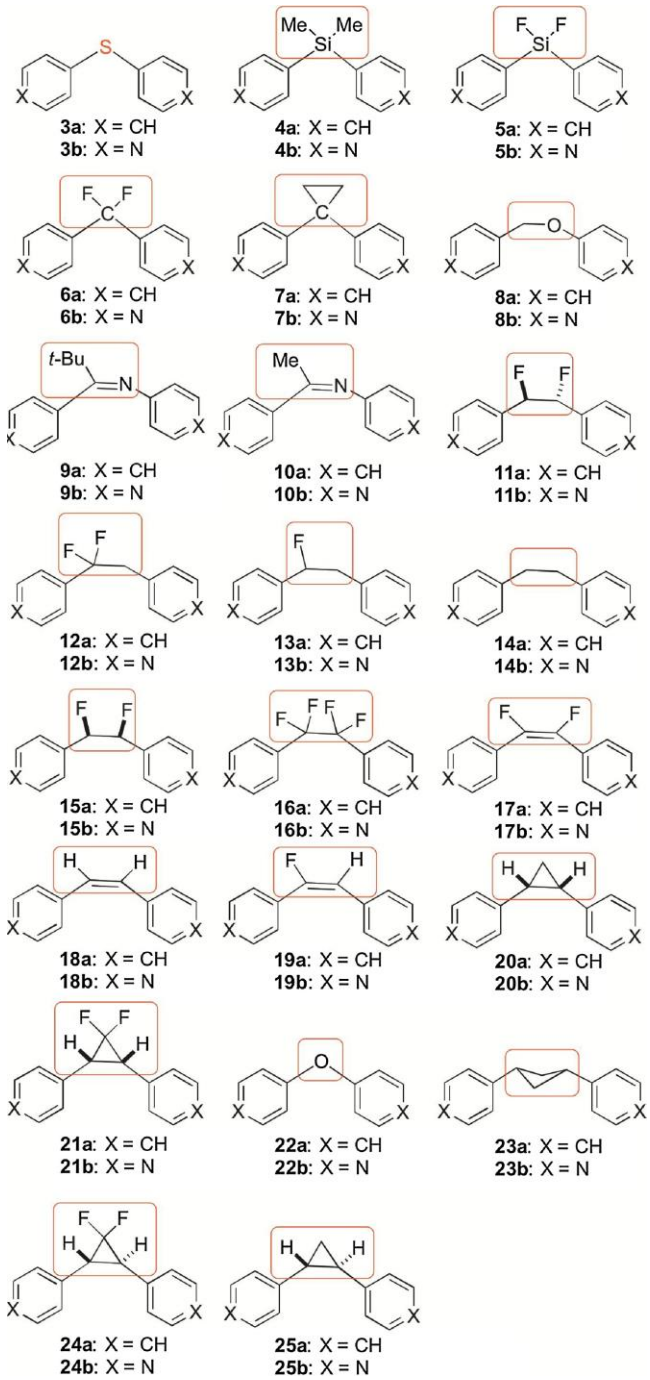


Figure 5-3: Proposed isosteres of diphenyl thioether (3a). The isostere bridging fragment intended to mimic thioether sulfur (S) is enclosed in a red box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5-1: Calculated structural parameters for diphenyl thioether (3a) and proposed isosteres 4a–25a.

		Angles		Dihedral Angle	Distances			E (distort'n)
Compd #	Isostere Fragment Name	C___C to C___C (degrees	Ar to Ar		C(ipso) dist.	C(ipso)–mid-pt	C(para) dist.	
		(degrees)	(degrees)	(avg., deg.)	(Å)	(Å)	Å)	(kcal/mol)
3a	thia-diyl	103.6	111.1	46.5	2.82	1.79	7.44	0.0
4a	dimethylsilylene	108.9	110.7	42.7	3.08	1.89	7.74	4.2
5a	difluorosilylene	114.8	116.8	39.0	3.13	1.85	7.92	5.4
6a	difluoromethylene	114.8	112.8	75.0	2.56	1.52	7.22	6.8
7a	1,1-cyclopropylidene	116.6	115.7	44.7	2.57	1.51	7.35	8.0
8a	oxymethylene	170.0	166.7	19.0	3.68	1.84	9.15	10.3
9a	(*Z*)-2,2-dimethylpropanimine-N,1-diyl	67.6	71.1	69.1	2.90	1.89	6.17	10.7
10a	(Z)-ethanimine-N,1-diyl	71.8	78.0	51.1	2.99	1.90	6.52	14.2
11a	(1*R*,2*R*)-1,2-	73.6	73.7	76.1	3.10	1.97	6.35	15.4

	difluoroethane-1,2-diyl							
12a	1,1-difluoroethane-1,2-diyl	79.4	78.9	81.2	3.18	1.97	6.62	15.7
13a	1-fluoroethane-1,2-diyl	174.9	173.7	77.5	3.90	1.95	9.34	16.3
14a	ethane-1,2-diyl	180.0	180.0	89.1	3.89	1.95	9.35	16.3
15a	(1*R*,2*S*)-1,2-difluoroethane-1,2-diyl	180.0	180.0	79.5	3.90	1.95	9.34	16.4
16a	1,1,2,2-tetrafluoroethane-1,2-diyl	180.0	180.0	89.1	3.93	1.96	9.34	17.7
17a	(*Z*)-1,2-difluoroethene-1,2-diyl	82.4	87.3	34.7	3.28	1.98	7.16	17.7
18a	(*Z*)-ethene-1,2-diyl	82.7	88.9	0.0	3.30	1.99	7.25	18.1
19a	(*E*)-1-fluoroethene-1,2-diyl	83.6	88.7	33.3	3.30	1.98	7.24	18.3
20a	cyclopropane-*cis*-1,2-diyl	68.1	71.9	58.3	3.21	2.03	6.52	21.6

21a	1,1-difluorocyclopropane-*cis*-2,3-diyl	67.9	72.2	57.8	3.25	2.04	6.57	22.6
22a	oxa-diyl	120.9	129.9	39.9	2.40	1.38	7.46	30.1
23a	(diequatorial) cyclobutane-*cis*-1,3-diyl	116.3	119.0	0.0	4.73	2.50	9.59	41.7
24a	cyclopropane-*trans*-1,2-diyl	142.6	140.3	42.7	3.89	2.00	9.08	47.8
25a	1,1-difluorocyclopropane-*trans*-2,3-diyl	138.0	133.6	44.3	3.91	2.03	8.98	48.0

In order to move the isostere comparison closer to a medicinal chemistry context, a separate set of calculations was run on bis(pyridin-4-yl) thioether **3b** and its analogues **4b**–**25b** (**Figure 5-4** and

Table 5-2). Here, pyridin-4-yl represents a generic attached heterocycle that can facilitate discussion of geometry and electrostatics.

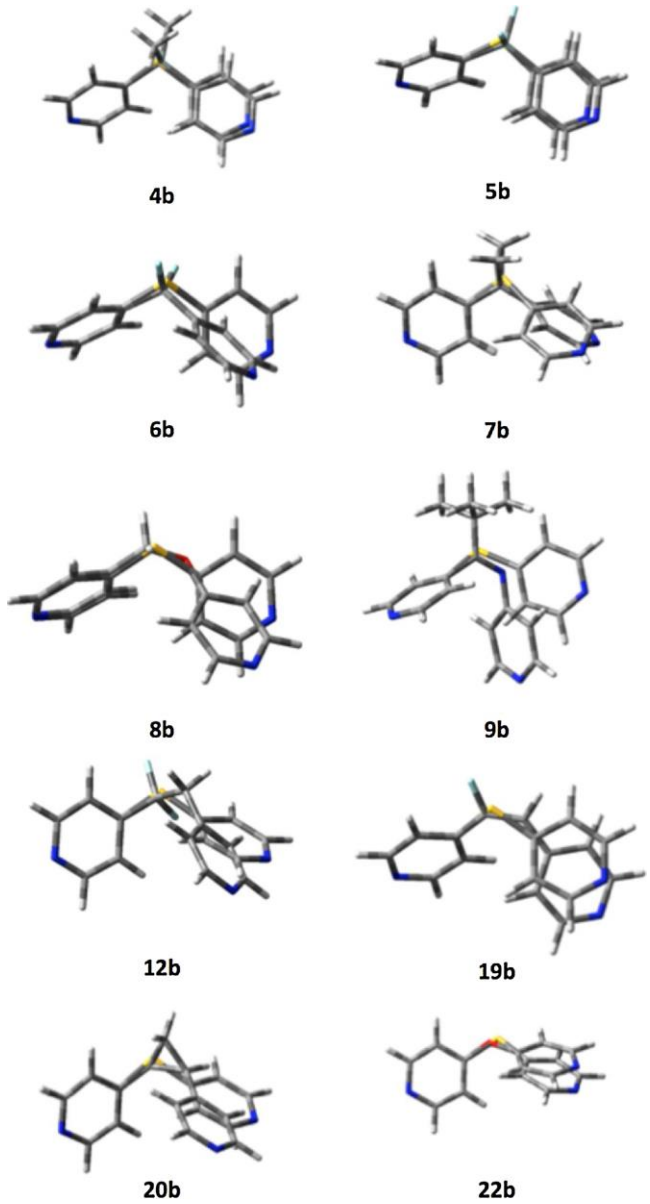


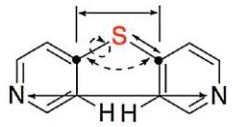
Figure 5-4: One-ring overlap models of selected isosteres (labeled), each with bis(pyrid-4-yl) thioether (3b). For each overlapped pair (thioether and isostere), the matched pyridine rings are on the left, and the comparison rings are on the right.

