DECIPHERING THE LINKS BETWEEN PROTEIN DYNAMICS AND EVOLUTION

by

Yiğit Kutlu

B.S., Chemical Engineering, Boğaziçi University, 2016

M.S., Chemical Engineering, Boğaziçi University, 2020

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Chemical Engineering

Boğaziçi University

2025

ACKNOWLEDGEMENTS

First and foremost, I would like to express my heartfelt gratitude to my thesis supervisor, Prof. Türkan Haliloğlu, for her unwavering guidance, support, and patience throughout this journey. It has been a privilege to work under her mentorship, and I have learned immensely from her deep knowledge, scientific insight, and generous wisdom.

I would also like to extend my sincere appreciation to the members of my thesis jury committee, Assoc. Prof. Burak Alakent and Asst. Prof. Hasan Demirci, for their time, thoughtful feedback, and constructive suggestions during my thesis progress evaluations. I am also deeply grateful to Assoc. Prof. Hilal Taymaz Nikerel and Prof. Özlem Keskin for serving as members of my defense committee and for their valuable contributions to the assessment of my work.

I am thankful to Prof. Nir Ben-Tal, Dr. Rachel Kolodny, and Gabriel Axel for their collaboration and significant contributions to our study titled *"Reused Protein Segments Linked to Functional Dynamics."* It was a great honor to work with Prof. Ben-Tal, from whom I have learned a great deal.

I am especially grateful to Özge Duman for being a constant source of support throughout our PhD journey. Her insightful ideas during countless brainstorming sessions, as well as her encouragement and camaraderie, made a lasting impact. I would also like to thank my lab mates, Dr. Burçin Acar, Dr. Erge Akbaş Buz, and Bengi Altıntel, for the enjoyable and collaborative atmosphere we shared. Special thanks to Yüksel Gülten, whose energy and positivity always brightened the PRC group.

I would like to thank all my friends in the department with whom I shared both challenges and laughter throughout this journey—especially Dr. Seymen İlke Kaykanat and Dr. Sercan Altundemir. İlke, thank you for always encouraging me to

stay on the “good side” and for your unwavering support and positive spirit. Sercan, I truly appreciated our open conversations where we could freely share our thoughts, particularly on our perspectives about academia. Your presence made this journey significantly easier and more enjoyable.

I also thank my friends Çağatay Göler, Emre Kamiloğlu, and Utku Özcan for their friendship and emotional support throughout my thesis period. Your presence made this journey more manageable, and I am truly grateful for having you by my side during this time.

Finally, I wish to thank my family for their endless love and support. I am deeply grateful to my father and brother for always doing everything they could for me. Most of all, I thank my mother, who encouraged me from an early age to ask “why,” nurturing the curiosity and inquisitiveness that shaped my path as a researcher.

It was a great opportunity for me to have my research supported by the NATO Science for Peace Multi-Year Project G5685, TÜBİTAK Project 119F392 and TÜSEB Project 24218.

ABSTRACT

DECIPHERING THE LINKS BETWEEN PROTEIN DYNAMICS AND EVOLUTION

Proteins are inherently dynamic, and their functional mechanisms often rely on collective internal motions. While the relationship between structure and dynamics has been widely studied, the connection between intrinsic dynamics and evolutionarily conserved sequence motifs remains poorly understood. This thesis addresses this gap by investigating the dynamic roles of themes—short conserved sequence segments shared across diverse proteins—through elastic network modeling and information-theoretic analysis. Using the Gaussian Network Model (GNM), dynamic elements (DEs) were defined as contiguous, co-fluctuating segments derived from low-frequency collective modes. To assess their correspondence with themes, mutual information (MI) was calculated between DEs and theme-based partitions across a curated dataset of 150 ECOD domains, with statistical validation performed against randomized controls. The results indicate that many themes align significantly with DEs, suggesting that reused segments are embedded in functional dynamic units. In a focused case study, previously identified adenine-binding themes were analyzed using a GNM-based Transfer Entropy (GNM-TE) method. The analysis revealed that these motifs frequently act as dynamic communication hubs, implicating them in long-range allosteric signaling. Building on this, the thesis classifies themes according to their information transfer roles and applies statistical analysis to evaluate enrichment patterns among binding themes. Together, these studies offer the first systematic framework for linking evolutionarily reused sequence segments with intrinsic dynamic architectures and communication properties. By introducing the concept of dynamic elements and integrating them with directional information flow, this work advances our understanding of protein evolution and function from a structure–dynamics perspective.

ÖZET

PROTEIN DİNAMİKLERİ İLE EVRİMSEL BAĞLANTILARIN DEŞİFRE EDİLMESİ

Proteinler doğası gereği dinamiktir ve işlevsel mekanizmaları çoğunlukla kolektif içsel hareketlere dayanır. Yapı ile dinamikler arasındaki ilişki geniş çapta incelenmiş olsa da, proteinlerin içsel dinamikleri ile evrimsel olarak korunan kısa dizilim motifleri arasındaki bağlantı halen yeterince anlaşılmamıştır. Bu tez, farklı proteinler arasında yalnızca dizi benzerliğine dayalı olarak paylaşılan kısa ve korunan dizilim segmentleri (temalar) ile ilgili dinamik rolleri, elastik ağ modelleme ve bilgi kuramına dayalı analizler yoluyla inceleyerek bu alandaki boşluğu doldurmayı amaçlamaktadır. Çalışmada, Gaussian Ağ Modeli (GNM) kullanılarak düşük frekanslı kolektif modlardan elde edilen, birlikte salınan art arda gelen kalıntılardan oluşan dinamik elementler (DE'ler) tanımlanmıştır. Bu DE'ler ile temalar arasındaki örtüşmeyi değerlendirmek için, 150 ECOD domeninden oluşan özenle seçilmiş bir veri kümesinde temalara dayalı kümeler ile DE'ler arasındaki karşılıklı bilgi (MI) hesaplanmış ve istatistiksel anlamlılık rastgele kontrol gruplarıyla test edilmiştir. Sonuçlar, birçok temanının işlevsel dinamik birimlerle anlamlı biçimde örtüştüğünü ortaya koymaktadır. Odaklı bir vaka çalışmasında, önceden tanımlanmış adenine-bağlayan temalar, GNM tabanlı Transfer Entropisi (GNM-TE) yöntemiyle analiz edilmiş ve bu temaların çoğunlukla uzun menzilli allosterik sinyalleşmeye aracılık eden dinamik iletişim merkezleri olarak davrandığı gösterilmiştir. Bu bulgulara dayanarak, temalar bilgi transferi rollerine göre sınıflandırılmış ve bağlanma temaları arasında zenginleşme örüntüleri istatistiksel olarak değerlendirilmiştir. Bu tez, evrimsel olarak yeniden kullanılan dizilim segmentlerini içsel dinamik mimariler ve iletişim özellikleriyle sistematik biçimde ilişkilendiren ilk çerçeveyi sunmakta; dinamik element kavramı ile yönlü bilgi akışının entegrasyonu sayesinde protein evrimi ve işlevine yapı-dinamik ekseninde yeni bir bakış açısı kazandırmaktadır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii
ABSTRACT v
ÖZET vi
LIST OF FIGURES xi
LIST OF TABLES xviii
LIST OF SYMBOLS xxiv
LIST OF ACRONYMS/ABBREVIATIONS xxv
1. INTRODUCTION 1
2. LITERATURE SURVEY 7
2.1. Protein Dynamics and Computational Approaches 7
2.1.1. Elastic Network Models 8
2.1.2. Gaussian Network Model (GNM) 9
2.1.3. Anisotropic Network Model (ANM) 10
2.1.4. Gaussian Network Model-Based Transfer Entropy (GNM-TE) . 11
2.1.5. Molecular Dynamics (MD) Simulations 13
2.1.6. Langevin Dynamics (LD) 14
2.1.7. ANM-driven Langevin Dynamics (ANM-LD) Method 15
2.1.8. Enhanced Sampling Methods 16
2.1.9. AlphaFold 17
2.2. Evolutionary Units in Proteins 18
2.2.1. Domains 19
2.2.1.1. CATH Domains 20
2.2.1.2. SCOP Domains 21
2.2.1.3. ECOD Domains 21
2.2.2. Sub-domain Units 22
2.2.2.1. Themes as Subdomain Evolutionary Units 24
2.2.2.2. Bridging Themes 24
2.2.2.3. Adenine-Binding Themes 25

2.3. Mechanisms of Protein Evolution and Links to Dynamics 26
3. MATERIALS AND METHODS 28
3.1. Themes: Identification and Dataset Construction 28
3.1.1. Adenine-binding themes 28
3.2. Gaussian Network Model (GNM) 29
3.3. Identification of Dynamic Elements from GNM Modes 30
3.4. GNM based Transfer Entropy Method 33
3.5. Anisotropic Network Model (ANM) 35
3.6. ANM-LD Simulations: ANM-Driven Langevin Dynamics 35
3.7. Mutual Information-Based Clustering Evaluation 37
3.7.1. Adjusted and Standardized Mutual Information Calculations . . 38
3.7.2. Monte Carlo-Based Statistical Validation 39
3.8. Fisher's Exact Test and Hypothesis Correction 40
3.9. Principal Component Analysis (PCA) of MD Correlation Maps 41
4. RESULTS AND DISCUSSION 42
4.1. Reused Protein Segments Linked to Functional Dynamics 42
4.1.1. Motivation and Background 42
4.1.2. Data and Approach 43
4.1.3. In-depth analysis of repeat domains 49
4.1.3.1. All-β architecture: Propellers 49
4.1.3.2. Themes, dynamics, and binding sites 54
4.1.3.3. All-α architecture: alpha helix bundles 58
4.1.3.4. α/β architecture: TIM barrels 62
4.1.4. In-depth analysis of non-repeat domains 66
4.1.4.1. In-depth set: Synthesis of observations 77
4.1.5. Mutual information analysis quantifies the statistical significance of correspondence between theme boundaries and DEs 78
4.1.6. P-value analysis for the statistical significance of correspondence between themes and DEs 85
4.1.7. Comparison of DEs with GNM and MD simulations 88
4.1.8. Correspondence between DEs and Pfam entries 94

4.1.9. Impact of protein conformational changes on GNM modes and theme correspondence 95
4.1.10. Conclusion 98
4.2. Intrinsic Dynamics and Allosteric Roles of Conserved Adenine-Binding Themes 100
4.2.1. Motivation and Background 100
4.2.2. Data and Approach 101
4.2.2.1. Dataset of Adenine-Binding Themes 101
4.2.2.2. GNM and GNM based dynamic elements 102
4.2.2.3. Overlap Score Calculation 103
4.2.2.4. GNM-Based Transfer Entropy Analysis 104
4.2.3. Localization of Adenine-Binding Residues at Dynamical Hinge Sites 106
4.2.4. Adenine-Binding Theme and DE Correspondence in Cluster 1 . 107
4.2.4.1. Representative Cases of Theme–Dynamic Element Correspondence 112
4.2.5. Allosteric Information Flow from Adenine-Binding Themes 896 and 1924 in Rio2 kinase 118
4.2.6. Allosteric Information Flow from Theme 1924 in BUD32 124
4.2.7. Expansion of Overlap Analysis Across Additional Adenine-Binding Theme Clusters 126
4.2.8. Conclusion 130
4.3. Systematic Identification and Dynamic Classification of Binding Themes 132
4.3.1. Identification of Binding Themes and Theme Pairs 132
4.3.2. Theme Communication Analysis Using GNM-TE 138
4.3.3. Case Studies of Allosteric Communication Through Binding Themes144
4.3.4. Conclusion 151
4.4. Additional Contributions 151
4.4.1. Extension of GNM Perturbation Framework 151
4.4.1.1. Poster Summary and Methodological Enhancements . 152
4.4.1.2. Case Study: ABC Transporter ErfCD 153

4.4.1.3. Outlook: Server Development 153
4.4.2. Optimization and Extension of the GNM-Based Transfer Entropy Method 154
4.4.3. Enhancements to the ANM-LD Simulation Framework 156
4.4.3.1. Preliminary Efforts to Incorporate Forward Flux Sampling into ANM-LD 158
4.4.4. Comparative Analysis of GNM-Derived Segments and Protein Language Model Representations 161
4.4.5. Summary of Additional Contributions 162
5. CONCLUSION 163
5.1. Recommendations 165
REFERENCES 166
APPENDIX A: Complete List of Binding Themes 186
APPENDIX B: Theme Classification Results for All Domains 221
APPENDIX C: FIGURE PERMISSION 315

LIST OF FIGURES

Figure 3.1. A toy model to represent decomposition of the structure into the dynamic parts from a correlation matrix of GNM. 31

Figure 3.2. Dynamic segments and DEs in the e2xyiA1 propeller. 32

Figure 3.3. The ten slowest elastic modes of the e2xyiA1 propeller. 33

Figure 4.1. Schematic representation of the approach used here to examine the correlation between themes and dynamic elements. 47

Figure 4.2. Correlation between shared protein segments and structural dynamics. 49

Figure 4.3. DEs and themes in the e3emhA1 propeller. 50

Figure 4.4. Correspondence between themes and DEs in the e2xyiA1 propeller. 55

Figure 4.5. The same theme (or themes combination) may correspond to DEs of different modes of motion. 56

Figure 4.6. Correspondence between binding, dynamics, and themes in the e2xyiA1 domain of the histone-binding protein. 57

Figure 4.7. Putative evolutionary roots of allostery. 58

Figure 4.8. DEs and themes in ARM-repeat (PDB ID: 2OF3, ECOD domain ID: e2of3A1. 59

Figure 4.9. DEs and themes in ARM-repeat (PDB ID: 1B3U, ECOD domain ID: e1b3uA1). 60

Figure 4.10. DEs and themes in proteasome/cyclosome repeat (PDB ID: 4ADY, ECOD domain ID: e4adyA2). 60

Figure 4.11. Correspondance of theme c180-36 and DEs in repetitive alpha hairpin protein domains. 61

Figure 4.12. Correspondance of theme c180-19 and DEs in repetitive alpha hairpin protein domains. 61

Figure 4.13. DEs and themes in TatD-related deoxyribonuclease (PDB ID: 1J6O, ECOD domain ID: e1j6oA1). 62

Figure 4.14. DEs and themes in Putative TatD related DNAse (PDB ID: 2GZX, ECOD domain ID: e2gzxA1). 63

Figure 4.15. DEs and themes in Tat-linked quality control protein TatD (PDB ID: 4P5U, ECOD domain ID: e4p5uA1). 63

Figure 4.16. Correspondence between themes and functional dynamics in the e2gzxA1 TIM barrel domain. 64

Figure 4.17. Themes and DEs in homologous TIM barrels 66

Figure 4.18. DEs and themes in rRNA N-glycosidase (PDB ID: 3CTK, ECOD domain ID: e3ctkA1). 70

Figure 4.19. DEs and themes in Ribosome-inactivating protein gelonin (PDB ID: 3KTZ, ECOD domain ID: e3ktzA1). 70

Figure 4.20. DEs and themes in Putative 3-oxoacyl-acyl carrier protein reductase (PDB ID: 1ULS, ECOD domain ID: e1ulsB1). 71

Figure 4.21. DEs and themes in Tropinpne reductase-II (PDB ID: 2AE2, ECOD domain ID: e2ae2A1). 71

Figure 4.22. DEs and themes in 3-Ketoacyl-(Acyl-Carrier-Protein) reductase (PDB ID: 3N74, ECOD domain ID: e3n74B1). 72

Figure 4.23. Correspondence between themes and dynamics in the e3ktzA1 domain of ribosome-inactivating protein gelonin. 73

Figure 4.24. The same theme may correspond to a DE or a DE combination of different modes of motion in RIP homologues. 73

Figure 4.25. Correspondence between themes and functional dynamics in the e1ulsB1 Rossmann-related domain. 75

Figure 4.26. Correspondence between themes and DEs in Rossmann homologous domains. 76

Figure 4.27. DEs and themes do not always correspond to secondary structure elements. 77

Figure 4.28. Ten sample possible theme combinations of the e2xyiA1 propeller structure. 80

Figure 4.29. An exemplary case of low AMI (AMI <0.50) correspondence between themes and DEs for the e2xyiA1 propeller. 82

Figure 4.30. Probability distribution of MI for the randomly generated themes of the e2xyiA1 propeller and DEs of mode 5. 86

Figure 4.31. The distribution of statistical significance per domain. 87

Figure 4.32. One-to-one correspondence between the slowest GNM mode and PC3 obtained from MD simulation. 89

Figure 4.33. One-to-one correspondence between the second slowest GNM mode and PC3 obtained from MD simulation. 90

Figure 4.34. One-to-one correspondence between the third slowest GNM mode and PC1 obtained from MD simulation. 90

Figure 4.35. One-to-one correspondence with relatively low correlation between the fourth slowest GNM mode and PC4 obtained from MD simulation. 91

Figure 4.36. One-to-one correspondence between the fifth slowest GNM mode and PC2 obtained from MD simulation. 91

Figure 4.37. One-to-one correspondence with relatively low correlation between the sixth slowest GNM mode and PC5 obtained from MD simulation. 92

Figure 4.38. One-to-one correspondence with relatively low correlation between the seventh slowest GNM mode and PC7 obtained from MD simulation. 92

Figure 4.39. DEs obtained with GNM and MD-PCA. 93

Figure 4.40. Themes and DEs of the first MD-PCA and the third slowest mode of GNM. 94

Figure 4.41. GNM mode comparisons of the ECOD domains in the dataset and their conformationally changed counterpart ECOD domains. . . . 97

Figure 4.42. Schematic representation of the methodological workflow used to investigate the intrinsic dynamics and allosteric roles of conserved adenine-binding themes. 105

Figure 4.43. GNM analysis of Rio2 kinase highlights hinge localization of adenine-binding residues. 107

Figure 4.44. Adenine binding themes identified in Archaeoglobus fulgidus Rio2 kinase. 113

Figure 4.45. Dynamic domains obtained from the slowest ten modes of 1ZAR together with the adenine-binding themes. 114

Figure 4.46. Multiple adenine-binding themes in Rio2 kinase correspond to distinct DEs from different GNM modes. 115

Figure 4.47. The same adenine-binding theme corresponds to distinct dynamic elements across different protein structures. 116

Figure 4.48. Cross-structure correspondence of theme 896 to different dynamic elements. 117

Figure 4.49. Dynamic mapping of theme 1924 in different structural contexts. . 118

Figure 4.50. Residue-by-residue 2D TE map of the 1ZAR for mode subset 1-4-5-8, with the theme 896 outlined and labeled with rectangular box. 119

Figure 4.51. GNM-TE analysis of theme 896 in Archaeoglobus fulgidus Rio2 kinase. 120

Figure 4.52. Residue-by-residue 2D TE map of the 1ZAR for mode subset 2 and 6, with the theme 1924 outlined and labeled with rectangular box. 122

Figure 4.53. GNM-TE analysis of theme 1924 in Archaeoglobus fulgidus Rio2 kinase. 123

Figure 4.54. GNM-TE analysis of theme 1924 in KEOPS complex subunit BUD32. 125

Figure 4.55. Schematic representation of the workflow used for the systematic identification and dynamic classification of binding themes 133

Figure 4.56. Enrichment of communication classes among binding themes. . . . 143

Figure 4.57. GNM-TE analysis of domain e2xyiA1 highlighting themes 14940 and 14950. 145

Figure 4.58. Information flow from global source residues in theme 14940 in domain e2xyiA1. 146

Figure 4.59. Information flow to global sink residues in theme 14950 in domain e2xyiA1. 147

Figure 4.60. GNM-TE analysis of domain e2ae2A1. 148

Figure 4.61. Information flow from source residues in theme 2939 in domain e2ae2A1. 149

Figure 4.62. Information flow to global sink residues in theme 2962 of domain e2ae2A1. 150

Figure 4.63. Thirteen mutation sensitive positions and related local/global minima positions of delta CRF on EfrCD. 154

Figure 4.64. Schematic overview of the low-resolution structural data driven ANM-LD hybrid simulation. 157

Figure 4.65. Representation of intermediate structures obtained from ANM-LD simulations using SAXS data for the AdK system on the free energy landscape. 158

Figure 4.66. Structure of adenylate kinase (AdK). 159

Figure 4.67. Collective variables calculated from intermediate structures obtained at each cycle of the standard ANM-LD simulation for the AdK system. 160

Figure 4.68. Rates of successful and unsuccessful simulations for the first step of FFS. 160

Figure 4.69. Rates of successful and unsuccessful simulations for the second step of FFS. 161

LIST OF TABLES

Table 3.1. Contingency table for Fisher's exact test. 40

Table 4.1. A set of eight ECOD domains from four ECOD H-groups. 45

Table 4.2. A set of five ECOD domains from RIP and Rossmann-related H-groups. 46

Table 4.3. Themes detected in e2xyiA1 propeller. 51

Table 4.4. Themes detected in e3emhA1 propeller. 52

Table 4.5. Themes of the e3ktzA1 and e3ctkA1 domains. 67

Table 4.6. Themes of Rossmann-related domains. 68

Table 4.7. AMI and SMI values with 3-residue overlap and 8-residue gap. . . 81

Table 4.8. AMI and SMI values with 5-residue overlap and 10-residue gap. . . 81

Table 4.9. AMI and SMI values with 5-residue overlap and 15-residue gap. . . 82

Table 4.10. AMI and SMI values for each ECOD domain. 83

Table 4.11. Average AMI values for DEs of GNM modes per domain. 84

Table 4.12. Maximum AMI values for DEs of GNM modes per domain. 84

Table 4.13. P-values and significant mode counts for each ECOD domain. . . . 87

Table 4.14. Pearson correlation between GNM modes and PCs. 89

Table 4.15. Pfam-DE correspondence via AMI in 4 ECOD domains. 95

Table 4.16. Max AMI values for e4qvhB1 and e4blaB3. 97

Table 4.17. Max AMI values for e1twfA10 and e3j1nA7. 98

Table 4.18. Protein structures and homology groups in Cluster 1. 108

Table 4.19. Summary of theme–DE correspondence in Cluster 1. 111

Table 4.20. Summary of theme–DE correspondence in Cluster 2. 127

Table 4.21. Summary of theme–DE correspondence in Cluster 4. 128

Table 4.22. Summary of theme–DE correspondence in Cluster 6B. 128

Table 4.23. Overlap of themes and peptide-binding residues in domain e2xyiA1. 134

Table 4.24. Binding theme pairs for peptide binding in domain e2xyiA1. 137

Table 4.25. Classification criteria with structural and dynamic conditions. . . . 139

Table 4.26. Enrichment analysis of communication classes among binding themes. 142

Table A.1. Complete list of binding themes. 186

Table B.1. Classification results for e1a7tA1 domain. 221

Table B.2. Classification results for e1dzkA1 domain. 223

Table B.3. Classification results for e1elrA1 domain. 224

Table B.4. Classification results for e1elwA1 domain. 224

Table B.5. Classification results for e1fmcA1 domain. 225

Table B.6. Classification results for e1h3dA2 domain. 226

Table B.7. Classification results for e1h5qA1 domain. 228

Table B.8. Classification results for e1m2xA1 domain. 229

Table B.9. Classification results for e1nh8A2 domain. 230

Table B.10. Classification results for e1o04A1 domain. 231

Table B.11. Classification results for e1p22A1 domain. 235

Table B.12. Classification results for e1so2A1 domain. 237

Table B.13. Classification results for e1uxtA1 domain. 239

Table B.14. Classification results for e1y2kA1 domain. 242

Table B.15. Classification results for e1zemA1 domain. 244

Table B.16. Classification results for e1zk4A1 domain. 245

Table B.17. Classification results for e1zklA1 domain. 246

Table B.18. Classification results for e2ae2A1 domain. 248

Table B.19. Classification results for e2ag5A1 domain. 249

Table B.20. Classification results for e2bcqA3 domain. 250

Table B.21. Classification results for e2dtxA1 domain. 251

Table B.22. Classification results for e2j13A1 domain. 253

Table B.23. Classification results for e2o23A1 domain. 254

Table B.24. Classification results for e2r25B1 domain. 255

Table B.25. Classification results for e2vd3A1 domain. 255

Table B.26. Classification results for e2wsbA1 domain. 257

Table B.27. Classification results for e2xyiA1 domain. 258

Table B.28. Classification results for e2yntB1 domain. 260

Table B.29. Classification results for e2zatA1 domain. 262

Table B.30. Classification results for e3b2dA4 domain. 263

Table B.31. Classification results for e3bwhA1 domain. 264

Table B.32. Classification results for e3e13X1 domain. 266

Table B.33. Classification results for e3ecmA1 domain. 269

Table B.34. Classification results for e3emhA1 domain. 271

Table B.35. Classification results for e3f11A1 domain. 273

Table B.36. Classification results for e3g7wA1 domain. 274

Table B.37. Classification results for e3ignA1 domain. 276

Table B.38. Classification results for e3j80g1 domain. 277

Table B.39. Classification results for e3kffA1 domain. 279

Table B.40. Classification results for e3m7nH2 domain. 280

Table B.41. Classification results for e3osrA1 domain. 281

Table B.42. Classification results for e3ozyA4 domain. 282

Table B.43. Classification results for e3pk0A1 domain. 283

Table B.44. Classification results for e3q25A1 domain. 284

Table B.45. Classification results for e3rg1A6 domain. 286

Table B.46. Classification results for e3rg1E1 domain. 288

Table B.47. Classification results for e3ro3A1 domain. 289

Table B.48. Classification results for e3v30A2 domain. 290

Table B.49. Classification results for e4c1dA1 domain. 292

Table B.50. Classification results for e4c1hA1 domain. 293

Table B.51. Classification results for e4fn4A1 domain. 294

Table B.52. Classification results for e4h4rA3 domain. 295

Table B.53. Classification results for e4i3vA1 domain. 296

Table B.54. Classification results for e4k5uB1 domain. 298

Table B.55. Classification results for e4kliA3 domain. 300

Table B.56. Classification results for e4lscA1 domain. 301

Table B.57. Classification results for e4mn8B1 domain. 303

Table B.58. Classification results for e4pf4A1 domain. 305

Table B.59. Classification results for e4psxB1 domain. 306

Table B.60. Classification results for e4qedA1 domain. 308

Table B.61. Classification results for e4qgeA1 domain. 309

Table B.62. Classification results for e4qvhB1 domain. 310

Table B.63. Classification results for e4urgA1 domain. 312

Table B.64. Classification results for e4wecA1 domain 313

LIST OF SYMBOLS

a_i	Acceleration of particle i
D_F	Deformation factor
d_i	Distance between the coordinates of atom i in two structures
D_i	Difference vector
F_i	Force exerted on particle i
H	Hessian matrix
k_B	Boltzmann constant
m_i	Mass of particle i
N	Number of atoms
O_{ij}	Overlap value of j^{th} eigenvector with the difference vector for cycle i
r_{cut}	Cut-off radius
R_i	Position vector of residue i
R_{ij}	Distance between residues i and j
t	Time
T	Absolute temperature
u_k	k^{th} eigenvector
V	Potential function
γ_f	Force constant for harmonic spring
Γ	Kirchhoff matrix
$\triangle R_i$	Fluctuation of residue i
λ_k	k^{th} eigenvalue
ν_i	Velocity of atom i

LIST OF ACRONYMS/ABBREVIATIONS

AMI	Adjusted Mutual Information
ANM	Anisotropic Network Model
ANM-LD	Anisotropic Network Model Guided Langevin Dynamics
CRF	Cumulative Residue Fluctuation
DE	Dynamic Element
ENM	Elastic Network Model
GNM	Gaussian Network Model
GNM-TE	GNM-Transfer Entropy
LD	Langevin Dynamics
MD	Molecular Dynamics
MI	Mutual Information
MSF	Mean Square Fluctuations
PC	Principal Component
PCA	Principal Component Analysis
PDB	Protein Data Bank
RMSD	Root Mean Square Deviation
SMI	Standardized Mutual Information
TMD	Transmembrane Domain

1. INTRODUCTION

Proteins are not merely static molecular machines; they are dynamic entities whose internal motions underlie essential biological functions. These motions, often large-scale and cooperative, enable proteins to switch between functional states, propagate allosteric signals, and interact selectively with other molecules. While many efforts have focused on the structure–function paradigm, a growing body of evidence highlights the importance of intrinsic dynamics in shaping protein functionality [1–3].

At the same time, studies in protein evolution have revealed that proteins often contain short sequence segments that are reused across different proteins, sometimes even among those with entirely distinct folds. This suggests that proteins evolve, at least in part, by recombining modular building blocks that confer evolutionary advantage [4–12]. Protein domains, as defined in classification systems such as SCOP, CATH, CDD, and ECOD [13–18]—are perhaps the most well-established examples of such recurring segments. However, accumulating evidence suggests that many domains themselves arose through the reuse and recombination of shorter subdomain fragments [19–31].

Recent work has focused on detecting these evolutionarily conserved subdomain-level segments. These segments are typically 20–80 residues in length and recur across structurally and evolutionarily unrelated proteins [25, 32–34]. Nepomnyachiy et al. introduced a sensitive HHsearch-based [35] approach that identifies statistically significant sequence similarities between distant proteins, uncovering thousands of "themes" across the ECOD structural space [33]. Themes are thus believed to represent ancestral peptide fragments that have been recombined over evolutionary time to give rise to modern proteins. Their recurrence across disparate protein families—and enrichment at functionally important sites such as ligand-binding pockets, metal-coordination motifs, or protein-protein interfaces—suggests that themes are not merely structural coincidences but potentially reflect evolutionary and functional units [27, 36–38].

However, despite extensive research into the sequence and structural roles of themes, their dynamic properties have remained largely unexplored. Do these conserved segments also function as dynamic units within the protein architecture? Are they preferentially positioned at hinge regions, or involved in long-range communication pathways such as allostery? Given their deep evolutionary conservation and functional significance, it is reasonable to hypothesize that themes may also exhibit characteristic dynamic behavior, contributing to protein function beyond what is evident from static structure alone.

Intrinsic dynamics, particularly the low-frequency, cooperative motions encoded by protein structure, play a central role in enabling proteins to function. These collective conformational changes, often captured by normal mode analysis, are robustly favored by a protein's topology and correspond to motions involved in binding, catalysis, and allosteric regulation [2, 39–42]. Numerous studies have shown that these intrinsic motions are functionally relevant [43–46] and that residues involved in dynamic coordination, especially those mediating long-range allosteric communication, tend to be evolutionarily conserved [47, 48]. Moreover, homologous proteins often share similar global dynamics, suggesting that intrinsic flexibility may itself be under evolutionary selection [49–51]. While these studies have largely focused on full-domain architectures, in this thesis we explore whether evolutionary signals at the subdomain level, specifically themes, are also dynamically meaningful.

To analyze these dynamics at scale, we employed the Gaussian Network Model (GNM) and its extensions as central tools throughout this thesis. GNM is a coarse-grained elastic network model that efficiently predicts residue-level fluctuations and cross-correlations based solely on the 3D topology of a protein [52, 53]. It captures collective, low-frequency modes that represent the intrinsic dynamic tendencies of the structure—precisely the types of motions involved in conformational transitions and functional switching. GNM and its anisotropic counterpart, ANM, have been applied to hundreds of proteins and shown to accurately recover experimental features such as B-factors in addition to identification of hinge residues, domain motions, and allosteric

couplings [2, 40, 54].

To go beyond equilibrium fluctuation correlations and capture directional communication, we also applied the GNM-based Transfer Entropy (GNM-TE) method [55,56]. GNM-TE extends the standard GNM framework by integrating time-delayed correlations into an information-theoretic approach, enabling the quantification of how fluctuations at one residue influence another—effectively capturing causal signal flow within the protein. This makes GNM-TE particularly well suited for identifying allosteric pathways, communication hubs, and source/sink residues in complex structures [56–58]. In this thesis, GNM-TE is used to assess whether themes participate in information-rich dynamic regions and to evaluate their possible roles in long-range communication and allosteric regulation.

The work presented in this thesis is organized into four main parts, each addressing a distinct but interrelated aspect of the relationship between evolutionary reuse and protein dynamics. Section 4.1 presents a large-scale analysis of theme reuse and intrinsic dynamics across structurally diverse ECOD domains, offering the first quantitative investigation of the correspondence between sequence-derived themes and dynamics-derived segments (published during the course of this thesis; Kutlu et al., 2024 [59]). Section 4.2 shifts focus to a specific functional context, adenine-binding proteins, and examines whether ligand-interacting themes also display dynamic coherence and participate in allosteric information flow. In Section 4.3, themes identified as ligand-binding across the full dataset are systematically classified, and their enrichment in dynamic communication classes is evaluated using GNM-TE. Finally, Section 4.4 summarizes several additional contributions made during the PhD, including methodological advances in GNM-TE, extensions of ANM-LD simulations, and collaborative applications of the dynamic analysis framework developed in this thesis.

The first part of the thesis (Section 4.1) addresses a fundamental and previously unexamined question: To what extent do evolutionarily conserved themes align with the dynamic architecture of proteins? To explore this relationship, we systematically

analyzed a curated set of protein domains and selected those with high theme coverage for further analysis. Two complementary datasets were constructed: an in-depth set comprising protein domains with diverse architectures selected for detailed inspection, and an expanded set of 150 ECOD domains spanning 26 homology groups, designed to enable broad and statistically meaningful comparisons. Dynamic elements (DEs) were defined from low-frequency GNM modes by identifying coherent, co-fluctuating residue segments. Short fragments below a threshold length were merged to eliminate noise and ensure the structural relevance of identified DEs. To assess the correspondence between these dynamics-derived segments and sequence-derived themes, we developed a pipeline to construct valid theme combinations over the full protein sequence, accounting for gaps and overlaps using a Monte Carlo sampling approach. This allowed for the generation of alternative theme decompositions under biologically reasonable constraints. The Mutual Information (MI) metric was used to quantify the alignment between themes and DEs, and a null model was created by sampling random segment sets with matched length and coverage properties. This enabled statistical validation of observed MI values through empirical p-value estimation. The results reveal that in a substantial fraction of proteins, reused themes overlap significantly with intrinsic dynamic elements, suggesting that evolutionary reuse is not purely structural or sequence-based, but is often dynamically encoded as well.

The second part of the thesis (Section 4.2) aims to determine whether adenine-binding themes, previously identified based on sequence similarity, also exhibit distinct dynamic signatures and participate in allosteric communication pathways. Building on the work of Narunsky et al. (2020) [36], we selected four representative theme clusters associated with different adenine-binding modes and constructed a dataset of high-resolution structures containing these themes. Using GNM-TE, we quantified the directional flow of information from residues within each theme to the rest of the protein. This allowed us to assess whether themes act as information sources, sinks, or dynamically central regions in the context of global protein motions. By integrating GNM-TE with mode-based dynamic segmentation and structural mapping, we evaluated the dynamic role of binding themes beyond static contacts, offering new

insights into how evolutionarily conserved binding elements contribute to long-range functional communication.

The third part of the thesis (Section 4.3) aims to identify ligand-binding themes across a large and structurally diverse set of proteins and to classify these themes based on their dynamic communication profiles, as characterized using GNM-TE. Starting from the expanded dataset of 150 ECOD domains, we systematically identified themes that are in direct contact with biologically relevant ligands. These binding themes were then analyzed using GNM-TE to determine their information transfer characteristics, including whether they act as senders, receivers, or bidirectional hubs of dynamic communication. Each theme was assigned to one of several communication classes, reflecting its role in the propagation of dynamic signals across the protein structure. Finally, we assessed whether specific communication classes were statistically enriched among binding themes using Fisher's exact test and applied multiple hypothesis correction to ensure robustness. This classification framework provides a new lens through which to interpret the functional significance of theme reuse in the context of protein dynamics and allostery.

The final part of the thesis (Section 4.4) presents a summary of additional contributions made during the PhD that extend beyond the main research questions. These include methodological enhancements to GNM-TE and ANM-LD, applications of the developed tools in collaborative studies, and early-stage integration of protein language models with dynamics-based segmentation. While these topics are only briefly discussed in the main body of the thesis, they reflect the broader impact and versatility of the computational framework developed throughout this work.

Together, the studies presented in this thesis establish the first systematic framework for integrating sequence-based evolutionary reuse with structure-based intrinsic dynamics and communication patterns. By introducing the concept of dynamic elements (DEs) and combining it with mutual information metrics, transfer entropy analysis, and rigorous statistical validation, this work demonstrates that many reused

protein segments are not only structurally and evolutionarily conserved, but also embedded within functionally significant dynamic architectures. These findings reveal a previously uncharacterized connection between modular reuse and intrinsic dynamics, offering new perspectives on the evolutionary design of proteins and providing a foundation for future investigations into allosteric regulation, functional innovation, and structure-based rational design.

2. LITERATURE SURVEY

2.1. Protein Dynamics and Computational Approaches

Understanding proteins solely through their static structures provides an incomplete picture, as biological function frequently depends on conformational flexibility, structural rearrangements, and dynamic interactions with other biomolecules. Proteins operate across multiple dynamic timescales, from fast side-chain fluctuations to slower domain motions, and these fluctuations are integral to processes such as enzyme catalysis, signal transduction, ligand binding, and allosteric regulation [60–62]. These phenomena underscore a conceptual shift from the traditional paradigm of "sequence → structure → function" to an expanded framework that includes dynamics as a key intermediate: "sequence → structure → dynamics → function" [2]. This revised dogma emphasizes that function often emerges from conformational ensembles and the accessibility of collective motions encoded by the protein's topology.

Experimental techniques such as nuclear magnetic resonance (NMR), cryogenic electron microscopy (cryo-EM), and single-molecule Förster resonance energy transfer (smFRET) have significantly advanced our ability to observe protein motion. However, their limitations in temporal or spatial resolution, as well as practical challenges in cost and throughput, restrict their applicability for large-scale or high-resolution studies. In this context, computational approaches have emerged as powerful and versatile tools, capable of probing protein dynamics at atomic detail and across biologically relevant timescales.

Among the most widely used computational strategies are elastic network models (ENMs), such as the Gaussian Network Model (GNM) and Anisotropic Network Model (ANM), which offer efficient and accurate predictions of collective motions by modeling residue-residue connectivity based on structural topology [2, 62]. These coarse-grained models have been instrumental in identifying global modes associated with ligand bind-

ing, allostery, and functional transitions. At finer temporal and spatial resolutions, molecular dynamics (MD) simulations provide detailed atomistic trajectories of proteins in solvent environments, while Langevin dynamics (LD) methods incorporate stochastic forces and friction to mimic thermal fluctuations and environmental damping. Enhanced sampling techniques, including metadynamics, accelerated MD, and replica exchange, have also been developed to overcome kinetic barriers and explore rare events or slow transitions that may be inaccessible to conventional MD.

In parallel, deep learning tools such as AlphaFold have revolutionized protein structure prediction and, more recently, are being adapted for modeling conformational variability [63–65]. These computational models, whether physics-based or data-driven, now serve as indispensable complements to experimental studies, offering mechanistic insight into dynamic behavior, revealing pathways of allosteric communication, and enabling structure-based drug design at an unprecedented scale.

In the following subsections, we review the theoretical foundations and representative applications of these methods, with a particular focus on those relevant to the present work.

2.1.1. Elastic Network Models

Proteins are not rigid structures; their function is often mediated by intrinsic motions that are encoded in their structural topology. Among computational approaches used to investigate these motions, Elastic Network Models (ENMs)—and in particular, Normal Mode Analysis (NMA)—have proven remarkably efficient for predicting cooperative, large-scale conformational changes that are often functionally relevant [2, 39–42]. These low-frequency collective motions, also known as soft modes, are robust features of a protein's three-dimensional fold and are crucial for enabling ligand binding, enzymatic catalysis, allosteric regulation, and substrate recognition [43, 44, 46, 66–68].

The value of ENMs lies in their simplified representation of protein structures, which enables efficient and scalable exploration of conformational dynamics. ENM-based NMA has been used to identify functional motions in a wide range of systems and has been shown to correlate well with experimentally observed conformational transitions. Moreover, these intrinsic dynamics are often conserved across homologous proteins, suggesting that they are under evolutionary selection [49–51, 69]. Residue positions that mediate these dynamics—particularly those involved in allosteric communication networks—also tend to exhibit strong sequence conservation [47, 48, 70]. In specific folds, even ancestrally reconstructed proteins have been analyzed via NMA to hypothesize how their dynamic behavior evolved over time [51, 71].

In the context of this study, ENMs serve as a core tool for probing whether evolutionarily conserved subdomain units (i.e., themes) are also dynamically relevant.

2.1.2. Gaussian Network Model (GNM)

The Gaussian Network Model (GNM) is one of the earliest and most widely used elastic network models. Introduced by Bahar and colleagues [52, 53], GNM represents a protein as a network of nodes, typically corresponding to the Cα atoms of residues, connected by harmonic springs if they lie within a cutoff distance in three-dimensional space. This model captures the topological connectivity of the protein and assumes that thermal fluctuations of the nodes follow a Gaussian distribution.