Table 5-2: Calculated intramolecular N–N' distances for bis(pyridin-4-yl) thioether (3b) and proposed isosteres 4b–25b, and intermolecular N'–N' distances in overlap of each of 4b–25b with 3b.

Compd #	Isostere Fragment Name	Isostere Structure	N-N' dist. (Å)	N'-N' dist. in overlap with Pyr-S-Pyr (Å)
3b	thia-diyl	Pyr–S–Pyr	7.52	0.0
4b	dimethylsilylene	Pyr–Si(Me2)–Pyr	7.73	0.3
5b	difluorosilylene	Pyr–Si(F2)–Pyr	7.92	0.4
6b	difluoromethylene	Pyr–CF2–Pyr	7.21	2.9
7b	1,1-cyclopropylidene	Pyr–C(–CH2CH2–)–Pyr	7.37	2.5
8b	oxymethylene	Pyr–OCH2–Pyr	9.18	1.2
9b	(Z)-2,2-dimethylpropanimine-N,1-diyl	Pyr–N = C(CMe3)–Pyr	6.11	4.3
10b	(Z)-ethanimine-N,1-diyl	Pyr–N = C(Me)–Pyr	6.50	2.7
11b	(1R,2R)-1,2-difluoroethane-1,2-diyl	(R,R) Pyr–CHFCHF–Pyr	6.31	3.6
12b	1,1-difluoroethane-1,2-diyl	Pyr–CF2CH2–Pyr	6.62	2.6
13b	1-fluoroethane-1,2-diyl	Pyr–CHFCH2–Pyr	9.36	4.0
14b	ethane-1,2-diyl	Pyr-CH2CH2–Pyr	9.36	4.5

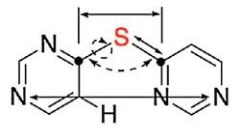
15b	(1R,2S)-1,2-difluoroethane-1,2-diyl	(R,S) Pyr–CHFCHF–Pyr	9.36	4.3
16b	1,1,2,2-tetrafluoroethane-1,2-diyl	Pyr–CF2CF2–Pyr	9.35	4.5
17b	(Z)-1,2-difluoroethene-1,2-diyl	(Z) Pyr–C(F) = C(F)–Pyr	7.21	1.6
18b	(Z)-ethene-1,2-diyl	(Z) Pyr–CH = CH–Pyr	7.21	1.6
19b	(E)-1-fluoroethene-1,2-diyl	(E) Pyr-CF = CH–Pyr	7.24	1.5
20b	cyclopropane-cis-1,2-diyl	(cis) Pyr–CH(–CH2–)CH–Pyr	6.49	2.9
21b	1,1-difluorocyclopropane-cis-2,3-diyl	(cis) Pyr–CH(–CF2–)CH–Pyr	6.54	2.8
22b	oxa-diyl	Pyr–O–Pyr	7.51	1.3
23b	(diequatorial) cyclobutane-cis-1,3-diyl	(cis) Pyr–CH(–CH2–)(–CH2–)CH–Pyr	9.57	4.5
24b	cyclopropane-trans-1,2-diyl	(trans) Pyr–CH(–CH2–)CH–Pyr	9.09	3.5
25b	1,1-difluorocyclopropane-trans-2,3-diyl	(trans) Pyr–CH(–CF2–)CH–Pyr	9.00	3.4

Furthermore, bis(pyridinyl) system **3b** was compared with the bis(pyrimidin-4-yl) thioether, **3c**. The key structural parameters for the most stable conformations for **3b** and **3c** are shown in **Figure 5-5**.



3b

104.3° C(•)–S–C(•): bond angle
1.79 Å C(•)–S: bond length
40.4° CC(•)SC(•): torsional angle
2.83 Å C(•)–C(•): ipso–ipso distance
7.52 Å N(para)–N(para): distance



3c

110.6° C(•)–S–C(•): bond angle
1.79 Å C(•)–S: bond length
0.0° CC(•)SC(•): torsional angle
2.94 Å C(•)–C(•): ipso–ipso distance
7.80 Å N(para)–N(para): distance

Figure 5-5: Comparison of the calculated structural parameters of bis(pyridin-4-yl) thioether (3b, left) and bis(pyrimidin-4-yl) thioether (3c, right).

Bis(pyridin-4-yl) thioether **3b** shows a slight widening of the central C—S—C angle compared with the corresponding diphenyl system (see **1**, **Figure 5-2**). The para ring atoms of **3b** are also moved slightly farther apart, and the inter-ring torsional angle is reduced from 46.5° to 40.4°. Introduction of the second nitrogen into the heteroaromatic rings, however, leads to a much more dramatic structural effect (**Figure 5-5**). An inter-ring steric interaction between the ortho-H's of **3b** is likely relieved by replacing them with N, as in **3c**. The central angle of **3c** is expanded to 110.6° from 104.3°, and the two heterocyclic rings of **3c** are completely coplanar. This suggests a modified hybridization at sulfur, and possible conjugation of a sulfur p lone pair with the pyrimidinyl rings of **3c**.

Two parameters are shown in

Table 5-2: the N—N' distance within the bis(pyridin-2-yl) isostere (compared to 7.52 Å for **3b**), and the intermolecular N'—N' distance in a hypothetical overlap of the isostere with **3b**. In the latter comparison, one pyridine ring of **3b** is made to exactly overlap with one ring of the isostere. The distal N' of **3b** can then be compared with the distal N' of the isostere, with their through-space separation given in angstroms. This

comparison might approximate a situation where both isostere N's are serving as H-bond acceptors. If one isostere H···N bond matches that of the thioether exactly, the isostere N' will be displaced to some extent from the thioether N', and that difference may make for a worse (or better) H···N' bond on the other end. Several of the overlapped structures are shown in **Figure 5-4**. While many of the isosteres perform poorly in this comparison, the dimethylsilylene and difluorosilylene isosteres (**4b** and **5b**) show excellent three-dimensional overlap. Qualitative next-best overlap is exhibited by the *E*-fluoroethene-1,2-diyl (**19b**), 1,1-cyclopropylidene (**7b**), and ether (**22b**) isosteres.

5.4.2 Electrostatic potentials for selected isosteres of bis(pyrid-4-yl) thioether

Determination of electrostatic potential surfaces for the various thioethers and their isosteres was undertaken in order to gain a qualitative assessment of the electrostatic environment near the bridging group and in the aryls. **Figure 5-6** shows the ESP diagrams for the key thioethers **3a**, **3b**, and **3c**, as well as the selected bis(pyridyl) isosteres **4b** (dimethylsilylene), **7b** (1,1-cyclopropylidene), and **19b** [(*E*)-1-fluoroethene-1,2-diyl].

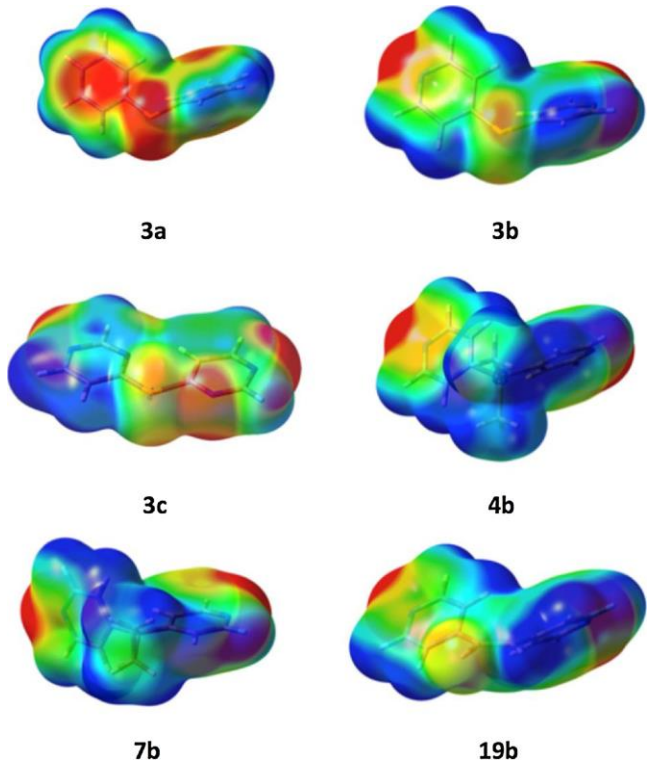


Figure 5-6: Chart4: Electrostatic potential diagrams for diphenyl thioether (3a), bis(pyrid-4-yl) thioether (3b), bis(pyrimid-4-yl) thioether (3c), and three isosteres.