GNM simplifies the problem by assuming isotropic fluctuations, meaning that each residue's motion is characterized by a single degree of freedom representing its fluctuation amplitude. Despite this simplicity, GNM has proven remarkably successful in identifying functionally relevant fluctuations and residue mobility profiles, particularly in predicting B-factors, hinge regions, and global modes of motion. These low-frequency modes often correspond to biologically significant conformational changes involved in enzyme activity, allosteric transitions, and molecular recognition [40, 72].

Recent studies have expanded the utility of GNM in novel directions. For example, sequence-based implementations of GNM now allow prediction of residue flexibility and mutational impacts without requiring 3D structural data [73, 74]. GNM has also been used to explain the dynamic coupling between distal residues and active sites, offering mechanistic insights into the pathogenicity of missense mutations [75]. Applications to protein–protein interfaces have demonstrated that GNM can distinguish between obligate and non-obligate dimers by comparing interfacial mobility [76], while recent work has confirmed its adequacy in approximating long-range information transfer using mutual information frameworks [77]. In protein design contexts, GNM has been employed to study mobility patterns and binding site residues in DARPin–virus protein complexes [78], and comparative modeling studies have shown GNM's limitations and complements when paired with differential geometry or machine learning-based predictors [79]. Finally, GNM has proven useful in chromatin dynamics by modeling genome-wide conformational correlations using Hi-C data, enabling insights into gene regulation and cell-type-specific expression [80].

2.1.3. Anisotropic Network Model (ANM)

While GNM provides a scalar measure of residue fluctuations, the Anisotropic Network Model (ANM) extends the approach to capture directional preferences in atomic motions. ANM models the protein as a three-dimensional elastic network, where each node is allowed to fluctuate in three Cartesian dimensions [54, 81]. This allows for the calculation of 3N–6 nontrivial modes (excluding rigid-body motions), each defined by a directional displacement vector.

The key advantage of ANM is its ability to generate structural deformations along the softest (lowest-frequency) modes, providing physically realistic pathways between functional states of a protein. These mode-driven deformations are frequently used to interpolate between crystallographic or cryo-EM structures, or to explore potential allosteric pathways and conformational transitions [2, 40].

Because ANM considers anisotropic motion vectors, it offers a more detailed picture of conformational dynamics than GNM, making it especially suitable for studies involving domain movements, hinge-bending motions, or coupled subunit rearrangements in multi-domain or oligomeric proteins.

Recent work continues to demonstrate the value of the ANM in capturing biologically relevant conformational dynamics of proteins and macromolecular assemblies. The updated ANM 2.0 web server extended accessibility by supporting protein–nucleic acid and protein–ligand systems with enhanced computational and visualization features [82]. Computational advances have enabled ANM-based normal mode analysis on massive molecular systems (more than 2 million atoms) by introducing GPU-accelerated eigensolvers capable of calculating low-frequency vibrational modes with unprecedented efficiency [83]. Experimental integration has also emerged: temperature-jump time-resolved crystallography revealed how enzyme motions unfold across timescales and how dynamic coupling is modulated by inhibitors [84]. Finally, ANM studies of tandem-repeat proteins confirmed that their elongated architectures are evolutionarily tuned for robust, functionally relevant global motions, with just a few soft modes sufficient to capture their structural variability [85]. Together, these findings highlight ANM's versatility for probing dynamic features encoded in both structure and sequence.

2.1.4. Gaussian Network Model-Based Transfer Entropy (GNM-TE)

The Gaussian Network Model-based Transfer Entropy (GNM-TE) framework integrates the principles of network dynamics with information theory to capture causal allosteric communication within proteins. Rooted in Schreiber's transfer entropy formulation, this method extends GNM by introducing a time-delayed mutual information component, enabling the distinction between driver and responder residues in fluctuation-based interactions [55]. Unlike mutual information, which only measures symmetric correlation, transfer entropy quantifies directionality, revealing how residue motions causally influence one another—a central aspect of long-range allosteric sig-

naling.

Altintel et al. (2022) refined the approach by introducing TECol scores, which couple transfer entropy with the collectivity of residue motions across different mode subsets [56]. This allowed for the identification of global information sources—residues that act as allosteric effectors by dispersing signals collectively to other parts of the protein [56]. Ersoy et al. (2023) applied this approach to CFTR, uncovering dynamic allosteric hubs that explain both disease-associated mutation sites and the mechanisms of action of FDA-approved CFTR modulators, offering unprecedented insight into latent allosteric regulation [57].

Duman et al. (2025) further demonstrated the power of GNM-TE in membrane protein systems by applying it to manganese ABC transporters [58]. Their study revealed bidirectional signal propagation between the transmembrane domains and asymmetric information flow between protomers—results that were validated experimentally through mutagenesis. These applications underscore the model's strength in capturing emergent communication patterns in proteins where traditional atomistic simulations are computationally expensive or limited in timescale [58].

Beyond specific case studies, the GNM-TE framework also provides a generalizable and scalable approach to investigate the causal underpinnings of protein dynamics. It enables the identification of dynamically central residues, quantification of cross-residue influence, and generation of hypotheses for mutation effects or allosteric drug targets. In the context of this thesis, GNM-TE is employed to examine whether evolutionarily reused segments act as dynamic information hubs, offering a new layer of interpretation that bridges sequence conservation, structural modularity, and functional communication.

2.1.5. Molecular Dynamics (MD) Simulations

Molecular dynamics (MD) simulation is a computational technique used to model the physical movements of atoms and molecules over time, based on principles of classical mechanics. The method iteratively solves Newton's equations of motion for a system of particles, providing atomic-level trajectories that reflect the system's dynamic behavior under defined thermodynamic conditions [86, 87]. The first successful application of MD to proteins was performed by McCammon et al. in 1977, simulating the internal dynamics of the bovine pancreatic trypsin inhibitor [88].

In classical MD simulations, the system's total energy is described by a potential energy function derived from empirical force fields. These force fields account for bonded interactions—such as bond lengths, bond angles, and torsion angles—as well as non-bonded interactions including van der Waals forces and electrostatics [89]. The dynamics of each atom are governed by the equation

$$m_i \frac{d^2 r_i}{dt^2} = -\frac{\partial V}{\partial r_i}, \tag{2.1}$$

where m is the mass of the particle, r is its position vector, t is time, and V is the potential energy of the system. This formalism enables the construction of time-resolved molecular trajectories, revealing conformational changes, binding events, and other dynamic phenomena that often underlie biological function.

MD simulations are widely used to study biomolecular processes such as protein folding, ligand binding, and allosteric transitions. However, they are limited by two key factors: the accuracy of the force field used to describe interatomic interactions and the high computational cost associated with simulating biologically relevant timescales [90]. Despite these challenges, advances in algorithms, hardware (e.g., GPU acceleration), and hybrid quantum-classical methods have significantly expanded the scope and applicability of MD simulations in recent years.

Recent developments have expanded the accessibility, accuracy, and scope of MD simulations across biological research. Shrimpton-Phoenix et al. [91] introduced drMD, a user-friendly pipeline that automates OpenMM simulations for non-experts, simplifying the setup, execution, and interpretation of MD workflows. Meanwhile, MDRepo, developed by Roy et al. [92], addresses the need for centralized infrastructure by providing an open-access platform for the storage, sharing, and reuse of community-contributed MD trajectories—paving the way for large-scale machine learning applications. Performance advances were also achieved by Persson et al., who demonstrated that fast multipole methods (FMM) can drastically accelerate gas-phase simulations for large protein complexes, extending the utility of MD in native mass spectrometry [93]. On the accuracy front, Wang et al. proposed AI^{2}BMD, an artificial intelligence–based system that achieves ab initio-level energy predictions with near-DFT accuracy using machine learning force fields, enabling microsecond-scale simulations and free energy calculations for proteins with over 10,000 atoms [94]. Together, these innovations highlight the continued relevance and evolution of MD simulations as a key tool for probing protein conformational dynamics, folding, and interaction landscapes.

2.1.6. Langevin Dynamics (LD)

Langevin Dynamics extends classical Molecular Dynamics by incorporating stochastic forces and frictional damping, effectively modeling the interaction of a system with an implicit solvent. In addition to the deterministic force derived from the potential energy function, LD introduces a frictional force proportional to velocity and a random force that captures thermal fluctuations. The equation of motion is modified as

$$m_i \frac{d^2 r_i}{dt^2} = -\frac{\partial v}{\partial r_i} - \xi_i \frac{dr_i}{dt} + R_i(t), \tag{2.2}$$

where, m is the mass, r is the position, ξ is the friction coefficient, and R(t) is a random force representing the solvent-induced noise. By adjusting the friction coefficient, LD can interpolate between traditional MD and overdamped diffusive motion. This makes LD particularly useful for exploring rugged energy landscapes characterized by deep

local minima—regions where classical MD simulations often become kinetically trapped over short timescales [90,95,96]. As such, LD has become an effective tool for enhancing conformational sampling, especially in studies of slow biomolecular transitions and allosteric mechanisms.

2.1.7. ANM-driven Langevin Dynamics (ANM-LD) Method

The ANM-LD method is a hybrid simulation approach that combines Anisotropic Network Model (ANM) dynamics with the stochastic framework of Langevin Dynamics (LD) to explore large-scale conformational transitions in biomolecules. This technique is particularly well suited for modeling transitions between experimentally determined structural states, offering both computational efficiency and physical interpretability [97, 98].

The method requires an initial and a target structure as input. At each iteration, the conformational update is guided by a selected ANM mode that best overlaps with the difference vector between the current and target structures. The overlap metric ensures that the chosen mode drives the system along the most relevant collective motion pathway. The intermediate structure is then perturbed along this mode using a deformation factor, and the process is repeated iteratively until convergence is achieved, typically assessed by the root-mean-square deviation (RMSD) between the evolving and target structures.

By leveraging both the directionality of ANM-derived modes and the stochastic perturbations from Langevin dynamics, ANM-LD efficiently samples functionally relevant conformational changes, especially for large protein complexes. Its ability to generate physically plausible transition pathways makes it a valuable tool in mechanistic studies of allosteric regulation, enzyme activation, and conformational switching in membrane proteins.

2.1.8. Enhanced Sampling Methods

The scope of atomistic molecular simulations has been greatly extended through the development of enhanced sampling methods, which enable the exploration of rare events and long-timescale phenomena that are otherwise inaccessible to conventional molecular dynamics (MD) due to energetic barriers and slow transitions. These techniques accelerate the sampling of relevant configurations by focusing on a reduced set of collective variables (CVs)—low-dimensional descriptors that capture the essential dynamics of a system [99].

Choosing appropriate CVs is central to the success of enhanced sampling. Most CV discovery strategies fall into two broad categories: those that identify slow CVs, which preserve the long-timescale dynamics, and those that identify high-variance CVs, which capture the largest structural fluctuations. Principal Component Analysis (PCA) is the most widely known high-variance approach, whereas time-lagged Independent Component Analysis (tICA), Markov State Models (MSMs), and the Variational Approach to Conformational Dynamics (VAC) are commonly used for identifying slow modes [100, 101].

Traditionally, CVs were chosen manually using chemical intuition and empirical knowledge. However, for complex biomolecular systems, where CVs may involve nontrivial combinations of atomic motions, this manual process becomes infeasible. In recent years, machine learning (ML) methods—both supervised and unsupervised—have revolutionized CV discovery by automating this process and extracting interpretable low-dimensional representations directly from simulation data [102–105].

Once identified, these data-driven CVs can be used to bias the sampling via methods such as metadynamics, variationally enhanced sampling (VES), umbrella sampling, or adaptive biasing force methods. Beyond their utility in accelerating sampling, CVs also provide a powerful lens for interpreting the free energy landscape and for guiding molecular design, allosteric site prediction, or transition path discovery.

2.1.9. AlphaFold

AlphaFold is a deep learning-based artificial intelligence system developed by DeepMind that revolutionized protein structure prediction by achieving atomic-level accuracy in predicting three-dimensional structures from amino acid sequences [63, 64]. The success of AlphaFold lies in its attention-based neural network architecture, which interprets protein folding as a graph problem where residues are nodes and edges represent spatial or evolutionary constraints. By incorporating information from multiple sequence alignments (MSAs), homologous sequences, and pairwise residue distances, AlphaFold iteratively refines its predictions through a structure module and evaluates their reliability using an internal confidence metric.

The breakthrough performance of AlphaFold2 in the CASP14 competition demonstrated, for the first time, a computational method capable of producing predictions comparable to experimental structures, even in the absence of close homologs [64, 106]. This has enabled the large-scale release of predicted structures for over 200 million proteins, expanding structural coverage across the protein universe and transforming fields such as drug design, functional annotation, and variant effect prediction [107].

Building on these advances, AlphaFold3 introduced a diffusion-based architecture capable of predicting the joint structure of biomolecular complexes, including proteins, nucleic acids, small molecules, and ions. This unified model outperforms specialized tools in areas like protein–ligand and protein–nucleic acid interactions and significantly improves antibody–antigen complex predictions [65]. These capabilities have opened new directions in structural biology and molecular interaction modeling.

In parallel, AlphaFold-Multimer has been effectively used to predict protein–protein complexes and even cross-kingdom plant-pathogen interactions, as shown by Homma et al., who identified secreted pathogen proteins targeting tomato defense enzymes. Such examples demonstrate the system's potential not only in structural prediction but also in interaction screening and functional inference [108].

Despite these advances, challenges remain, AlphaFold does not explicitly model protein dynamics, disorder, or alternate conformational states. Nonetheless, its predictive accuracy and accessibility have made it a cornerstone of modern structural bioinformatics.

2.2. Evolutionary Units in Proteins

The protein universe, shaped by billions of years of evolution, spans an immense range of sequences, structures, and functions. With millions of species and thousands of protein-coding genes per species, the diversity of protein structures is vast—yet beneath this diversity lies a surprising degree of structural and sequence similarity. To navigate and interpret this complexity, systematic classification schemes have been developed that reflect evolutionary relationships. These hierarchical frameworks are essential not only for understanding how proteins evolve but also for enabling structure and function prediction based on homology [13, 109].

Traditionally, such classifications have focused on protein domains, which are structurally and functionally autonomous units often conserved across evolutionary lineages. Prominent databases like SCOP [14, 15], CATH [16, 17], and ECOD [13, 18] have cataloged domains based on similarities in fold, sequence, and function, providing a foundational view of protein evolution at the domain level. However, a growing body of research has shifted attention to subdomain units—recurrent segments smaller than domains that appear to be reused across disparate protein contexts. These studies suggest that the evolutionary building blocks of proteins extend beyond domains and may involve short, conserved segments that have been recombined over time to generate the modern protein repertoire [25, 33, 110].

This emerging perspective challenges the long-standing domain-centric view by proposing that many proteins are mosaics of ancestral segments. These subdomain evolutionary units—sometimes referred to as themes or building blocks—may represent an even more ancient layer of protein evolution, predating the emergence of whole

domains. In the following sections, we explore both the classical concept of domains and recent discoveries concerning these smaller conserved fragments, discussing their evolutionary significance and their role in shaping the structure and function of modern proteins.

2.2.1. Domains

The term domain in protein science refers to a recurring structural and evolutionary unit, though its precise definition can vary across contexts. Generally, a protein domain is described as a compact, semi-independent region within a polypeptide chain that folds into a stable structure and can often carry out specific functions on its own [6, 111]. Some proteins consist of a single domain, while others are multidomain proteins, in which several domains combine to enable more complex or regulated functions. From an evolutionary perspective, domains are considered fundamental units of protein architecture, capable of recombination and reuse across different proteins and species.

Domains represent the most well-studied level of modularity and reuse in protein evolution. Despite the lack of a universal definition, they are typically recognized as structural segments that fold independently and occur in various combinations throughout the protein universe [10]. Their conservation across diverse proteins highlights their evolutionary success as reusable building blocks.

To manage the vast diversity of domains, several domain classification databases have been developed, each employing distinct criteria based on sequence, structure, and evolutionary relationships. Among the most widely used are SCOP, CATH and ECOD. SCOP (Structural Classification of Proteins), which categorizes domains through expert visual inspection and fold hierarchy [14,15], CATH (Class, Architecture, Topology, Homologous superfamily), which applies structural and sequence-based clustering using algorithms like SSAP and HMMER3 [16, 17], ECOD (Evolutionary Classification of Protein Domains), which emphasizes remote homology and is continuously updated

with new structural data from the PDB [13, 18].

These databases serve as essential resources for comparative studies in protein evolution, structure-function analysis, and annotation of newly determined protein structures. Consequently, when an entity is referred to as a domain in literature, it is often defined according to one or more of these curated classification systems.

2.2.1.1. CATH Domains. The CATH database is a widely used, openly accessible resource that classifies protein domains based on their three-dimensional structures, sequence similarity, and inferred evolutionary relationships. It identifies structural domains within protein chains and organizes them into homologous superfamilies when sufficient evidence of common ancestry is detected [16, 17].

To uncover these relationships, CATH employs a combination of structure- and sequence-based comparison tools, including the Sequential Structure Alignment Program (SSAP) and profile Hidden Markov Models implemented via HMMER3. Following these analyses, domains are categorized into a hierarchical scheme consisting of four levels: Class, based on secondary structure content; Architecture, reflecting the overall spatial arrangement of secondary structures; Topology, which captures fold-level structural connectivity; and Homologous Superfamily, indicating domains with a shared evolutionary origin.

This hierarchical classification allows increasingly specific groupings as one moves from broader structural features (Class) toward detailed evolutionary relationships (Homologous Superfamily). Furthermore, for each superfamily, CATH provides structural superpositions of representative protein domains using SSAP, enabling precise structural comparisons and supporting functional inference across protein families. The CATH system thus serves as a powerful framework for understanding the organization and evolution of protein structures in a biologically meaningful way.

2.2.1.2. SCOP Domains. The SCOP (Structural Classification of Proteins) database organizes proteins with known three-dimensional structures based on their structural and evolutionary relationships. Unlike automated systems, SCOP classification has traditionally relied on manual curation, involving visual inspection and expert analysis to identify meaningful groupings of protein domains [14, 15].

SCOP assigns protein domains to a hierarchical scheme that reflects both structural similarity and evolutionary proximity. At its core are the family and superfamily levels: families cluster domains with clear evolutionary relationships and high sequence similarity, while superfamilies group together domains that share structural and functional features despite limited sequence identity, suggesting more distant evolutionary connections. Superfamilies are further organized into folds, which represent distinct combinations of secondary structure topology, architecture, and overall structural layout. Folds, in turn, are grouped into broader classes based on general secondary structure composition (such as all-alpha, all-beta, or alpha/beta classes).

An important distinction in SCOP is that while a fold typically characterizes a superfamily, some families within the same superfamily may diverge structurally to the extent that they are assigned to different folds. This reflects the evolutionary plasticity of protein structures and underscores the limitations of defining rigid fold boundaries.

Unlike other classification systems, SCOP does not include optional intermediate levels; each protein domain is unambiguously assigned to a single family, superfamily, fold, and class. This rigid yet interpretable hierarchy has made SCOP a foundational resource for structural bioinformatics, particularly in studies of distant homology, fold evolution, and structural genomics.

2.2.1.3. ECOD Domains. The ECOD (Evolutionary Classification of Protein Domains) database provides a hierarchical classification of protein domains based on their inferred evolutionary relationships. Unlike SCOP or CATH, which emphasize fold topology or structural similarity, ECOD places a stronger focus on distant homology as the main

organizing principle. Originally limited to proteins with experimentally determined structures from the Protein Data Bank (PDB), ECOD has recently expanded its scope to include AlphaFold-predicted structures, thereby substantially increasing the number of analyzable domains and offering deeper coverage across the proteome [13, 18].

One of ECOD's distinguishing features is its tight synchronization with the PDB, maintained through an automated classification pipeline that processes new entries on a weekly basis. This pipeline enables rapid and consistent incorporation of both experimental and predicted structures, making ECOD uniquely current among structural classification resources.

2.2.2. Sub-domain Units

While domains have long been regarded as the primary evolutionary units of proteins, growing evidence suggests that smaller building blocks—subdomain-level fragments—have played a foundational role in shaping protein structure and function. These fragments, often referred to as *themes*, *motifs*, or *ancient peptides*, may predate domains and persist across structurally diverse protein families, pointing to a deeper evolutionary origin.

One of the earliest systematic efforts to trace such units was conducted by Farías-Rico et al., who used HMM-based sequence searches to identify a 40-residue fragment shared between TIM-barrel and flavodoxin-like folds—suggesting a hidden evolutionary link between structurally distinct proteins [32]. Building on this concept, Alva et al. expanded the search to a larger dataset and identified a recurring vocabulary of 40 ancient peptides—subdomain fragments up to 38 residues long—found across a wide range of fold families [25]. These fragments were proposed to be remnants of primordial peptides that may have served as modular precursors to modern domains.

A more scalable and quantitative approach was introduced by Nepomnyachiy et al., who defined themes as sequence–structure segments of at least 35 amino acids,

reused across otherwise unrelated proteins [33]. Their pipeline used HHsearch to detect remote homology based on both sequence and structural similarity, uncovering thousands of theme instances in a dataset of over 20,000 domains. Unlike domains—which are non-overlapping and typically 100 residues long—these themes can overlap and vary in size, offering a more flexible and fine-grained view of evolutionary reuse.

Building on this, Ferruz et al. introduced additional filtering and clustering strategies to identify a broader range of reused fragments, resulting in the creation of the Fuzzle database, which comprises over eight million reuse instances and provides one of the most extensive catalogs of subdomain-level reuse in the protein universe [110].

Collectively, these studies reveal a striking pattern: reuse frequency increases as fragment length decreases [112]. This trend suggests that shorter reused segments—such as themes—may be vestiges of a pre-domain evolutionary layer, representing modular building blocks that were recombined and elaborated to generate modern domain architectures [25, 33, 112]. As further highlighted by Romero-Romero et al. and others [34, 112], the recurrence of these themes across distinct folds may reflect conserved principles of protein folding, function, and evolutionary fitness.

Among these, bridging themes—shared segments found across structurally unrelated domain folds—offer particularly compelling evidence of deep evolutionary reuse. Some bridging themes have been shown to retain ligand-binding functions, underscoring their potential role as functionally conserved modules. Studies by Kolodny et al. propose that these conserved fragments may be molecular remnants of early peptide-based ancestors, retained across evolutionary timescales due to their structural adaptability or biochemical utility [29, 112].

In this study, we adopt the theme definition introduced by Nepomnyachiy et al. (2017) as our operational concept of subdomain units [33]. These themes, defined by their sequence and structure similarity and identified across diverse domain contexts, provide a flexible and evolutionarily grounded framework for investigating the modular

architecture and dynamic behavior of proteins. A detailed analysis of the themes used in this work and their dynamic significance is provided in the sections that follow.

2.2.2.1. Themes as Subdomain Evolutionary Units. A major conceptual advancement in identifying subdomain-level evolutionary relationships is the introduction of themes, short protein segments shared across otherwise unrelated proteins. These segments offer a more granular and flexible view of homology than the traditional domain-centric framework. While domains partition the protein sequence into non-overlapping, functionally independent regions, themes permit residue-level overlap, meaning that a single amino acid may simultaneously belong to multiple themes of varying length, origin, or evolutionary depth [33,112]. This non-exclusive assignment allows for a richer depiction of evolutionary reuse, capturing both localized and distributed patterns of conservation across protein space.

In their foundational study, Nepomnyachiy et al. (2017) used HHsearch [35] to detect shared segments of at least 35 amino acids, exhibiting statistically significant similarity in both sequence and structure [33]. These themes are not identical repetitions like reused domains but represent homologous motifs that vary in sequence while retaining core structural and evolutionary features. Their recurrence across disparate domain and fold contexts suggests that many themes may represent relics of ancient peptide fragments, predating the emergence of fully folded domains. Moreover, the ability of a theme to appear in multiple contexts—ranging from closely related proteins to structurally unrelated folds—underscores the layered complexity of protein evolution and highlights the modular nature of these reused segments.

2.2.2.2. Bridging Themes. A particularly compelling subset of themes is the class of bridging themes, first characterized by Kolodny et al. (2021) [29]. These are homologous protein segments identified across domains that belong to distinct folds, including some of the most ancient structural classes such as the Rossmann, TIM-barrel, P-loop NTPase, and flavodoxin folds. The presence of these segments in domains that are

otherwise thought to have evolved independently suggests either a deep evolutionary origin or convergent structural reuse. Kolodny and colleagues identified 525 bridging themes distributed across 73 different folds, highlighting the widespread nature of this phenomenon [29].

Interestingly, approximately half of these bridging themes exhibit structural variation, indicating that while the core sequence similarity is retained, the segments undergo conformational adaptation depending on their structural context. This plasticity suggests that bridging themes are not merely passive remnants of common ancestry, but actively integrated elements capable of functional or structural accommodation.

To explain their evolutionary emergence, two plausible scenarios have been proposed. In one, a short ancestral peptide fragment may have been independently recruited into different structural contexts via accretion or fusion, giving rise to distinct modern domains that nevertheless share a common segment. Alternatively, a segment from an existing domain may have been duplicated or transferred into a new protein environment, where it subsequently evolved into part of a new fold. In either case, the modular and adaptable character of themes allows them to function as architectural bridges between protein lineages—hence the term bridging themes [29, 112].

2.2.2.3. Adenine-Binding Themes. Beyond their structural recurrence, some themes have been linked to specific biochemical functions. A prominent example is the set of themes involved in binding to adenine, a molecule that plays a central role in metabolism and signaling and is considered one of the most ancient ligands in biology. In a comprehensive study, Narunsky et al. (2020) investigated the distribution and characteristics of themes located at adenine-binding sites across a wide range of protein families [36].

Their analysis revealed that similar themes often recur in ligand-interacting regions, and that these themes exhibit conserved interaction patterns at the atomic level, particularly in their hydrogen-bonding configurations with adenine. Notably,

these themes were found in proteins belonging to structurally distinct folds, suggesting that they have been reused across independent evolutionary trajectories to mediate a shared functional role.

The implications of these findings are twofold. First, they indicate that some themes are not only structurally conserved but also act as functionally conserved modules, retaining specific interaction capabilities that are preserved through evolutionary time. Second, the recurrence of these adenine-binding themes across unrelated proteins supports the notion that they may have originated very early in protein evolution, serving as foundational elements in the development of modern biochemical functions. As such, these themes may be regarded as molecular fossils, preserving a trace of the biochemical logic that shaped early protein-ligand interactions. [36]

2.3. Mechanisms of Protein Evolution and Links to Dynamics

Understanding how proteins evolve—both structurally and functionally—requires integrating insights from sequence conservation, structural adaptation, and conformational dynamics. While early views of protein evolution focused primarily on sequence divergence and domain recombination, recent studies have revealed a more nuanced landscape, emphasizing the roles of fragment reuse, structural metamorphosis, and dynamical constraints in shaping the protein universe [112].

For example, Longo et al. (2022) proposed a "peptide budding" mechanism to explain the emergence of the β-trefoil fold from an immunoglobulin-like β-sandwich, illustrating how ancestral fragments can be modularly transformed into new folds [113]. Similarly, Alvarez-Carreño et al. (2022) described a process of "creative destruction" to account for the diversification of β-barrel folds within the translation machinery, suggesting a shared ancestral origin. These mechanisms support the view that structural innovation often results from the rearrangement and repurposing of ancient fragments, rather than de novo formation of entirely new topologies [114].

Complementing these structural models, increasing attention has been given to the dynamic properties of evolving proteins. Studies by Campitelli et al. (2020) and Zhang et al. (2020) suggest that conformational dynamics are not passive consequences of structure, but may be evolutionarily optimized features [51, 69]. Indeed, patterns of flexibility and long-range allosteric communication have been observed across homologous proteins with diverged sequences, implying that dynamics can be conserved even in the absence of strict structural or sequence similarity. This notion was already anticipated by Maguid et al. (2005), who demonstrated that dynamical features are preserved among globin homologs despite significant evolutionary divergence in structure [49].

Collectively, these studies underscore the evolutionary importance of conserved dynamics—especially in the context of reused structural fragments. In the following chapters, this concept is explored further through a comparative analysis of subdomain themes and their dynamic behavior across diverse protein families.

3. MATERIALS AND METHODS

3.1. Themes: Identification and Dataset Construction

Themes are evolutionarily conserved protein segments shared across diverse proteins, identified solely through sequence similarity without requiring global structural or functional similarity [33]. These segments, typically ranging from 35 to 100 amino acids, are curated in the Themes Database of Reuse in Proteins [115], which currently includes 2,195 themes mapped to 3,798 ECOD domains. Each protein domain can host multiple themes, and in some cases, a theme may appear more than once within the same domain; such repeated occurrences are uniquely indexed (e.g., 14815-1 through 14815-4). An accompanying interactive resource, developed as part of this thesis [116], was used to link structural reuse to dynamic signatures.

To systematically investigate the dynamic and functional relevance of theme reuse, two datasets were constructed: an in-depth set of 13 ECOD domains (8 repeat and 5 non-repeat domains) selected for structural diversity and detailed analysis, and an expanded set of 150 ECOD domains drawn from 26 ECOD H-groups. Domains in the expanded set were chosen based on high theme coverage and structural non-redundancy (RMSD <1.25 Å), and filtered to ensure compatibility with dynamic decomposition criteria. These datasets formed the structural basis for analyses throughout this study. Full dataset composition and selection criteria are described in Section 4.1.2.

3.1.1. Adenine-binding themes

Adenine-binding themes are conserved protein segments involved in ligand recognition of adenine-containing molecules such as ATP, NAD, or FAD. These themes were identified by Narunsky et al. (2020) through a large-scale network analysis based on sequence and structural similarities across adenine-binding sites in the Protein Data Bank. The resulting adenine-binding network includes 756 themes, of which 537 are

reused across multiple binding regions, suggesting evolutionary conservation and potential allosteric relevance. To extract biologically meaningful patterns, the authors grouped these themes into 10 major clusters, each representing distinct binding motifs—such as the direct, reverse, or Asp motif—and ligand interaction geometries. [36] In this study, we focused on four representative clusters from the original classification: Cluster 1 (direct motif), Cluster 2 (direct motif with conserved lysine interaction), Cluster 4 (reverse motif), and Cluster 6B (NAD-binding "Asp" motif). After applying consistent filtration criteria to ensure structural quality, we selected 76 structures and their associated adenine-binding themes across these clusters for detailed analysis. The resulting dataset includes 9 themes from Cluster 1, 15 themes from Cluster 2, 16 from Cluster 4, and 7 from Cluster 6B. These themes were structurally mapped to corresponding PDB entries and served as input for GNM-based dynamic analysis and transfer entropy calculations, enabling evaluation of their dynamic coherence, information flow potential, and possible allosteric roles. The full dataset generation and selection procedure is described in detail in Section 4.2.2.

3.2. Gaussian Network Model (GNM)

To investigate intrinsic dynamics of protein structures, we employed the Gaussian Network Model (GNM) [52, 53], which models a protein as a three-dimensional elastic network. In this model, residues (represented by their Cα atoms) are treated as nodes connected by springs if they lie within a cutoff distance (r_c). The network captures collective motions by analyzing residue fluctuations around the equilibrium conformation.

The model describes equilibrium cross-correlations between the fluctuations of residues i and j as

$$\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j \rangle = \left(\frac{3k_bT}{\gamma} \right) [\Gamma^{-1}]_{ij}, \tag{3.1}$$

where Γ is a symmetric matrix known as a Kirchhoff (connectivity) matrix. γ is

the force constant of the Hookean pairwise potential function, which represents the interactions between the residues in the folded structure. T is the absolute temperature in Kelvin degrees, and k_b is the Boltzmann constant. The matrix Γ is defined as

$$\Gamma_{ij} = \begin{cases} -1 & \text{if } i \neq j \text{ and } R_{ij} \leq r_c, \\ 0 & \text{if } i \neq j \text{ and } R_{ij} > r_c, \\ -\sum\limits_{i,i\neq j} \Gamma_{ij} & \text{if } i = j, \end{cases} \tag{3.2}$$

where R_{ij} is the distance between the Cα atoms of i-th and j-th amino acids, and the r_c threshold defines whether they are close enough to interact. r_c =10 Å is used here.

The inverse of Γ, excluding the zero eigenvalue, is obtained through its eigen decomposition as

$$\Gamma^{-1} = U(\Lambda^{-1})U^T, \tag{3.3}$$

where U is the matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues λ_k. The correlation of fluctuations can thus be decomposed into contributions from individual modes as

$$\langle \Delta R_i \cdot \Delta R_j \rangle = \left(\frac{3k_bT}{\gamma} \right) \sum_{k=1}^{n-1} \left[\lambda_k^{-1} u_k u_k^T \right]_{ij}, \tag{3.4}$$

where, n is the number of residues, and the summation excludes the trivial zero-frequency mode (k = 1) corresponding to global translation. This formulation allows us to analyze both the total and mode-specific dynamics of protein structures.

3.3. Identification of Dynamic Elements from GNM Modes

Each GNM mode describes a distinct pattern of residue-residue correlations, reflecting the concerted motion of structural elements. For an individual mode, a sym-

metric correlation matrix is generated, where each entry indicates whether residue i and j fluctuate in the same sense (positive correlation, +1) or opposite sense (negative correlation, -1) [72]. Thus, each mode naturally decomposes the structure into two dynamically coherent parts, one moving in the positive direction and the other in the negative, with residues fluctuating in opposite directions around hinge regions. These dynamic parts can be visualized and mapped across the protein's architecture, as shown in a simplified toy model (Figure 3.1).

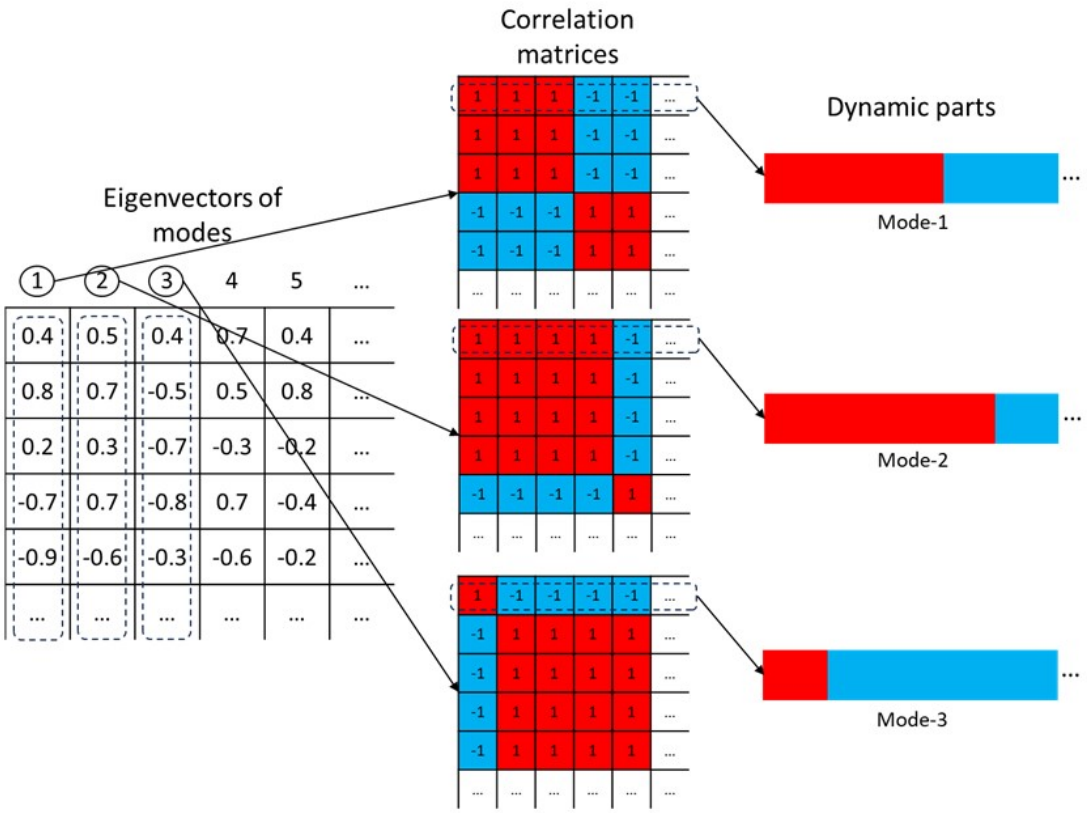


Figure 3.1. A toy model to represent decomposition of the structure into the dynamic parts from a correlation matrix of GNM.

To define robust structural units, residue-wise mode-based correlation patterns were projected onto the protein sequence. Continuous stretches of residues belonging to the same dynamic part were grouped into dynamic segments. To avoid noise from small-scale fluctuations, segments shorter than 15 residues were merged with their adjacent longer segments. The resulting larger structural units are termed dynamic elements (DEs) (Figure 3.2) [59].

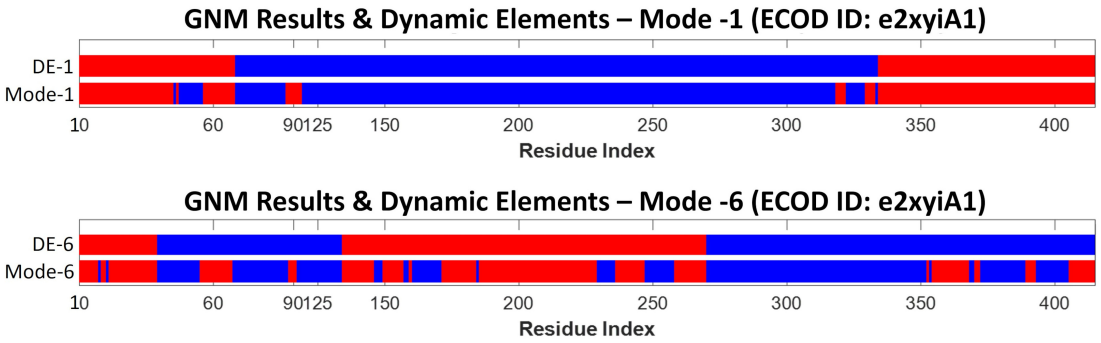


Figure 3.2. Dynamic segments and DEs in the e2xyiA1 propeller (from the CAF1 protein, PDB ID: 2XYI). For representative purposes only two of the slow modes of motion are shown: The first and sixth.

Since the slowest modes (i.e., those with the smallest eigenvalues) capture the most cooperative and global structural changes, we restricted our DE extraction to the seven slowest non-trivial modes. Beyond this threshold, modes tend to contribute less to global dynamics and yield fragmented segments that fail to meet the 15-residue criterion (Figure 3.3). This cutoff ensures that the identified DEs reflect biologically meaningful and statistically tractable structural dynamics. [59]

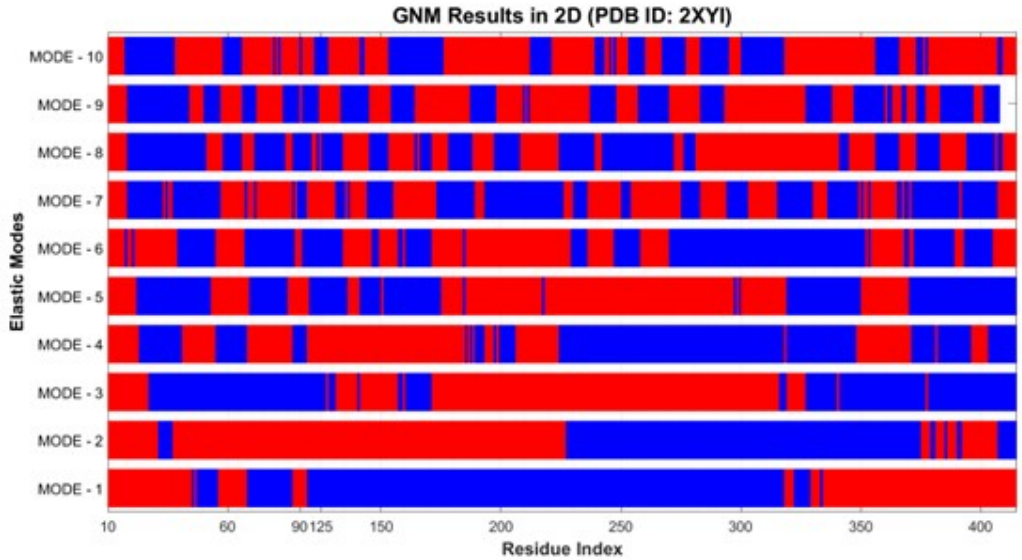


Figure 3.3. The ten slowest elastic modes of the e2xyiA1 propeller. We observe that the prevalence of larger segments decays with progression through the dynamic spectrum

3.4. GNM based Transfer Entropy Method

To investigate directional communication between residues, we employed the GNM-based Transfer Entropy (GNM-TE) framework [55, 56]. This method integrates transfer entropy (TE), an information-theoretic measure of causality, into the GNM formalism to estimate how fluctuations at one residue affect the future fluctuations of another. For any residue pair (i, j), the directional TE at a given time delay τ quantifies the extent to which knowledge of the past motion of residue i reduces the uncertainty about the future motion of residue j, beyond what is already predictable from j's own past.