Replacing the para carbons of **3a** with nitrogens, as in **3b**, has a dramatic effect in lowering the negative electrostatic potential at the sulfur atom. Replacing a second ring carbon with nitrogen [the bis(pyrimidyl) thioether, **3c**] has less of an additional electrostatic effect at sulfur, despite the change in geometry and apparent re-hybridization at S. These changes, discussed in **Section 5.4.1**, are therefore likely more attributable to removal of the ortho-H in **3c** (vs **3b**), rather than additional conjugation or electron withdrawal imposed by the second ring nitrogen. However, simply replacing phenyl with pyrimid-4-yl (**3a** vs **3b**) may well result in less susceptibility of the sulfur atom to metabolic oxidation. Steric and other effects involving the ortho position of the aromatic rings of isosteres are further discussed in Section 2.3.

The ESP determinations for the various bis(pyridyl) isosteres are exemplified by **4b**, **7b**, and **19b** (**Figure 5-6**). In each case, the apparent qualitative negative electrostatic potential at the bridging group is reduced compared with the thioether **3b**, each relative to the pyridine nitrogen. The expectation is that these isosteres would be less susceptible to metabolic oxidation on the bridging group than **3b**, which after all possesses non-bonded sulfur lone pairs. The pyridyl nitrogens of isosteres remain sites of potential metabolic oxidation[57].

5.4.3 Isosteres of Axitinib and their computed interactions with crystallographically characterized Axitinib binding sites in VEGFR and ABL1 proteins

We selected, based on their structural and electronic features, four isostere fragments (1,1-cyclopropylidene (**CP-Ax**), (E)-fluoroethene-1,2-diyl (**FE-Ax**), difluorosilylene (**DFS-Ax**), and dimethylsilylene (**DMS-Ax**)) for computational evaluation in known protein binding sites. From among the aromatic thioethers shown in **Figure 5-1**, we chose Axitinib (**Figure 5-1:A**), the FDA approved drug for treatment of renal cell carcinoma, as a good candidate for isostere evaluation, since the crystal structures of **A** in two different protein binding sites are available from the PDB (**Figure 5-8**)[58]. Additionally, extensive in vitro hit-to-lead information on this class of compounds is available, including the binding data for commonly used analogues wherein —O—, —CH_2—, —C(=O)—, and —NH— fragments replace the thioether sulfur of **A** and various related compounds. While the choice of **A** as a model for evaluating isosteres is arbitrary, many of the anticipated issues that might complicate protein binding do, in fact, arise here and can be evaluated in this context. The four proposed isosteres of **A** are shown in **Figure 5-7**.

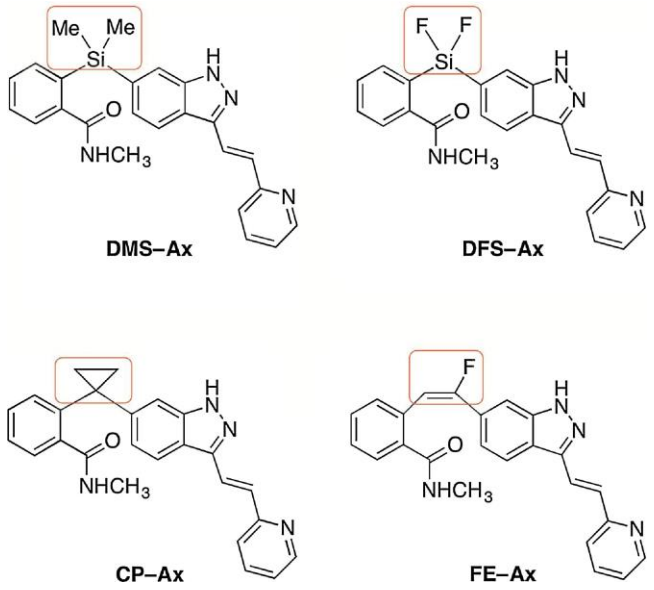


Figure 5-7: Proposed Axitinib isosteres. The isosteric fragment that replaces S is shown in the red box. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

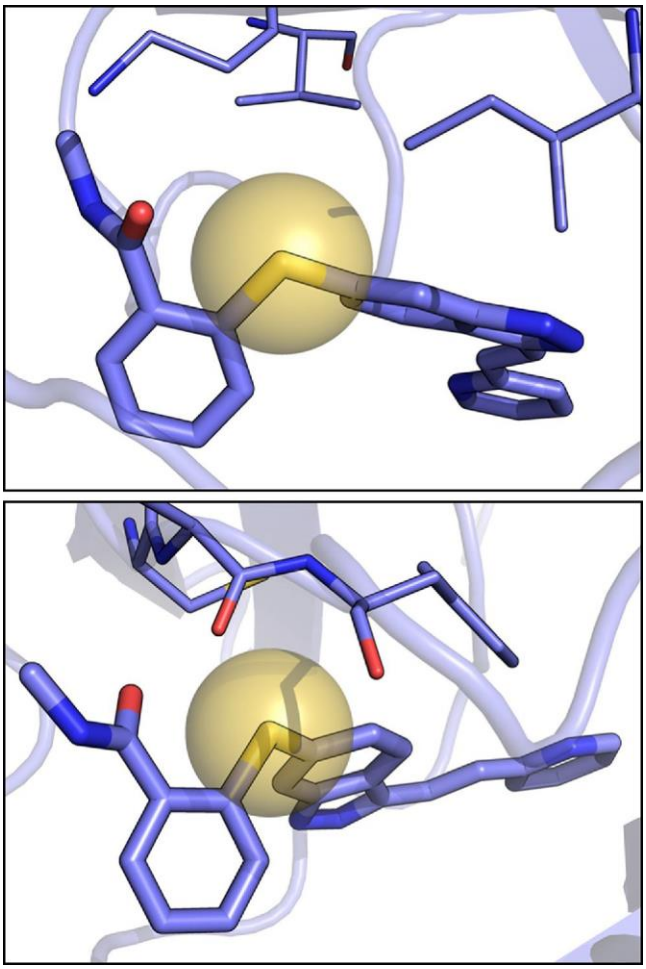


Figure 5-8: Axitinib shown in a PDB rendering in the respective binding sites of ABL1 (top) and VEGFR (bottom). The translucent gold sphere represents the thioether sulfur atom at its default van der Waals radius, 1.8 Å.

We used a RosettaDock[59–63] based method to investigate the goodness of fit of these four isosteres in two kinase active sites, Abl kinase mutant T315I (PDB code: 4TWP7), and VEGFR2 kinase (PDB code: 4AG8)[64]. In each case the isostere model was superimposed onto the bound conformation of **A** based on the pyridine and azaindole rings, and these were held fixed in space in the subsequent modeling. Rotations around other bonds in each isostere were performed to evaluate energetically favorable conformations in the context of each kinase active site. Energy minimizations and evaluations were performed with the Rosetta software as described in the Methods section

(page 230). Protein backbone and sidechain conformations were held fixed in these calculations in order to evaluate the ability of each isostere to bind the same kinase conformation as **A** in the crystal structures. In the binding of small molecules to protein active sites, a key parameter to consider is the energy difference between the global minimum-energy conformation of the isostere and the bound conformation. This energy difference – the reorganization penalty – is an indicator of the free energy cost paid upon binding due to adoption of a higher-energy conformation by the isostere. To evaluate the reorganization penalty – expressed as the energy difference between the bound conformation and the global minimum-energy conformation – we performed energy evaluations of the isosteres by using the Amber molecular mechanics software[65] in the bound and global minimum conformations as identified by quantum mechanics calculations.

Both types of calculations described above suggest that for each kinase active site, at least one isostere is able to provide a reasonable alternative to **A** (**Table 5-3**). For the Abl kinase active site, which features the larger binding pocket, all isosteres adopt conformations in which the benzamide group is placed in the vicinity of its position in the original bound structure of **A**. The key binding interactions between the benzamide group of **A** and the protein, i.e., hydrogen bonds between (1) the backbone carbonyl group of Tyr253 and the small molecule amide nitrogen, and (2) sidechain amino group of Lys271 and the small molecule carbonyl group, are perturbed to different extents. Similarly, the differing steric requirements introduced in the isostere at the —S— position lead to perturbations of interactions between the small molecules and the protein residues 314–316, including a close contact with the Ile315 sidechain. Based on the calculated interaction energies and hydrogen bond geometries, we find that the isosteres CP-Ax and FE-Ax, followed by DFS-Ax and DMS-Ax, provide the closest mimicry of the

bound **A** conformation. The unusually close approach (3.0 Å) of the carbonyl group from the benzamide portion of **A** with the sulfur atom suggests a possible sulfur hole effect. This effect, the weak attraction of a heteroatom lone pair to a C—S sigma star orbital, is often seen in sulfur containing protein-ligand and small molecule crystal structures[66]. It is most common for sulfur containing heterocycles such as thiophene and thiazole, but is observed occasionally in aryl thioethers[67–71]. This attraction is, in any case, not possible with the isosteres, and may place some limitations on the conformational mimicry exhibited by these compounds. Nonetheless, reasonable hydrogen bonds with the protein and steric fits are observed with all four isosteres (**Table 5-3**). The energies of DMS-Ax and DFS-Ax are less favorable compared to CP-Ax and FE-Ax due to increased steric repulsion with hydrophobic groups (e.g., Ile316) in the active site. These minor clashes may result from the greater steric bulk of the group replacing the sulfur atom and may be relieved by allowing the protein backbone and sidechains to move away slightly. Thus, in these simulations, all four isosteres can reasonably mimic **A** in bound conformation and energetics when interacting with the Abl kinase active site.

Table 5-3: Calculated distances and energies for Axitinib and four isosteres in two protein binding sites.

Isostere	Binding site	Nearest distance O···group (Å)	H-bond distance O···donor (Å)	H-bond distance HN·· acceptor (Å)	Rosetta energy (R. E. U.)	AMBER reorg'n (kcal/mol)	AMBER RMSD (Å)
Axitinib	ABL1	3.0	2.7	2.9	−10.0	1.418	0.639
DMS-Ax	ABL1	2.8	3.5	3.2	−6.8	0.000144	0.731
DFS-Ax	ABL1	2.4	2.8	2.5	−7.8	6.7430	0.825
CP-Ax	ABL1	3.1	2.7	2.7	−10.8	0.0944	2.078
FE-Ax	ABL1	2.9	3.2	3.0	−10.2	0.000236	0.486

Axitinib	VEGFR	3.1	3.0	2.7	−12.0	0.139	0.716
DMS-Ax	VEGFR	3.2	2.2	2.9	51.8	−0.00301	0.709
DFS-Ax	VEGFR	2.1	3.0	2.9	−9.1	1.555	1.507
CP-Ax	VEGFR	2.5	3.4	2.8	−5.0	3.440	1.621
FE-Ax	VEGFR	4.1	NA	NA	NA	0.271, −0.456	0.546, 0.695

Compared with the Abl kinase binding site, the VEGFR kinase site features a more restrictive binding pocket. In the bound structure of **A**, no sulfur hole effect involving the carboxamide C=O is apparent, since the alignment of the carbonyl O with the relevant C—S sigma star orbital is poor (the C—S—O angle is 134°). The sulfur atom from **A** is tightly packed against backbone atoms of residues Asp1046 and Phe1047, as well as the sidechain of Cys1045. A small change in shape in this region leads to significant clashes, which can be resolved to varying extents by minimization of the isostere torsion angles (**Table 5-3**). The DFS-Ax isostere can adopt configurations in which the hydrogen bonding interactions between the carbonyl oxygen and amide groups of the benzamide moiety with backbone amide of Asp1046 and Glu885, respectively, can be preserved; the small size of the fluorine atoms allows the difluorosilylene moiety to be accommodated in the place of the sulfur of **A** without any significant clashes or structural rearrangements. On the other hand, DMS-Ax, with its larger methyl groups, makes several steric clashes with the protein backbone. The altered shape of FE-Ax imposed by the two-atom bridge places the benzamide ring in a position such that it cannot make hydrogen bond interactions with the pocket. Finally, the CP-Ax isostere can be accommodated, but the altered geometry at the cyclopropyl center compared with the sulfur atom, and the steric restrictions of the pocket, make its fit less favorable compared to that of DFS-Ax.

While our initial examinations did not allow the kinase conformation to change, we extended our investigation to see whether thermal movements of the active site pocket and loops might allow better fits of the modeled isosteres. Thus, unrestrained implicit-solvent molecular dynamics simulations of each isostere for both the Abl kinase and VEGFR binding pockets were performed. These simulations reveal varying degrees of accommodating flexibility in nearby active site residues and loop regions that, in some cases, can stabilize the Rosetta-docked conformations of the isosteres. Ligand **A**, as modeled in the active sites of Abl kinase and VEGFR, was simulated under implicit solvent conditions for over 100 nanoseconds, while CP-Ax, FE-Ax, DFS-Ax, and DMS-Ax were simulated inside each binding pocket for 50 nanoseconds. Details of the simulation settings can be found in the Methods section (page 230).

In the Abl kinase binding pocket, **A** is stabilized mainly by hydrogen bond interactions with residues K39 and E79 (**Figure 5-9:**A), and a low internal energy (**Figure 5-11:A**). Accommodated by the less restrictive Abl binding pocket, **A** also undergoes a conformational switch that is explored briefly by the molecule and is then later reversed to the crystal structure form for the remainder of the trajectory (**Figure 5-9:B**). Similarly, isosteres CP-Ax and FE-Ax were shown to utilize hydrogen bond interactions with K39 and E79 as well as the backbone and sidechain of Y21 (**Figure 5-9:A**). The initial Rosetta-docked CP-Ax, which had a high RMSD after implicit solvent minimization (**Table 5-3**), underwent a conformational change during the initial thermalization steps (**Figure 5-9:B**), and could remain bound in the active site. The flexibility of the surrounding protein allows for hydrogen bonds similar to those of Axitinib (**Figure 5-9:C**). The binding conformation of FE-Ax, on the other hand, does not undergo a large conformational change during implicit solvent minimization (**Table 5-3**) or in the context of the Abl binding pocket (**Figure 5-9:B and C**), and finds more hydrogen bonding partners to the protein than does **A** (**Figure**

5-9:A). Adding flexibility to the protein in simulations also led to improvement in the ranking of DMS-Ax in the Abl binding pocket. The Rosetta-docked molecule can remain bound to the protein throughout the trajectory due to Axitinib-like hydrogen bond energies between DMS-Ax and protein residues K39 and E79 (**Figure 5-11:D,E**). This isostere, however, has higher internal energies than CP-Ax and FE-Ax, making the relative ranking FE-Ax, CP-Ax > DMS-Ax. DFS-Ax, which does not undergo a large conformational change during implicit solvent minimization (**Table 5-3**) or during molecular dynamics simulations in the protein active site (**Figure 5-11:D**), is not a good Axitinib isostere candidate, according to the Amber metrics reflecting its high internal energy and inability to remain in the active site. The molecule eventually undergoes a rotation about the Si-phenyl bond to a less sterically congested conformation, which causes all hydrogen bonds to break and the molecule to be released.

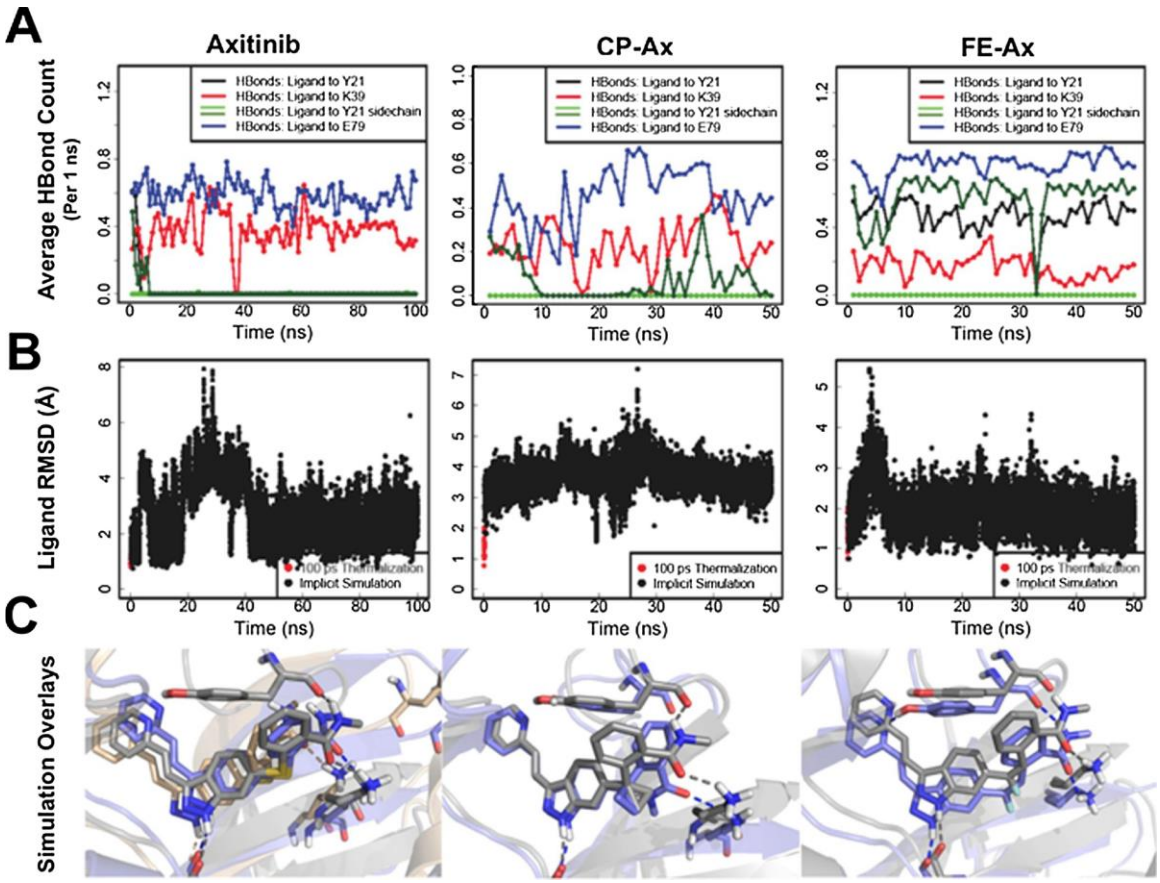


Figure 5-9: Implicit molecular dynamics simulations of Abl kinase with Axitinib (first column), CP-Ax (middle column), and FE-Ax (third column).