Mathematically, TE is computed from the difference in conditional entropies as

$$T_{i\to j}(\tau) = S(\Delta R_j(t+\tau) \mid \Delta R_i(t)) - S(\Delta R_j(t+\tau) \mid \Delta R_i(t), \Delta R_j(t)), \tag{3.5}$$

where S represents the conditional entropy. To account for directionality, the net transfer entropy between residue pairs computed with the following equation

$$\Delta T_{i\to j}(\tau) = T_{i\to j}(\tau) - T_{j\to i}(\tau). \tag{3.6}$$

To evaluate the collective impact of a residue's information transfer across the structure, a collectivity measure inspired by Bruschweiler (1995) is adopted [117]. This metric, computed over a selected subset of GNM modes, reflects how broadly a residue disseminates dynamic influence. The collectivity of residue i, K_i is calculated as

$$K_{i,s} = \frac{1}{N}\exp\left(-\sum_{j=1}^{N}\alpha(\Delta T_{i\to j,s}(\tau))^2\log\left[\alpha(\Delta T_{i\to j,s}(\tau))^2\right]\right), \tag{3.7}$$

where s refers to the chosen set of slow GNM modes, and α is a normalization factor which ensures the following equation.

$$\sum_{j=1}^{N}\alpha(\Delta T_{i\to j,s}(\tau))^2 = 1. \tag{3.8}$$

Finally, to integrate both the magnitude and collectivity of information transfer, the TEcol score for is defined for each residue with the following equation

$$\text{TEcolScore}_{i,s} = K_{i,s}\cdot\sum_{j=1}^{N}\Delta T_{i\to j,s}. \tag{3.9}$$

This score highlights residues that act as globally collective information hubs, and is used to identify candidate sites for allosteric regulation. This allows identification of residues displaying collective net positive information transfer as global information sources and residues displaying collective net negative information transfer as global information sinks [56–58].

3.5. Anisotropic Network Model (ANM)

To capture direction-specific collective motions in protein structures, we used the Anisotropic Network Model (ANM), a three-dimensional extension of elastic network models that incorporates anisotropic fluctuations [54, 81]. Unlike the Gaussian Network Model (GNM), which assumes isotropic fluctuations and yields N - 1 modes for a protein of N residues, ANM considers directional flexibility and produces 3N - 6 non-trivial normal modes by accounting for the full three-dimensional position vectors of Cα atoms. The ANM potential energy function is defined as

$$V_{ANM} \;=\; \frac{1}{2}\sum_{ij}\gamma_{f,ij(R_{ij}-R^0_{ij})}, \tag{3.10}$$

where R_{ij} is the equilibrium distance between residues i and j, R^0_{ij} is the instantaneous distance between residues i and j and γ_f is the force constant. The global motion behavior is then derived by diagonalizing the Hessian matrix, composed of second derivatives of the potential with respect to atomic coordinates. The correlation of positional fluctuations between atoms i and j is computed as

$$< \Delta R_i.\Delta R_j > = \;\frac{3k_BT}{\gamma}tr\lfloor H^{-1}\rfloor \;=\; \frac{3k_BT}{\gamma_f}\sum_{k=1}^{N}tr\lfloor\lambda_k^{-1}u_k u_k^T\rfloor_{ij}, \tag{3.11}$$

here λ_k and u_k are the eigenvalues and eigenvectors of the Hessian matrix, respectively, and the summation excludes the six trivial modes (three translational and three rotational). ANM is especially useful for modeling large-scale conformational changes, as it enables the perturbation of structures along low-frequency collective modes to generate plausible alternate conformations [2, 40].

3.6. ANM-LD Simulations: ANM-Driven Langevin Dynamics

The ANM-LD (Anisotropic Network Model–Langevin Dynamics) method is a hybrid simulation approach that integrates the collective, low-frequency motions pre-

dicted by ANM with the stochastic framework of Langevin dynamics to generate physically plausible transition pathways between two experimentally determined protein conformations [97, 98]. By iteratively perturbing the structure along selected ANM modes, ANM-LD efficiently samples large-scale functional transitions while maintaining computational tractability.

The procedure requires both an initial and a target structure. At each iteration, the method computes the difference vector D between the current structure and the target and identifies the ANM mode uj that best aligns with this vector. The optimal mode is selected based on its overlap score with the following equation

$$O_{i,j} = \frac{\mid u_j.D_i \mid}{\mid u_j \mid . \mid D_i \mid}. \tag{3.12}$$

Once the best-overlapping mode is selected, the structure is perturbed along this mode by a user-defined deformation factor, producing a new intermediate conformation. This updated structure becomes the starting point for the next cycle. The convergence of the pathway is evaluated using the Root-Mean-Square Deviation (RMSD) between the intermediate and target structures, calculated as

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}}, \tag{3.13}$$

where di is the Euclidean distance between the coordinates of residue i in the current and target structures. The simulation continues iteratively until the RMSD falls below a pre-defined threshold, ensuring a continuous and directionally meaningful transition pathway. By constraining the system to deform along functionally relevant ANM modes, ANM-LD generates physically plausible intermediates that capture the essence of global, cooperative motions.

3.7. Mutual Information-Based Clustering Evaluation

To quantify the agreement between theme-based and dynamics-based residue groupings, we applied Mutual Information (MI) and its two derived metrics: Adjusted Mutual Information (AMI) and Standardized Mutual Information (SMI). MI captures the amount of information shared between two clusterings, whereas AMI normalizes MI by accounting for the expected similarity between random clusterings [118], and SMI standardizes MI based on its variance under the null model, expressing deviation in terms of standard deviations [119].

To evaluate the overlap between dynamic elements and themes, we constructed a k by l contingency matrix, where k and l represent the number of clusters in the dynamic (D) and theme-based (T) segmentations, respectively. Each matrix entry, nij, denotes the number of residues shared between the ith dynamic element and the jth theme. This matrix forms the basis for computing Mutual Information and its adjusted and standardized variants. All clustering comparisons were performed on residue-level datasets of equal length, ensuring that both D and T partitioned the same set of residues.

In our analysis, we compare dynamic elements (D) defined from the slow modes of motion in GNM with themes (T) which are identified solely through sequence similarity. Because dynamic elements span the entire residue sequence of a protein domain, it is essential that the themes collectively cover as much of the protein structure as possible to allow for meaningful clustering comparisons. However, themes often exhibit gaps and overlaps, which complicates their direct use as cluster partitions. To address this, we applied coverage heuristics using variable threshold values: overlaps of 3 and 5 residues, and gaps of 8, 10, and 15 residues. These criteria were used to construct sets of theme sequences that approximate full coverage of the domain. We employed a Monte Carlo-like sampling algorithm to explore all valid combinations under each threshold scheme. For instance, applying the three overlap-gap combinations to the CAF1 domain (2XYI) produced 147, 755, and 2,218 unique theme arrangements, respectively.

The mutual information between two clusterings (D) and T (themes) is computed using entropy-based formulation as

$$MI(D,T) \;=\; H(D) \;+\; H(T) \;-\; H(D,T), \tag{3.14}$$

where H(D) and H(T) are the marginal entropies of D and T, and H(D,T) is the joint entropy of D and T. Adjusted Mutual Information (AMI) and Standardized Mutual Information (SMI) are calculated using expectations and variance under a random clustering model [119].

3.7.1. Adjusted and Standardized Mutual Information Calculations

To account for the mutual information expected by chance, two corrected forms are used: Adjusted Mutual Information (AMI) and Standardized Mutual Information (SMI). These metrics enable robust evaluation of the similarity between two clusterings when the number and size of clusters differ or overlap partially.

The AMI is calculated as

$$AMI(D,T) \;=\; \frac{MI(D,T) \;-\; E\{MI(D,T)\}}{\sqrt{H(D)\cdot H(T)} \;-\; E\{MI(D,T)\}}, \tag{3.15}$$

where $E\{MI(D,T)\}$ is the expected mutual information under a null model of random clusterings with fixed marginals.

The SMI further standardizes MI by its standard deviation under the null model as

$$SMI(D,T) \;=\; \frac{MI(D,T) \;-\; E\{MI(D,T)\}}{\sqrt{Var\{MI(D,T)\}}}, \tag{3.16}$$

where $Var\{MI(D,T)\}$ is the variance of mutual information under random clustering. In our analysis, these values are computed using the MATLAB implementation

provided by Romano et al. (2014) [119], ensuring statistically meaningful comparison between clusterings of dynamic elements and themes.

3.7.2. Monte Carlo-Based Statistical Validation

To statistically validate the observed mutual information (MI) values between dynamic elements and themes, we implemented a Monte Carlo-based randomization strategy. For each protein domain, 1,000 valid sets of randomly sampled consecutive amino acid segments were generated, where each set contained on average 30 segments—matching the average number of themes per domain. Segment lengths were uniformly sampled between 30 and 175 residues, consistent with the observed range of theme lengths in our dataset. A randomly generated set was accepted only if at least one combination of its segments could fully cover the domain, ensuring feasibility for MI calculations.

Using the MATLAB implementation by Romano et al. (2014) [119], MI values were computed for each of the 1,000 random sets per domain. The maximum MI value obtained from the actual theme decomposition was subsequently compared to the empirical distribution of maximum MI values derived from the randomly generated sets. One-tailed p-values were computed as the probability of observing a mutual information value greater than that obtained from the actual theme decomposition, based on the empirical distribution of MI values as

$$p\text{-value} \;=\; P(X > x) \;=\; 1 - P(X < x) \;=\; 1 - F(x), \tag{3.17}$$

where x is the maximum MI obtained with the real themes and F(x) is the cumulative distribution function of MI values from the random sets. This procedure was repeated for each individual mode per domain, enabling a robust statistical assessment of the observed correlations.

3.8. Fisher's Exact Test and Hypothesis Correction

To assess the statistical enrichment of communication classes among biologically defined binding themes, we performed Fisher's exact test independently for each communication class. This test, suitable for small sample sizes and categorical data, evaluates the non-random association between two binary variables [120], in this case, theme type (binding vs. non-binding) and communication class membership. For each test, a 2 by 2 contingency table was constructed (Table 3.1).

Table 3.1. Contingency table used in Fisher's exact test to evaluate the enrichment of binding themes within a specific communication class.

	In Communication Class	**Not in Communication Class**
Binding Theme	a	b
Non-Binding Theme	c	d

The p-value for each 2×2 contingency table was computed using MATLAB's built-in fishertest function, which calculates the hypergeometric probability as

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,(a+b+c+d)!}. \tag{3.18}$$

To correct for multiple hypothesis testing, we applied:

- Bonferroni correction, which controls the family-wise error rate by multiplying each p-value by the number of tests. This approach is conservative and minimizes the chance of false positives (Type I errors) [121].
- Benjamini–Hochberg false discovery rate (FDR) correction [122], which controls

the expected proportion of false positives among significant results and offers improved sensitivity in exploratory analyses.

A communication class was deemed significantly enriched if the adjusted p-value (Bonferroni or FDR) was below 0.05.

3.9. Principal Component Analysis (PCA) of MD Correlation Maps

To extract dynamic segments from MD simulations, Principal Component Analysis (PCA) was applied to the residue-residue correlation matrices derived from trajectory data [123]. PCA enabled the identification of major modes of fluctuation by projecting the correlation data onto its principal components. Residue clusters along the dominant principal axes were interpreted as coherent dynamic segments, which were then compared to those obtained from GNM-derived cross-correlation analysis. This approach allowed for a consistent evaluation of structural dynamics across different modeling frameworks.

4. RESULTS AND DISCUSSION

4.1. Reused Protein Segments Linked to Functional Dynamics

4.1.1. Motivation and Background

Proteins evolve not only through point mutations but also through the reuse and recombination of modular sequence segments. Such evolutionarily conserved fragments—referred to as themes—are short sequences that recur across structurally and functionally diverse proteins [33]. Themes are thought to represent ancestral building blocks that have been repurposed throughout evolution to contribute to essential biochemical functions such as ligand binding, catalysis, or structural scaffolding [12,28,29]. While the recurrence of themes in sequence and structure has been well documented, their possible link to intrinsic protein dynamics remains largely unexplored.

If proteins emerge via evolutionary processes that mix and match segments from different sources, the resulting structural arrangements must satisfy not only spatial compatibility but also dynamic compatibility. In other words, distinct parts must exhibit coordinated motion to ensure functional integration. Prior studies have shown that such dynamic coupling plays an essential role in long-range regulation—for example, between kinase domains and their regulatory regions [124], or across the membrane-spanning and ATPase domains in ABC transporters [98]. These examples reflect dynamic integration at the domain level; however, whether similar principles apply at the sub-domain level, such as for reused sequence fragments like themes, is still an open question.

Intrinsic dynamics, as captured by low-frequency normal modes, offer a powerful way to investigate this hypothesis. These soft modes represent collective motions encoded by the native structure and are often associated with conformational flexibility, substrate binding, or allosteric signaling [2, 39, 40, 68]. Importantly, such dynamics are

evolutionarily conserved: proteins within the same family often exhibit similar low-frequency motions [49, 51], and residues critical for allosteric regulation tend to be conserved across homologs [47, 48]. These findings suggest that functional dynamics may be under evolutionary selection, just as much as sequence or structure.

This chapter investigates the relationship between evolutionarily reused segments and intrinsic dynamics, focusing on whether themes correspond to dynamic elements (DEs) which are compact, cooperatively moving segments identified through the GNM [52, 53]. By analyzing a curated set of 150 ECOD domains enriched in themes, we assess whether the boundaries of these reused segments align with regions involved in coherent motion. If so, this would suggest that themes are not only sequence-based evolutionary modules, but also dynamically coherent units, supporting the idea that modularity in protein evolution extends into the dynamic realm.

4.1.2. Data and Approach

We compiled two datasets specifically for this work: an in-depth set and an expanded set (which encompasses the in-depth set). The in-depth set consists of eight repeat and five non-repeat ECOD domains used for detailed inspection. The expanded set consists of 150 ECOD domains belonging to 26 different ECOD H-groups used for systematic statistical analysis.

In-depth set: The set includes both repeat and non-repeat domains to cover maximal structural diversity. The repeat domains include eight large symmetrical protein domains from the database of reuse in proteins [33]. These domains, presented in Table 4.1, include obvious examples of repeated elements that result in symmetrical geometry: two homologous seven-blade beta-propellers—histone-binding protein CAF1 (PDB ID: 2XYI, ECOD Domain ID: e2xyiA1) and WD repeat-containing protein 5 (WDR5, PDB ID: 3EMH, ECOD Domain ID: e3emhA1); two homologous ARM repeats—ZYG-9 (PDB ID: 2OF3, ECOD Domain ID: e2of3A1) and protein phosphatase PP2A (PDB ID: 1B3U, ECOD Domain ID: e1b3uA1); a repetitive alpha hair-

pin—26S proteasome subunit Rpn2 (PDB ID: 4ADY, ECOD Domain ID: e4adyA2); and three homologous TIM barrels—TatD-related deoxyribonuclease (PDB ID: 1J6O, ECOD Domain ID: e1j6oA1), Putative TatD related DNAse (PDB ID: 2GZX, ECOD Domain ID: e2gzxA1), and Tat-linked quality control protein TatD (PDB ID: 4P5U, ECOD Domain ID: e4p5uA1). This sample includes protein domains of three very different architectures, all-beta, all-alpha and alpha/beta. Our sample of non-repeat domains includes five domains from two different ECOD H-groups. These domains, presented in Table 4.2, include two homologous RIP (ribosome-inactivating proteins) domains—ribosome-inactivating protein gelonin (PDB ID: 3KTZ, ECOD Domain ID: e3ktzA1) and rRNA N-glycosidase (PDB ID: 3CTK, ECOD Domain ID: e3ctkA1); and three homologous Rossmann-related domains—putative 3-oxoacyl-acyl carrier protein reductase (PDB ID: 1ULS, ECOD Domain ID: e1ulsB1), tropinone reductase-II (PDB ID: 2AE2, ECOD Domain ID: e2ae2A1), and 3-ketoacyl-(acyl-carrier-protein) reductase (PDB ID: 3N74, ECOD Domain ID: e3n74B1).

Expanded set: We systematically generated a significantly larger collection of domains characterized by enhanced structural diversity to facilitate a comprehensive statistical analysis. A paramount criterion guiding the selection of this set was the maximal coverage of domain sequences by themes, thereby enabling a meaningful investigation of their correlation with the dynamic components. To achieve this objective, we conducted a systematic scan of the database of reuse in proteins [33], and domains conforming to the specified criterion were identified. Among the domains meeting this criterion, elements with very high structural similarity from those in the same homology group (RMSD <1.25 Å) were removed from the set to eliminate redundancy. Finally, if the dynamic segments within multiple modes of a given domain were too short to allow decomposition into DEs with our threshold of 15 amino acids, that domain was removed from the set, to ensure robustness of the statistical analysis. We ended up with 150 ECOD domains belonging to 26 different ECOD H-groups.

Table 4.1. A set of eight ECOD domains from four ECOD H-groups, and three very different architectures, all-beta, all-alpha and alpha/beta (in-depth set repeats).

PDB ID	**Domain ID**	**H-Group Name**	**T-Group Name**	**Protein Name**
2XYI	e2xyiA1	Beta-propeller	7-bladed	Probable Histone-binding Protein CAF1
3EMH	e3emhA1			WD Repeat-containing Protein 5
2OF3	e2of3A1	ARM repeat	ARM repeat	ZYG-9
1B3U	e1b3uA1			Protein Phosphatase PP2A
4ADY	e4adyA2	Proteasome / cyclosome (PC) repeat	Proteasome / cyclosome (PC) repeat	26s Proteasome Regulatory subunit RPN2
1J6O	e1j6oA1	TIM barrels	TIM barrels	TatD-related deoxyribonuclease
2GZX	e2gzxA1			Putative TatD related DNAse
4P5U	e4p5uA1			Tat-linked quality control protein TatD

Table 4.2. A set of five ECOD domains from ribosome-inactivating proteins (RIP) and Rossmann-related ECOD H-groups (in-depth set non-repeats).

PDB ID	**Domain ID**	**H-Group Name**	**T-Group Name**	**Protein Name**
3KTZ	e3ktzA1	Ribosome inactivating proteins (RIP)	Ribosome inactivating proteins (RIP)	Ribosome-inactivating protein gelonin
3CTK	e3ctkA1			rRNA N-glycosidase
1ULS	e1ulsB1	Rossmann-related	NAD(P)-binding Rossmann-fold domains	Putative 3-oxoacyl-acyl carrier protein reductase
2AE2	e2ae2A1			Tropinone reductase-II
3N74	e3n74B1			3-ketoacyl-(acyl-carrier-protein) reductase

A flowchart of the approach used is provided in Figure 4.1. As noted above, our analysis focused on a set of 150 ECOD domains. All the calculations in this study are performed on these ECOD domain regions (rather than whole PDB structures). We conducted two main sets of analyses:

- an in-depth analysis on a small subset of 13 ECOD domains (the "in-depth set"), in which we examined possible correlations between themes and dynamics on a case-by-case basis; this dataset comprised 8 repeat domains (Table 4.1) and 5 non-repeat domains (Table 4.2).
- a broader statistical analysis of the complete set of 150 ECOD domains (the

"expanded set"), in which we sought to examine the statistical significance of the correlations observed in our in-depth set.

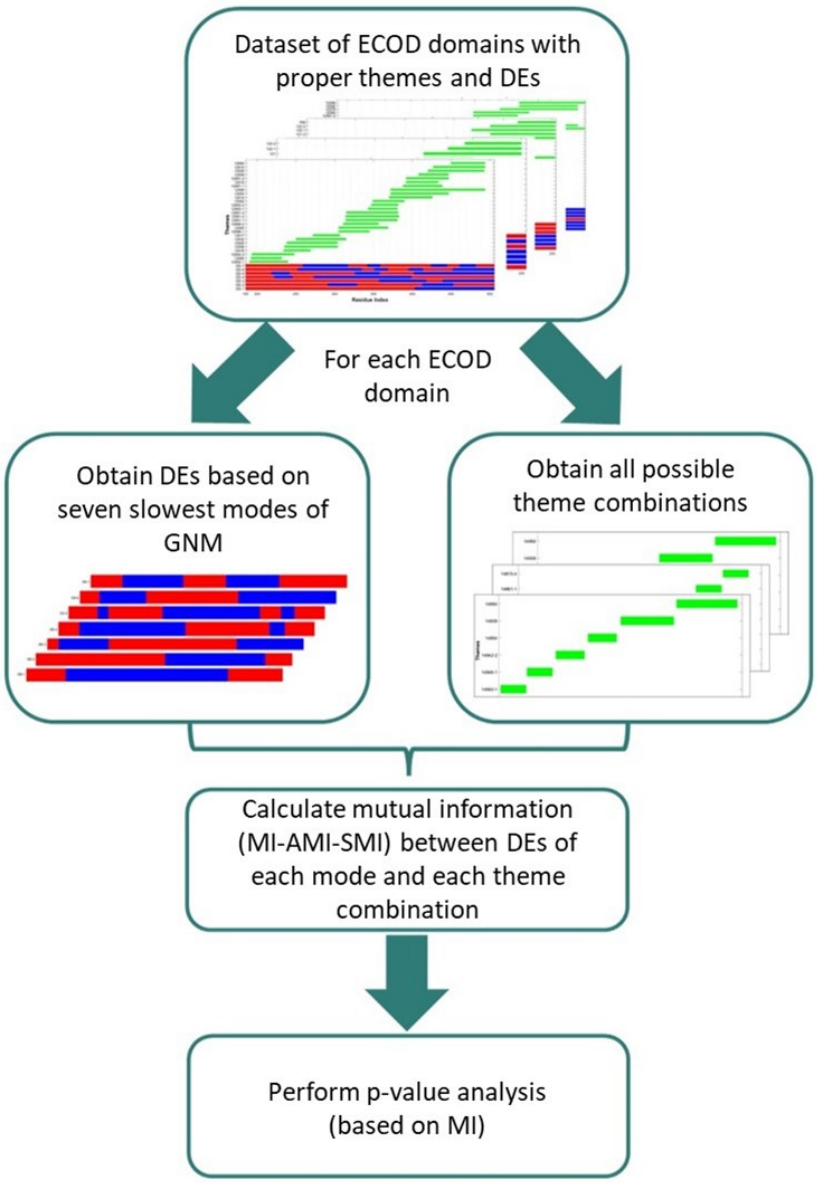


Figure 4.1. Schematic representation of the approach used here to examine the correlation between themes and dynamic elements.

For each domain in our dataset, we used GNM to approximate the domain's equilibrium dynamics based on its 3D structure [52,53]. In this model, the dynamics of

a protein of n amino acids is represented as a spectrum of (n - 1) orthogonal modes of motion. The modes of motion are sorted based on their contributions, from the slowest (the so-called soft modes) to the fastest, such that mode 1 corresponds to the most collective global motion and mode n - 1 to the most local fluctuations. We focused on the seven slowest modes of collective motion, as explained in the Materials and Methods section. Each mode of motion partitions the amino acids in the domain into two dynamic parts, which move in opposite "senses"—meaning that the motions of the amino acids in one dynamic part are positively correlated with each other (within that specific mode) and are negatively correlated with the motions of the amino acids in the other dynamic part [72]. As described in the Materials and Methods section, we applied a filter to the results to obtain segments of at least 15 consecutive amino acids with correlated motion. These segments are predicted to move in a coordinated fashion, and we refer to them as "dynamic elements" (DEs). Namely, DEs are consecutive subsets of the structural domains. Figure 4.2 shows the DEs for the seven slowest modes of the e2xyiA1 propeller projected along its amino acid sequence. Each DE (DE-1 through DE-7) is color-coded by its direction of motion—blue and red corresponding to opposite senses. Recurring sequence segments, referred to as "themes," are aligned on the y-axis and highlighted in green. For comparison, mappings of Pfam domains based on sequence data are also shown in cyan.

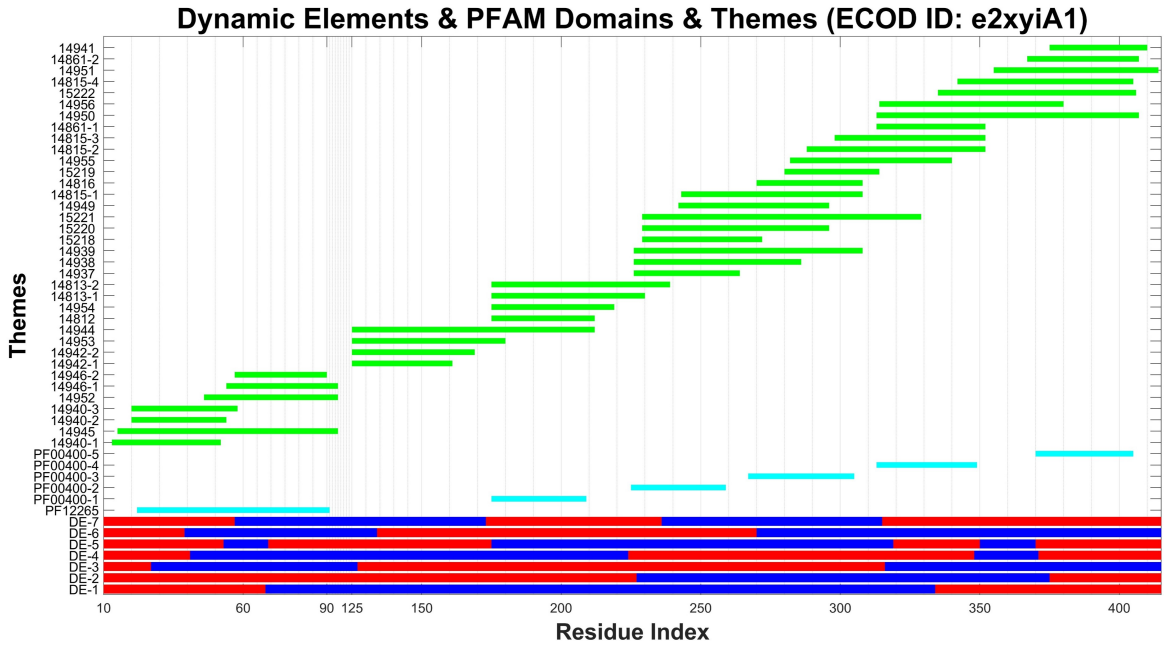


Figure 4.2. Correlation between shared protein segments and structural dynamics. DEs (dynamic elements) for the seven slowest modes of the e2xyiA1 propeller (CAF1; PDB ID: 2XYI) are shown in red and blue along the residue index. Recurring sequence segments ("themes") and Pfam entries are highlighted in green and cyan, respectively.

4.1.3. In-depth analysis of repeat domains

We first conducted in-depth analysis for the set of eight ECOD domains with symmetrical repeat architectures (Table 4.1). This set comprised two (homologous all-β-) propeller structures, two (homologous all-α) ARM repeats, a repetitive alpha hairpin, and three (homologous α/β) TIM barrels.

4.1.3.1. All-β architecture: Propellers. First, we considered two homologous seven-bladed beta-propellers, sharing 24% sequence identity (of similar structures: superimposition RMSD of 3.26 Å): ECOD domain e2xyiA1 from histone-binding protein CAF1 (Chromatin Assembly Factor 1, Nurf55 in complex with an H4 peptide; PDB

ID: 2XYI); and ECOD domain e3emhA1 from WD repeat-containing protein 5 (PDB ID: 3EMH). Figures 4.2 and 4.3 show the themes and predicted DEs in the seven slowest modes of motion of the two ECOD domains; Tables 4.3 and 4.4 list the themes for e2xyiA1 and e3emhA1, respectively. Each mode of motion comprises two dynamic parts (marked by red and blue) that move in opposite senses around hinges. Each of the two slowest modes of motion in e2xyiA1 and e3emhA1 feature at most three relatively long DEs. The higher modes feature more DEs, which, in turn, typically stretch over shorter segments. Within a given mode, each amino acid can belong only to one DE. However, when considering all (seven) slowest modes, the same amino acid belongs to many (seven) DEs. In this respect, the modes of motion are intertwined, just like the themes.

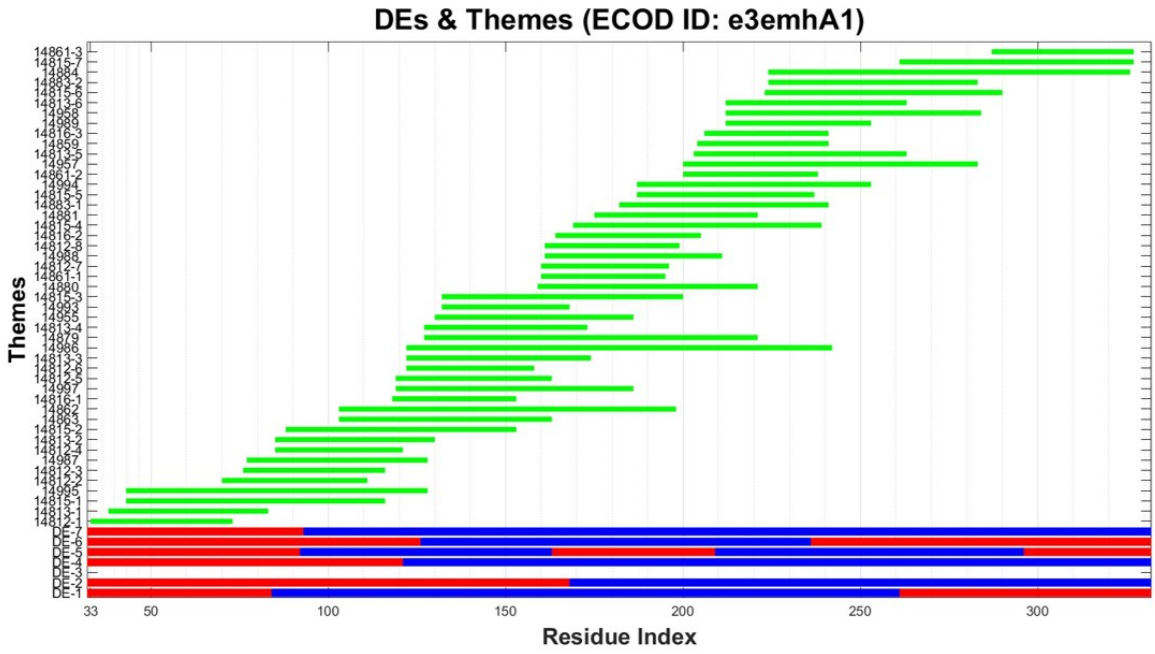


Figure 4.3. DEs and themes in the e3emhA1 propeller. The DEs of the e3emhA1 propeller in each of the seven slowest modes (DE1-through-DE7) are marked along the Residue Index (x-axis). The dynamic segments of mode 3 are too short to yield DEs. The recurring sequences (themes) are marked along the y-axis and their positions are highlighted in green.

Table 4.3. The themes detected in the e2xyiA1 propeller. In cases of nearly overlapping themes only one representative is used to minimize the computational work and confusion in figures.

Theme ID	**Position on e2xyiA1**
14940	14-53 / 21-55 / 21-59
14945	16-120
14952	47-120
14946	55-120 / 58-91 / 60-119
14942	126-162 / 126-170
14953	126-181
14944	126-213
14812	176-213
14954	176-220
14813	176-231 / 176-240 / 227-287
14937	227-265
14938	227-309
15218	230-273
15220	230-297
15221	230-330
14949	243-297
14815	244-309 / 289-353 / 299-353 / 343-406
14816	271-309
15219	281-315
14955	283-341
14861	314-353 / 368-408
14950	314-408
14956	315-381

Table 4.3. Themes of the e2xyiA1 domain. (cont.)

Theme ID	**Position on e2xyiA1**
15222	336-407
14951	356-415
14941	376-411

Table 4.4. The themes detected in the e3emhA1 propeller. Themes that are shared with the e2xyiA1 propeller are listed first.

Theme ID	**Position on e3emhA1**	**Position on e2xyiA1**
14812	38–77 / 75–115 / 81–120 / 90–125 / 124–167 / 154–193	176–213
14813	–	176–231 / 176–240 / 227–287
14815	48–120 / 93–157 / 137–204 / 174–243 / 192–241 / 206–260	244–309 / 289–353 / 299–353 / 343–406
14816	123–157 / 127–160 / 169–209 / 211–245	271–309
14861	165–199 / 205–242 / 292–331	314–353 / 368–407 / 368–408 / 371–407
14955	48–132	283–341
14862	108–202	–
14863	108–167	–
14997	124–190	–
14986	127–246	–
14993	137–172	–
14880	164–225	–

Table 4.4. Themes in e3emhA1 and e2xyiA1. (cont.)

14988	166–215	–
14881	180–225	–
14883	187–245 / 229–287	–
14994	192–257	–
14957	205–287 / 217–288	–
14859	209–245	–
14989	217–257	–
14958	217–288	–
14884	229–330	–
14995	48–132	–
14987	82–132	–

In the propellers e2xyiA1 and e3emhA1, most of the DEs align with either a single theme or a combination of several reused themes. The dynamic segments of the third mode of e3emhA1 are an exception, as they are too short to allow decomposition into DEs with our 15 amino-acid threshold (as explained in the Materials and Methods section). We see different ways to combine themes, which correspond to the DEs of the different modes. This suggests a (possibly hierarchal) rewiring of the themes to facilitate functional motions. Figure 4.4 shows an example of selections of non-overlapping themes that collectively cover, in essence, the entire e2xyiA1 propeller and correspond to the DEs of the fifth (Figure 4.4a) and seventh (Figure 4.4b) modes. In these representations, themes are projected on the 3D structure of the β-propeller domain (left panels) and also mapped onto the sequence (right panels). DEs are color-coded in shades of red and blue by their order of appearance along the sequence using an arbitrarily chosen palette, which is used consistently across subsequent figures. Themes are shown in various distinct colors. Notably, while a one-to-one correspondence between a theme and a DE exists in some cases, in others, multiple themes may collectively align with a single DE, such as the fourth DE in mode 5 or the second DE in mode 7. Likewise, the same theme pair could align together with a single DE in one mode while

each of them separately aligns with other DEs in another mode. In fact, to partition the protein domain e2xyiA1 so that the (sufficiently long) segments would fall along all the boundaries of DEs in both the fifth and seventh modes, one must use at least six segments. Indeed, the reused themes shown in the figure for e2xyiA1 are exactly these segments.

Sometimes, equivalent themes in the two propeller domains correspond to DEs from different modes of motion. For example, theme 14815 aligns with a DE from mode 1 in e2xyiA1 and with a DE from mode 5 in e3emhA1, as illustrated in Figure 4.5a. Variations of this theme, indexed as 14815-1 through 14815-4, appear at multiple positions in both structures. Specifically, variation 14815-4 corresponds to a DE of mode 1 in e2xyiA1, while 14815-2 and 14815-4 correspond to DEs of mode 5 in e3emhA1. Another example is shown in Figure 4.5b, where a combination of themes 14813 and 14815 jointly corresponds to a DE in mode 4 of e2xyiA1 and a DE in mode 2 of e3emhA1.

4.1.3.2. Themes, dynamics, and binding sites. For e2xyiA1, which is the histone binding domain of CAF1, we know the histone peptide H4 binding site, as the protein's crystal structure is of the bound structure [125]. The dynamic dissection of e2xyiA1 shows that the H4 binding site is at the interfaces of the DEs of several slow modes of motion (Figure 4.6), demonstrating the importance of the DEs to H4 binding. The relationship between the boundaries of these DEs and themes at the histone binding site suggests evolutionary links between dynamics and histone binding in CAF1. The H4 binding site residues are encompassed by two themes (14940 and 15222) in the slowest mode, three themes (14945, 14956, and 14941) in the second slowest mode, two themes (14940 and 14951) in the fifth slowest mode, and two themes (14940 and 14950) in the seventh slowest mode (Figure 4.6). Themes 14940 and 14945 repeatedly appear and partner with different themes in different modes to accompany the motion defined by each slow mode. Similarly, another histone binding site (histone H3) projected from a different crystal structure (PDB ID: 2YBA, ECOD Domain ID: e2ybaA1) is also located at the DE interfaces of e2xyiA1. That the same DEs mediate both H3

and H4 histone binding suggests allosteric communication between the corresponding binding sites. Some of the themes are shared between both binding sites (Figure 4.7), suggesting possible evolutionary roots of the allostery in this case. The correspondence between DEs, themes, and binding sites supports the idea that dynamics is evolutionarily advantageous (perhaps by promoting allostery; see also [2, 44, 45, 68, 126–132]).

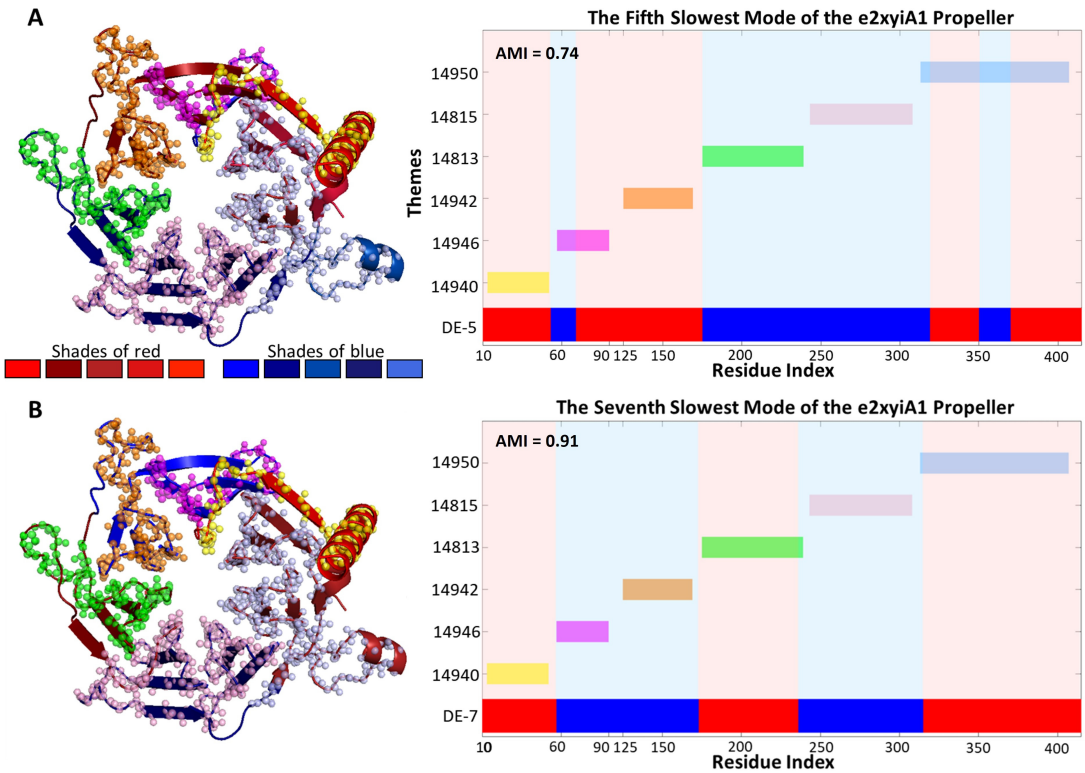


Figure 4.4. Correspondence between sequence-based themes and dynamic elements (DEs) in the e2xyiA1 propeller (CAF1, PDB ID: 2XYI). (Left) Themes and DEs from modes 5 (a) and 7 (b) projected on the 3D structure. (Right) Projection on the sequence using the same DE coloring scheme applied throughout the study.