Due to the more restricted nature of the VEGFR2 kinase binding pocket, Amber trajectory analyses of the four isosteres result in a reverse-ordered ranking compared to the Rosetta-docked conformations. In this binding pocket, **A** is stabilized mainly through hydrogen bond interactions to E102 and K53 (**Figure 5-10:A**), and exhibits a stable binding conformation (**Figure 5-10:B,C**) with a low internal energy (**Figure 5-11:F**). This is mostly likely due to the steric buttressing against protein sidechains that would affect the probability of other conformations of **A**. In comparing the trajectory of **A** with that of the four isosteres, one finds that FE-Ax and DMS-Ax remain stable in the binding pocket due to similar hydrogen bond interactions with K53 and E102 (**Figure 5-10:A**). Furthermore, both binding conformations of FE-Ax and DMS-Ax have low implicit solvent minimization

RMSDs, as well as low reorganization penalties (**Table 5-3**), indicating a higher probability that the binding conformation can be adopted. Molecular dynamics simulations also suggest stable conformations for binding (**Figure 5-10:B,C**). CP-Ax and DFS-Ax, which were ranked highly by Rosetta energies and docked H-bond distances, have unstable trajectories in which both molecules have high implicit solvent minimization RMSDs and high reorganization penalties (**Table 5-3**), indicating that the binding conformations of these molecules may be relatively improbable. Furthermore, CP-Ax and DFS-Ax are unable to make Axitinib-like hydrogen bonds (**Figure 5-11:G,I**), leading to release of the molecule in the case of CP-Ax, and frequent conformational switching in the case of DFS-Ax. Thus, the predicted order of post-binding stability in the VEGFR2 kinase pocket is FE-Ax > DMS-Ax >> CP-Ax > DFS-Ax.

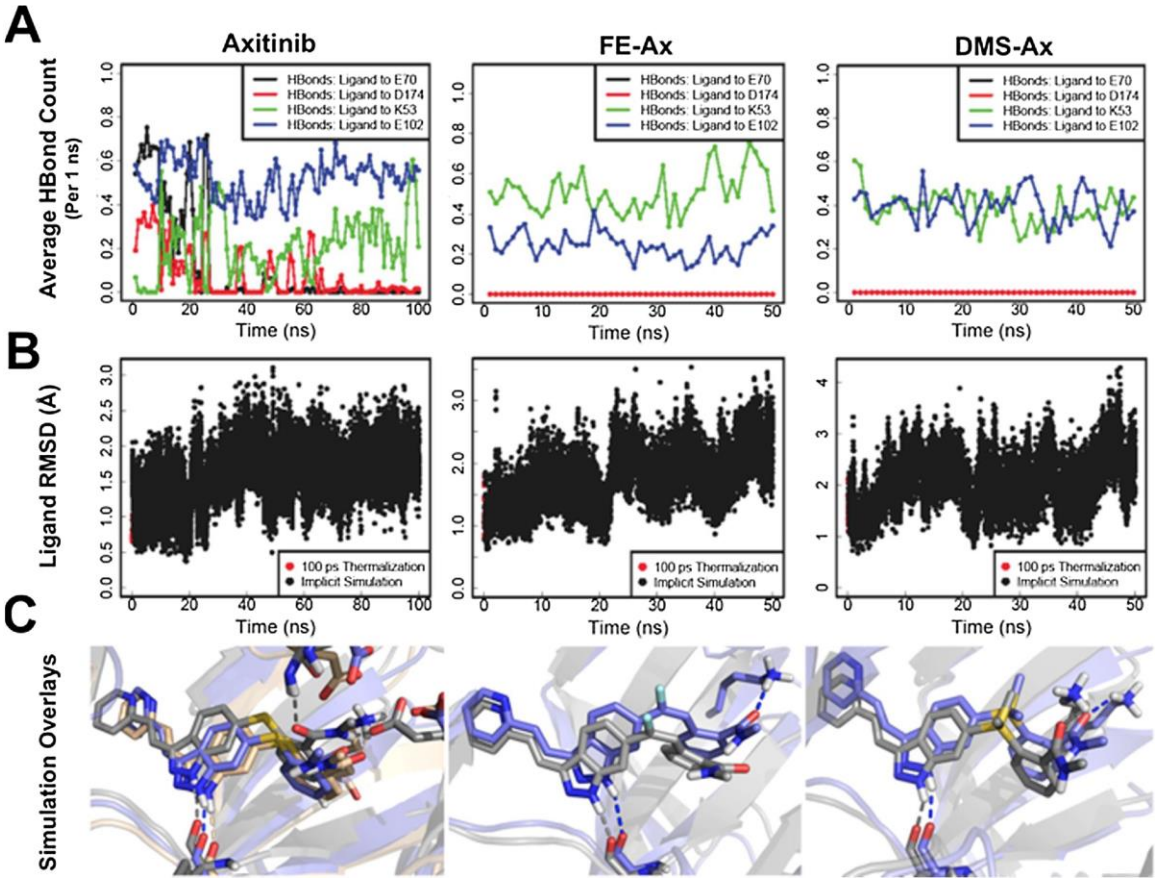
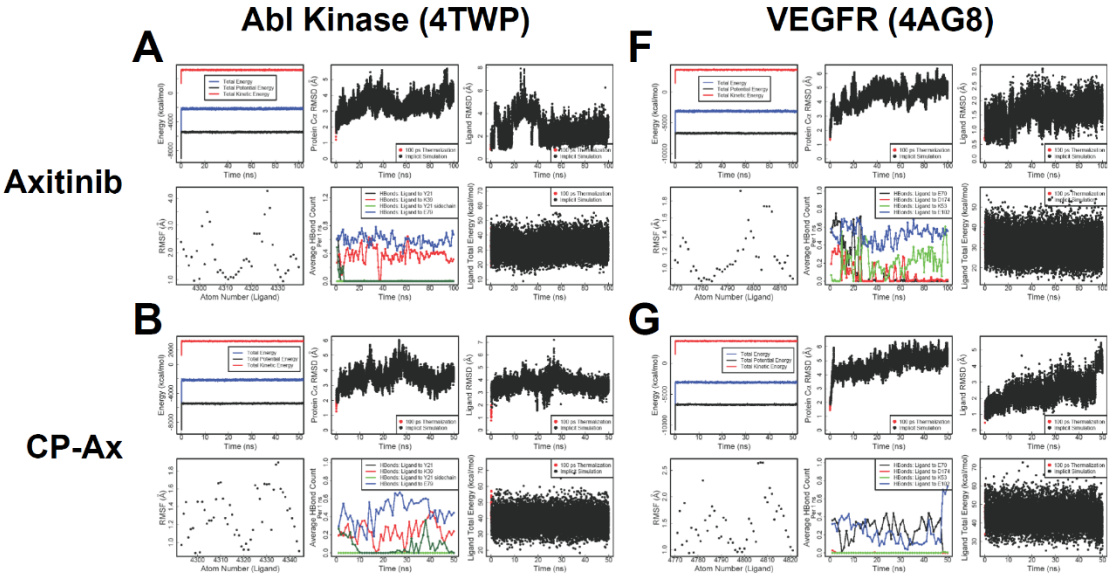


Figure 5-10: Implicit molecular dynamics simulations of the VEGFR2 kinase with Axitinib (first column), FE-Ax (middle column), and DMS-Ax (third column).