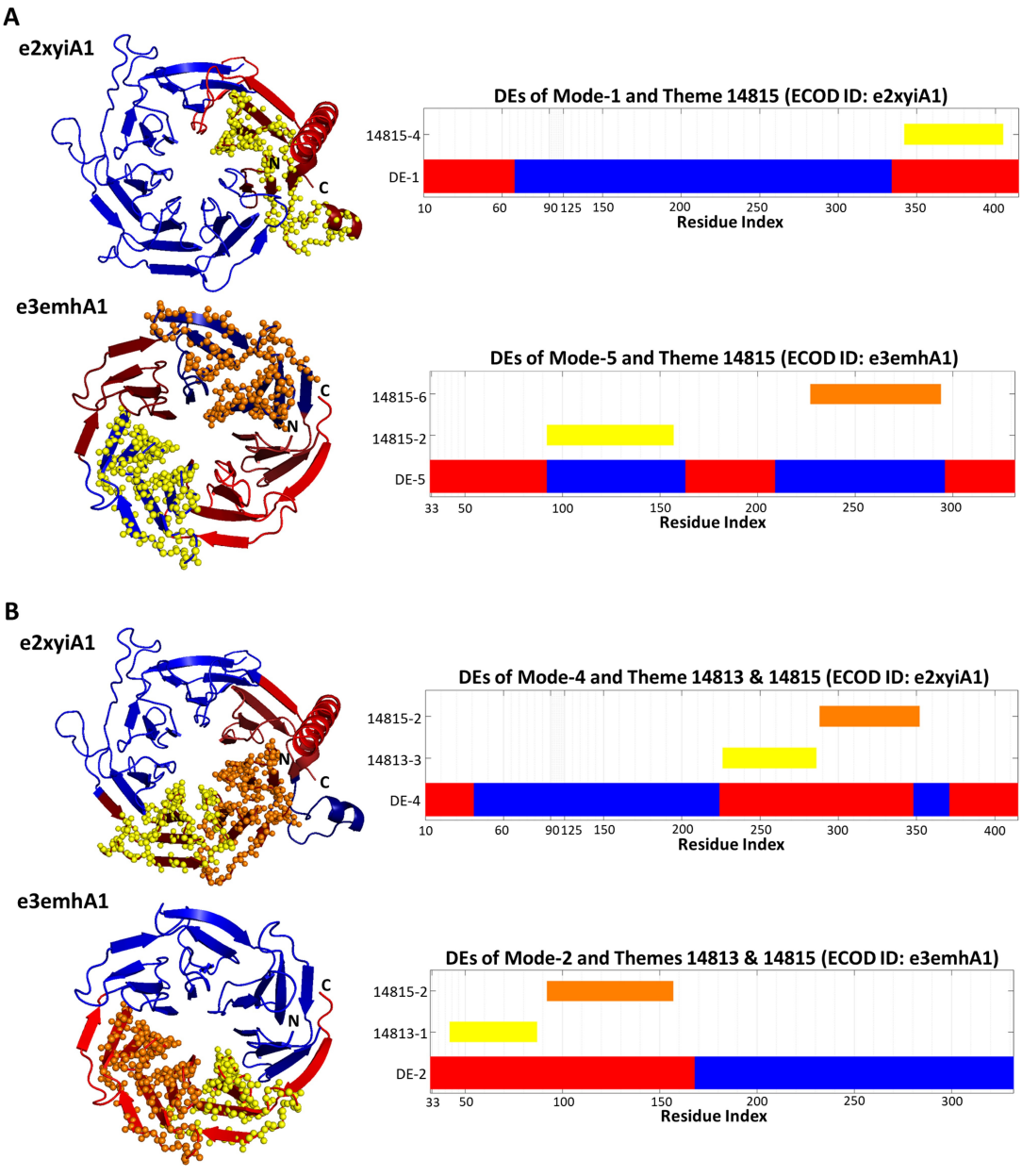


Figure 4.5. The same theme or theme combination may correspond to DEs from different modes of motion. (a) Variations of theme 14815 align with DEs in e2xyiA1 (mode 1) and e3emhA1 (mode 5). (b) Themes 14813 and 14815 together match DEs from mode 4 (e2xyiA1) and mode 2 (e3emhA1).

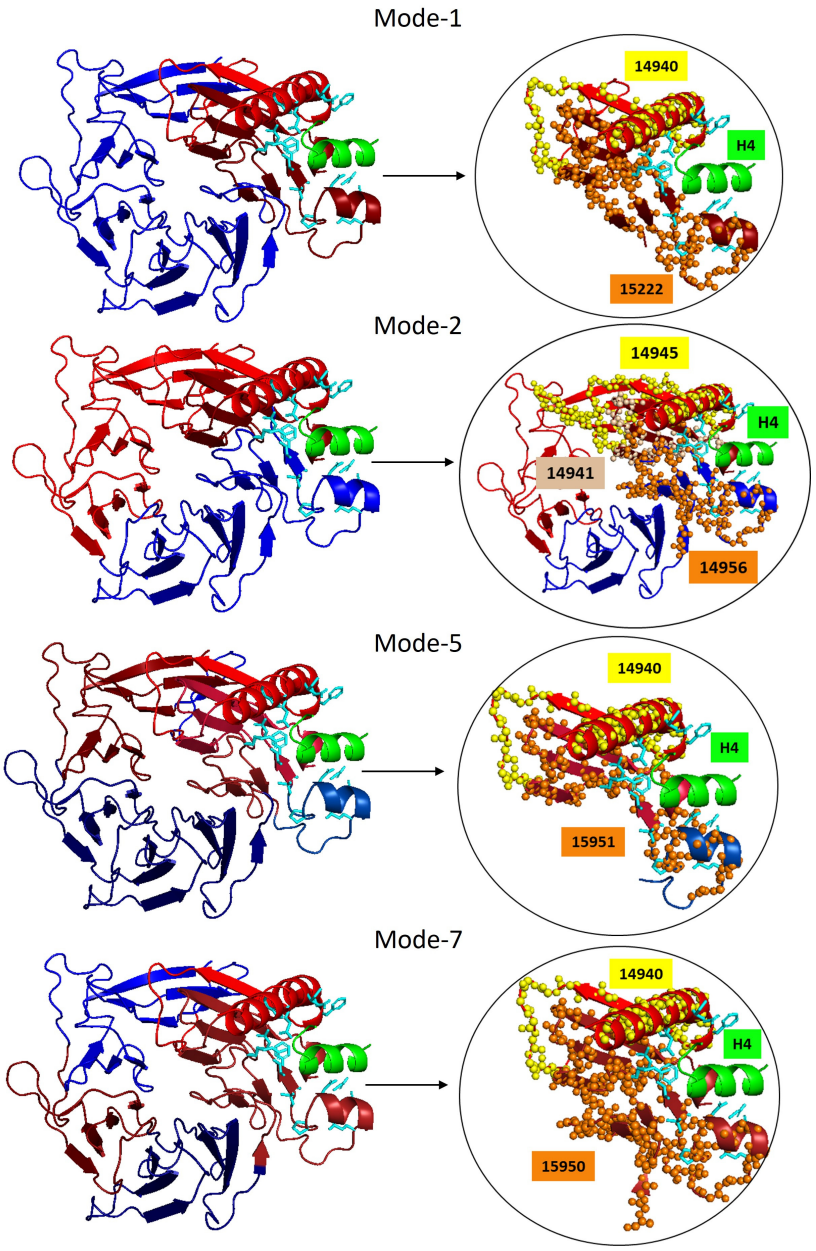


Figure 4.6. Correspondence between binding, dynamics, and themes in the e2xyiA1 β-propeller (CAF1, PDB ID: 2XYI) in complex with histone H4. DEs from the first, second, fifth, and seventh slowestmodes are shown in red and blue; H4 in green. Histone-binding residues (cyan sticks) and binding-related themes (yellow/orange spheres) are highlighted with zoom-in views.

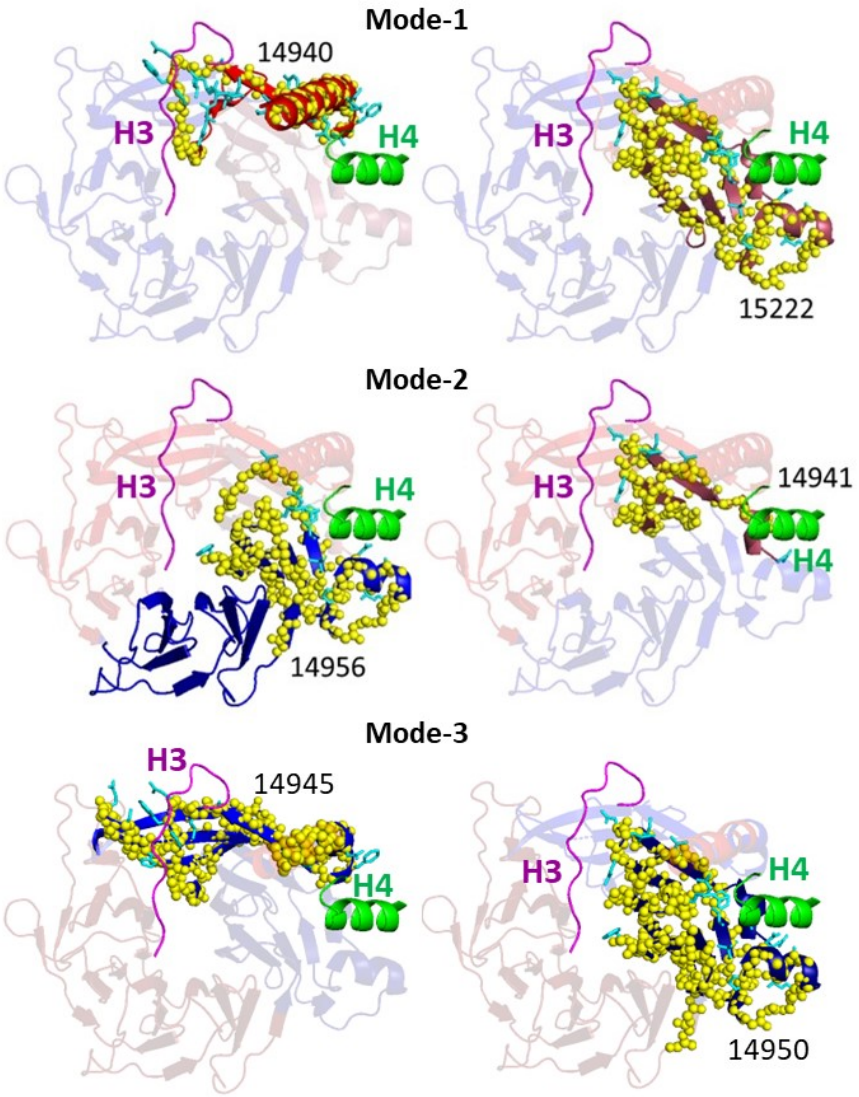


Figure 4.7. Putative evolutionary roots of allostery in e2xyiA1 (CAF1, PDB ID: 2XYI) β–propeller. DEs from the 3 slowest modes are shown in red and blue. Histone H3 (magenta) and H4 (green) peptides bind at DE interfaces. Binding residues are shown as cyan sticks; shared themes implicated in binding are highlighted in yellow spheres.

4.1.3.3. All-α architecture: alpha helix bundles. We next examine the correlation between DEs and themes in repetitive alpha hairpin protein domains (ECOD X-group 109): two homologs of the homology group (H-group) ARM-repeat (ECOD H-group 109.4): e2of3A1 from PP2A (PDB ID: 2OF3) and e1b3uA1 from ZYG-9 (PDB ID:

1B3U) (with 50% sequence identity), and a domain from the H-group proteasome / cyclosome repeat (ECOD H-group 109.35) e4adyA2 from 26S proteasome subunit Rpn2 (PDB ID: 4ADY) (sequence identity of 27% and 26% to e2of3A1 and e1b3uA1, respectively) [13]. Figures 4.8 to 4.10 compare these themes to the DEs of the seven slowest modes of each protein domain. For Figures 4.8 to 4.10, themes shared between all three helix-bundle structures (2OF3, 1B3U and 4ADY) are marked green, themes shared between 1B3U and 4ADY are in magenta and themes that are unique to one of the three proteins are in cyan Here, too, hinge points and themes' edges often overlap, further demonstrating correlation between the DEs and the themes. For example, theme c180-36 approximately corresponds to DEs in mode 3 of e2of3A1, mode 4 of, and mode 5 of e4adyA2 (Figure 4.11). Similarly, the theme c180-19 perfectly aligns with DEs of mode 5 in e2of3A1, with DEs of modes 4, 5 and 6 of e1b3uA1, and with mode 6 of e4adyA2, all of which form alpha-hairpins (Figure 4.12).

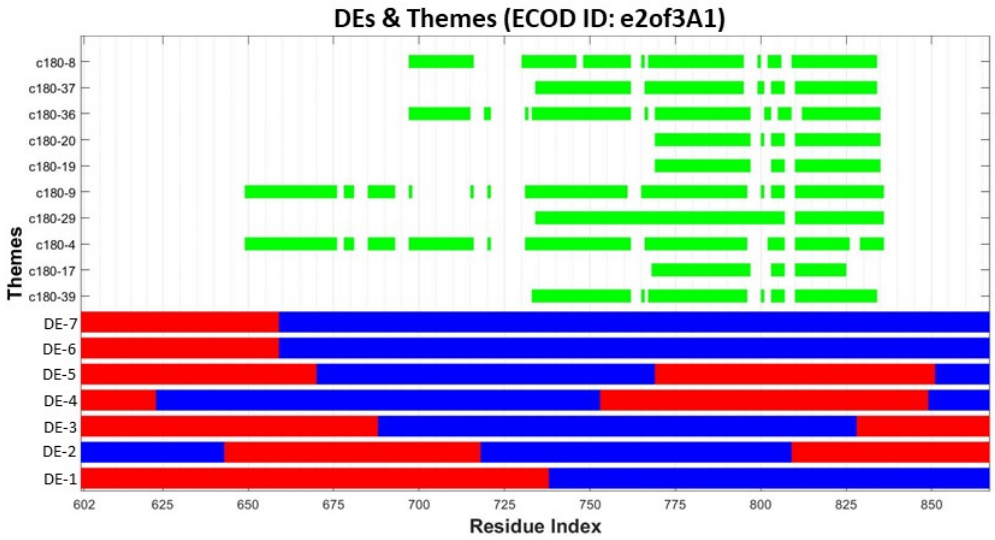


Figure 4.8. DEs and themes in ARM-repeat (PDB ID: 2OF3, ECOD domain ID: e2of3A1.

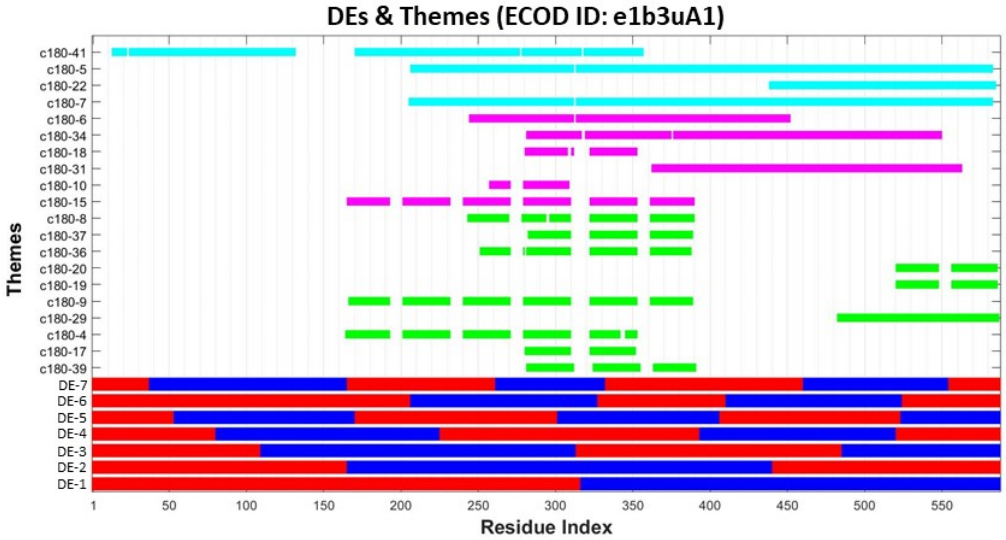


Figure 4.9. DEs and themes in ARM-repeat (PDB ID: 1B3U, ECOD domain ID: e1b3uA1).

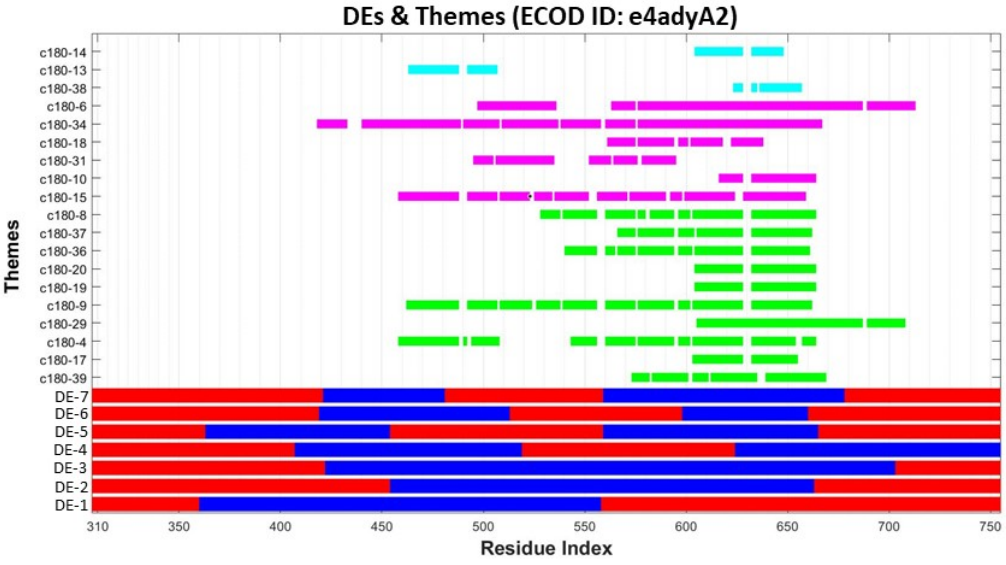


Figure 4.10. DEs and themes in proteasome/cyclosome repeat (PDB ID: 4ADY, ECOD domain ID: e4adyA2).

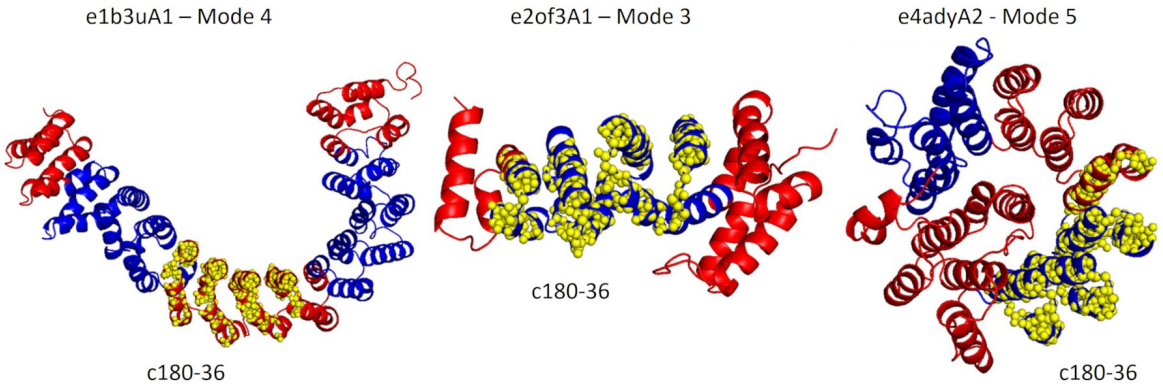


Figure 4.11. Theme c180-36 approximately corresponds to DEs in the fourth mode of e1b3uA1 (PP2A, PDB ID: 1B3U), the third mode of e2of3A1 (ZYG-9, PDB ID: 2OF3), and the fifth mode of e4adyA2 (26S proteasome subunit Rpn2, PDB ID: 4ADY). The theme is represented as yellow spheres, and the DEs in blue and red.

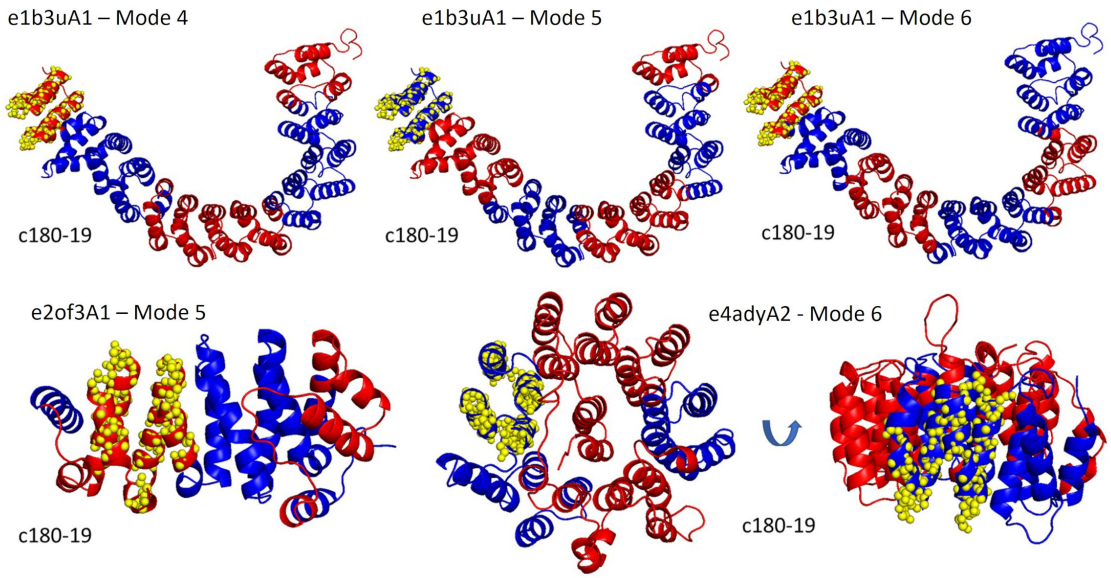


Figure 4.12. Theme c180-19 corresponds to DEs of the fourth, fifth, and sixth modes of e1b3uA1 (PP2A, PDB ID: 1B3U), and approximately to the fifth mode of e2of3A1 (ZYG-9, PDB ID: 2OF3) and sixth mode of e4adyA2 (26S proteasome subunit Rpn2, PDB ID: 4ADY).

4.1.3.4. α/β architecture: TIM barrels. The TIM barrel, one of the oldest folds, also made of repeats, is represented here with (domains from) three homologous proteins, namely, TatD-related deoxyribonuclease (PDB ID: 1J6O, ECOD Domain ID: e1j6oA1), Putative TatD related DNAse (PDB ID: 2GZX, ECOD Domain ID: e2gzxA1), and Tat-linked quality control protein TatD (PDB ID: 4P5U, ECOD Domain ID: e4p5uA1). The themes detected in these three protein domains are presented with the DEs of seven slow modes in Figures 4.13-15. As in the cases of the propellers and alpha helix bundles, here, too, hinge points and themes' edges often overlap, demonstrating the correlation between the themes and DEs.

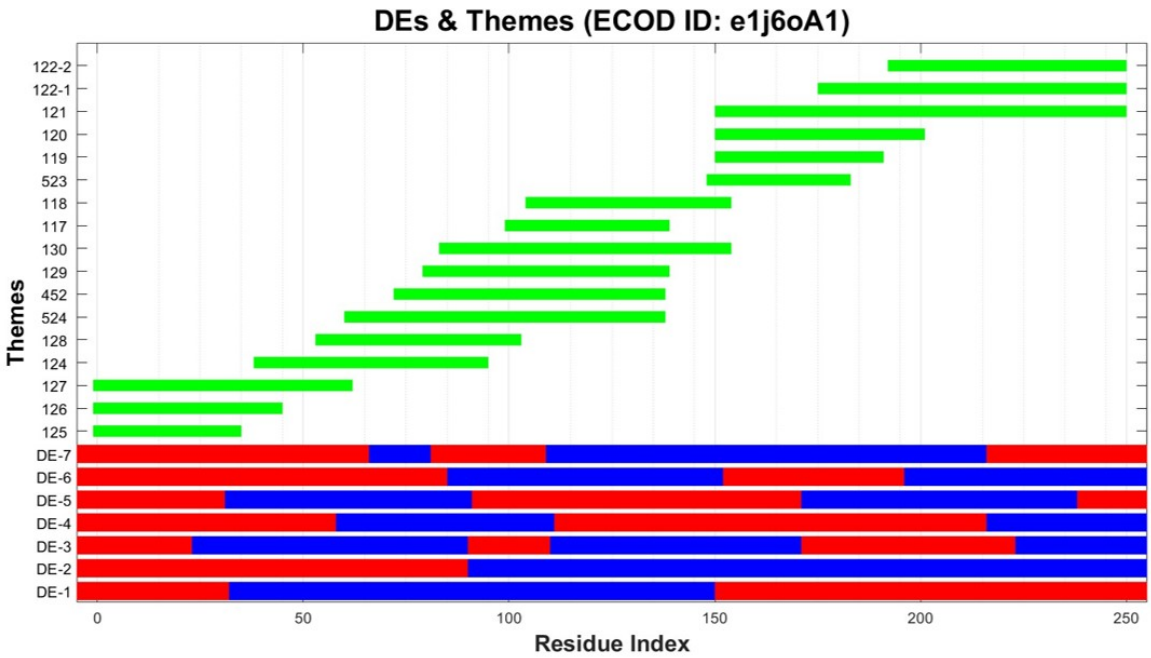


Figure 4.13. DEs and themes in TatD-related deoxyribonuclease (PDB ID: 1J6O, ECOD domain ID: e1j6oA1).

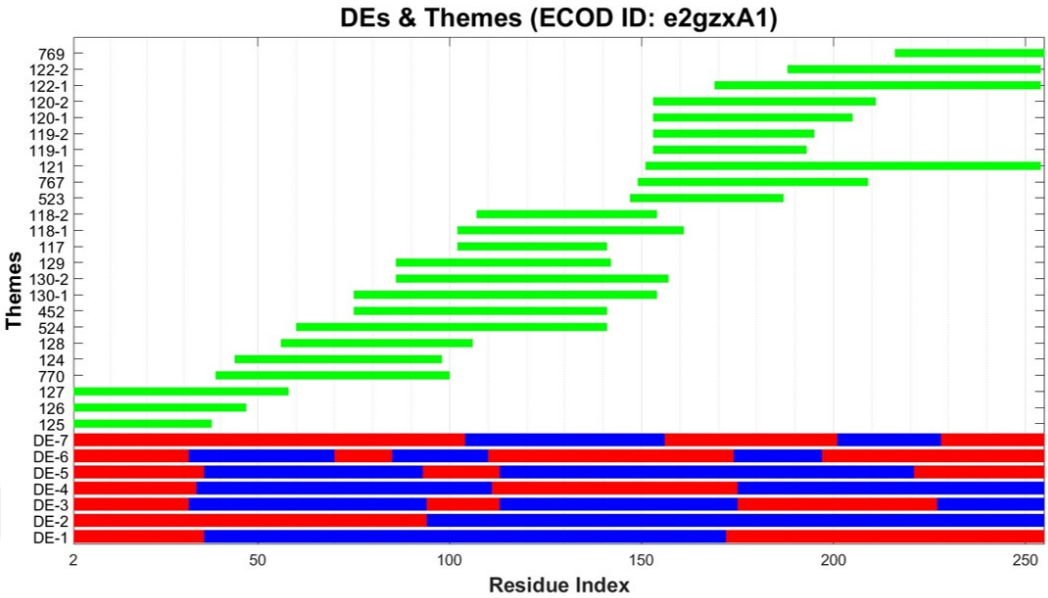


Figure 4.14. DEs and themes in Putative TatD related DNAse (PDB ID: 2GZX, ECOD domain ID: e2gzxA1).

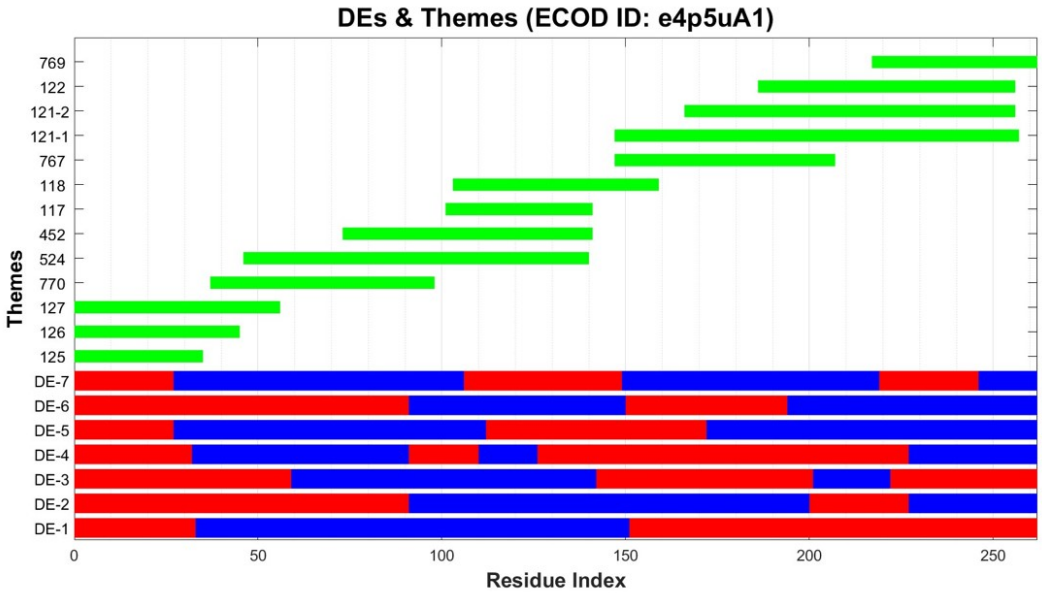


Figure 4.15. DEs and themes in Tat-linked quality control protein TatD (PDB ID: 4P5U, ECOD domain ID: e4p5uA1).

As with the e2xyiA1 propeller, an anthology of non-overlapping themes can collectively cover the entire sequence and correspond to all the DEs of specific modes. Figure 4.16 shows that the DEs of mode 1 and mode 4 of e2gzxA1 are entirely covered by the same four themes. However, in mode 4 there is one-to-one correspondence between the DEs and themes, while in mode 1 two themes (124 and 118) combine to cover a single DE.

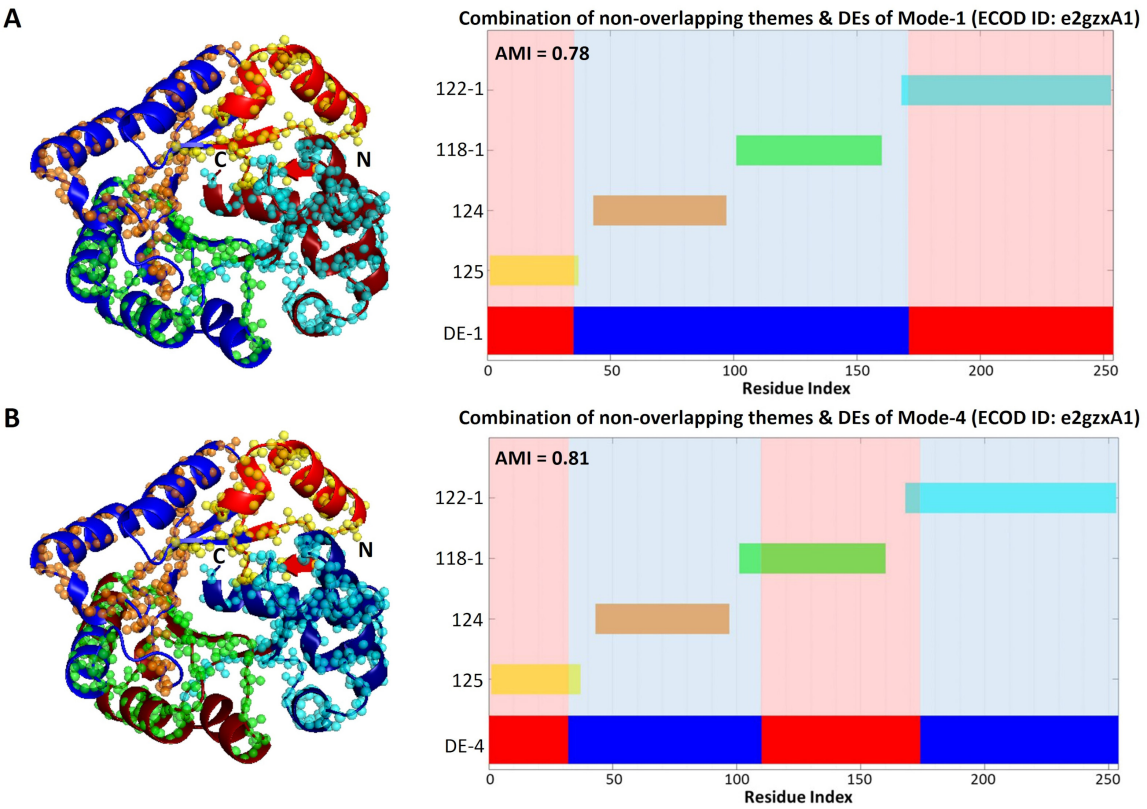


Figure 4.16. Correspondence between themes and functional dynamics in the e2gzxA1 TIM barrel (TatD-related DNase, PDB ID: 2GZX). (Left) Themes and DEs from modes 1 (a) and 4 (b) projected on the 3D structure. (Right) Corresponding sequence view. In mode 4, each DE aligns with a single theme; in mode 1, one DE spans themes 124 and 118-1.

Figure 4.17 compares the themes and DEs across the homologous TIM barrels. Variations of theme 121 cover a single DE from mode 1 in the C-terminus of each of the three homologous domains. It is noteworthy that the partitioning into DEs of e2gzxA1 is somewhat different from that of the other two homologs, where the DE in the C-terminus includes an additional α helix (α6), marked with an arrow in Figure 4.17a), which alters its sense of correlations (red vs. blue) and becomes part of another DE in e1j6oA1 and e4p5uA1. Interestingly, this difference is also reflected in the variations of theme 121 vs. 121-2. Additionally, this DE of e1j6oA1 is also covered by a combination of two themes: 119 and 122. Theme 119 covers helix α6 (with the adjacent helix α7).Thus, theme 119 in combination with theme 122 does not align with this C-terminal DE in e2gzxA1, and it does not exist in e4p5uA1. On the other hand, the cooperative motion described by the slowest mode is likely the key mode to be associated with DNA binding (Figure 4.17B). Here, the interface of the three DEs of the slow modes is probably adjusting the DNA binding mode. The DEs of e2gzxA1 display a difference compared to those of e1j6oA1 and e4p5uA1. The DNA segment is projected here from another homologous protein (PDB ID: 4PE8) to show the likely position of the trinucleotide DNA located at the interface of DEs in this most cooperative motion.

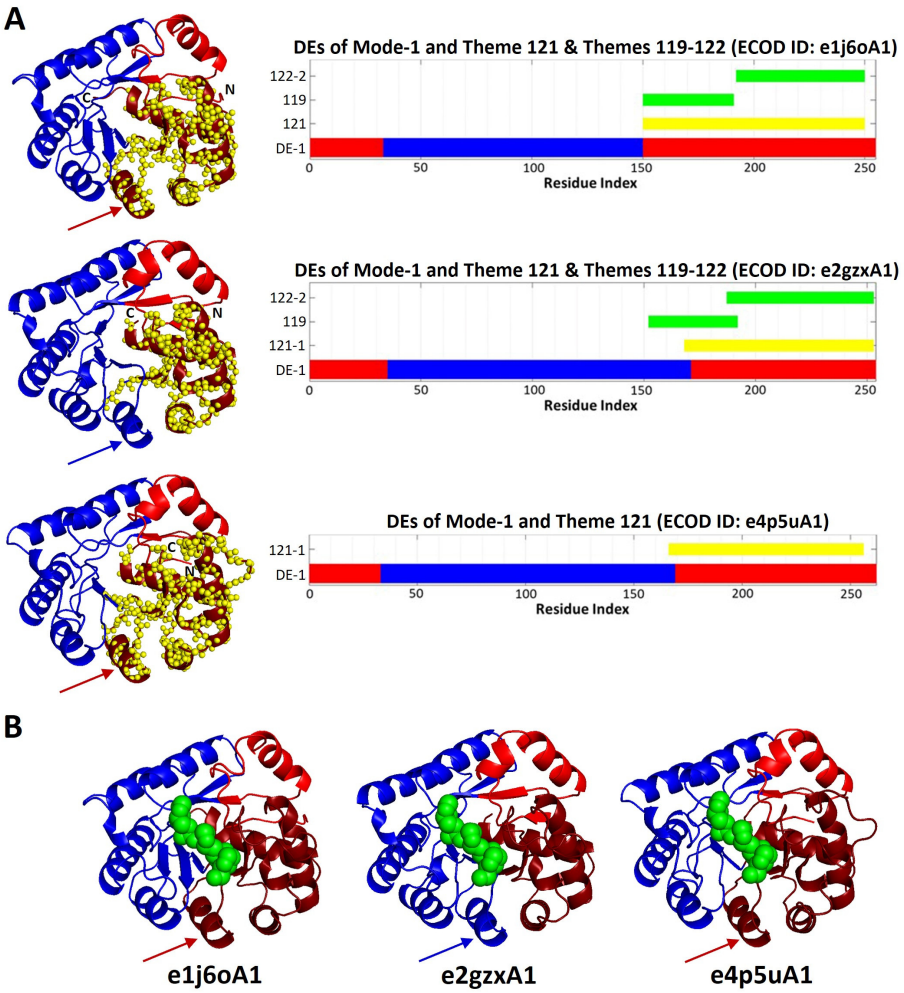


Figure 4.17. Themes and DEs in homologous TIM barrels. (a) Mode 1 DEs of e1j6oA1, e2gzxA1, and e4p5uA1 are similar, but the orientation of helix $\alpha 6$ (arrow) differs in e2gzxA1. Theme 121 (yellow spheres) captures this variation. (b) Trinucleotide DNA (green spheres, from PDB ID: 4PE8) maps to DE interfaces in all three domains.

4.1.4. In-depth analysis of non-repeat domains

Next, we analyze the five non-repeat domains in our in-depth set; of these, two belong to ribosome-inactivating proteins (RIP), and three belong to Rossmann-related homology groups (Table 4.2). The themes detected in these domains and their sequence

positions are listed in Tables 4.5 and 4.6. As in the repeat domains introduced above, here, too, hinge points often overlap with themes' edges. Most of the DEs align with a single theme or with a combination of several themes, and/or themes align with either a single or several DEs (Figures 4.18-22).

Table 4.5. Themes of the e3ktzA1 and e3ctkA1 domains from ribosome-inactivating proteins (RIP) 3KTZ and 3CTK and their respective sequence positions. Variations of many themes are shared between all three proteins. Some themes are shared only between two of the proteins.

Theme ID	**Position on e3ktzA1**	**Position on e3ctkA1**
10078	3–42	3–37 / 3–43
10081	4–69	3–75
10103	45–80	45–80
10079	49–84 / 51–91	43–91 / 45–84
10109	51–140	46–142 / 47–142
10104	62–103	58–104
10095	62–130 / 66–140	58–132
10105	66–116	62–117
10107	81–177	80–176
10096	84–140	84–142
10097	98–140	98–142
10098	106–177	107–176
10090	118–177 / 123–181	119–176
10099	133–174	135–173
10100	133–188	135–188
10075	148–187	150–188 / 153–188
10076	152–196	150–195 / 151–196
10077	152–245	150–246
10091	154–233	153–234

Table 4.5. Themes of the e3ktzA1 and e3ctkA1 domains. (cont.)

Theme ID	**Position on e3ktzA1**	**Position on e3ctkA1**
10089	190–245 / 192–245	180–246
10092	191–234	191–234
10080	–	3–53
10087	–	3–89
10101	–	175–209

Table 4.6. Themes of the e1ulsB1, e2ae2A1 and e3n74B1 Rossmann-related domains and their respective sequence positions. Variations of many themes are shared between all three proteins. Some themes are shared only between two of the proteins.

Theme ID	**Position on e1ulsB1**	**Position on e2ae2A1**	**Position on e3n74B1**
2939	1-45 / 2-36 / 2-66 / 3-57 / 9-48	6-60 / 12-46	1-50 / 2-56 / 3-67 / 4-38 / 6-50
2967	46-86	54-98	48-92 / 55-89 / 56-91
2968	54-93 / 54-95	57-106 / 71-125	50-116 / 53-102 / 56-96 / 58-112 / 65-124
2969	64-139	83-149	72-143 / 76-146
2960	93-139	105-146	102-146
2962	99-167 / 104-183 / 111-169 / 130-185	105-171 / 109-196 / 131196	95-213 / 96-176 / 98-173 / 99-163 / 102-167 / 102-213 / 104-176 / 106-213 / 109-213 / 114-171 / 114-213 / 120-213 / 122-214 / 126-187

Table 4.6. Themes of Rossmann-related domains. (cont.)