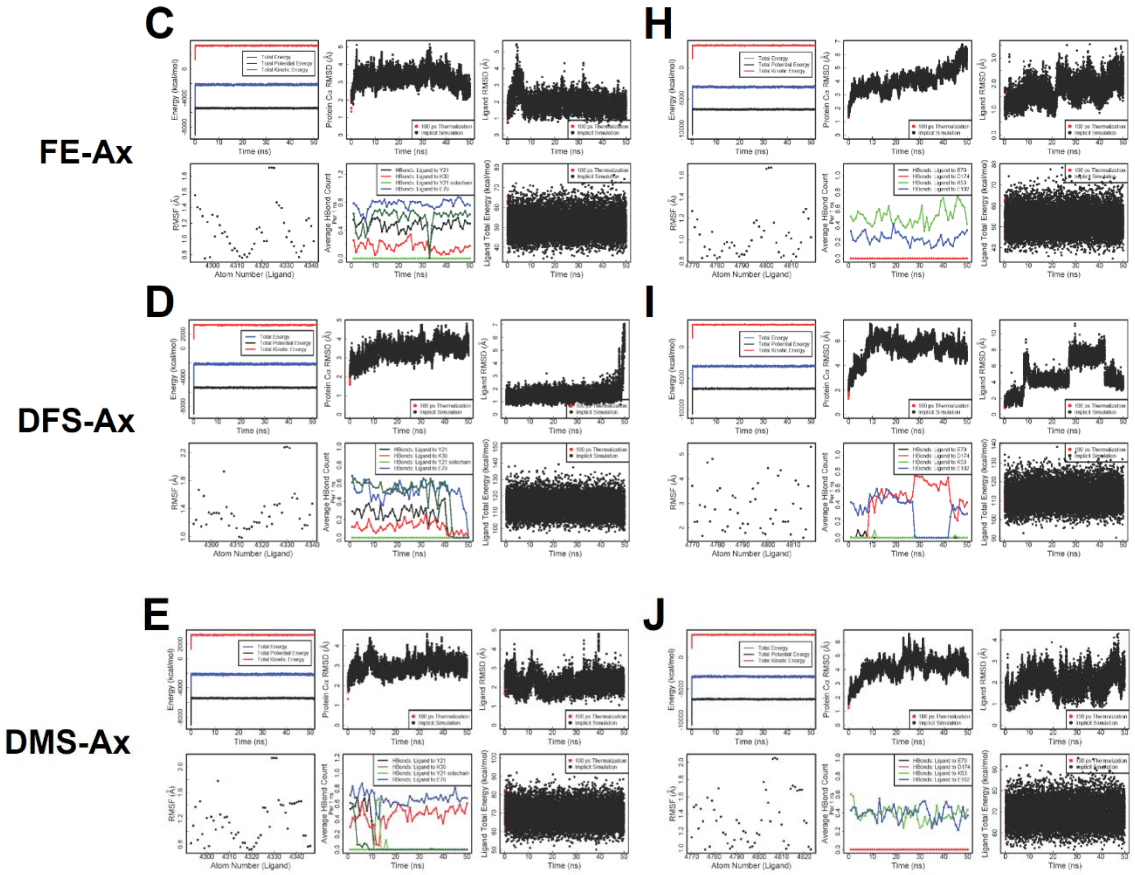


Figure 5-11: Molecular dynamics simulation analyses of CP-Ax, FE-Ax, DFS-Ax, and DMS-Ax in the Abl Kinase (left side) and VEGFR (right side) binding pockets. Each isostere in each binding pocket has six graphs associated with it in the following order: (Top row, left) Energies (kcal/mol) *versus* Time (ns) for Total Energy (blue line), Total Potential Energy (black line), and Total Kinetic Energy (red line). (Top row, middle) Whole-protein Cα RMSD (Å) *versus* Time (ns). (Top row, right) All-atom ligand RMSD (Å) *versus* Time (ns). (Bottom row, left) All-atom ligand RMSF (Å). (Bottom row, middle) Average hydrogen bond count (averaged over every 1 ns block) *versus* Time (ns). (Bottom row, right) Ligand total energy (kcal/mol) *versus* Time (ns).

5.5 Conclusions

A number of potential isosteric bridging groups replacing the sulfur atom of diaryl and heteroaryl thioethers have been evaluated according to their geometric (**Table 5-1**, **Table 5-2**) and electrostatic (**Figure 5-6**) features. While the more obvious choices, such as —O— and —CF_2—, might suffice in some situations, other, more unorthodox,

isosteres show promise for possible application where a closer geometric match or other feature is desired. Among these are 1,1-cyclopropylidene (**7**), (*E*)-fluoroethene-1,2-diyl (**19**), dimethylsilylene (**4**), and difluorosilylene (**5**). The substitution of one or more nitrogen atoms for carbons in the aromatic rings can also alter both the electronics and geometry at the sulfur atom (**Figure 5-6**).

Conformational exploration of four proposed Axitinib isosteres (**Figure 5-7**) in the active sites of two kinase targets indicates that the larger size of the proposed bridging groups compared to the thioether sulfur atom, and the unusually close approach of these groups to the *ortho*-carboxamide group, are features can that lead to suboptimal fits in the static crystallographic kinase structures. In some cases, a lower number of hydrogen bonds between the isostere and the protein active site results. Nevertheless, reasonable binding fits can be identified (**Table 5-3**), particularly for the ABL1 site (**Figure 5-8**). The flexibility of the kinase active sites can compensate for the increased steric demand, and full hydrogen bonding between the isostere and protein can sometimes be restored according to molecular dynamics trajectories. Thus, in addition to the overall geometric differences between diaryl thioether and isostere, the size of the bridging group and its interactions with protein residues in the binding pocket, and with nearby, especially *ortho*-aryl, isostere substituents, are important features to consider when designing and evaluating candidate isosteres.

5.6 Methods

5.6.1 QM calculations

The Gaussian 09 package was used to carry out all electronic structure calculations[72]. Geometries were optimized by using the 6-31G(d) basis set and the

B3LYP hybrid density functional[73]. Structures were optimized without geometric constraint by using tight convergence criteria, and were verified as minima through subsequent calculation of vibrational frequencies.

Conformational searching was conducted manually, and the lowest energy conformation was selected in cases where more than one possibility existed.
The electrostatic potential plots were generated by using GaussView, with the electron density contour set to 0.001 and the color scale set to span from $-2.0 \times 10-2$ to $+2.0 \times 10-2$, as illustrated in **Figure 5-12**.

Figure 5-12: Electron density from total SCF density (isoval = 0.0001).

The distortion energies were computed by forcing the C(ipso)–C' (ipso), C(para)–C' (para), C(ipso)–C' (para), C(para)–C' (ipso) distances to be the same as in diphenylthioether, while allowing all other geometric paramaters to relax. The difference in energy between this constrained structure, and the completely unrestrained structure, is the distortion energy.

The "N'–N' distance in overlap (

Table 5-2) was obtained by manually overlaying one pyridyl ring of the candidate structure as closely as possible with one of the pyridyl rings of 4,4′-dipyridylthioether **3b**, and then observing the distance between the "other" pyridyl nitrogen atoms. For unsymmetrical structures, each pyridyl ring was independently overlaid with a pyridyl ring of **3b**, and the lower of the two distances thereby obtained is the one reported in

Table 5-2.

5.6.2 Rosetta calculations

The Ax-bound PDB structures (codes 4TWP and 4AG8)[7,64] were used as starting points for energy calculations with Ax- and isosteres. Rosetta FastRelax[74–77] algorithm was used with coordinate constraints to generate the initial models, and ligand residue energy for Ax in the crystallographically observed conformation was calculated with the talaris2014 energy function[78,79] and was used for comparison with the assessed isosteres. For generating models with isosteres bound in the active sites, QM-generated lowest energy conformations of the isosteres were used. The azaindole and pyridinium groups, which are common between Ax and the isosteres, were maintained in the geometry generated for Ax-bound kinases upon FastRelax. The positions of other atoms in the ligand, including the thioether sulfur and the substituted aryl ring were obtained from QM-generated geometries. Rosetta FastRelax simulations were then performed while allowing free rotation around the isostere dihedrals, and free rigid-body movement of the isostere relative to the protein. The protein sidechains were either kept fixed in their crystallographic conformations or allowed to repack in the Rosetta simulations. The resulting conformations were visually examined for hydrogen bond formation, and energies of the isosteres were calculated by using the talaris2014 energy function.

5.6.3 Amber calculations

The Rosetta-docked configurations of CP-Ax, FE-Ax, DFS-Ax, and DMS-Ax were prepared for implicit-solvent minimization by using the antechamber and LEaP modules[65] in Amber16 with the General Amber Force Field, version 2 (GAFF2)[80].

Silicon parameters were approximated by substituting the sp^3 carbon atom type 'c3' in place of silicon where possible from either the GAFF[80] or GAFF2 set of parameters. The equilibrium bond distances for aromatic, sp^2 carbon ('ca')-to-silicon ('Si') and Si-to-fluorine ('f') were set as the initial bond lengths present in the original DFS or DMS configurations, while the harmonic force constants were set as the ca-c3 and c3-f GAFF values, respectively. Angle equilibrium values and harmonic force constants were both taken from the Si to c3 GAFF approximations. Dihedral values were approximated by using the Si to c3 GAFF2 approximations. Nonbonded terms were obtained from the DOCK 6.5 software suite[81–83].

Minimizations were carried out for a maximum of 1000 steps under the Limited-memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm[84] with a convergence criterion of 0.01 kcal/mol-Å. Solvent effects were treated with a generalized Born implicit solvent model (GB-Neck2, mbondi3 radii)[85] implemented in the Amber16 package with a cutoff value of 999 Å for nonbonded interactions.

Total potential energies of the minimized structures were obtained by using the pytraj 2.0.0 interactive molecular dynamics simulation data analysis Python package[86], which is a Python wrapper for cpptraj[87].

MD simulations were carried out on the Rosetta-docked configurations of CP-Ax, FE-Ax, DFS-Ax, and DMS-Ax in either the VEGFR2 or ABL1 binding sites for 50 nanoseconds. Simulations of Rosetta-docked Axitinib in both binding sites were carried out for 100 nanoseconds. Structures were prepared as before, and all simulations used the GPU implementation of the pmemd program in Amber16 with the combination of GB-Neck2 model, mdbondi3 radii, and ff14SBonlySC[88].

Analyses of the implicit solvent molecular dynamics simulations were performed with cpptraj, where the average hydrogen bond count was determined by evaluating the existence of a hydrogen bond between specific atoms at each frame in the trajectory and averaging the count over 1 ns time frames. Protein Cα RMSD and Ligand RMSD were obtained by aligning all protein residue Cα atoms to the original Rosetta-docked conformation followed by the superimposition of all ligand atoms without alignment. Ligand RMSF values were evaluated after fitting the protein Cα atoms to the original Rosetta-docked conformation, obtaining the average protein structure based on Cα atoms only, fitting each frame of the trajectory to the average structure, and calculating the RMSF of all ligand atoms.