Theme ID	**Position on e1ulsB1**	**Position on e2ae2A1**	**Position on e3n74B1**
2963	116-237 / 123-237	135-253	102-249
2964	141-185 / 148-183	152-196	149-213 / 151-213 / 154-214 / 156-213
2965	147-237 / 158-237 / 170-237	155-253 / 173-253	166-249
2966	188-237 / 191-240 / 192-237 / 195-240	194-253	214-249
3034	53-102	–	–
3380	61-137	–	63-144
3397	93-127	–	–
3395	128-181	–	–
3396	139-182	–	–
3379	146-237	–	156-249
2972	148-192	–	–
3493	181-237	–	–
3117	–	–	44-96
3193	–	63-98	56-90
3314	–	–	99-133
3315	–	–	100-186
3472	–	–	133-142
3313	–	–	138-186
3316	–	–	146-486

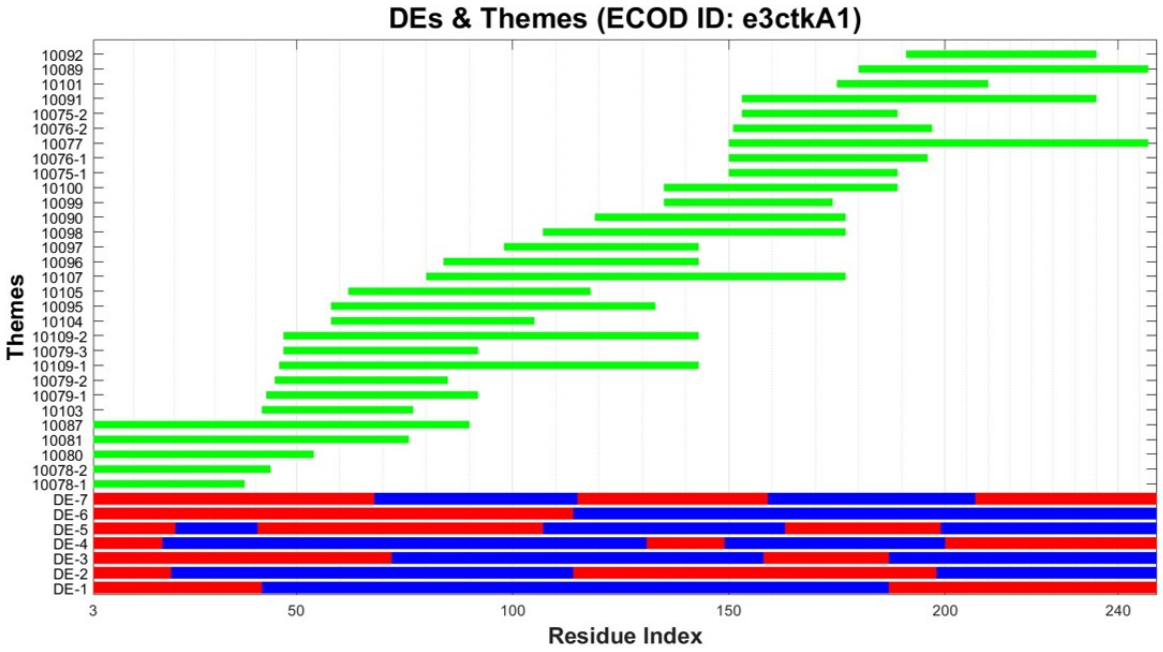


Figure 4.18. DEs and themes in rRNA N-glycosidase (PDB ID: 3CTK, ECOD domain ID: e3ctkA1).

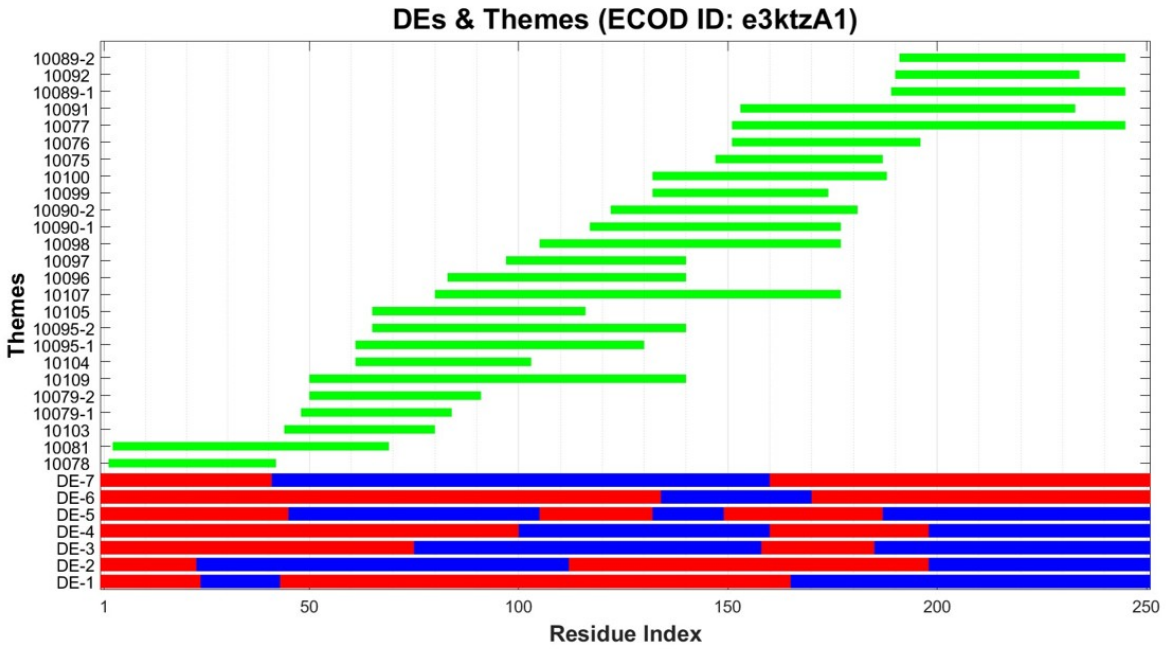


Figure 4.19. DEs and themes in Ribosome-inactivating protein gelonin (PDB ID: 3KTZ, ECOD domain ID: e3ktzA1).

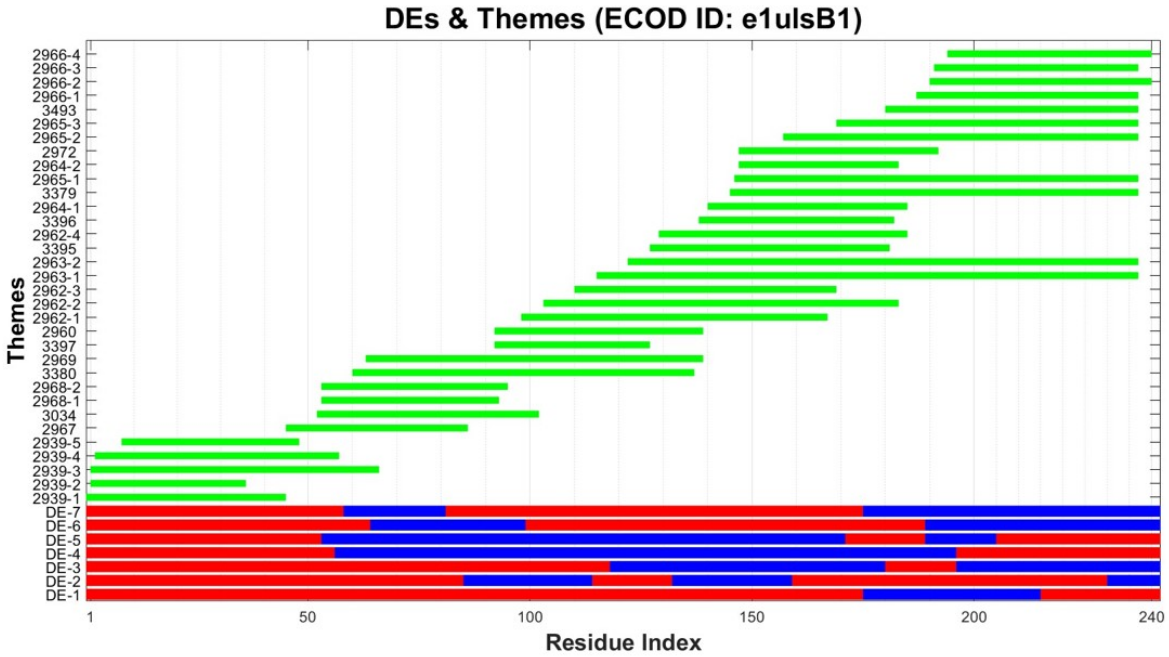


Figure 4.20. DEs and themes in Putative 3-oxoacyl-acyl carrier protein reductase (PDB ID: 1ULS, ECOD domain ID: e1ulsB1).

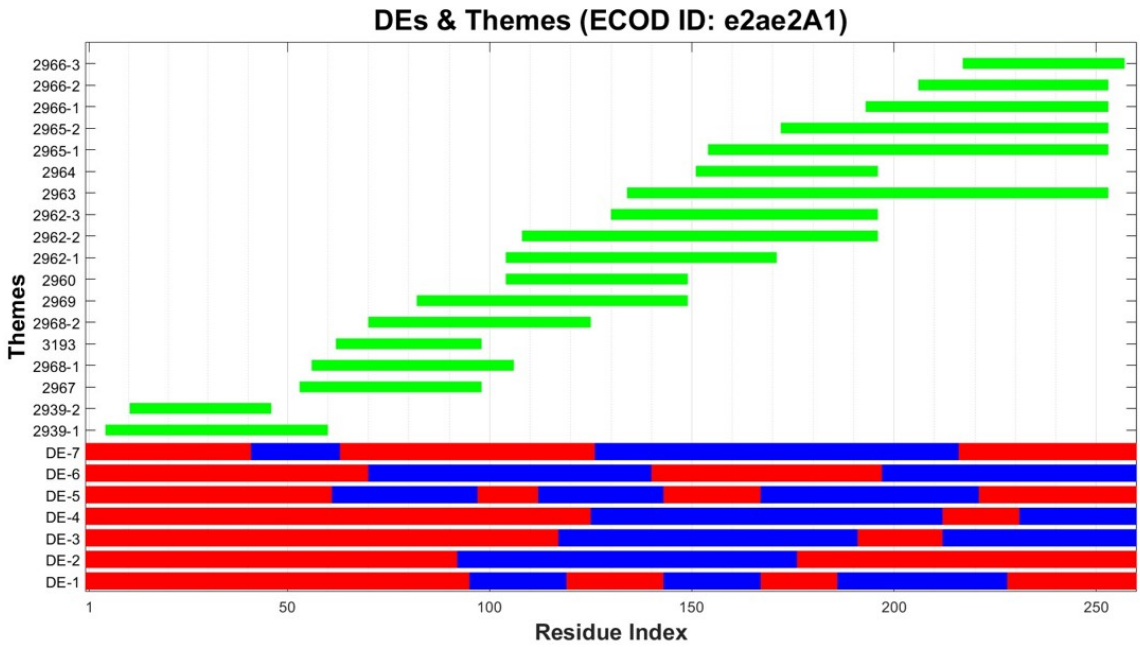


Figure 4.21. DEs and themes in Tropinpne reductase-II (PDB ID: 2AE2, ECOD domain ID: e2ae2A1).

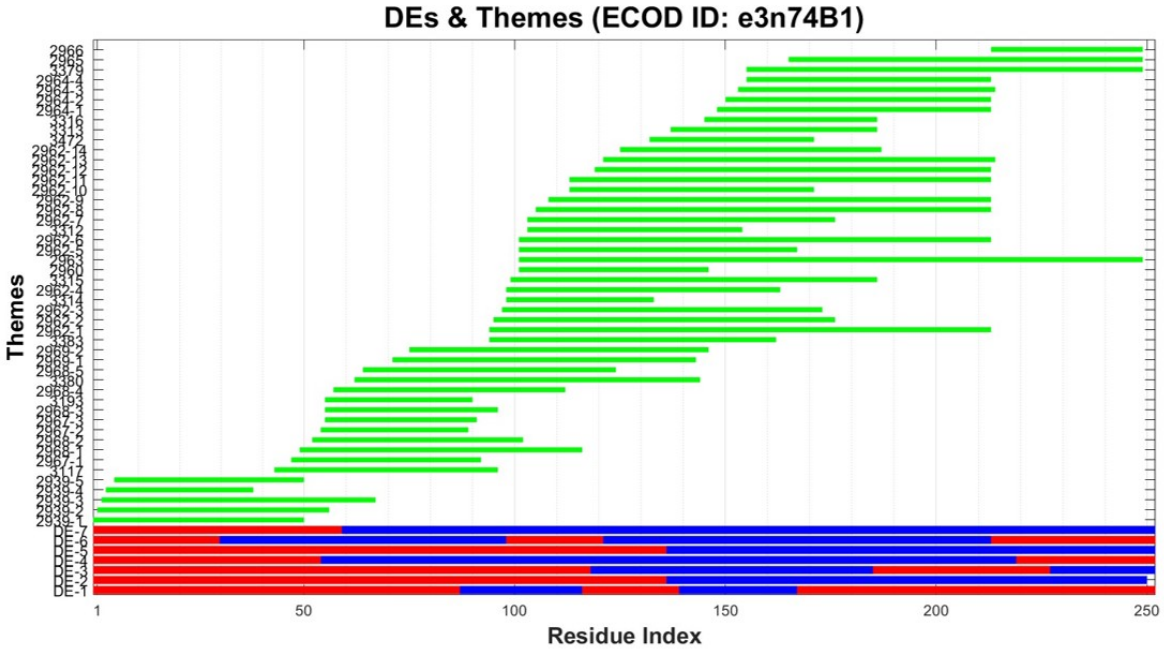


Figure 4.22. DEs and themes in 3-Ketoacyl-(Acyl-Carrier-Protein) reductase (PDB ID: 3N74, ECOD domain ID: e3n74B1).

In the RIP homology group, for example, we observe that in domain e3ktzA1 (from ribosome-inactivating protein gelonin; PDB ID: 3KTZ), themes 10078 and 10079 correspond to different DEs of mode 5, whereas the same themes combined correspond to a single DE of mode 4 (Figure 4.23). Additionally, in domain e3ctkA1 (from rRNA N-glycosidase; PDB ID: 3CTK), variations of theme 10078 correspond to a single DE of mode 1 (10078-1) and to a combination of two DEs of mode 5 (10078-2) (Figure 4.24a). Likewise, in e3ktzA1, the same theme (10078) corresponds to a combination of two DEs in mode 1 as well as to a single DE of mode 7 (Figure 4.24b). This example illustrates the flexibility in the alignment of DEs with variations of the same theme.

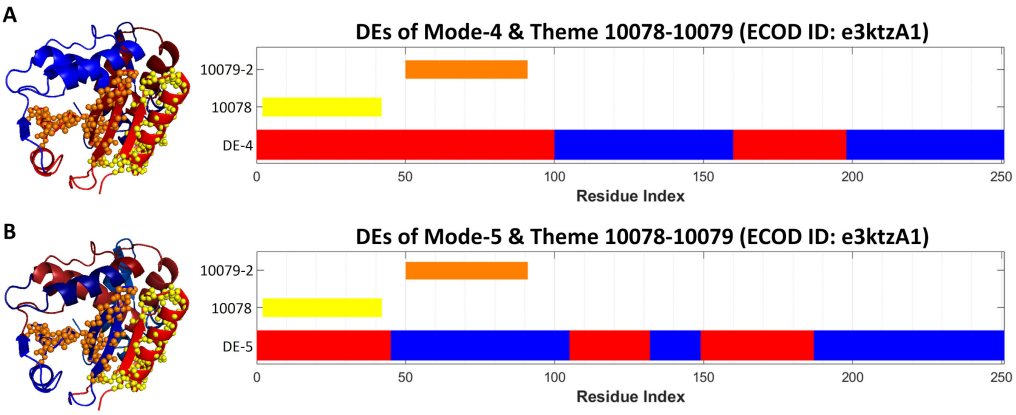


Figure 4.23. Correspondence between themes and dynamics in the e3ktzA1 domain (gelonin, PDB ID: 3KTZ). (a) Themes 10078 and 10079 together align with one DE in mode 4. (b) The same themes correspond to separate DEs in mode 5. Themes are shown as yellow/orange spheres; DEs in red and blue.

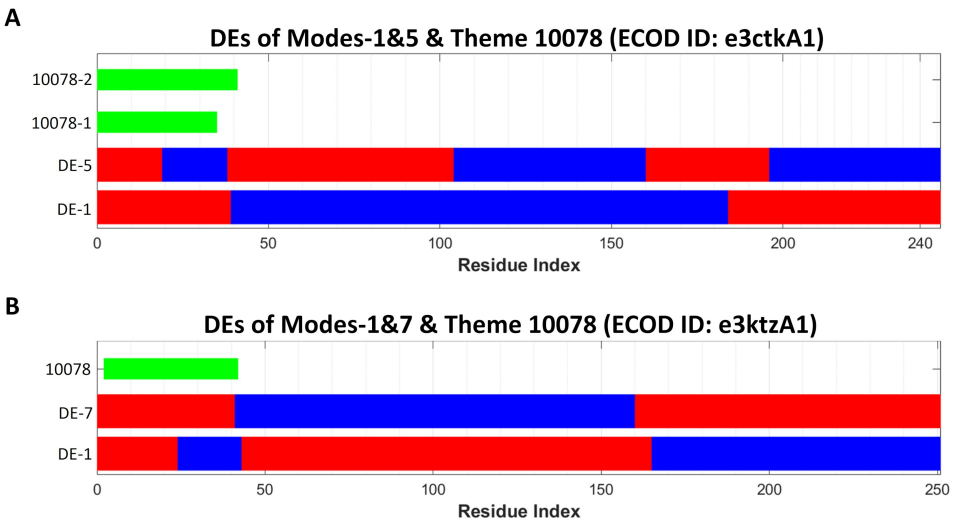


Figure 4.24. The same theme may correspond to a DE or a DE combination of different modes of motion in RIP homologues. (a) In e3ctkA1 (PDB ID: 3CTK), theme 10078 maps to one DE in mode 1 and two DEs in mode 5. (b) In e3ktzA1 (PDB ID: 3KTZ), the same theme (green) aligns with one DE in mode 7 and two DEs in mode 1. DEs are shown in red and blue.

For the Rossmann-related homology group, Figure 4.25 shows that a selection of non-overlapping themes can collectively cover the entire sequence and correspond to all the DEs from a specific mode, as shown above in the e2xyiA1 propeller and e2gzxA1 TIM barrel. DEs of mode 4 and mode 6 for putative 3-oxoacyl-acyl carrier protein reductase (PDB ID: 1ULS, ECOD Domain ID: e1ulsB1) are entirely covered by the same combination of non-overlapping themes. As seen, a theme alone or in combination with other themes may correspond to a DE. For example, theme 2962 and theme 2968, in combination, align with a single DE (mode 4), yet theme 2962 may also align with a single DE individually (mode 6).

In another example, for each of the three proteins with Rossmann-related homology, variations of theme 2939 correspond to a DE of mode 5 in e1ulsB1, to a DE of mode 5 for the e2ae2A1 domain in tropinone reductase-II (PDB ID: 2AE2), and a DE of mode 4 for the e3n74B1 domain in 3-ketoacyl-(acyl-carrier-protein) reductase (PDB ID: 3N74) (Figure 4.26a). Variations of the same 2939 theme together with theme 2967 correspond to a DE of mode 2 in e1ulsB1, a DE of mode 1 in e2ae2A1, and a DE of mode 1 in e3n74B1 (Figure 4.26b). As seen, variations of the same themes among homologs may compensate for variations in the dynamics between them, which further reinforces the correspondence between themes and DEs.

DEs may correspond to secondary structure elements, with the hinges between them residing in loops that connect the elements. However, the Rossmann-related domain e3n74B1 (from the 3-ketoacyl-(acyl carrier protein) reductase, FabG) provides counter-examples. As seen in Figure 4.27, the dynamic dissection of the slowest mode includes hinges in the middle of two long helices. Interestingly, the hinges correspond to the termini of themes 2962-15, 2965-2, and 2967-2. This illustrates that the sequence and dynamic dissection occur at a level different from what secondary structures may imply. The slowest mode may facilitate FabG binding to its NADP cofactor, which is key to the protein's function (Hou et at., 2015).

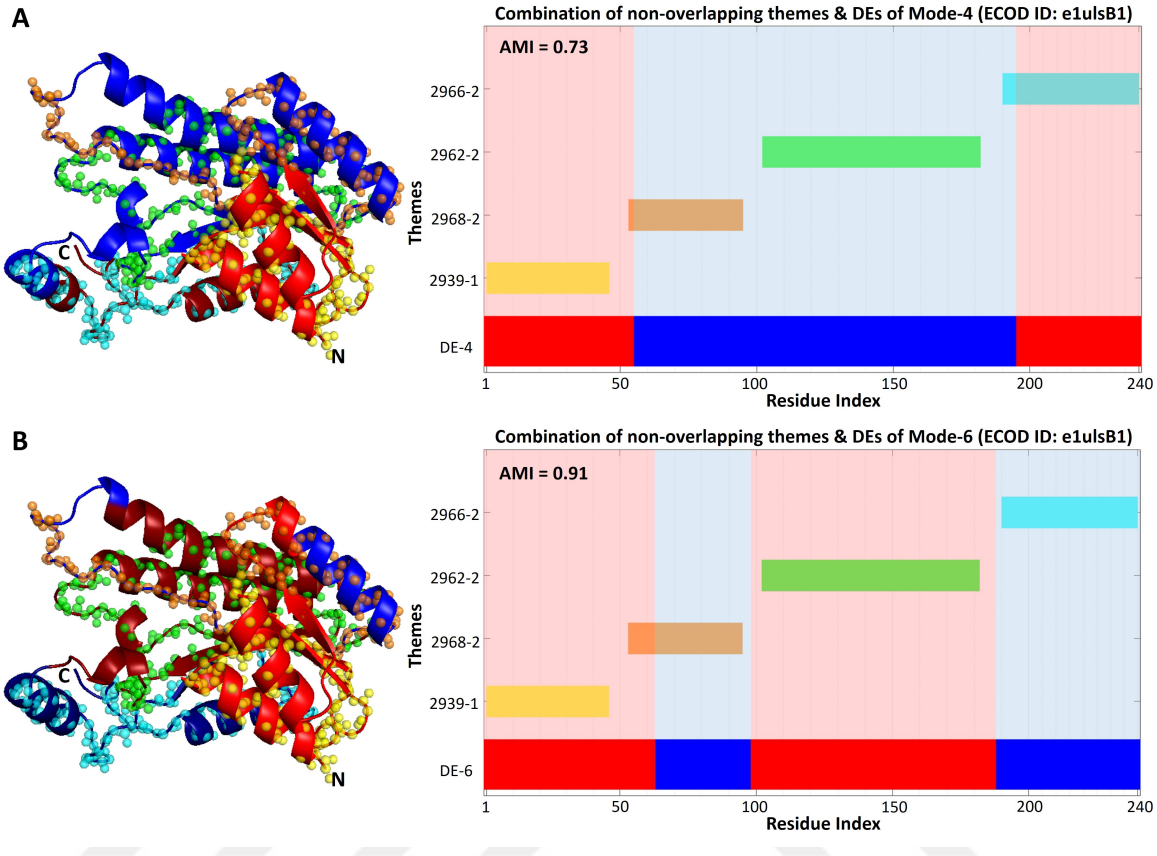


Figure 4.25. Correspondence between themes and dynamics in the e1ulsB1 Rossmann-related domain (PDB ID: 1ULS). (Left) Themes and DEs from modes 4 (a) and 6 (b) projected on the 3D structure. (Right) Sequence mapping. Themes are shown in various colors; DEs in red and blue.

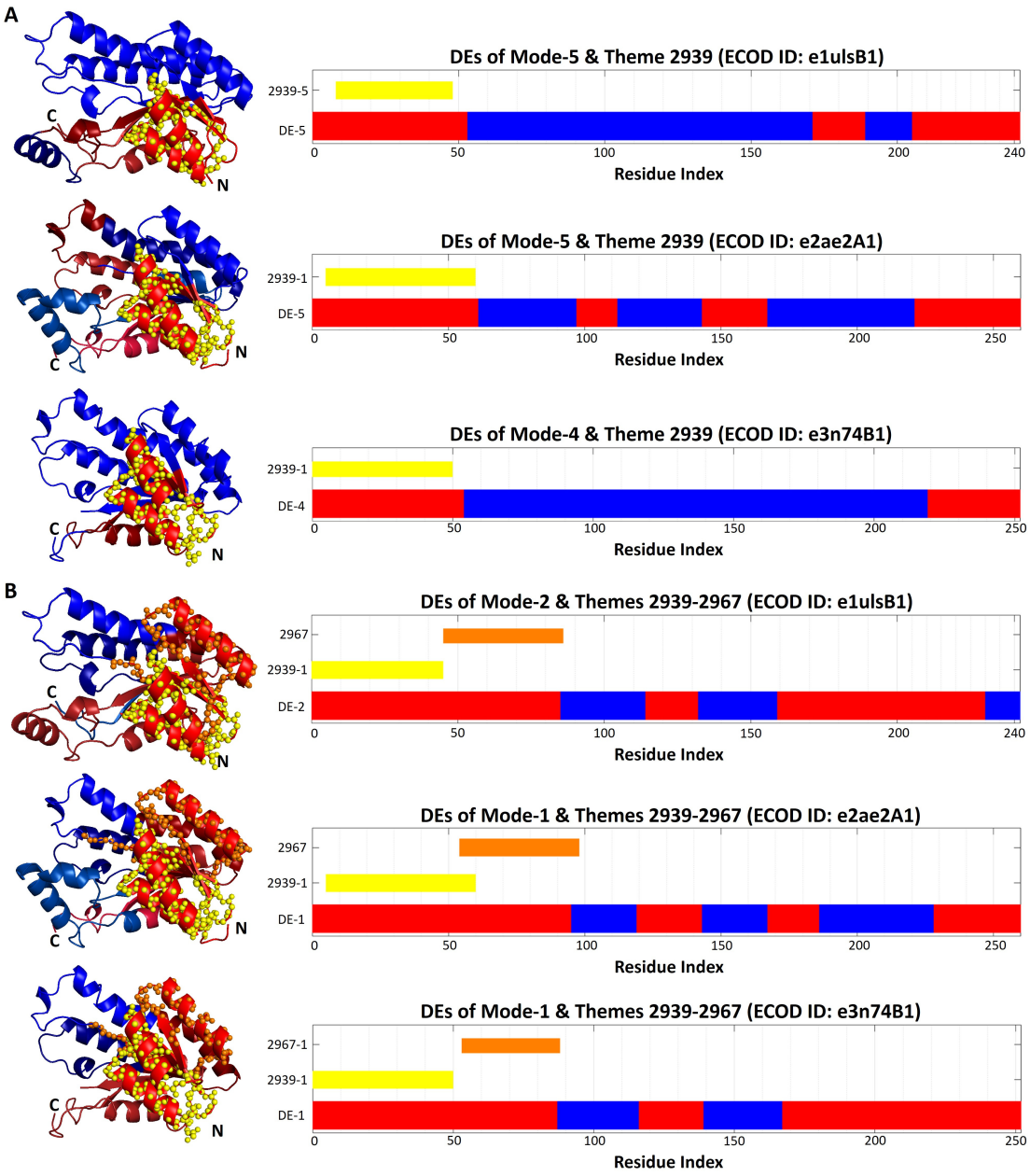


Figure 4.26. Correspondence between themes and DEs in Rossmann homologous domains. (a) Theme 2939 variations align with DEs in e1ulsB1, e2ae2A1, and e3n74B1. (b) A combination of themes 2939 and 2967 corresponds to DEs in all three domains. Themes are shown as yellow/orange spheres; DEs in red and blue.

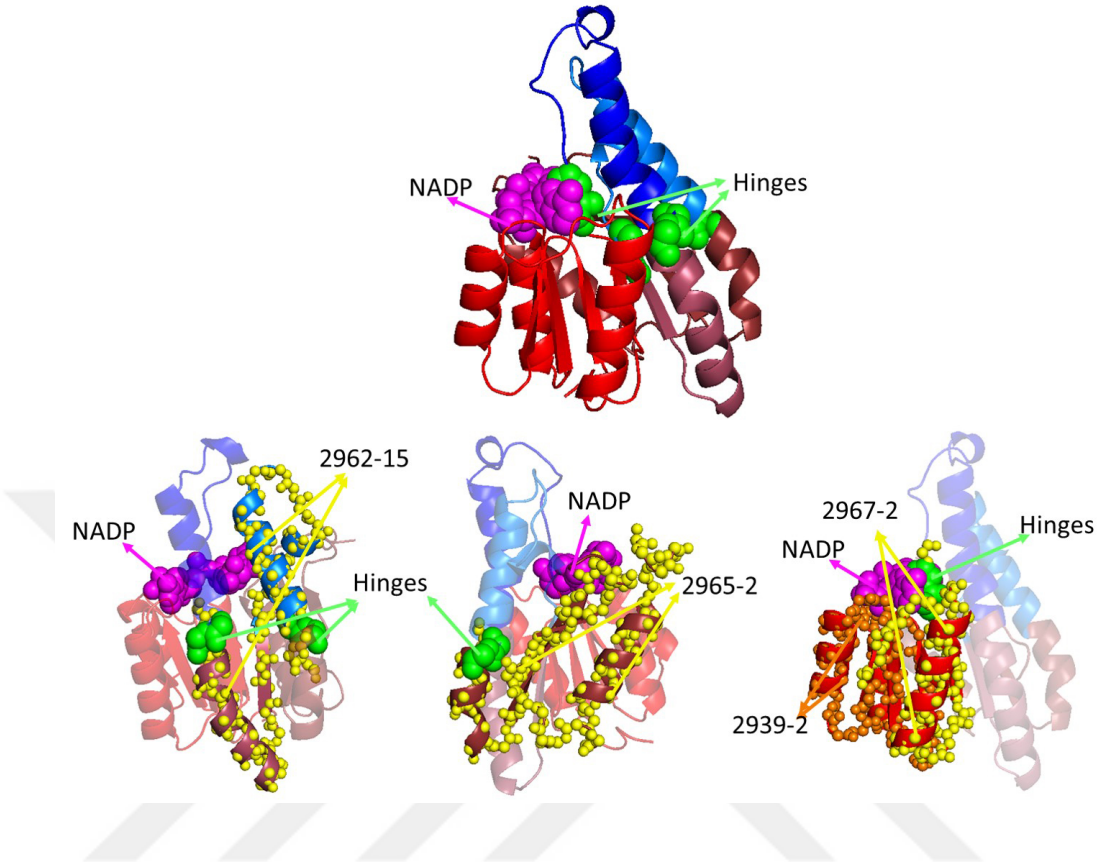


Figure 4.27. DEs and themes in e3n74B1 (PDB ID: 3N74) do not always align with secondary structures. Shown are DEs and hinges from the slowest mode, with themes near hinges (2962-15, 2965-2, 2967-2) in yellow. NADP is shown as magenta spheres; hinges as green spheres.

4.1.4.1. In-depth set: Synthesis of observations. Taking together our observations from our in-depth analysis, we observe that, in some cases, the same theme combination may correlate with more than one dynamic mode, i.e., the same sequence of themes complies with several inherent motions that are embedded in the structure. This occurs in the cases of the e2xyiA1 propeller, the e2gzxA1 TIM barrel, and the e1ulsB1 Rossmann-related domain (Figure 4.4, Figure 4.16, Figure 4.25). For example, themes 2939, 2968, 2962 and 2966 correspond to DEs of the fourth and sixth modes in e1ulsB1. On the other hand, between homologous domains, there may be a shift in mode space (e.g., the fifth dynamic mode of e1ulsB1 corresponds to the fourth dynamic mode of its homolog e3n74B1; Figure 4.26). Additionally, comparison of homologous protein domains may

reveal some alterations (differences in mode shape and DEs) across homologs' corresponding dynamic modes. These alterations might be minor (e.g., when comparing the first dynamic modes of Rossmann-related homologs e1j6oA1, e2gzxA1 and e4p5uA1; Figure 4.17a) or major (e.g., when comparing the first dynamic modes of RIP homologs 3CTK and 3KTZ; Figure 4.24). Yet, these dynamic modes are still correlated with the themes and their combinations. Thus, the same themes (and combinations of themes) or their variations may align with the DEs in the corresponding modes between homologs as in Rossmann-related domains (Figure 4.26). Moreover, new themes (and combinations of themes) may appear to align with more significant variations in slow modes as in TIM barrels (themes 119 and 122) (Figure 4.17), adding viability to the correlation of the themes and DEs. Themes may appear in multiple positions along the protein domain (indexed as xxx-1, xxx-2, etc.). Unlike some themes in the propellers (Figure 4.2, Figure 4.3, Figure 4.5), which are made of repeats, these positions are in similar segments of the protein in non-repeat domains (Figure 4.18-22). As a result, although specific themes cover relatively similar DEs among homologous protein domains in the corresponding modes and/or in different modes, their variations enable the theme to comply with the dynamics and thus function (Figure 4.24 and Figure 4.26).

4.1.5. Mutual information analysis quantifies the statistical significance of correspondence between theme boundaries and DEs

To quantify the apparent relationship between the DEs and the themes, we used mutual information analysis. Conventional mutual information does not provide any measure of statistical significance. Thus, we used two variants that innately consider the randomness factor by including the expected mutual information in the calculations, and provide some assessment of statistical significance [119]: Adjusted Mutual Information (AMI), which ranges between 0 and 1 (1 being perfectly identical), and Standardized Mutual Information (SMI), which measures the distance in standard deviations from dissimilarity (the larger the value the more similar the distributions). Complicating this analysis is the fact that DEs and themes are fundamentally different

entities. In particular, the DEs of each mode always cover the whole structure, but themes typically do not. Also, DEs of a given mode do not overlap, while the themes sometimes do. To minimize gaps and overlaps, we sampled various combinations of themes, as described in the Materials and Methods section. A detailed description of our application of AMI and SMI to themes and DEs is provided in the Materials and Methods section.

To illustrate our analysis, we describe it in detail for the e2xyiA1 propeller. Figure 4.28 shows examples of 10 theme combinations out of 147 possible theme combinations of e2xyiA1 with (an arbitrarily chosen) overlap of 3 residues and gap restriction of 8 residues. The AMI values calculated between each of the seven slowest modes of e2xyiA1 and all possible theme combinations range between 0.49 and 0.91, and the corresponding SMI values range between 79 and 160 (Table 4.7). Mutual information results using other combinations of overlap and gap restrictions give AMI values of 0.48-0.94 and SMI values of 76-160 (Tables 4.8 and 4.9). These quantify the strong statistical correlation between the DEs and the themes, being particularly strong in the third-through-seventh modes. In the exemplary case of Figure 4.4b, themes corresponding to the DEs of the seventh mode of the e2xyiA1 propeller are assigned a high AMI score of 0.91. For comparison, Figure 4.29 typifies a case with low similarity.

Ten Sample Combinations of Themes for AMI-SMI Analysis (e2xyiA1)									
Comb-1	**Theme**	**14940-1**	**14946-1**	**14942-2**	**14954**	**14937**	**14816**	**14861-1**	**14951**
	Residue	14-53	55-120	126-170	176 220	227-265	271-309	314-353	356-415
Comb-2	**Theme**	**14940-1**	**14946-1**	**14942-2**	**14954**	**14939**	**14950**		
	Residue	14-53	55-120	126-170	176 220	227-309	314-408		
Comb-3	**Theme**	**14940-1**	**14946-2**	**14942-2**	**114813-2**	**14815-1**	**14861-1**	**14951**	
	Residue	14-53	58-91	126-170	176-240	244-309	314-353	356-415	
Comb-4	**Theme**	**14945**	**14942-2**	**14813-1**	**15218**	**15219**	**14861-1**	**14951**	
	Residue	16-120	126-170	176-231	230-270	281-315	314-353	356-415	
Comb-5	**Theme**	**14945**	**14942-2**	**14813-1**	**15221**	**15222**			
	Residue	16-120	126-170	176-231	230-330	336-407			
Comb-6	**Theme**	**14940-2**	**14946-1**	**14942-2**	**14813-1**	**14937**	**14816**	**14861-1**	**14951**
	Residue	21-55	55-120	126-170	176-231	227-265	271-309	314-353	356-415
Comb-7	**Theme**	**14940-2**	**14946-1**	**14942-2**	**14813-1**	**14939**	**14950**		
	Residue	21-55	55-120	126-170	176-231	227-309	314-408		
Comb-8	**Theme**	**14940-3**	**14946-2**	**14942-2**	**14813-1**	**14938**	**14955**	**14815-4**	
	Residue	21-59	58-91	126-170	176-231	227-287	283-341	343-406	
Comb-9	**Theme**	**14940-3**	**14946-2**	**14942-2**	**14813-1**	**15218**	**14816**	**14861-1**	**14951**
	Residue	21-59	58-91	126-170	176-231	230-270	271-309	314-353	356-415
Comb-10	**Theme**	**14940-3**	**14946-2**	**14942-2**	**14813-1**	**15221**	**15222**		
	Residue	21-59	58-91	126-170	176-231	230-330	336-407		

Figure 4.28. Ten sample possible theme combinations of the e2xyiA1 propeller structure, filtered with 3-residue overlap and 8-residue gap restriction.

Table 4.7. AMI and SMI values for the correlation between the DEs of each of the seven slowest GNM modes of e2xyiA1 and theme combinations, filtered with thresholds of a 3-residue overlap and an 8-residue gap.

AMI							
	Mode1	**Mode2**	**Mode3**	**Mode4**	**Mode5**	**Mode6**	**Mode7**
Minimum	0.49	0.52	0.61	0.54	0.65	0.58	0.72
Maximum	0.67	0.64	0.73	0.71	0.80	0.71	0.91
Average	0.55	0.57	0.66	0.63	0.73	0.64	0.82
SMI							
	Mode1	**Mode2**	**Mode3**	**Mode4**	**Mode5**	**Mode6**	**Mode7**
Minimum	86	92	98	79	87	94	123
Maximum	127	128	133	109	111	124	160
Average	99	106	112	94	99	108	137

Table 4.8. AMI and SMI values for the correlation between the DEs of each of the seven slowest GNM modes of e2xyiA1 and theme combinations, filtered with thresholds of a 5-residue overlap and a 10-residue gap.

AMI							
	Mode1	**Mode2**	**Mode3**	**Mode4**	**Mode5**	**Mode6**	**Mode7**
Minimum	0.48	0.52	0.60	0.54	0.62	0.58	0.69
Maximum	0.67	0.69	0.73	0.71	0.80	0.72	0.91
Average	0.55	0.59	0.66	0.63	0.71	0.64	0.79
SMI							
	Mode1	**Mode2**	**Mode3**	**Mode4**	**Mode5**	**Mode6**	**Mode7**
Minimum	88	92	99	78	92	97	125
Maximum	130	129	133	108	110	125	160
Average	101	108	112	93	102	111	139

Table 4.9. AMI and SMI values for the correlation between the DEs of each of the seven slowest GNM modes of e2xyiA1 and theme combinations, filtered with thresholds of a 5-residue overlap and an 15-residue gap.

AMI							
	Mode1	**Mode2**	**Mode3**	**Mode4**	**Mode5**	**Mode6**	**Mode7**
Minimum	0.48	0.52	0.59	0.54	0.58	0.57	0.64
Maximum	0.68	0.73	0.77	0.76	0.81	0.74	0.94
Average	0.54	0.60	0.66	0.65	0.72	0.64	0.79
SMI							
	Mode1	**Mode2**	**Mode3**	**Mode4**	**Mode5**	**Mode6**	**Mode7**
Minimum	91	97	99	89	91	98	126
Maximum	131	131	132	117	121	126	166
Average	104	111	113	101	106	112	142

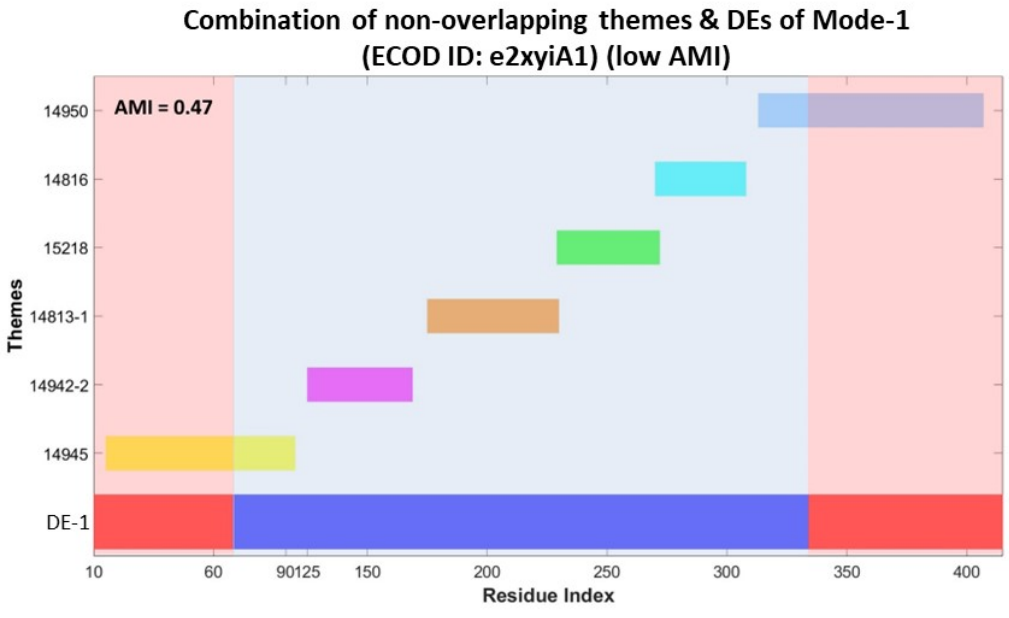


Figure 4.29. An exemplary case of low AMI (AMI <0.50) correspondence between themes and DEs for the e2xyiA1 propeller.