5.7 References

1. Brown, N. *Bioisosteres in Medicinal Chemistry*. *Methods and Principles in Medicinal Chemistry* (WILEY-VCH, 2012).

2. Meanwell, N. A. Synopsis of some recent tactical application of bioisosteres in drug design. *J. Med. Chem.* **54,** 2529–2591 (2011).

3. Malik, M. A., Wani, M. Y., Al-Thabaiti, S. A. & Shiekh, R. A. Tetrazoles as carboxylic acid isosteres: Chemistry and biology. *J. Incl. Phenom. Macrocycl. Chem.* **78,** 15–37 (2014).

4. Ostrovskii, V. A., Trifonov, R. E. & Popova, E. A. Medicinal chemistry of tetrazoles. *Russ. Chem. Bull.* **61,** 768–780 (2012).

5. Hu-Lowe, D. D. *et al.* Nonclinical antiangiogenesis and antitumor activities of axitinib (AG-013736), an oral, potent, and selective inhibitor of vascular endothelial growth factor receptor tyrosine kinases 1, 2, 3. *Clin. Cancer Res.* **14,** 7272–7283 (2008).

6. Kania, R. S. *et al.* INDAZOLE COMPOUNDS AND PHARMACEUTICAL COMPOSITIONS FOR INHIBITING PROTEIN KINASES, AND METHODS FOR THEIR USE. **1,** (2003).

7. Pemovska, T. *et al.* Axitinib effectively inhibits BCR-ABL1(T315I) with a distinct binding conformation. *Nature* **519,** 102–105 (2015).

8. La Regina, G. *et al.* New Indole Tubulin Assembly Inhibitors Cause Stable Arrest of Mitotic Progression, Enhanced Stimulation of Natural Killer Cell Cytotoxic Activity,

and Repression of Hedgehog-Dependent Cancer. *J. Med. Chem.* **58,** 5789–5807 (2015).

9. Hasegawa, T., Eiki, J. I. & Chiba, M. Interindividual variations in metabolism and pharmacokinetics of 3-(6-methylpyridine-3-yl-sulfanyl)-6-(4H-[1,2,4]triazole-3-yl-sulfanyl)- N-(1,3-thiazole-2-yl)-2-pyridine carboxamide, a glucokinase activator, in rats caused by the genetic polymorphism of . *Drug Metab. Dispos.* **42,** 1548–1555 (2014).

10. Delfín, D. A., Morgan, R. E., Zhu, X. & Werbovetz, K. A. Redox-active dinitrodiphenylthioethers against Leishmania: Synthesis, structure-activity relationships and mechanism of action studies. *Bioorganic Med. Chem.* **17,** 820–829 (2009).

11. Grime, K. *et al.* An S-warfarin and AZD1981 interaction: in vitro and clinical pilot data suggest the N-deacetylated amino acid metabolite as the primary perpetrator. *Br. J. Clin. Pharmacol.* **83,** 381–392 (2017).

12. Luker, T. *et al.* Substituted indole-1-acetic acids as potent and selective CRTh2 antagonists-discovery of AZD1981. *Bioorganic Med. Chem. Lett.* **21,** 6288–6292 (2011).

13. Zhou, N. *et al.* 3,4-Disubstituted indole acylsulfonamides: A novel series of potent and selective human EP3 receptor antagonists. *Bioorganic Med. Chem. Lett.* **19,** 123–126 (2009).

14. Nishimura, T. *et al.* Identification of novel and potent 2-amino benzamide derivatives as allosteric glucokinase activators. *Bioorganic Med. Chem. Lett.* **19,** 1357–1360 (2009).

15. Berger, D. M. *et al.* 4-Anilino-7-alkenylquinoline-3-carbonitriles as potent MEK1 kinase inhibitors. *Bioorganic Med. Chem.* **16,** 9202–9211 (2008).

16. Miller, L. M. *et al.* Structure-activity relationships of small molecule autotaxin inhibitors with a discrete binding mode. *J. Med. Chem.* **60,** 722–748 (2017).

17. Llauger, L. *et al.* Evaluation of 8-Arylsulfanyl, and 8-Arylsulfonyl Adenine Derivatives as Inhibitors of the Heat Shock Protein 90. *J. Med. Chem* **48,** 2892–2905 (2005).

18. Smith, B. J. *et al.* Pharmacokinetics, Metabolism, and Excretion of [14C]Axitinib, a Vascular Endothelial Growth Factor Receptor Tyrosine Kinase Inhibitor, in Humans. *Drug Metab. Dispos.* **42,** 918–931 (2014).

19. Patel, P. D. *et al.* Paralog-selective Hsp90 inhibitors define tumor-specific regulation of HER2. *Nat. Chem. Biol.* **9,** 677–84 (2013).

20. Carry, J. C. *et al.* SAR156497, an exquisitely selective inhibitor of Aurora kinases. *J. Med. Chem.* **58,** 362–375 (2015).

21. Galatsis, P. *et al.* Kinase domain inhibition of leucine rich repeat kinase 2 (LRRK2) using a [1,2,4]triazolo[4,3-b]pyridazine scaffold. *Bioorganic Med. Chem. Lett.* **24,** 4132–4140 (2014).

22. Tari, L. W. *et al.* Pyrrolopyrimidine inhibitors of DNA gyrase B (GyrB) and topoisomerase IV (ParE). Part I: Structure guided discovery and optimization of dual targeting agents with potent, broad-spectrum enzymatic activity. *Bioorganic Med. Chem. Lett.* **23,** 1529–1536 (2013).

23. Edlin, C. D. *et al.* Identification and in-vitro ADME assessment of a series of novel anti-malarial agents suitable for hit-to-lead chemistry. *ACS Med. Chem. Lett.* **3,** 570–573 (2012).

24. Weidner, T. *et al.* Antiplasmodial dihetarylthioethers target the coenzyme A synthesis pathway in Plasmodium falciparum erythrocytic stages. *Malar. J.* **16,** 192 (2017).

25. Patel, H. J. *et al.* Structure-activity relationship in a purine-scaffold compound series with selectivity for the endoplasmic reticulum Hsp90 paralog Grp94. *J. Med. Chem.* **58,** 3922–3943 (2015).

26. Vernier, W. F. *et al.* Erratum: Thioether benzenesulfonamide inhibitors of carbonic anhydrases II and IV: Structure-based drug design, synthesis, and biological evaluation (Bioorganic & Medicinal Chemistry 18 (2010) (3307)). *Bioorganic Med. Chem.* **18,** 7144 (2010).

27. Stewart, A. O. *et al.* Discovery of Inhibitors of Cell Adhesion Molecule Expression in Human Endothelial Cells. 1. Selective Inhibition of ICAM-1 and E-Selectin Expression. *J. Med. Chem.* **44,** 988–1002 (2001).

28. La Regina, G. *et al.* Toward highly potent cancer agents by modulating the C-2 group of the arylthioindole class of tubulin polymerization inhibitors. *J. Med. Chem.* **56,** 123–149 (2013).

29. La Regina, G. *et al.* Design and Synthesis of 2-Heterocyclyl-3-arylthio-1 H -indoles as Potent Tubulin Polymerization and Cell Growth Inhibitors with Improved Metabolic Stability. *J. Med. Chem.* **54,** 8394–8406 (2011).

30. Stierand, K., Maaß, P. C. & Rarey, M. Molecular complexes at a glance: Automated generation of two-dimensional complex diagrams. *Bioinformatics* **22,** 1710–1716 (2006).

31. Fricker, P. C., Gastreich, M. & Rarey, M. Automated drawing of structural molecular formulas under constraints. *J. Chem. Inf. Comput. Sci.* **44,** 1065–1078 (2004).

32. Usmani, K. A., Karoly, E. D., Hodgson, E. & Rose, R. L. In vitro sulfoxidation of thioether compounds by human cytochrome P450 and flavin-containing monooxygenase isoforms with particular reference to the CYP2C subfamily. *Drug Metab. Dispos.* **32,** 333–339 (2004).

33. Silverman, R. B. *The Organic Chemistry of Drug Design and Drug Action*. (Academic Press, 2004).

34. Bates, C. G., Gujadhur, R. K. & Venkataraman, D. A General Method for the Formation of Aryl – Sulfur Bonds Using Copper (I) Catalysts. *Org. Lett.* **4,** 2803–2806 (2002).

35. Kwong, F. Y. & Buchwald, S. L. A General , Efficient , and Inexpensive Catalyst System for the Coupling of Aryl Iodides and Thiols. *Org. Lett.* **4,** 3517–3520 (2002).

36. Itoh, T. & Mase, T. A General Palladium-Catalyzed Coupling of Aryl Bromides / Triflates and Thiols. *Org. Lett.* **6,** 4587–4590 (2004).

37. Kunz, K., Scholz, U. & Ganzer, D. Renaissance of Ullmann and Goldberg Reactions – Progress in Copper Catalyzed C–N-, C–O-and C–S-Coupling. *Synlett* **2,** 2428–24390 (2003).

38. Ley, S. V. & Thomas, A. W. Modern Synthetic Methods for Copper-Mediated C(aryl)-O, C(aryl)-N, and C(aryl)-S Bond Formation. *Angew. Chemie - Int. Ed.* **42,** 5400–5449 (2003).