We also applied mutual information analysis to all domains examined in detail above (the in-depth set). For this set, we used thresholds of a 3-residue overlap and an 8-residue gap. We note that the mutual information analysis excluded the alpha-helix bundle domains studied in detail above (e2of3A1, e1b3uA1, and e4adyA2) because themes only partially cover each of these protein domains, preventing a meaningful calculation of the correlation with the DEs. Mutual information analysis results for the in-depth set present equally significant correlations (Table 4.10, Table 4.11 and Table 4.12).

Table 4.10. Mean, standard deviation, and maximum values of AMI and SMI for each ECOD domain in the selected set.

Domain ID	**Mean AMI**	**Std Dev (AMI)**	**Max AMI**	**Mean SMI**	**Std Dev (SMI)**	**Max SMI**
e2xyiA1	0.66	0.09	0.91	108	15	160
e3emhA1	0.59	0.12	0.92	99	20	178
e1j6oA1	0.64	0.05	0.74	90	12	113
e2gzxA1	0.66	0.07	0.81	89	12	120
e4p5uA1	0.66	0.09	0.81	94	18	140
e3ktzA1	0.68	0.09	0.89	87	10	114
e3cktA1	0.63	0.08	0.84	83	12	126
e1ulsB1	0.64	0.10	0.91	78	16	126
e2ae2A1	0.64	0.08	0.80	85	17	130
e3n74B1	0.58	0.09	0.84	86	11	119
Average	0.64	0.09	0.85	90	14	133

Table 4.11. Average AMI values for the DEs of each of the seven slowest GNM modes for every ECOD domain.Italic: AMI less than 0.5; Bold: AMI greater than 0.6.

DomainID	**Mode1**	**Mode2**	**Mode3**	**Mode4**	**Mode5**	**Mode6**	**Mode7**
e2xyiA1	0.55	0.57	**0.66**	**0.63**	**0.73**	**0.64**	**0.82**
e3emhA1	0.57	0.59	N/A	0.55	**0.76**	**0.68**	*0.41*
e1j6oA1	**0.68**	0.54	**0.64**	**0.67**	**0.64**	**0.70**	**0.62**
e2gzxA1	**0.62**	0.56	**0.68**	**0.71**	**0.67**	**0.67**	**0.71**
e4p5uA1	**0.78**	0.51	**0.64**	**0.66**	**0.69**	**0.66**	**0.70**
e3kyzA1	**0.64**	**0.61**	**0.75**	**0.66**	**0.83**	**0.62**	**0.68**
e3cktA1	**0.66**	0.58	**0.70**	**0.64**	**0.68**	0.51	**0.64**
e1ulsB1	*0.44*	**0.64**	**0.62**	**0.63**	**0.66**	**0.74**	**0.70**
e2ae2A1	**0.65**	**0.69**	0.59	0.51	**0.75**	**0.68**	**0.64**
e3n74B1	**0.63**	0.51	0.56	**0.63**	0.51	**0.71**	0.54

Table 4.12. Maximum AMI values for the DEs of each of the seven slowest GNM modes for every ECOD domain.Italic: AMI less than 0.6; Bold: AMI greater than 0.7.

DomainID	**Mode1**	**Mode2**	**Mode3**	**Mode4**	**Mode5**	**Mode6**	**Mode7**
e2xyiA1	0.67	0.64	**0.73**	**0.71**	**0.80**	**0.71**	**0.91**
e3emhA1	**0.73**	0.67	N/A	0.66	**0.92**	**0.83**	*0.55*
e1j6oA1	**0.72**	*0.56*	0.65	0.68	0.64	**0.74**	0.62
e2gzxA1	**0.78**	0.64	**0.75**	**0.81**	**0.77**	0.70	**0.80**
e4p5uA1	**0.81**	*0.57*	**0.72**	**0.71**	**0.78**	0.69	**0.74**
e3ktzA1	**0.73**	0.69	**0.79**	**0.73**	**0.89**	0.65	**0.76**
e3ctkA1	**0.81**	0.67	**0.84**	**0.73**	**0.76**	0.66	**0.75**
e1ulsB1	*0.51*	0.69	0.68	**0.80**	**0.76**	**0.91**	**0.79**
e2ae2A1	0.68	**0.79**	0.69	*0.56*	**0.80**	**0.77**	0.66
e3n74B1	**0.75**	0.61	**0.70**	**0.76**	0.61	**0.84**	0.69

Table 4.10 lists the mean AMI and SMI values of the mutual information analysis for the domains in the in-depth set, together with the maximum AMI and SMI values observed in each domain. The averages of the mean AMI and SMI values over the ten domains are 0.64 and 90, respectively, with corresponding standard deviations of 0.09 and 14. When the maximum AMI and SMI values for each domain are considered, the analysis gives 0.85 as the average of maximum AMI values and 133 as the average of the maximum SMI values (Table 4.10). The correlations at the resolution of individual modes are listed in Table 4.11 (calculated mean AMI values) and Table 4.12 (calculated maximum AMI values). DEs of the vast majority of the slow modes show significant correlation with the themes. Overall, our results for representatives of different protein families strongly support the correlation between the DEs and the themes. Furthermore, we observe that DEs and themes show a high correlation (AMI > 0.70) in multiple modes of motion for all cases, suggesting that themes contribute to protein dynamics in multiple wirings.

Lastly, we performed mutual information analysis on the expanded set of 150 domains to further support our findings. In the expanded set, we observed a high correlation (AMI > 0.70) for an average of 3.9 modes in the 150 domains examined. For 131 of the 150 domains, high AMI values were observed for two or more modes of motion.

4.1.6. P-value analysis for the statistical significance of correspondence between themes and DEs

To further consolidate and reinforce our findings, we used conventional mutual information analysis to assign p-values to the correlations between themes and DEs. This analysis provides an additional independent measure of significance beyond the AMI and SMI values, which inherently consider the expected mutual information.

We generated random sets of continuous protein segments resembling the themes identified in each domain. We calculated the mutual information (MI) between random

set combinations and the DEs, and between the real theme combinations and the DEs in each individual mode. Then, we conducted p-value calculations based on the distribution of the MI of the random set combinations and the DEs. Probability distribution of MI for the randomly generated themes of the e2xyiA1 propeller and DEs of mode 5 is provided in Figure 4.30, with the MI and p-value obtained for the real themes marked on the distribution graph. Details regarding the random generation process and p-value calculations are provided in the Materials and Methods section.

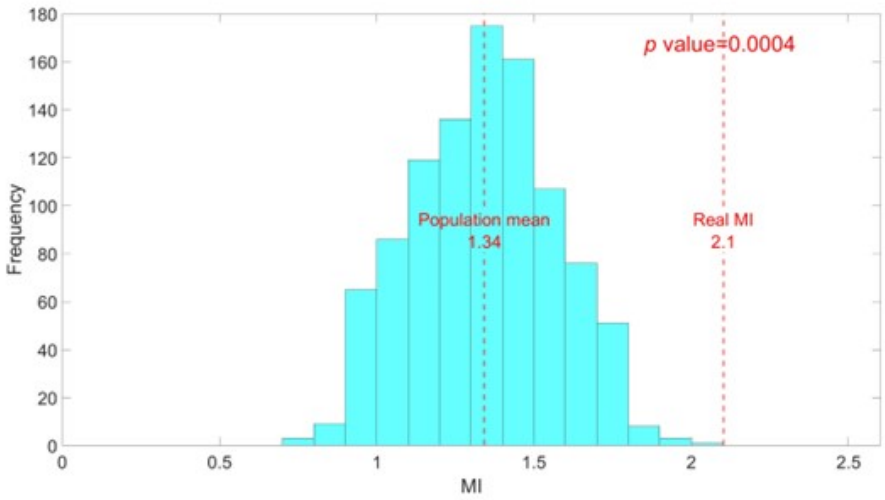


Figure 4.30. Probability distribution of MI for the randomly generated themes of the e2xyiA1 propeller and DEs of mode 5.

Table 4.13 lists the p-values assigned to the calculated MI between the themes and DEs for each of the seven slowest modes of motion in each of the 10 domains considered from the in-depth set. Significant p-values (< 0.05) were assigned to multiple individual modes of motion in each of the domains. Figure 4.31 shows a histogram of the number of modes with significant p-values (< 0.05) per domain in the expanded set of 150 domains. The distribution averages at 3.8 modes of motion with a standard deviation of 2.1 modes. We observed at least one mode with significant p-value (< 0.05) for 136 domains out of 150, and multiple modes with significant p-value (< 0.05) for 122 domains out of 150.

Table 4.13. P-values for the DEs of each of the seven slowest GNM modes for every ECOD domain. The last column reports the number of modes with p-values < 0.05.

Domain ID	**Mode-1**	**Mode-2**	**Mode-3**	**Mode-4**	**Mode-5**	**Mode-6**	**Mode-7**	# $p <$ 0.05
e2xyiA1	0.0165	0.0224	0.0133	0.0022	0.0004	0.0132	0.0001	7
e3emhA1	0.0330	0.1131	N/A	0.0865	0.0013	0.0913	0.1929	2
e1j6oA1	0.0143	0.2355	0.0116	0.0186	0.0165	0.0046	0.0437	6
e2gzxA1	0.0040	0.1107	0.0074	0.0049	0.0013	0.0217	0.0023	6
e4p5uA1	0.0015	0.1558	0.0890	0.0133	0.0013	0.0680	0.0052	4
e3ktzA1	0.0187	0.0354	0.0081	0.0247	0.0004	0.1228	0.0177	6
e3cktA1	0.0089	0.0719	0.0199	0.0288	0.0108	0.1850	0.0756	4
e1ulsB1	0.1621	0.0669	0.0736	0.0102	0.0227	0.0073	0.0437	4
e2ae2A1	0.0363	0.0574	0.0286	0.0167	0.0138	0.0660	0.0095	6
e3n74B1	0.1936	0.1212	0.0388	0.0212	0.1212	0.0063	0.1111	3

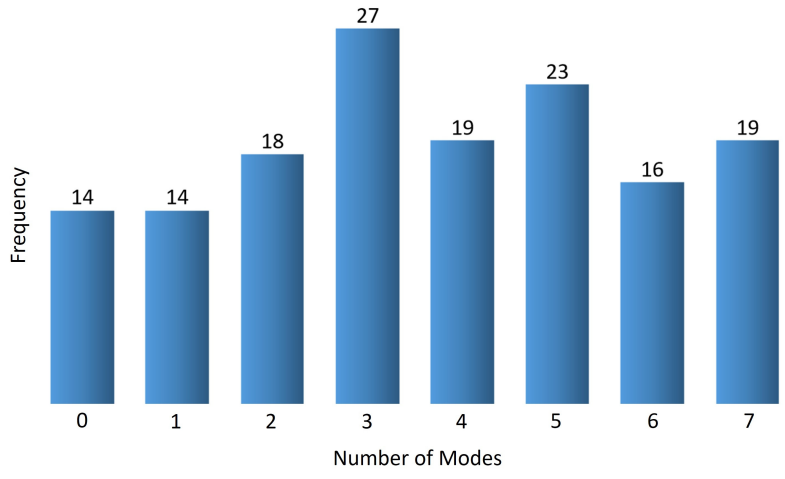


Figure 4.31. The distribution of statistical significance per domain. The x axis is the number of modes with statistically significant (P-value < 0.05) correlations with themes in each domain, and the y axis is the domain frequency.

It is important to note that domains lacking modes associated with a significant p-value typically manifest fewer themes and smaller domain sizes compared to their counterparts. Overall, the p-value analysis shows that the correspondences between themes and DEs in multiple levels shown by MI are statistically significant.

4.1.7. Comparison of DEs with GNM and MD simulations

GNM, in spite of its simplicity and use of approximated linear dynamics, can reveal long-time behavior [1, 62, 133–135] and allow us to observe protein dynamics on multiple levels, in various slow modes of motion; we utilize these properties in the present work. Nevertheless, to illustrate the comparison of the dynamic dissection into simpler components by GNM and molecular dynamics (MD) simulations, we analyzed type-2 angiotensin 2 receptor (AT2) (PDB ID: 5UNH). MD trajectory, topology, and model files were obtained from the GPCRmd dataset [136]. The accumulated simulation time of the MD simulation is 1.5 µs. We calculated the correlations between residue fluctuations and performed principal component analysis (PCA) on the covariance matrix to obtain eigenvalues and eigenvectors. We compared the seven slowest GNM modes individually with the MD principal components (PCs) that capture the largest variations. Each PC is binarized to 1 for positive and -1 for negative correlations between residues. We observe one-to-one correspondence between GNM modes and PCs (Table 4.14, Figures 4.32-38), and the DEs obtained with GNM and MD are observed to be very similar (Figure 4.39).

Table 4.14. Pearson correlation coefficient between GNM modes and principal components (PCs).

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**
GNM 1	0.13	0.11	0.86	-0.03	0.00	0.17	-0.27
GNM 2	-0.19	0.09	0.59	0.33	0.21	-0.08	0.03
GNM 3	0.82	-0.11	-0.04	0.02	0.09	0.01	-0.05
GNM 4	-0.35	0.03	0.16	0.52	0.14	0.04	0.30
GNM 5	0.21	-0.04	0.00	-0.17	0.58	0.69	-0.28
GNM 6	0.12	0.10	-0.02	-0.41	0.59	-0.66	-0.13
GNM 7	0.23	-0.96	0.00	0.11	-0.08	0.02	0.00

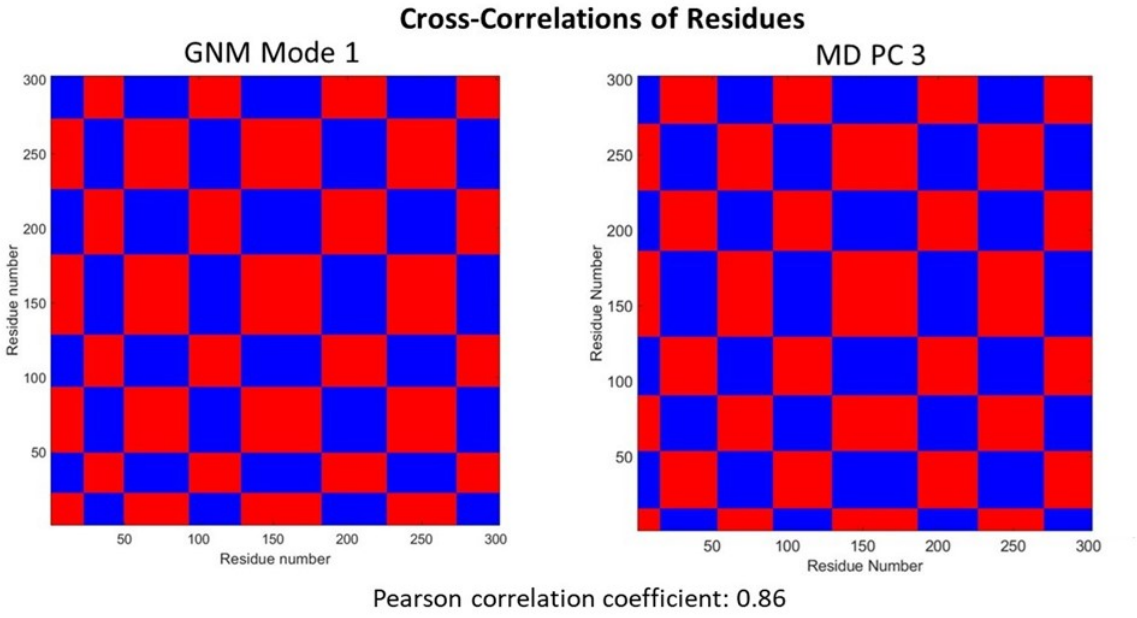


Figure 4.32. One-to-one correspondence between the slowest GNM mode and PC3 obtained from MD simulation.

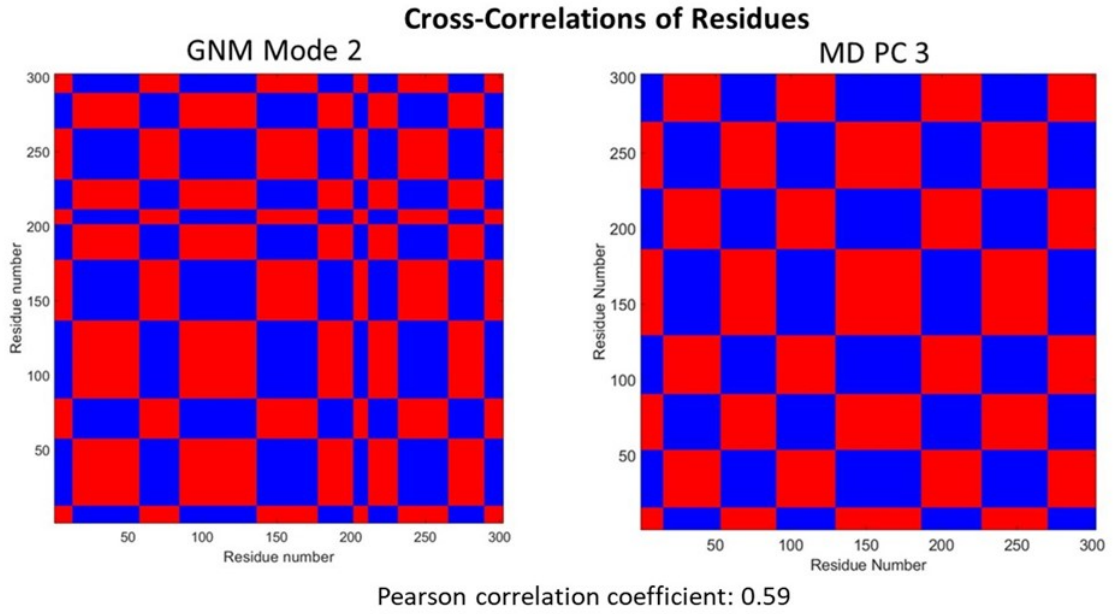


Figure 4.33. One-to-one correspondence between the second slowest GNM mode and PC3 obtained from MD simulation.

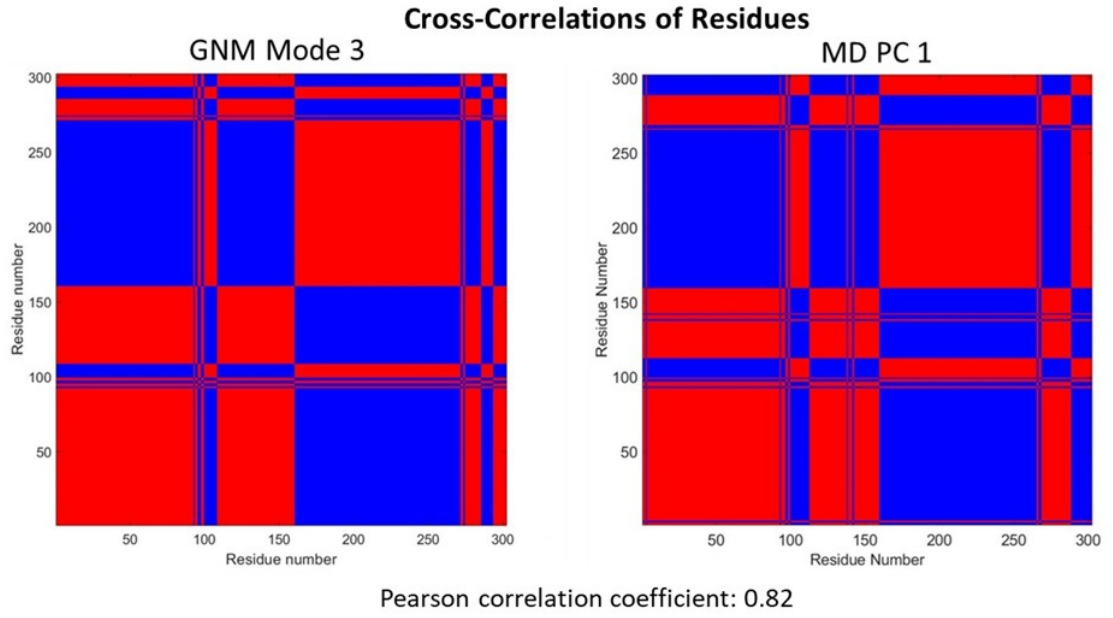


Figure 4.34. One-to-one correspondence between the third slowest GNM mode and PC1 obtained from MD simulation.

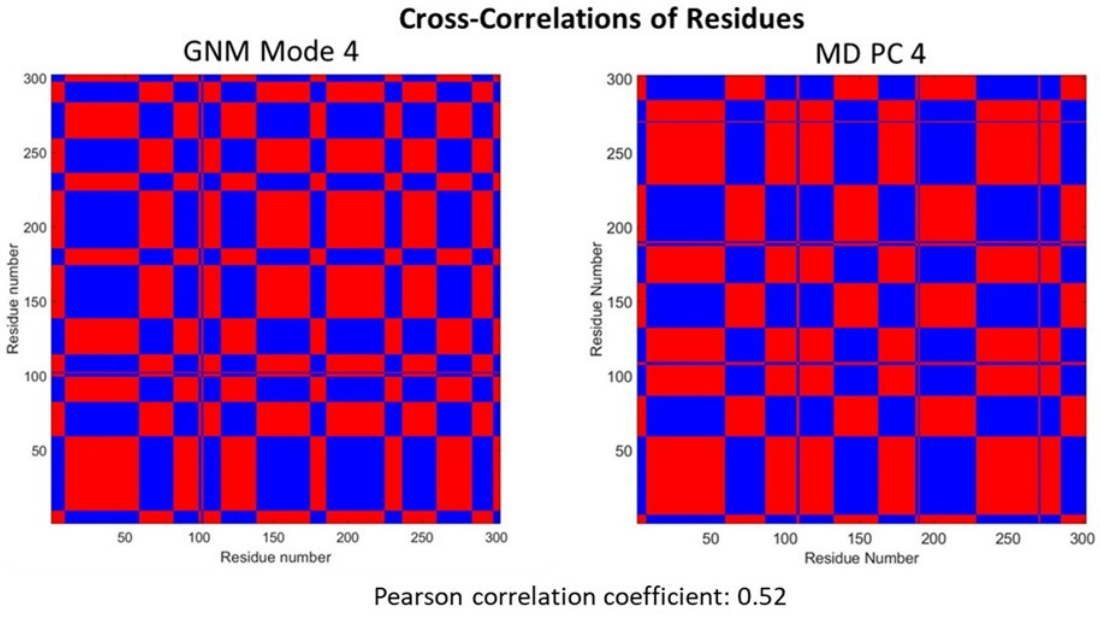


Figure 4.35. One-to-one correspondence with relatively low correlation between the fourth slowest GNM mode and PC4 obtained from MD simulation.

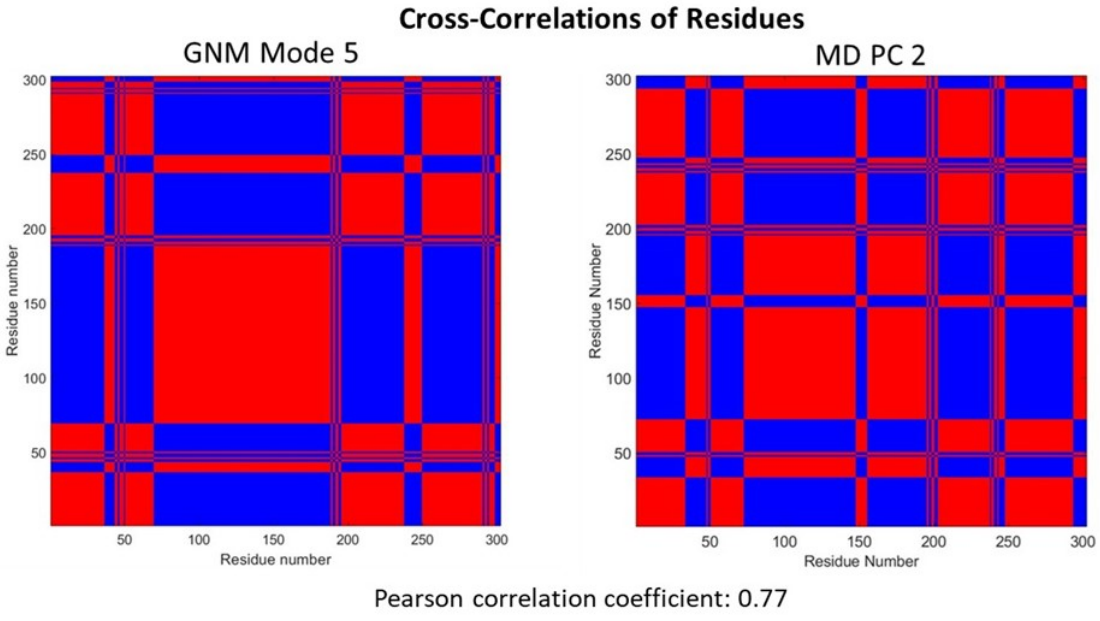


Figure 4.36. One-to-one correspondence between the fifth slowest GNM mode and PC2 obtained from MD simulation.

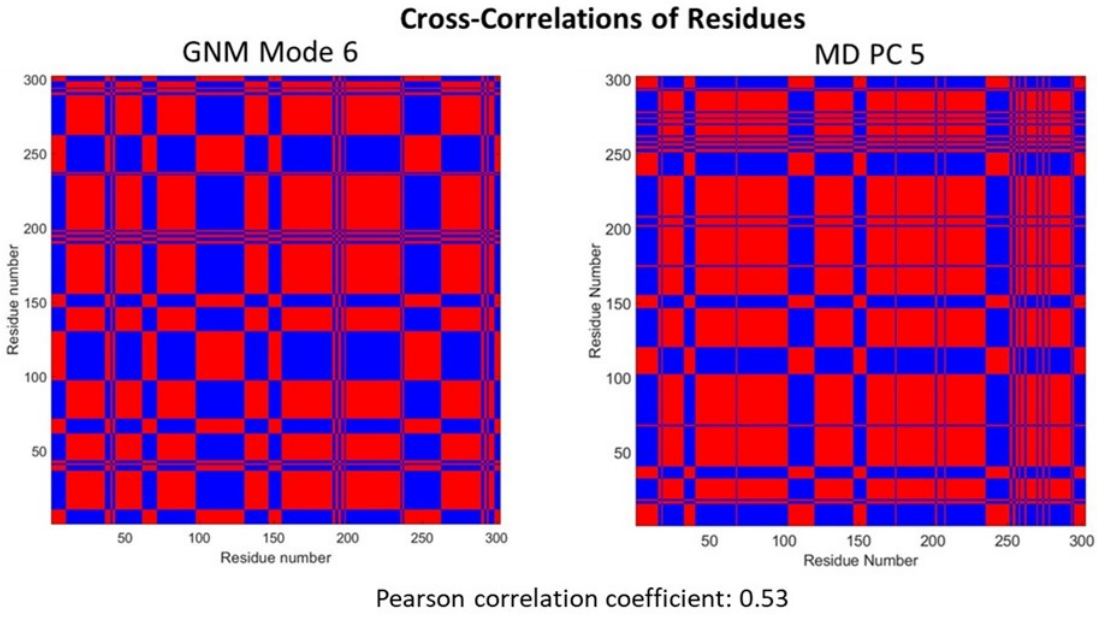


Figure 4.37. One-to-one correspondence with relatively low correlation between the sixth slowest GNM mode and PC5 obtained from MD simulation.

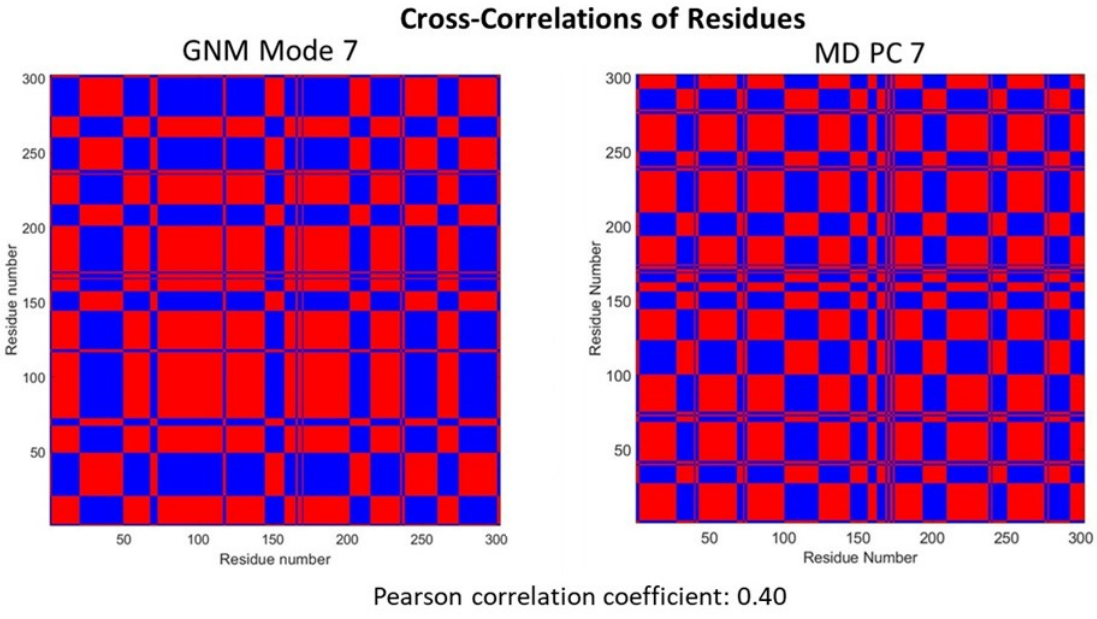


Figure 4.38. One-to-one correspondence with relatively low correlation between the seventh slowest GNM mode and PC7 obtained from MD simulation.

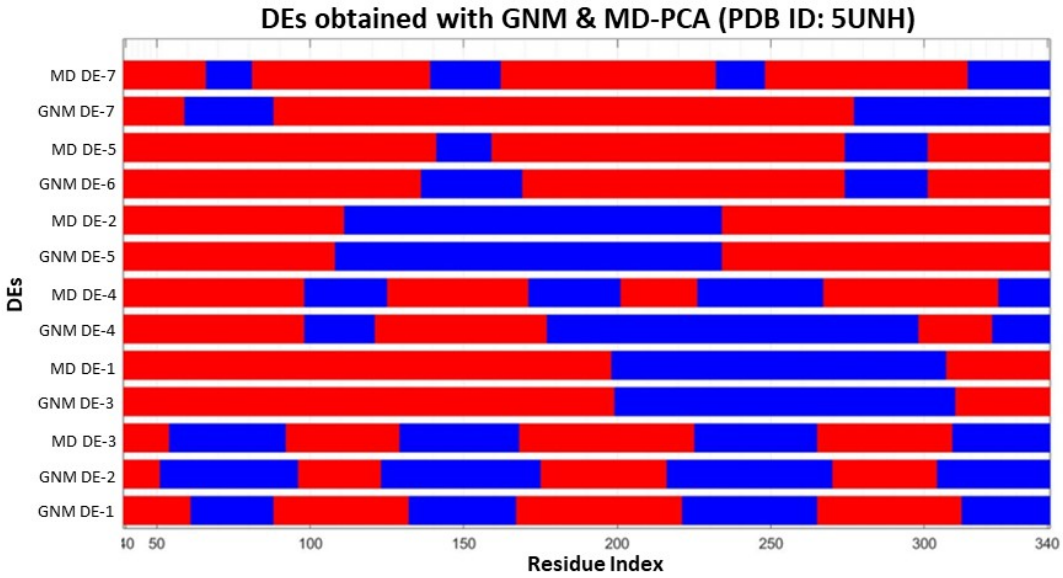


Figure 4.39. DEs obtained with GNM and MD-PCA. DEs obtained via two different methods are observed to be very similar, save differences between the seventh mode of GNM and PC7 of MD.

To note, the slowest and second slowest GNM modes correspond to the same PC (PC3), and PC6 does not correspond to any of the slow modes. The variations (eigenvalues) captured by each PC may change with the length of the MD simulation. However, the essential motions represented by the corresponding PCs seem to be well accounted for with slow modes of GNM.

We also compare here the DEs with themes of GPCRs (ECOD X-group 5001) via the HHsearch engine. Since the conservation levels among GPCRs are high, we obtain relatively long themes compared to the ones in our in-depth set. Yet, we observe that the correspondence between themes and DEs (obtained with both GNM and MD simulations) still holds. Theme-1 corresponds to two DEs of the third slowest mode of GNM and PC1. Furthermore, Theme-2 and Theme-3 correspond to a single DE (first and second DEs) of the same GNM mode and the same PC (Figure 4.40).

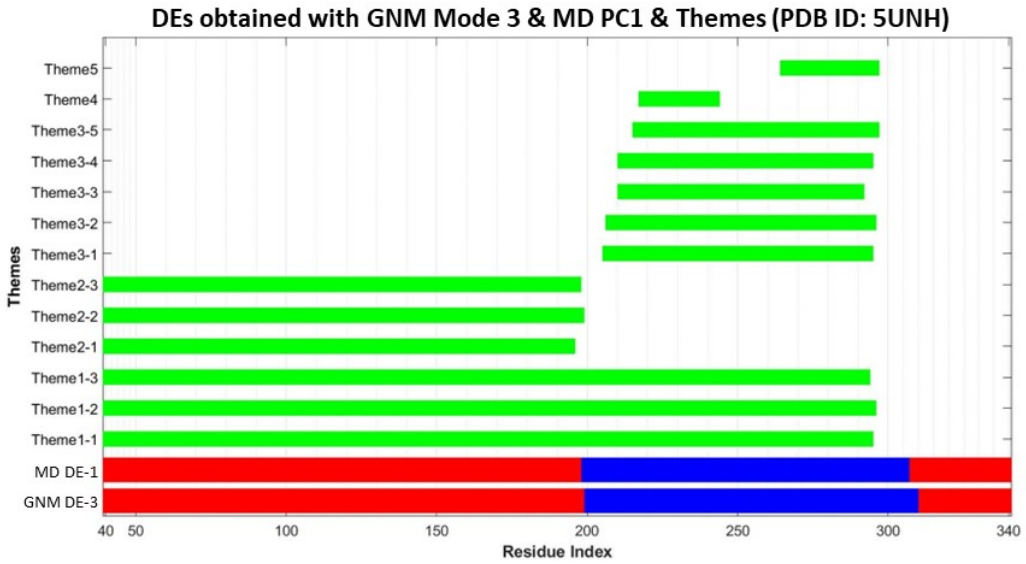


Figure 4.40. Themes and DEs of the first MD-PCA and the third slowest mode of GNM. Theme-1 corresponds to two DEs of the third slow mode GNM and first mode of MD-PCA. Theme-2 corresponds to the first DE and Theme-3 corresponds to the second DE of the same modes.

4.1.8. Correspondence between DEs and Pfam entries

Of the 150 ECOD domains in our dataset, only 61 can be defined with Pfam entries [137] using the approach that we used with the themes. Unfortunately, 55 of these domains are defined by a single Pfam entry and 2 of them are defined by repetition of the same Pfam entry. This leaves us with 4 ECOD domains in which the same definition of overlap and gap residues can be used to examine the correspondence between Pfam entries and DEs. These four ECOD domains are e4kliA3, e2bcqA3, e4drxF1, and e3ro3A1, and interestingly, in each of these four cases, the Pfam entry combinations correspond to DEs of one of the seven slowest modes. AMI analysis on these four domains shows that the correspondence of the partitioning into Pfam entries with the DEs is as high as that of the themes (Table 4.15).

Table 4.15. Correspondence between Pfam entries and DEs in four possible ECOD domains via AMI analysis.

Domain ID	**DEs Mode-1**	**DEs Mode-2**	**DEs Mode-3**	**DEs Mode-4**	**DEs Mode-5**	**DEs Mode-6**	**DEs Mode-7**
e4kliA3	0.89	0.70	0.71	0.55	0.31	0.28	0.32
e2bcqA3	1.00	0.57	0.44	0.25	0.70	0.50	0.62
e4drxF1	0.30	0.21	0.53	0.76	0.41	N/A	0.84
e3ro3A1	1.00	0.38	0.52	0.46	0.62	0.61	0.59

4.1.9. Impact of protein conformational changes on GNM modes and theme correspondence

The GNM analysis, used to detect the DEs, depends on the particular conformation used for the domain. Starting from various conformations of the same domain would yield various partitions into DEs, which may differ from each other in their correlations with the themes of the domain. (Obviously, the themes, derived from sequence analysis, would be exactly the same for all the conformations.) To examine the effect of different conformations on the DE-themes correlation, we conducted a thorough analysis of our ECOD domain dataset to identify and evaluate any significant structural changes. First, we scanned the dataset to identify other ECOD domains belonging to the same protein (as indicated by at least 99% sequence identity) that exhibit at least a moderate structural change, defined as an RMSD greater than 2.5 Å. Out of the 150 ECOD domains in our dataset, 145 do not have a counterpart ECOD domain showing significant structural changes. This leaves only five ECOD domains with counterparts exhibiting moderate structural changes. That only a handful of domains undergo conformational changes suggests that for most proteins, comprising multiple domains, conformational alterations are based on motions of the domains as rigid parts. (However, it is important to note that even relatively minor conformational changes can be important for the physiological function of the domain.) Among the

five domains with alternative conformations, three have counterparts with significant differences in sequence lengths due to missing or uncaptured parts in the 3D structures. In these cases, the themes cover regions larger than the length of the sequence, and the differences in sequence lengths result in the segments obtained from GNM and the DEs defining different regions. Consequently, it is not feasible to compare the GNM modes or the correspondence between themes and DEs for these three domains.

This leaves us with two domains that can be meaningfully compared: e4qvhB1 and e1twfA10. Their counterparts, which exhibit moderate structural changes, are e4blaB3 and e3j1nA7, with RMSD values of 3.96 Å and 3.15 Å respectively. We conducted the exact same analysis using the two structurally altered counterparts, and compared the results to the results obtained for the corresponding conformations.

Figure 4.41 illustrates the correlation matrices of the GNM modes for these domain pairs. For ECOD domain e4qvhB1, we found that six out of seven GNM modes have corresponding modes (correlation coefficient >0.75) in its conformationally changed counterpart, e4blaB3, and vice-versa. For ECOD domain e1twfA10, all seven GNM modes have corresponding modes in its conformationally changed counterpart, e3j1nA7, and vice-versa. Like the observation with homologous protein domains (Figure 4.26), we observed some shifts in mode space (e.g., the second dynamic mode of e1twfA10 corresponds to the third dynamic mode of e3j1nA7).

Subsequently, we projected the themes from the ECOD domains in our dataset onto their conformationally changed counterparts and calculated the AMI values to assess the correspondence between themes and DEs in these domains. We documented the maximum AMI values obtained for each mode in Tables 4.16 and 4.17, similarly to our treatment of ECOD domains within the dataset.

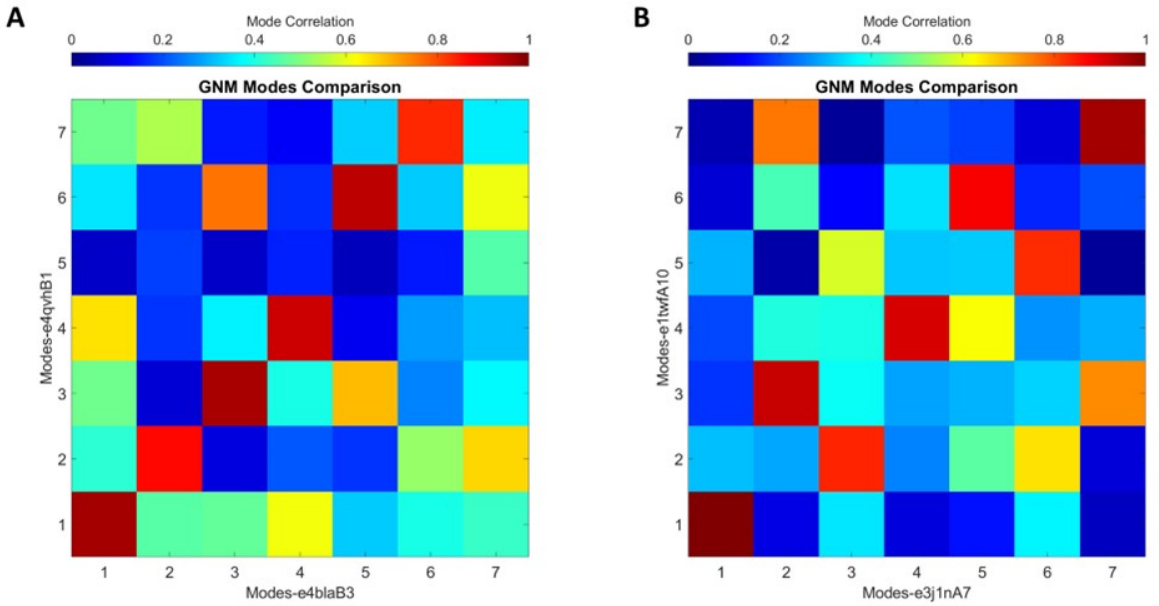


Figure 4.41. GNM mode comparisons of the ECOD domains in the dataset and their conformationally changed counterpart ECOD domains. The correlation coefficient between modes is represented with colors where dark red is 1, green is 0.5 and dark blue is 0.