39. Beletskaya, I. P. & Ananikov, V. P. Transition-metal-catalyzed C-S, C-Se, and C-te bond formation via cross-coupling and atom-economic addition reactions. *Chem. Rev.* **111,** 1596–1636 (2011).

40. Shen, C. *et al.* Recent advances in C–S bond formation via C–H bond functionalization and decarboxylation. *Chem. Soc. Rev.* **44,** 291–314 (2015).

41. Chekal, B. *et al.* Development of the Commercial Route for the Manufacture of a 5-Lipoxygenase Inhibitor PF-04191834. *Org. Process Res. Dev.* **19,** 1944–1953 (2015).

42. De Koning, P. D. *et al.* Development an efficient route to the 5-lipoxygenase inhibitor PF-04191834. *Org. Process Res. Dev.* **15,** 1046–1051 (2011).

43. Tudge, M. *et al.* Development of a kilogram-scale asymmetric synthesis of a potent DP receptor antagonist. *Org. Process Res. Dev.* **14,** 787–798 (2010).

44. Kim, J., Hewitt, G., Carroll, P. & Sieburth, S. M. N. Silanediol inhibitors of angiotensin-converting enzyme. Synthesis and evaluation of four diastereomers of Phe[Si]Ala dipeptide analogues. *J. Org. Chem.* **70,** 5781–5789 (2005).

45. Szmant, H. H. & Brost, G. A. The Base-Catalyzed Conversion of Triphenylfluorosilane and Triphenylsilanol to Hexaphenyldisiloxane. *J. Am. Chem. Soc.* **72,** 5763 (1950).

46. Showell, G. A. & Mills, J. S. Chemistry challenges in lead optimization: Silicon isosteres in drug discovery. *Drug Discov. Today* **8,** 551–556 (2003).

47. Franz, A. K. & Wilson, S. O. Organosilicon molecules with medicinal applications. *J. Med. Chem.* **56,** 388–405 (2013).

48. Waelchli, R. *et al.* DIPEPTIDE MIMETICS CAN SUBSTITUTE FOR THE RECEPTOR ACTIVATION DOMAIN RESULTING IN HIGHLY POTENT ANALOGUES OF hPTH(1-36) FRAGMENT. *Bioorganic Med. Chem. Lett.* **6,** 1151–1156 (1996).

49. Otaka, A. *et al.* New access to α-substituted (Z)-fluoroalkene dipeptide isosteres utilizing organocopper reagents under 'reduction-oxidative alkylation (R-OA)' conditions. *Tetrahedron Lett.* **42,** 5443–5446 (2001).

50. Dutheuil, G., Couve-Bonnaire, S. & Pannecoucke, X. Diastereomeric fluoroolefins as peptide bond mimics prepared by asymmetric reductive amination of a-fluoroenones. *Angew. Chemie - Int. Ed.* **46,** 1290–1292 (2007).

51. Narumi, T. *et al.* Diastereoselective synthesis of highly functionalized fluoroalkene dipeptide isosteres and its application to Fmoc-based solid-phase synthesis of a cyclic pentapeptide mimetic. *Tetrahedron* **64,** 4332–4346 (2008).

52. Watanabe, D. *et al.* Copper mediated defluorinative allylic alkylation of difluorohomoallyl alcohol derivatives directed to an efficient synthetic method for (Z)-fluoroalkene dipeptide isosteres. *J. Fluor. Chem.* **132,** 327–338 (2011).

53. Futatsugi, K. *et al.* Discovery and Optimization of Imidazopyridine-Based Inhibitors of Diacylglycerol Acyltransferase 2 (DGAT2). *J. Med. Chem.* **58,** 7173–7185 (2015).

54. Hassam, M. *et al.* Novel cyclopropyl-indole derivatives as HIV non-nucleoside reverse transcriptase inhibitors. *ACS Med. Chem. Lett.* **3,** 470–475 (2012).

55. Li, C. *et al.* Design, synthesis, and biological evaluation of novel imidazo[1,2-a]pyridine derivatives as potent c-met inhibitors. *ACS Med. Chem. Lett.* **6,** 507–512 (2015).

56. Johnson, J. K. *et al.* Small Molecule Antagonists of the Nuclear Androgen Receptor for the Treatment of Castration-Resistant Prostate Cancer. *ACS Med. Chem. Lett.* **7,** 785–790 (2016).

57. Kempf, D. J. *et al.* ABT-538 is a potent inhibitor of human immunodeficiency virus protease and has high oral bioavailability in humans. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 2484–2488 (1995).

58. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

59. Gray, J. J. *et al.* Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* **331,** 281–299 (2003).

60. Wang, C. H. U. & Schueler-furman, O. R. a. Improved side-chain modeling for protein – protein docking. *Protein Sci.* **14,** 1328–1339 (2005).

61. Chaudhury, S. *et al.* Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS One* **6,** (2011).

62. Wang, C., Bradley, P. & Baker, D. Protein-Protein Docking with Backbone Flexibility. *J. Mol. Biol.* **373,** 503–519 (2007).

63. Meiler, J. & Baker, D. ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility. *Proteins Struct. Funct. Bioinforma.* **65,** 538–548 (2006).

64. McTigue, M. *et al.* Molecular conformations, interactions, and properties associated with drug efficiency and clinical performance among VEGFR TK inhibitors. *Proc Natl Acad Sci USA* **109,** 18281–18289 (2012).

65. D.A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. G., A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C., Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I., Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. S. & R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. X. and P. A. K. Amber 2016 Reference Manual. (2016). doi:10.1021/ct200909j

66. Beno, B. R., Yeung, K. S., Bartberger, M. D., Pennington, L. D. & Meanwell, N. A. A Survey of the Role of Noncovalent Sulfur Interactions in Drug Design. *J. Med. Chem.* **58,** 4383–4438 (2015).

67. Dubey, R., Polaske, N. W., Nichol, G. S. & Olenyuk, B. Enantioselective organocatalytic α-sulfenylation of substituted diketopiperazines. *Tetrahedron Lett.* **50,** 4310–4313 (2009).

68. Nagao, Y., Iimori, H., Nam, K. H., Sano, S. & Shiro, M. Synthesis and antibacterial activity of new 1beta-methylcarbapenems having the potential for intramolecular nonbonded S...O interactions. *Chem. Pharm. Bull. (Tokyo).* **49,** 1660–1661 (2001).

69. Murray, J. S., Lane, P. & Politzer, P. Simultaneous _Sigma-Hole and Hydrogen Bonding by Sulfur- and Selenium-Containing Heterocycles. *Int. J. Quantum Chem.* **108,** 2770–2781 (2008).

70. Tomoo, T. *et al.* Design, synthesis, and biological evaluation of 3-(1-aryl-1 h -indol-5-yl)propanoic acids as new indole-based cytosolic phospholipase a2α inhibitors. *J. Med. Chem.* **57,** 7244–7262 (2014).

71. Shultz, M. D. *et al.* [1,2,4]Triazol-3-ylsulfanylmethyl)-3-phenyl-[1,2,4]oxadiazoles: Antagonists of the wnt pathway that inhibit tankyrases 1 and 2 via novel adenosine pocket binding. *J. Med. Chem.* **55,** 1127–1136 (2012).

72. Frisch, M. J. *et al.* Gaussian. (2009).

73. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98,** 5648–5652 (1993).

74. Misura, K. M. S. & Baker, D. Progress and challenges in high-resolution refinement of protein structure models. *Proteins Struct. Funct. Genet.* **59,** 15–29 (2005).

75. Das, R. & Baker, D. Macromolecular Modeling with Rosetta. *Annu. Rev. Biochem.* **77,** 363–382 (2008).

76. Tyka, M. D. *et al.* Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405,** 607–618 (2011).

77. Nivón, L. G., Moretti, R. & Baker, D. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLoS One* **8,** e59004 (2013).

78. Leaver-Fay, A. *et al.* Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol.* **523,** 109–143 (2013).

79. O'Meara, M. J. *et al.* Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theory Comput.* **11,** 609–622 (2015).

80. Wang, J. M., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25,** 1157–1174 (2004).

81. Ewing, T. J. A., Makino, S., Skillman, A. G. & Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided. Mol. Des.* **15,** 411–428 (2001).

82. Moustakas, D. T. *et al.* Development and validation of a modular, extensible docking program: DOCK 5. *J. Comput. Aided. Mol. Des.* **20,** 601–619 (2006).

83. Lang, P. T. *et al.* DOCK 6: Combining techniques to model RNA-small molecule complexes. *Rna* **15,** 1219–1230 (2009).

84. Dong C. Liu & Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **45,** 503–528 (1989).

85. Nguyen, H., Roe, D. R. & Simmerling, C. Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* **9,** 2020–2034 (2013).

86. Nguyen, H., Roe, D. R., Swails, J. & Case, D. A. PYTRAJ v1.0.0.dev1: Interactive data analysis for molecular dynamics simulations. (2016). doi:http://dx.doi.org/10.5281/zenodo.44612

87. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9,** 3084–3095 (2013).

88. Nguyen, H., Maier, J., Huang, H., Perrone, V. & Simmerling, C. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J. Am. Chem. Soc.* **136,** 13959–13962 (2014).