Table 4.16. The maximum AMI values for individual modes of e4qvhB1 and its counterpart domain e4blaB3.

Domain ID	**DEs Mode-1**	**DEs Mode-2**	**DEs Mode-3**	**DEs Mode-4**	**DEs Mode-5**	**DEs Mode-6**	**DEs Mode-7**
e4qvhB1	0.73	0.69	0.79	0.68	0.70	0.75	0.64
e4blaB3	0.73	0.73	0.74	0.68	0.78	0.77	0.75

Table 4.17. The maximum AMI values for individual modes of e1twfA10 and its counterpart domain e3j1nA7.

Domain ID	**DEs Mode-1**	**DEs Mode-2**	**DEs Mode-3**	**DEs Mode-4**	**DEs Mode-5**	**DEs Mode-6**	**DEs Mode-7**
e1twfA10	0.34	0.53	0.66	0.53	0.66	0.73	0.40
e3j1nA7	0.31	0.67	0.52	0.53	0.63	0.78	0.67

The fifth mode of domain e4qvhB1 lacks a corresponding mode in its counterpart domain e4blaB3, and conversely, the seventh mode of e4blaB3 has no counterpart in e4qvhB1. Despite these discrepancies, the DEs derived from these modes exhibit high correspondence with the themes. Conversely, the seventh modes of e1twfA10 and e3j1nA7 display a strong correlation, yet the resulting DEs differ, leading to a change in the correspondence with the themes. Notably, the seventh mode of e3j1nA7 exhibits a higher AMI value than e1twfA10. Despite minor differences between the GNM modes of ECOD domains and of their conformationally changed counterparts, resulting in slight modifications in DEs, our observations suggest that these differences do not significantly alter the correspondence between themes and DEs, at least not to an extent that would impact the findings of our study

4.1.10. Conclusion

This study provides systematic evidence for a non-trivial correspondence between evolutionarily reused protein segments (themes) and intrinsic dynamic elements (DEs) derived from GNM analysis. Themes and DEs are two independently defined types of sub-domain segments, one based on sequence similarity and evolutionary recurrence, the other on structural dynamics and collective motion. Despite their distinct origins, we observed frequent and statistically significant boundary overlaps between these two segment types across multiple proteins.

In the 13 domains examined in-depth, we found that themes often coincide with DEs, suggesting that the reused segments are not only conserved in sequence but also dynamically coherent. This observation was reinforced by mutual information and p-value analyses across a larger dataset of 150 ECOD domains, indicating that the correspondence is unlikely to be due to chance alone. Comparisons with Pfam domains and molecular dynamics (MD) simulations further validated the robustness of the DEs, while structural variation analysis confirmed the persistence of this correspondence across conformational states.

These findings parallel earlier work showing that structural domains correspond to moving parts in normal modes and participate in long-range allosteric communication [48,69,98]. While domains are known to fold independently, themes are typically shorter and unlikely to form autonomously folded units. Yet, their recurrence in ancient protein families [28, 29, 31] and their association with functional roles such as ligand binding reinforce their evolutionary and functional significance.

Our results support the hypothesis that domains may have emerged through the dynamic "mixing and matching" of themes. For such combinations to persist evolutionarily, not only geometric but also dynamic compatibility is likely required. The observed alignment between themes and DEs offers insight into how modularity in evolution may extend into protein dynamics and function.

While the results are compelling, caution is warranted. Our mutual information analyses do not account for all inter-residue dependencies such as secondary structure or packing effects, and the observed correlation should not be interpreted as direct causality. Nevertheless, the results open new avenues for applications in protein design, where identifying reused, dynamically coherent segments may aid in grafting or modular engineering strategies [138, 139]. More broadly, the work highlights the importance of integrating protein dynamics into models of evolutionary emergence, suggesting that functionally meaningful motion may itself be an evolutionary constraint.

4.2. Intrinsic Dynamics and Allosteric Roles of Conserved Adenine-Binding Themes

4.2.1. Motivation and Background

Proteins often contain sequence and structural segments that recur across unrelated folds and evolutionary lineages. These conserved elements are thought to represent evolutionary building blocks—segments that have diverged from a common ancestor and were reused across diverse proteins due to their functional and structural advantages [4, 5, 12]. While protein domains are well-known examples of such reusable units, more recent studies have identified shorter, subdomain segments referred to as "themes" [33]. Unlike domains, themes can overlap with one another and are not mutually exclusive; they represent flexible, often functionally important units that span 20–80 residues and appear across diverse protein folds.

These themes have been associated with various biochemical functions. For instance, prior research identified recurring themes involved in electron transfer, metal binding, and phosphate recognition [27, 37, 38]. Particularly notable are adenine-binding themes, which were characterized in a large-scale study by Narunsky et al. (2020) [36]. These themes are located in the binding pockets of proteins that interact with adenine-containing ligands such as ATP and NAD. Although the proteins may share low sequence identity and different folds, they frequently reuse the same theme sequences to mediate adenine recognition, supporting the notion that these segments may be functionally convergent or ancestrally conserved.

Despite their functional significance, the dynamic roles of such themes remain underexplored. Traditional functional annotation focuses on static structural features, yet many protein functions—including ligand recognition, allostery, and regulation—depend on intrinsic dynamics. In this context, the Gaussian Network Model (GNM) provides a powerful method for identifying slow, collective motions that are intrinsic to protein structure and often correlated with biological activity [41, 52, 53].

Furthermore, recent advances have extended GNM to analyze directional information flow between residues using transfer entropy (GNM-TE), allowing researchers to detect not only correlations but causal couplings between residue fluctuations [55–58].

In the preceding chapter (Section 4.1; Kutlu et al. (2024) [59]), we investigated the relationship between reused protein themes and intrinsic dynamics across a wide range of structural families. Using GNM analysis, we demonstrated that many sequence-derived themes align with DEs —coherent units of motion extracted from slow GNM modes—supporting a general correspondence between evolutionary reuse and collective protein motions. While that study provided the first large-scale statistical evidence for such a link, its treatment of functional roles was limited to geometric correspondence, with allosteric signaling explored in only a single illustrative case. The extent to which theme-residing residues actively participate in long-range communication and dynamic regulation remains an open question. In this chapter, we specifically address that gap by focusing on adenine-binding themes and investigating their potential roles as embedded allosteric units using GNM and GNM-TE analysis.

4.2.2. Data and Approach

4.2.2.1. Dataset of Adenine-Binding Themes. This chapter focuses on adenine-binding themes as defined by Narunsky et al. (2020) [36], who systematically analyzed protein–adenine interactions across the Protein Data Bank. From their evolutionary analysis, they identified 10 distinct clusters of adenine-binding motifs based on sequence and structural similarity. Each cluster consists of one or more themes—short sequence segments recurrently found in adenine-binding regions—and associated protein structures.

To investigate the relationship between these themes and intrinsic protein dynamics, we selected Cluster 1, which includes 20 adenine-binding sites, the majority of which adopt the well-characterized direct motif. After filtering for structural quality and consistency, we retained 22 high-quality structures and 9 distinct adenine-binding

themes from this cluster for further analysis.

To test the broader applicability of our findings beyond Cluster 1, we extended our analysis to three additional clusters described by Narunsky et al. Cluster 2, similar to Cluster 1, is dominated by the direct binding motif but also features an additional interaction between adenine's N7 atom and a conserved lysine residue. Cluster 4 represents the reverse motif, characterized by a distinctive N6–backbone interaction. In contrast, Cluster 6B consists of NAD-binding proteins and is associated with the "Asp" motif, which includes a notable interaction between the N3 atom of adenine and the protein. These clusters offer structural and functional variation in terms of binding modes, ligand classes, and residue-level interactions. [36]

A consistent filtration procedure was applied to all three clusters, similar to that used for Cluster 1. Following this refinement, the final dataset used in this chapter comprises 15 structures with 15 themes from Cluster 2, 7 structures with 16 themes from Cluster 4, and 32 structures with 7 themes from Cluster 6B. Together, these clusters provide a structurally and functionally diverse benchmark for evaluating the dynamic behavior of adenine-binding themes and their potential roles in allosteric communication.

4.2.2.2. GNM and GNM based dynamic elements. We used the Gaussian Network Model (GNM) to characterize intrinsic protein dynamics based on residue–residue interactions in native structures [52, 53]. GNM enables the identification of cooperative motions by analyzing the low-frequency modes of the network's fluctuation spectrum. In each mode, residues that exhibit coordinated fluctuations—either in the same direction (positive correlation) or opposite directions (negative correlation)—are grouped into dynamic segments [72].

To define meaningful dynamic units, we applied a previously developed procedure (Section 4.1; Kutlu et al., 2024 [59]), which projects mode-specific correlated regions onto the sequence of the protein. Contiguous stretches of residues moving

coherently within a single mode are first identified as dynamic segments. To avoid over-interpreting minor fluctuations, segments shorter than 15 residues are merged into their neighboring larger segments. The resulting merged units are referred to as dynamic elements (DEs).

In line with previous studies showing that the lowest-frequency modes dominate global protein motion [40, 41, 59], we restricted our DE analysis to the ten slowest nontrivial GNM modes. This cutoff ensures that DEs reflect functionally relevant, large-scale cooperative motions, while excluding high-frequency local fluctuations. The DE assignment procedure was applied individually to each protein structure in the dataset, ensuring consistent structural decomposition across all analyzed cases.

4.2.2.3. Overlap Score Calculation. To quantify the correspondence between adenine-binding themes and GNM derived dynamic elements (DEs), we implemented an overlap score metric. Unlike in prior studies that used mutual information-based co-occurrence statistics [59], the current analysis focuses on a single functional segment per structure (i.e., one adenine-binding theme), precluding such mutual metrics. Instead, we measured the extent of residue-level overlap between each dynamic element and the adenine-binding theme.

For each pairwise comparison between a dynamic element (from a specific GNM mode) and a theme, two directional overlap ratios were computed; Theme-to-DE overlap, the proportion of theme residues shared with the DE and DE-to-theme overlap, the proportion of DE residues shared with the theme. The final overlap score was calculated as the average of these two ratios.

The overlap score ranges from 0, indicating no shared residues, to 1, representing complete overlap. For instance, consider a dynamic element spanning residues 1–40 and a theme spanning residues 20–45, resulting in 21 overlapping residues (residues 20–40). The overlap relative to the dynamic element (Overlap_{DE}) is $21/40 = 0.525$, and the overlap relative to the theme (Overlap_{Theme}) is $21/26 = 0.808$. Taking the

average of these two values yields an overall overlap score of 0.666.

This metric was computed for each DE–theme pair across the dataset, enabling identification of theme–DE correspondences above a chosen threshold (e.g., >0.85) for further analysis

4.2.2.4. GNM-Based Transfer Entropy Analysis. To investigate directional information flow between residues and assess the allosteric communication capacity of adenine-binding themes, we applied a Transfer Entropy (TE)-based extension of the Gaussian Network Model (GNM), referred to as GNM-TE. This framework builds upon earlier developments in harmonic network models by quantifying causality between residue fluctuations using information-theoretic transfer entropy [55, 56]. Rather than simply identifying correlated motions, GNM-TE allows us to determine the extent to which fluctuations in one residue predict or influence fluctuations in another, thereby uncovering directional coupling between dynamically linked sites.

We followed the subset-based implementation described by Altintel et al. (2022), which restricts the TE computation to the lowest-frequency GNM modes, capturing collective motions associated with functional transitions. For each protein, a mode subset was selected based on prior DE-theme overlap analysis (modes with high correspondence to themes are selected). The net transfer entropy (netTE) was computed for each residue, reflecting the difference between information sent and received. We take the calculated the average netTE values for residues belonging to the themes and capture information flow through the theme. In addition, TEcol values – metric based on netTE and collectivity of the residues [56] – are calculated for each residue and global information source and sink residues that have broad influence across the structure [56–58], identified. Additional netTE flow analysis performed for the global source and sink residues that reside at the theme.

This analysis enables the identification of residues within or adjacent to conserved themes that participate in long-range communication, even in the absence of direct lig-

and contact. By comparing the dynamic communication patterns from theme residues and from global source/sink residues located within themes, we were able to identify both directly involved and allosterically connected sites potentially contributing to adenine recognition and regulation.

A flowchart of the approach used is provided in Figure 4.42. This integrated approach enables us to assess whether adenine-binding themes—identified purely by sequence conservation—also correspond to dynamic elements and serve as embedded hubs for long-range allosteric communication.

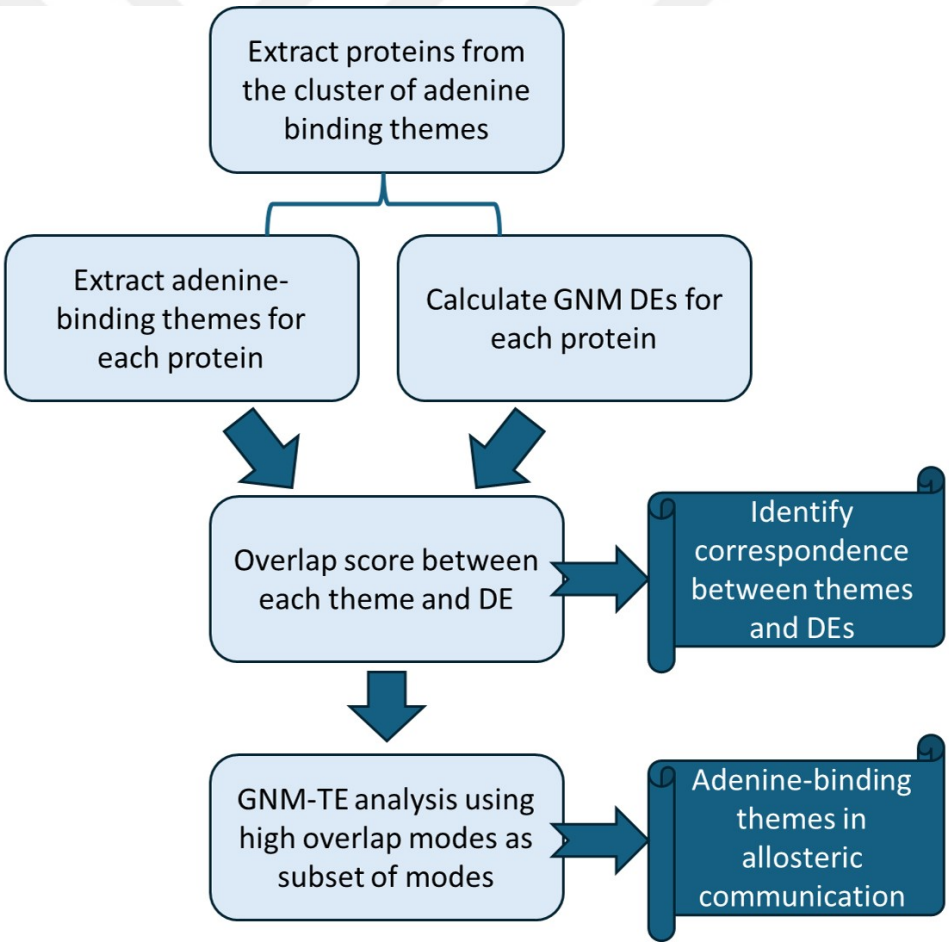


Figure 4.42. Schematic representation of the methodological workflow used to investigate the intrinsic dynamics and allosteric roles of conserved adenine-binding themes.

4.2.3. Localization of Adenine-Binding Residues at Dynamical Hinge Sites

To assess whether adenine-binding residues occupy structurally sensitive regions, we analyzed their localization with respect to GNM hinge sites. Previous studies have shown that functionally important residues, including those involved in catalysis or ligand binding, frequently reside at slow mode hinge points or at the interfaces of dynamic domains. Here, we also observed that the adenine binding residues often reside either at the hinge sites identified via GNM or domain interfaces. As a representative case, we examined the GNM slow modes of Archaeoglobus fulgidus Rio2 kinase (PDB ID: 1ZAR) (Figure 4.43).

In Figure 4.43, 3D structure is colored according to residue correlations in each mode as blue and red. Adenine binding residues are shown with magenta spheres and adenine containing molecule ADP is represented as green ribbon. Here, we observed that residues directly contacting adenine tend to cluster near hinge points of GNM slow modes. This suggests a potential coupling between adenine recognition and the global dynamic architecture of the protein.

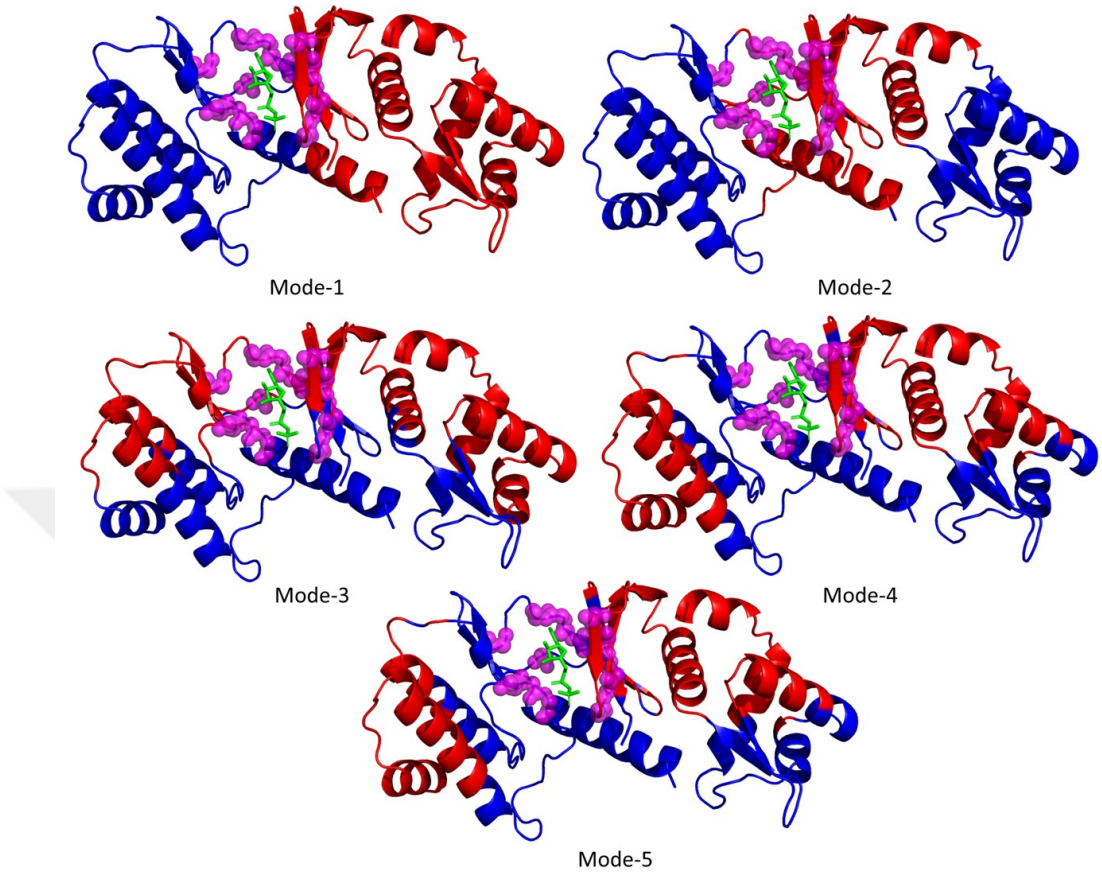


Figure 4.43. GNM analysis of Rio2 kinase (PDB ID: 1ZAR) highlights hinge localization of adenine-binding residues. The structure is colored by dynamic domains from the five slowest modes (red/blue). ADP is shown as a green ribbon; adenine-binding residues as magenta spheres.

4.2.4. Adenine-Binding Theme and DE Correspondence in Cluster 1

To investigate the relationship between evolutionarily conserved adenine-binding themes and intrinsic protein dynamics, we focused our analysis on Cluster 1 from the adenine-binding theme dataset. Proteins in this cluster primarily engage adenine via the direct motif, as previously described [36]. After filtering out structures with chain inconsistencies or other anomalies, we retained 22 high-quality protein structures for analysis for analysis (Table 4.18).

Table 4.18. Protein structures and homology group annotations in Cluster 1. List of PDB IDs, protein names, and their associated ECOD homology group (H-group) names for all Cluster 1 structures included in this study.

PDB ID	**Protein Name**	**H-Group Name**
1ZAR	Rio2 kinase	Protein kinase / SAICAR synthase / ATP-grasp
1ZTH	Rio1 serine protein kinase	Protein kinase / SAICAR synthase / ATP-grasp
2PML	Ser/Thr protein kinase	Protein kinase / SAICAR synthase / ATP-grasp
3A99	Proto-Oncogene Serine/Threonine-Protein Kinase Pim-1	Protein kinase/SAICAR synthase / ATP-grasp
3FEG	Choline/ethanolamine kinase	Protein kinase / SAICAR synthase / ATP-grasp
3GNI	Strad Alpha	Protein kinase / SAICAR synthase / ATP-grasp
3HKO	Calcium/calmodulin-dependent protein kinase with a kinase domain and 2 calmodulin-like EF hands	Protein kinase / SAICAR synthase / ATP-grasp
3PFQ	Protein kinase C beta type	Protein kinase / SAICAR synthase / ATP-grasp
3W0O	Hygromycin-B 4-O-kinase	Protein kinase / SAICAR synthase / ATP-grasp
4AZS	Methyltransferase WBDD	Protein kinase / SAICAR synthase / ATP-grasp
4GYI	Rio2 kinase	Protein kinase / SAICAR synthase / ATP-grasp

Table 4.18 Protein structures and homology groups in Cluster 1. (cont.)

PDB ID	**Protein Name**	**H-Group Name**
4OAV	Protein RNAse L	Protein kinase / SAICAR synthase / ATP-grasp
4OCV	N-acetylhexosamine 1-phosphate kinase	Protein kinase / SAICAR synthase / ATP-grasp
4R78	Choline kinase	Protein kinase / SAICAR synthase / ATP-grasp
4TND	G protein-coupled receptor kinase 5	Protein kinase / SAICAR synthase / ATP-grasp
4WW7	EKC/KEOPS complex subunit BUD32	Protein kinase / SAICAR synthase / ATP-grasp
4Y0X	Serine/threonine-protein kinase PknG	Protein kinase / SAICAR synthase / ATP-grasp
5CKW	LegK4	Protein kinase / SAICAR synthase / ATP-grasp
5DZC	cGMP-dependent protein kinase, putative	Protein kinase / SAICAR synthase / ATP-grasp
5GZA	Protein O-mannose kinase	Protein kinase / SAICAR synthase / ATP-grasp
5HNV	PpkA N terminal	Protein kinase / SAICAR synthase / ATP-grasp
5LVO	3-phosphoinositide-dependent protein kinase 1	Protein kinase / SAICAR synthase / ATP-grasp

While all 22 structures belong to the same homology group — Protein kinase / SAICAR synthase / ATP-grasp (according to ECOD classification [18]) — they display notable structural divergence. Pairwise RMSD comparisons reveal an average backbone RMSD of 4.96 Å, with values ranging from 1.77 Å to 6.86 Å, highlighting the conformational variability within this cluster.

For each structure, we computed the 10 slowest modes from the Gaussian Network Model (GNM) and extracted corresponding dynamic elements (DEs). We next quantified the degree of overlap between GNM-derived dynamic elements (DEs) and previously annotated adenine-binding themes in each structure. In all analyzed proteins, we identified multiple high-overlap theme–DE pairs, indicating a widespread correspondence between adenine-binding motifs and collective motions (Table 4.19).

Notably, each protein contains at least one theme with >85% positional overlap to dynamic elements (DEs) from more than one GNM mode. To assess whether this correspondence arises by chance, we performed a statistical analysis across Cluster 1. For each protein, we generated random segments that mimic the length distribution of its adenine-binding themes. The number of random segments matched the number of observed themes in that protein. We then calculated how many of these random segments exhibited high overlap (>85%) with DEs from more than one mode. This process was repeated 10,000 times to construct a null distribution, and a p-value was computed based on the empirical likelihood of observing the actual number of high-overlap themes under the null model. Full methodological details are provided in the Materials and Methods section. This cluster-wide analysis yielded a p-value of $5.25 \times 10^{-}7$, confirming that the observed theme–DE correspondences are statistically significant and not attributable to random overlap.

Table 4.19. Summary of theme–DE correspondence in Cluster 1 proteins. For each protein structure in Cluster 1, the table reports the number of adenine-binding themes present, the maximum and average of maximum overlap score between each theme and dynamic elements (DEs), and the number of themes with high correspondence (overlap >85%) to at least one or at least two GNM modes.

PDB ID	**# of Themes**	**Max Overlap (%)**	**Avg Overlap (%)**	**# Themes >85% in ≥1 Modes**	**# Themes >85% in ≥2 Modes**
1ZAR	9	97.66	90.4	8	4
1ZTH	8	96.5	90.16	6	6
2PML	8	93.64	86.13	5	3
3A99	11	94.24	87.74	7	4
3FEG	5	89.73	83.89	2	2
3GNI	10	85.71	82.24	2	1
3HKO	3	91.03	89.61	3	1
3PFQ	10	99.48	89.72	7	4
3W0O	5	98.42	91.91	4	3
4AZS	10	96.25	86.82	6	1
4GYI	7	99.27	92.31	6	3
4OAV	7	99.21	93.4	7	6
4OCV	4	96.08	87.68	2	2
4R78	5	98.39	86.16	2	2
4TND	9	97.96	90.86	8	3
4WW7	8	98.35	87.49	4	3
4Y0X	7	92.37	82.33	2	1
5CKW	9	97.37	89.4	7	3

Table 4.19 Summary of theme–DE correspondence in Cluster 1. (cont.)

PDB ID	**# of Themes**	**Max Overlap (%)**	**Avg Overlap (%)**	**# Themes >85% in ≥1 Modes**	**# Themes >85% in ≥2 Modes**
5DZC	11	96.85	87.05	6	1
5GZA	7	94.69	86.37	4	4
5HNV	11	95.55	88.83	7	5
5LVO	9	91.37	85.73	5	3

Given the statistical support for theme–DE correspondence, we examined the data more closely to identify broader trends. This analysis revealed several consistent patterns across the dataset. First, different adenine-binding themes within the same protein frequently correspond to DEs derived from different GNM modes, implying dynamic specialization of overlapping sequence regions. Additionally, the same adenine-binding theme, when found across different protein structures, may correspond to DEs from distinct modes, suggesting structural and dynamical adaptation despite sequence-level conservation. Finally, in some cases, a single theme aligns with DEs from multiple modes within the same protein, reflecting the multi-mode dynamical potential of a given sequence segment. Together, these observations underscore a deep interdependence between sequence conservation and dynamic specialization in adenine recognition. In the next section, we present representative case studies that illustrate each of these findings in detail.

4.2.4.1. Representative Cases of Theme–Dynamic Element Correspondence. The Archaeoglobus fulgidus Rio2 kinase (PDB ID: 1ZAR) was selected as an exemplary case, which contains multiple adenine-binding themes with distinct dynamic profiles. Adenine-binding themes observed in this structure are shown in Figure 4.44, along with the residues involved in adenine binding.

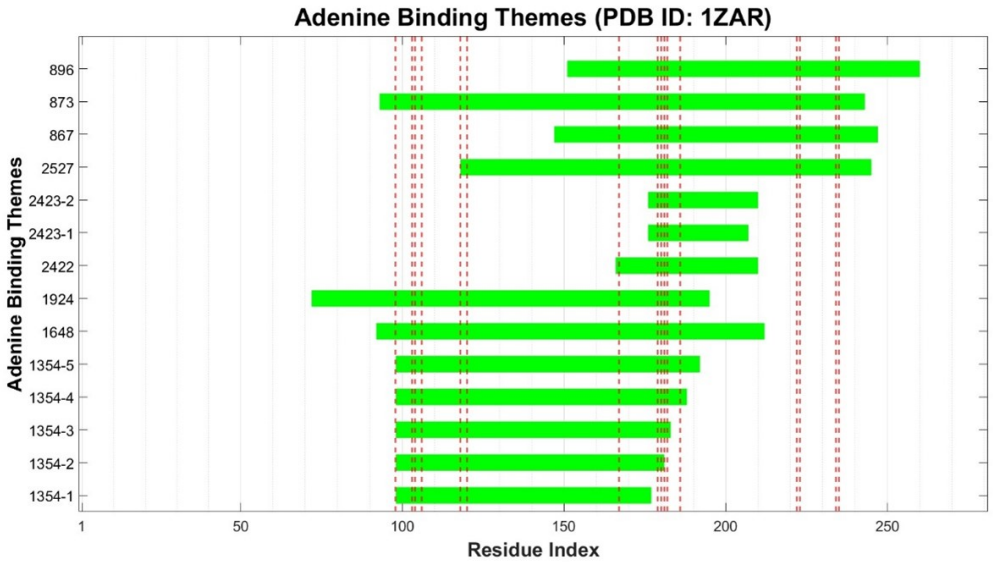


Figure 4.44. Adenine binding themes identified in Archaeoglobus fulgidus Rio2 kinase (PDB ID: 1ZAR). Themes shown as green stripes and adenine binding residues marked with red dashed lines.

Figure 4.45 illustrates the dynamic elements obtained from the slowest ten modes of 1ZAR alongside the mapped adenine-binding themes. In this analysis, the adenine-binding themes were found to correspond to dynamic elements across different slow modes. Since these varying themes reflect distinct layers of evolutionary conservation, they provide insights into the dynamic aspects of adenine-binding evolution.

Specifically, theme 2527 corresponds to DE-4 of mode 4, theme 896 maps to DE-3 of mode 5 and theme 1924 aligns with DE-3 of mode 6 (Figure 4.46). These three themes share partial sequence overlap, yet each associate with a different collective motion. The bar graphs and structural visualizations illustrate this divergence, highlighting how different subsets of the structure contribute to adenine-binding through distinct dynamic elements. This example underscores a key finding of the study: a single protein may employ multiple adenine-binding themes, each functionally coupled to a different GNM-derived dynamic domain, suggesting that the protein's ability to bind adenine is dynamically modular and distributed.

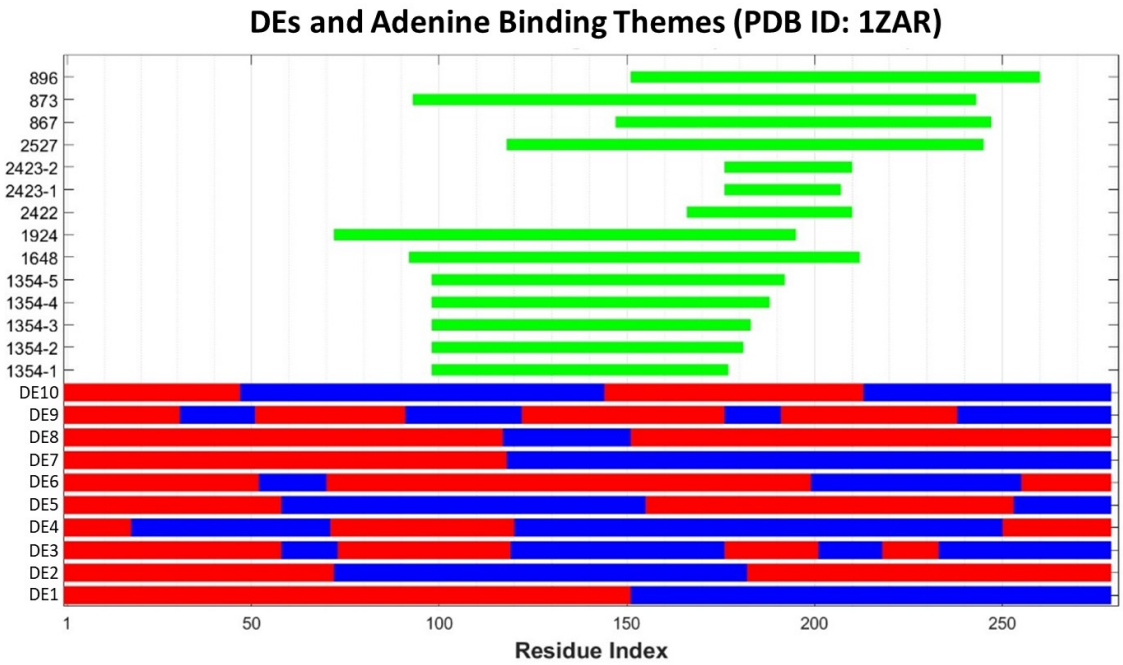


Figure 4.45. Dynamic domains obtained from the slowest ten modes of 1ZAR together with the adenine-binding themes.

These three themes share partial sequence overlap, yet each associate with a different collective motion. The bar graphs and structural visualizations illustrate this divergence, highlighting how different subsets of the structure contribute to adenine-binding through distinct dynamic elements. This example underscores a key finding of the study: a single protein may employ multiple adenine-binding themes, each functionally coupled to a different GNM-derived dynamic domain, suggesting that the protein's ability to bind adenine is dynamically modular and distributed.

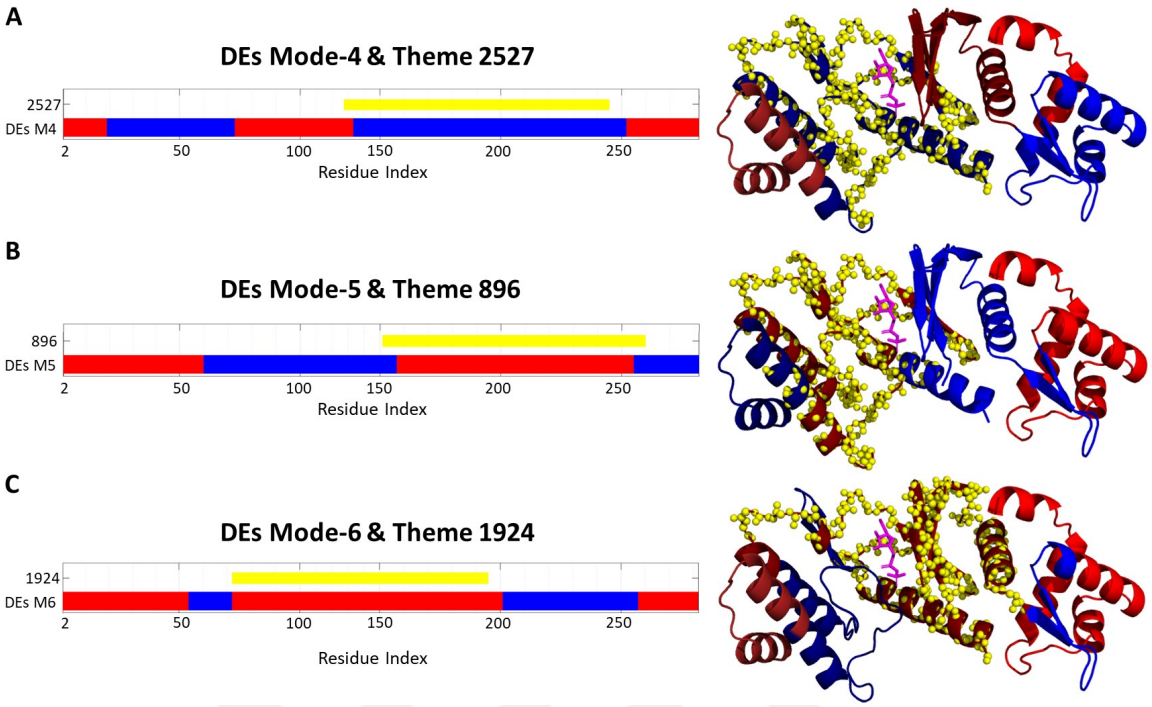


Figure 4.46. Multiple adenine-binding themes in Rio2 kinase (PDB ID: 1ZAR) correspond to distinct DEs from different GNM modes. (Left) Bar plots show DEs from modes 4–6 (red/blue); themes 2527, 896, and 1924 (yellow) each align with a separate DE. (Right) 3D view shows theme residues as yellow spheres.

Figure 4.47, Figure 4.48 and Figure 4.49 expand on this observation by illustrating how the same adenine-binding theme can correspond to different DEs in different protein structures. Figure 4.47 focuses on theme 2527, which maps to distinct DEs in 1ZAR, RNase L (PDB ID: 4OAV) and O-mannose kinase (PDB ID: 5GZA), each arising from different modes. Similarly, Figure 4.48 shows theme 896 aligning with DEs in 1ZAR, Rio kinase (PDB ID: 1ZTH) and Protein kinase C beta type (PDB ID: 3PFQ), all from mode 5 but occupying different regions of the structure. Figure 4.49 highlights theme 1924, which maps to mode 6 or mode 1 DEs across 1ZAR, EKC/KEOPS complex subunit BUD32 (PDB ID: 4WW7) and PpkA N terminal (PDB ID: 5HNV).

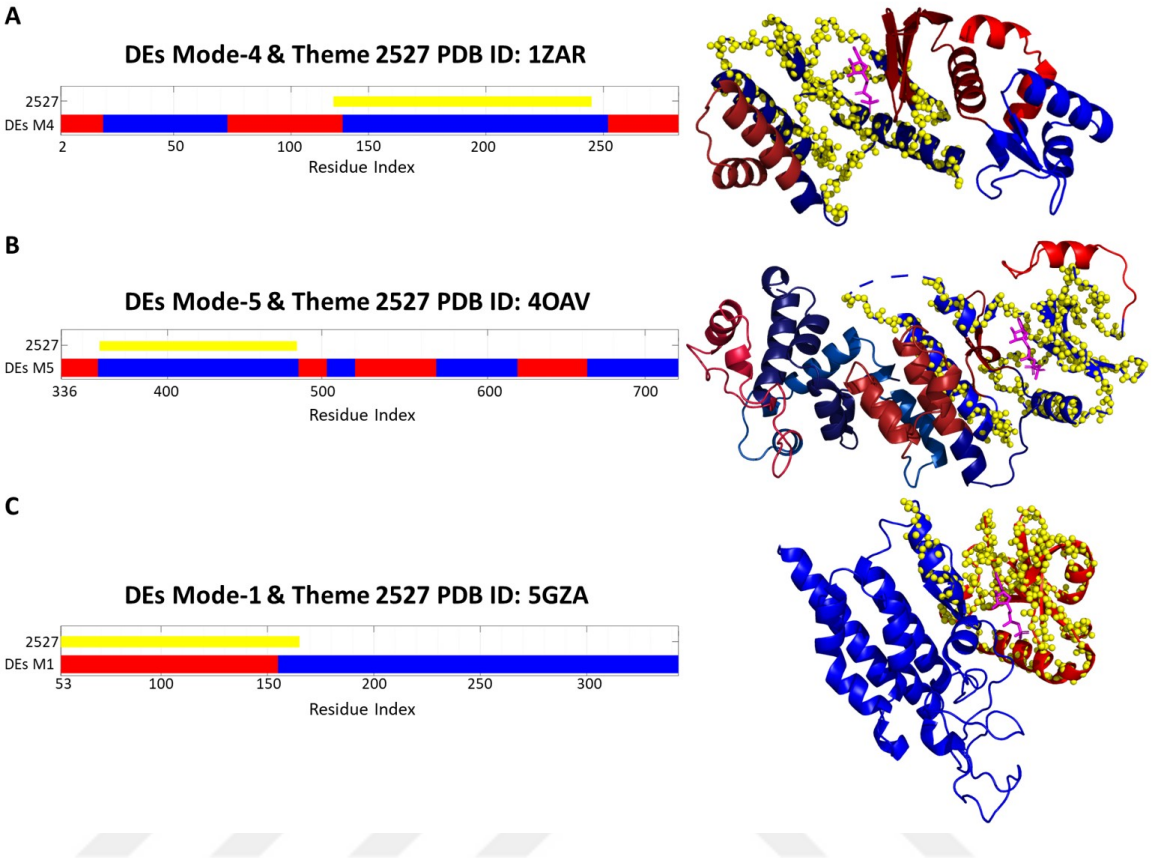


Figure 4.47. The same adenine-binding theme (2527) maps to distinct DEs across kinase-like proteins. (a) 1ZAR: Mode 4–DE4; (b) 4OAV: Mode 5–DE2; (c) 5GZA: Mode 1–DE1. DEs (red/blue) and theme (yellow) are shown in bar plots and 3D views. Structures are superimposed on 1ZAR for comparison.

These results emphasize that adenine-binding themes may be dynamically reused across proteins, adapting to local fold and motion constraints while maintaining their ligand-binding role. This flexibility further supports the notion that themes function as modular dynamic elements, whose evolutionary utility lies in their ability to be repurposed across both sequence and dynamic contexts.

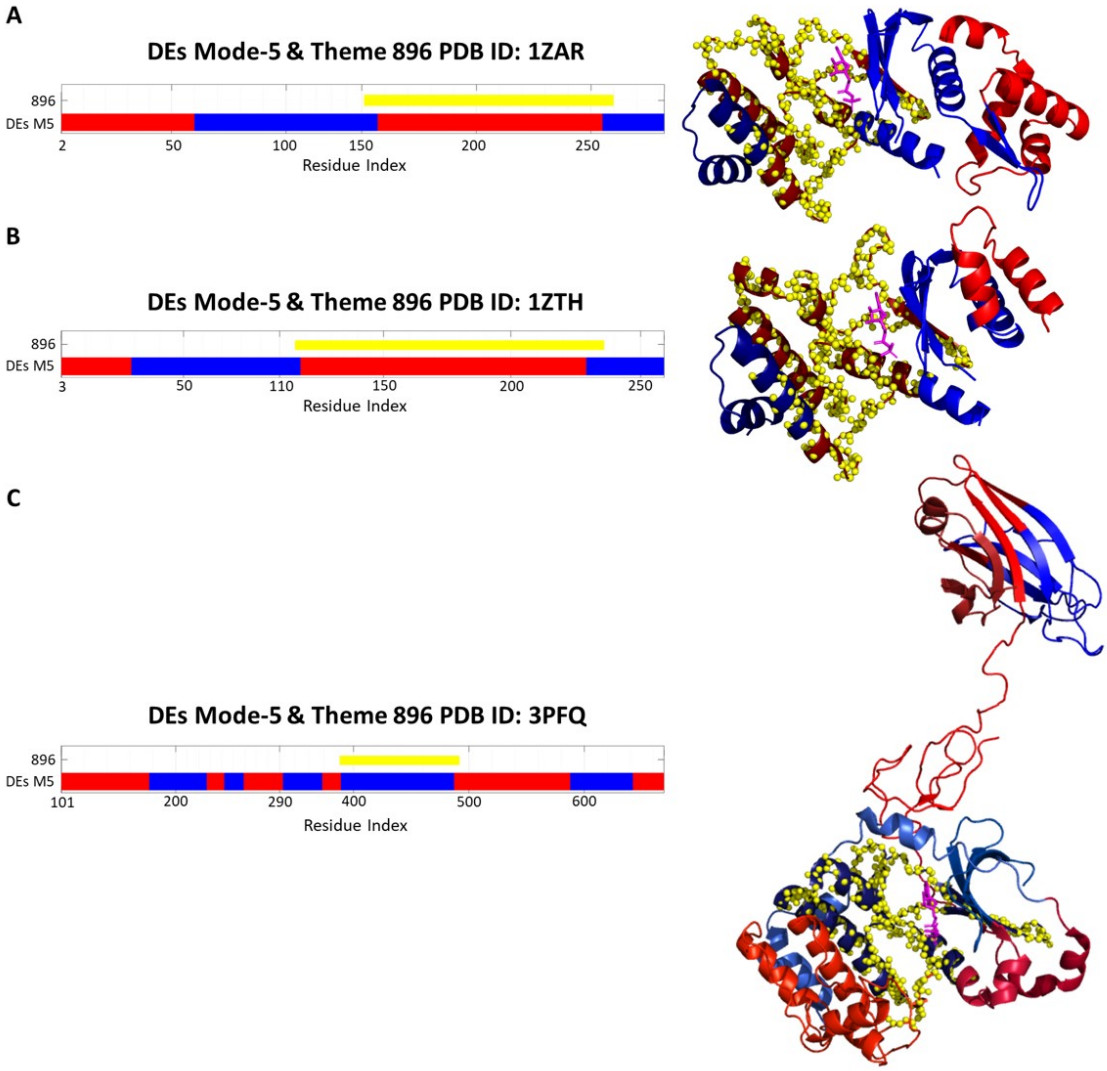


Figure 4.48. Cross-structure mapping of theme 896 to different DEs. (a) 1ZAR: Mode 5–DE3; (b) 1ZTH: Mode 5–DE3; (c) 3PFQ: Mode 5–DE2. Bar plots show DEs in red/blue and theme 896 in yellow. 3D structures are aligned to 1ZAR for comparability.

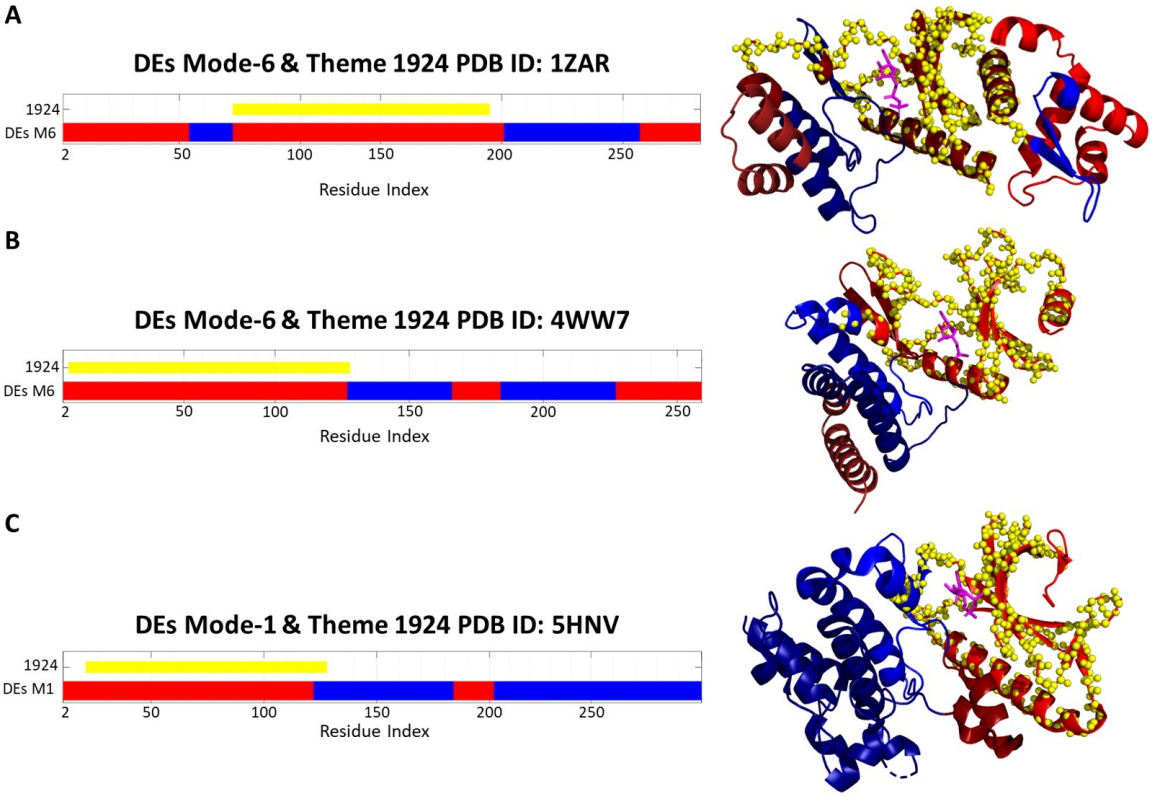


Figure 4.49. Dynamic mapping of theme 1924 across distinct proteins. (a) 1ZAR: Mode 6–DE3; (b) 4WW7: Mode 6–DE1; (c) 5HNV: Mode 1–DE1. DEs are shown in red/blue, theme 1924 in yellow. Structures are superimposed on 1ZAR in 3D views.

4.2.5. Allosteric Information Flow from Adenine-Binding Themes 896 and 1924 in Rio2 kinase

To investigate whether dynamic regions overlapping with adenine-binding themes participate in allosteric communication, we applied the Gaussian Network Model Transfer Entropy (GNM-TE) approach to the 1ZAR structure. First, we selected a subset of GNM modes (Modes 1, 4, 5 and 8) which displayed high-overlap theme-DE pairs with theme 896. Residue by residue 2D TE map of the structure, together with the position of theme 896 are given in Figure 4.50.

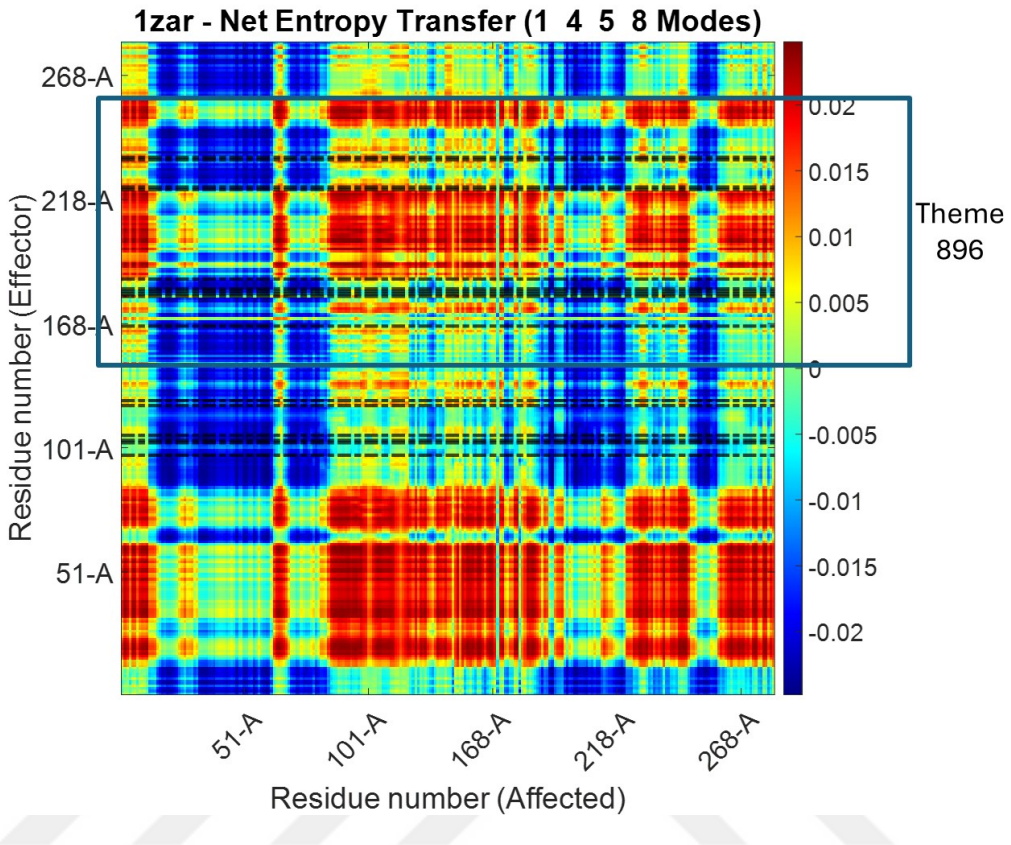
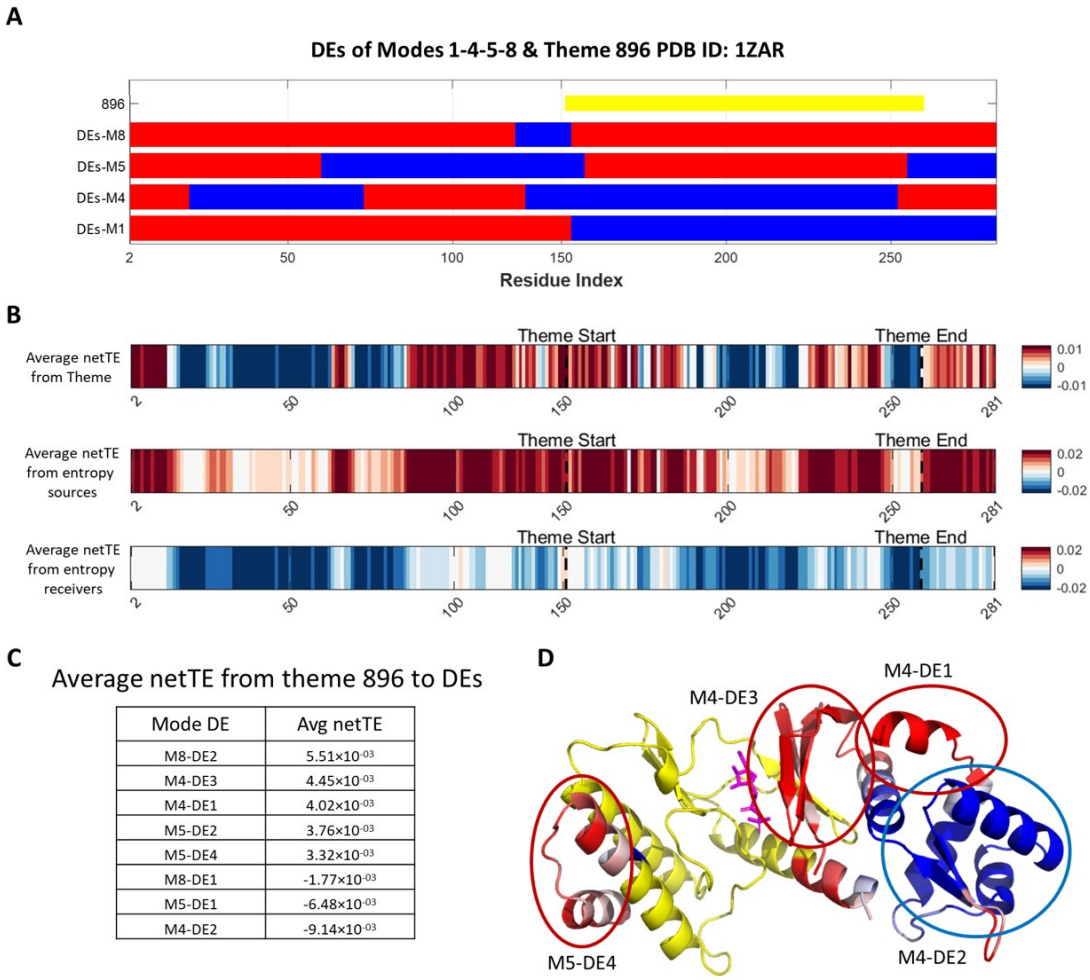


Figure 4.50. Residue-by-residue 2D TE map of the 1ZAR for mode subset 1-4-5-8, with the theme 896 outlined and labeled with rectangular box.

Figure 4.51a displays the DEs of these modes alongside the location of theme 896, confirming its primary overlap with DEs from these modes. Using this subset, we calculated the net transfer entropy (netTE) from the residues of theme 896 and assessed the directionality of information flow between the theme and other dynamic segments (Figure 4.51b). Additionally, based on TECol score we identified seven global information source residues: V18, D60, D72, V192, M201, S220 and D254; notably, four of these (V192, M201, S220, D254) reside within theme 896. We also detected four global receiver residues (D66, E117, A171 and E244), with two (A171 and E244) located within theme 896. The information flow profiles of these global sources and sinks within the theme were consistent with the netTE distribution across theme 896, supporting their roles in allosteric signaling. (Figure 4.51b).



Mode DE	Avg netTE
M8-DE2	5.51×10^{-03}
M4-DE3	4.45×10^{-03}
M4-DE1	4.02×10^{-03}
M5-DE2	3.76×10^{-03}
M5-DE4	3.32×10^{-03}
M8-DE1	-1.77×10^{-03}
M5-DE1	-6.48×10^{-03}
M4-DE2	-9.14×10^{-03}

Figure 4.51. GNM-TE analysis of theme 896 in Rio2 kinase (PDB ID: 1ZAR). (a) Theme overlaps with DEs from modes 1, 4, 5, and 8 (yellow). (b) netTE from theme residues, global sources, and sinks shown in bar plots and red–white–blue scale. (c) Top receiver/sender DEs listed by mode and average netTE. (d) 3D view with theme (yellow); top DEs circled in red (receivers) and blue (senders).

Several DEs received substantial information from the theme, including M8-DE2, M4-DE3, M4-DE1 and M5-DE4 (Figure 4.51c). Among these, M8-DE2 and M4-DE3 are spatially adjacent to the theme and lie near the ADP binding site, while M4-DE1 and M5-DE4 are located more distally, suggesting long-range information flow that may support allosteric regulation. Conversely, the segment that sends the most information to the theme was identified as M4-DE2, which is also distal to the binding site, providing

further evidence of dynamic crosstalk between distant regions. Figure 4.51d visualizes these results on the structure, with the theme in yellow and top sender/receiver DEs highlighted with red and blue circles.

Strikingly, the global information source residues within theme 896 are not in direct contact with ADP and are located away from the binding pocket. Despite being spatially distal from the binding pocket, these residues exhibited substantial information transfer toward the ADP-binding region. As expected, global information receiver residues (D66, E117, A171 and E244) are among the most significant recipients of information from these sources. Additionally, residues A105, K120, E154, L158 and A184, located at the ADP binding groove [140, 141], were identified as major information recipients. These observations suggest that residues V192, M201, S220 and D254 may act as allosterically active residues within the adenine-binding theme, functionally contributing to ligand binding from a distance.

Notably, residue S220 is situated within the catalytic loop of Rio2 kinase, a region critical for enzymatic activity [141, 142]. This positioning underscores its potential role in allosteric regulation. Like the global source residues, global receiver residues are also located away from the ADP site but show significant entropy reception from distant structural regions and global source residues, implying they may integrate information as part of an allosteric response. The co-localization of multiple global sources and sinks within theme 896 supports its dual role as both a ligand-binding motif and a communication hub in the protein's dynamic network.

In addition to analysis on theme 896, we also conducted a similar analysis on the theme 1924 for 1ZAR with the subset of modes 2 and 6. Residue by residue 2D TE map of the structure, together with the position of 1924 are given in Figure 4.52.

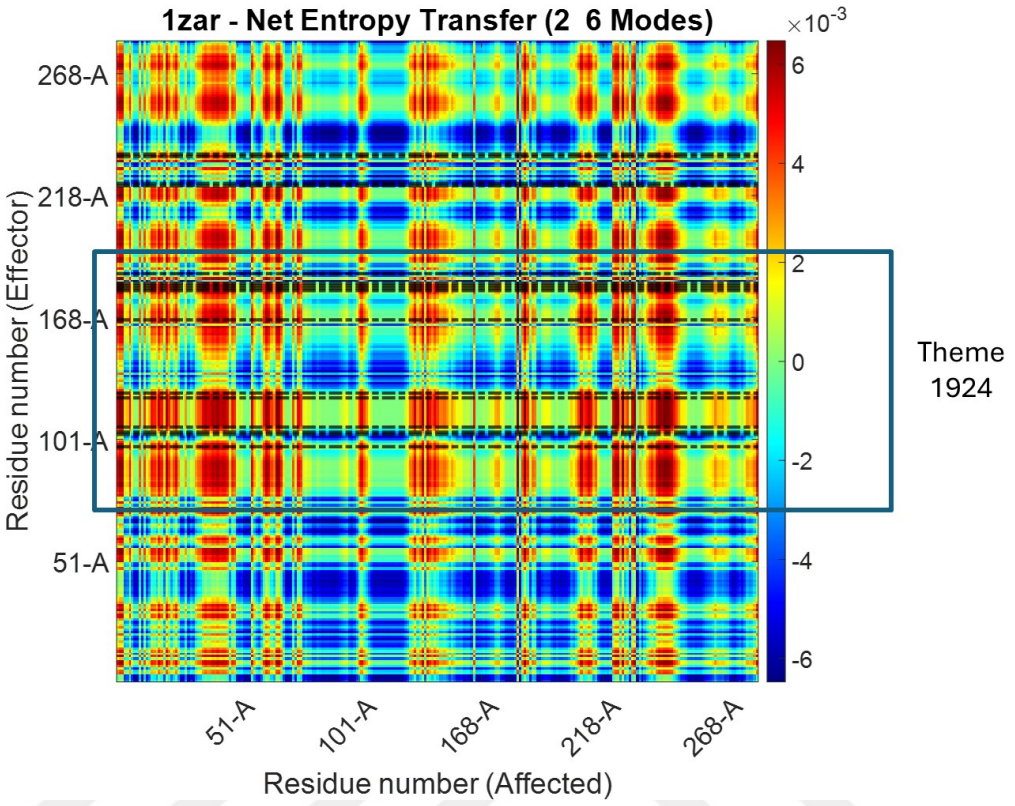


Figure 4.52. Residue-by-residue 2D TE map of the 1ZAR for mode subset 2 and 6, with the theme 1924 outlined and labeled with rectangular box.

Figure 4.53a displays the DEs of these modes alongside the location of theme 1924, confirming its primary overlap with DEs from these modes. Using this subset, we calculated the net transfer entropy (netTE) from the residues of theme 1924 and assessed the directionality of information flow between the theme and other dynamic segments (Figure 4.53b). Additionally, based on TECol score we observed that seven out of seventeen global information source residues identified for this subset of modes reside within the theme. We also detected six out of fifteen global information sink residues located within the theme 1924. The information flow profiles of these global sources and sinks within the theme were consistent with the netTE distribution across 1924, supporting their roles in allosteric signaling. (Figure 4.53b). Additionally, these residues are located at or near the adenine binding sites.

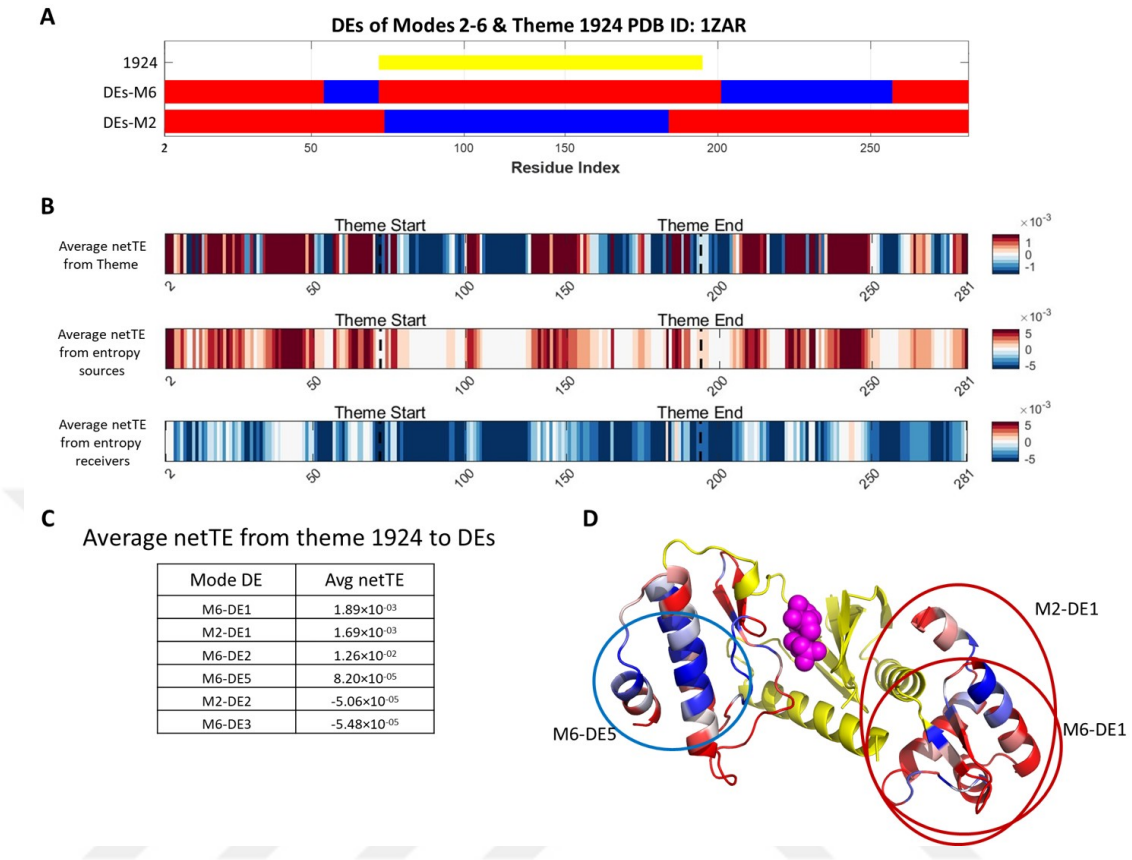


Figure 4.53. GNM-TE analysis of theme 1924 in Rio2 kinase (PDB ID: 1ZAR). (a) Theme aligns with DEs from modes 2 and 6 (yellow in bar plots). (b) netTE from theme residues, global sources, and sinks shown in bar plots and red–white–blue gradient. (c) Top sender/receiver DEs listed by mode and average netTE. (d) 3D view with theme (yellow); key DEs circled in red (receivers) and blue (senders).

In contrast to theme 896, theme 1924 displays a distinct information transfer pattern when analyzed using a different mode subset. While the overall information flow toward the C-terminal region remains relatively stable, notable shifts are observed in the N-terminal region and within theme 1924 itself. Specifically, residues in the N-terminal region, which acted as strong information sources in the theme 896 case, now emerge as information receivers. For example, in the analysis of theme 896, M5-DE1—a dynamic element located at the N-terminus—exhibits the second most negative netTE value, indicating that it serves as a major information source. However, in the analysis

of theme 1924, a highly similar region (M6-DE1) displays the highest positive netTE value, suggesting a reversal in directionality, now acting as a dominant information sink (Figure 4.53c-d).

A similar inversion is observed at the C-terminal region. In the case of theme 896, M5-DE4, located distally from the theme, displays a strongly positive netTE, highlighting it as a downstream information receiver. Conversely, in the theme 1924 analysis, the corresponding region (M6-DE5) shows negative netTE, indicating that it instead acts as an information sender toward the theme (Figure 4.53d). These opposing information flow patterns suggest that different themes within the same protein can engage in distinct allosteric communication routes, depending on their mode-specific dynamic context. Such findings highlight the flexibility and specificity of theme-mediated signaling, reinforcing the idea that conserved sequence motifs can be dynamically repurposed based on their structural and energetic environment.

4.2.6. Allosteric Information Flow from Theme 1924 in BUD32

To further investigate the relationship between adenine-binding themes and long-range dynamic communication, we conducted a GNM-TE analysis for theme 1924 in the 4WW7 structure. Modes 1, 6 and 8 were selected for the analysis, as theme 1924 showed high overlap with M1-DE1, M6-DE1 and M8-DE2 (Figure 4.54a). Although M8-DE2 provides the highest overlap in terms of coverage, we observed that theme 1924 better aligns with three distinct sub-regions in mode 8 — specifically, M8-DE1, M8-DE2 and M8-DE3 — rather than a single dominant DE.

Outside the theme, M6-DE3 and M6-DE4 were identified as information recipients from the theme, despite not being located within the AMP-binding groove (Figure 4.54d). In contrast, M6-DE5 displayed strong source-like behavior, with netTE patterns directed toward global information receivers within the theme (Figure 6C). Notably, M6-DE5 is located distally from the binding site, supporting the possibility of long-range dynamic input toward the theme region (Figure 4.54d).

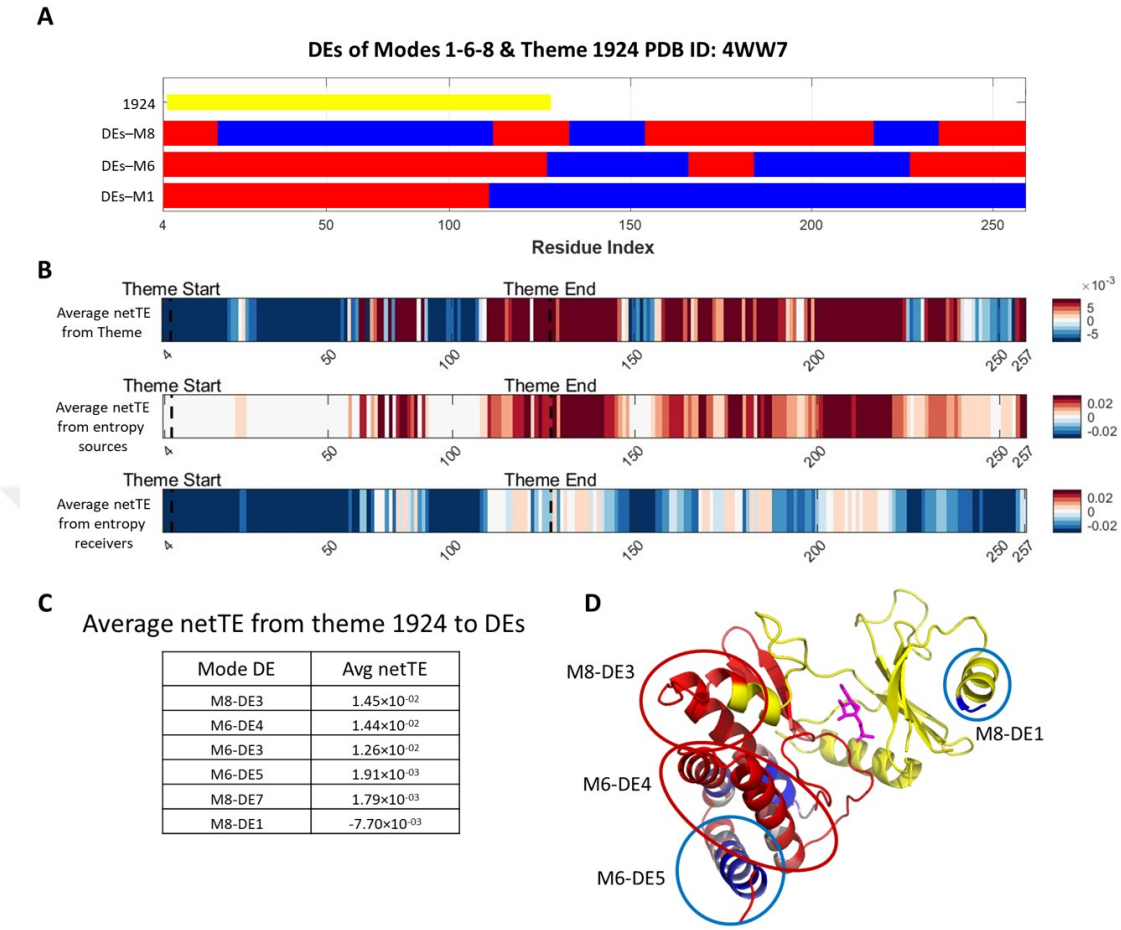


Figure 4.54. GNM-TE analysis of theme 1924 in BUD32 (PDB ID: 4WW7). (a) Theme overlaps with DEs from modes 1, 6, and 8 (yellow in bar plots). (b) netTE from theme residues, sources, and sinks shown in bar plots and red–white–blue scale. (c) Top receiver/sender DEs listed by mode and average netTE. (d) 3D view with theme (yellow); major receivers (red) and senders (blue) circled.

We also identified five global information source residues across the selected mode subset: S77, D98, G150, E226 and V250. Among these, S77 and D98 reside within theme 1924, although none of the source residues directly contact AMP. Global information receiver residues include N75, A81, L83, C90, N121, M128, Y134, L169, P178, D182, S187 and K216 — with six of these (N75 to M128) also belonging to theme 1924. Notably, D182 is a known AMP-binding and catalytic residue [143], supporting the functional significance of the entropy transfer directed toward the active site. In

addition, K216 participates in sulphate coordination, while L83 and C90 are located at the protein–protein interaction interface [144]. Beyond theme 1924, several strong information source residues — K52, E107 and L109 — were identified; although not part of the theme, these residues transmit significant information to the theme region and are known to contribute to AMP coordination and catalytic function [144]. This pattern suggests that, similar to the case observed in 1ZAR, theme 1924 includes residues involved in both receiving allosteric input and propagating information toward distant functional elements.

Together, these results highlight the dual role of theme 1924 as both a recipient and conduit of dynamic information. While not all components of the theme are directly involved in AMP binding, their integration into allosteric communication pathways further supports the notion that evolutionary themes function as dynamically embedded regulatory modules.

4.2.7. Expansion of Overlap Analysis Across Additional Adenine-Binding Theme Clusters

To assess whether the theme–DE correspondence observed in Cluster 1 is conserved across other adenine-binding motifs, we extended our analysis to three additional clusters defined by Narunsky et al [36]. Cluster 2, like Cluster 1, features a direct binding motif but includes an additional interaction between adenine's N7 and a conserved lysine residue. Cluster 4 represents a reverse binding motif with a distinctive N6–backbone interaction. Cluster 6B includes NAD-binding proteins and is characterized by the "Asp" motif, accompanied by a unique interaction between adenine's N3 and the protein. These clusters introduce variation in motif architecture, ligand class and residue-level interactions. In each case, we observed multiple high-overlap theme–DE pairs, including themes that mapped to more than one dynamic mode, similar to patterns observed in Cluster 1 (Table 4.20, Table 4.21 and Table 4.22).

Table 4.20. Summary of theme–DE correspondence in Cluster 2 proteins. For each protein structure in Cluster 2, the table reports the number of adenine-binding themes present, the maximum and average of maximum overlap score between each theme and dynamic elements (DEs), and the number of themes with high correspondence (overlap >85%) to at least one or at least two GNM modes.

PDB ID	**# Themes**	**Max Overlap (%)**	**Avg Overlap (%)**	**# Themes >85% in ≥1 Modes**	**# Themes >85% in ≥2 Modes**
1GSA	13	95.93	89.13	12	5
1M0W	3	95.93	95.19	3	2
1Z2N	10	95.30	91.44	9	2
2IO8	8	99.43	92.30	7	4
2P0A	12	97.24	88.32	8	4
2R85	12	95.58	87.51	8	4
3T7A	6	94.03	87.45	4	2
3VPB	12	99.40	87.68	6	4
3WNZ	6	95.37	86.27	3	2
4C5C	14	98.01	89.19	12	3
4FFL	7	99.38	91.75	5	4
4I4T	2	94.87	93.97	2	0
4IWX	13	98.95	91.48	13	6
4MV4	12	92.02	85.46	6	2
4YLR	4	92.85	86.41	3	0

Table 4.21. Summary of theme–DE correspondence in Cluster 4 proteins. For each protein structure in Cluster 4, the table reports the number of adenine-binding themes present, the maximum and average of maximum overlap score between each theme and dynamic elements (DEs), and the number of themes with high correspondence (overlap >85%) to at least one or at least two GNM modes.

PDB ID	**# Themes**	**Max Overlap (%)**	**Avg Overlap (%)**	**# Themes >85% in ≥1 Modes**	**# Themes >85% in ≥2 Modes**
2UV4	6	100.00	96.69	6	6
2V8Q	14	97.00	93.50	14	11
2YZQ	11	99.12	87.07	5	5
3DDJ	14	99.11	87.47	9	6
3KH5	15	98.36	92.45	14	10
3LRF	9	98.77	89.46	5	5
4HG0	9	96.95	93.03	9	5

Table 4.22. Summary of theme–DE correspondence in Cluster 6B proteins. For each protein structure in Cluster 6B, the table reports the number of adenine-binding themes present, the maximum and average of maximum overlap score between each theme and dynamic elements (DEs), and the number of themes with high correspondence (overlap >85%) to at least one or at least two GNM modes.

PDB ID	**# Themes**	**Max Overlap (%)**	**Avg Overlap (%)**	**# Themes >85% in ≥1 Modes**	**# Themes >85% in ≥2 Modes**
1E6U	5	92.97	87.45	3	1
1EQ2	5	90.90	80.35	1	1

Table 4.22 Summary of theme–DE correspondence in Cluster 6B.(cont.)

PDB ID	**# of Themes**	**Max Overlap (%)**	**Avg Overlap (%)**	**# Themes >85% in ≥1 Modes**	**# Themes >85% in ≥2 Modes**
1HDO	4	90.06	83.88	2	0
1HYH	3	85.50	78.99	1	0
1I24	6	92.65	85.75	3	0
1OAA	3	99.62	93.41	3	3
1R6D	6	92.86	83.10	3	2
1SBY	5	98.32	84.55	1	1
1Y1P	5	82.88	76.97	0	0
1Z45	5	93.22	86.60	2	0
1ZK4	4	94.68	84.07	2	2
2B69	6	93.48	87.03	3	1
2BD0	4	97.14	85.93	2	2
2C5A	5	98.79	85.96	2	0
2O7S	3	88.99	78.36	1	0
2WM3	3	85.15	78.87	1	0
2YUT	5	97.81	80.41	1	1
3L77	3	95.05	82.50	1	1
3PVZ	5	92.76	80.39	1	1
4GVX	3	91.67	87.40	2	1
4ID9	5	94.12	89.90	5	0
4WUV	2	93.94	90.56	2	0
4X79	2	82.14	80.04	0	0

Table 4.22 Summary of theme–DE correspondence in Cluster 6B.(cont.)

PDB ID	**# of Themes**	**Max Overlap (%)**	**Avg Overlap (%)**	**# Themes >85% in ≥1 Modes**	**# Themes >85% in ≥2 Modes**
4ZJU	4	84.39	82.01	0	0
5BT9	2	92.86	92.86	2	1
5D2E	4	99.15	92.12	3	1
5IN4	6	85.17	78.72	1	0
5IZ4	3	85.57	83.73	1	0
5L4L	2	86.03	81.28	1	0
5MSO	6	90.54	84.61	3	1
5U4Q	5	92.89	80.49	2	1
5UZX	3	94.80	89.71	2	0

These findings support the broader generalizability of the dynamic integration of adenine-binding themes across structurally and functionally diverse protein families. Additionally, these findings collectively demonstrate the interplay between evolutionary conservation, structural dynamics and functional communication — offering a modular and transferable model for analyzing allosteric behavior in diverse protein systems.

4.2.8. Conclusion

In this section, we investigated the dynamic and allosteric properties of evolutionarily conserved adenine-binding themes using GNM and GNM-TE analysis. Our focus was on a structurally diverse set of 22 kinase-like proteins, where we demonstrated that adenine-binding themes frequently overlap with dynamic elements (DEs) extracted from low-frequency GNM modes. These overlaps were not incidental: themes often corresponded to multiple DEs across orthogonal modes or mapped to different

DEs in structurally distinct proteins. This indicates that themes, although conserved at the sequence level, are dynamically repurposed to participate in context-specific collective motions.

Representative case studies—such as themes 896 and 1924 in Rio2 kinase and BUD32—revealed that residues within themes frequently act as sources or sinks of information in allosteric signaling pathways. Notably, many residues dynamically coupled to the adenine-binding region were not directly involved in ligand interaction, suggesting the presence of latent, yet potentially regulatory, allosteric elements embedded within conserved sequence motifs.

To assess the broader relevance of these findings, we extended our analysis to three additional clusters (Clusters 2, 4, and 6B) defined by distinct motif architectures and ligand types. Across all clusters, we observed consistent patterns of high overlap between themes and DEs, as well as examples of themes participating in multivalent dynamic coupling. A statistical analysis comparing observed theme–DE overlaps to random expectation confirmed that the correspondence is highly significant, supporting the non-random nature of this relationship.

Taken together, our results highlight the dual role of adenine-binding themes as both structural motifs for ligand recognition and dynamic modules for long-range signaling. These insights may inform protein design strategies by emphasizing the need to consider both direct binding residues and distal dynamic couplings. Conversely, the inclusion of such motifs in designed proteins could offer a platform for experimentally validating their functional and allosteric roles. More broadly, this framework provides a generalizable approach for linking evolutionary conservation to protein dynamics and may be extended to other ligand classes and motif types to explore allosteric regulation at the subdomain level.

4.3. Systematic Identification and Dynamic Classification of Binding Themes

In the previous section, the analysis of adenine-binding themes demonstrated the challenge of performing a comprehensive information flow analysis due to incomplete sequence coverage by themes. Many protein domains exhibited only partial theme representation, limiting our ability to characterize allosteric communication across the entire structure.

To address this limitation, we turned to the 150-domain dataset analyzed in Section 4.1 [59], where themes fully span the sequences of each domain. This dataset provides a robust framework for systematically investigating information transfer between themes using Gaussian Network Model-based Transfer Entropy (GNM-TE) calculations. In particular, we focused on identifying and characterizing binding themes — those overlapping with biologically relevant ligand-binding regions — and evaluating their communication properties across the dataset. Figure 4.55 provides a schematic overview of the workflow presented in this section.

4.3.1. Identification of Binding Themes and Theme Pairs

To systematically investigate theme-based binding events, we surveyed the 150-domain dataset used in Part 1 [59]. In order to ensure biological relevance, ligand-binding sites were identified using the BioLiP2 database [145], which provides curated biologically relevant ligand–protein interactions along with annotated binding site residues.

A theme was classified as a binding theme if it overlapped with at least half of the ligand-binding site residues. In cases where the binding site residues were distributed across distant regions of the sequence — a situation common for non-continuous binding sites — we additionally identified binding theme pairs. A binding theme pair was defined as two non-overlapping themes, each covering at least a quarter of the binding

site residues, whose combined coverage provided a more comprehensive representation of the binding site than either theme individually.

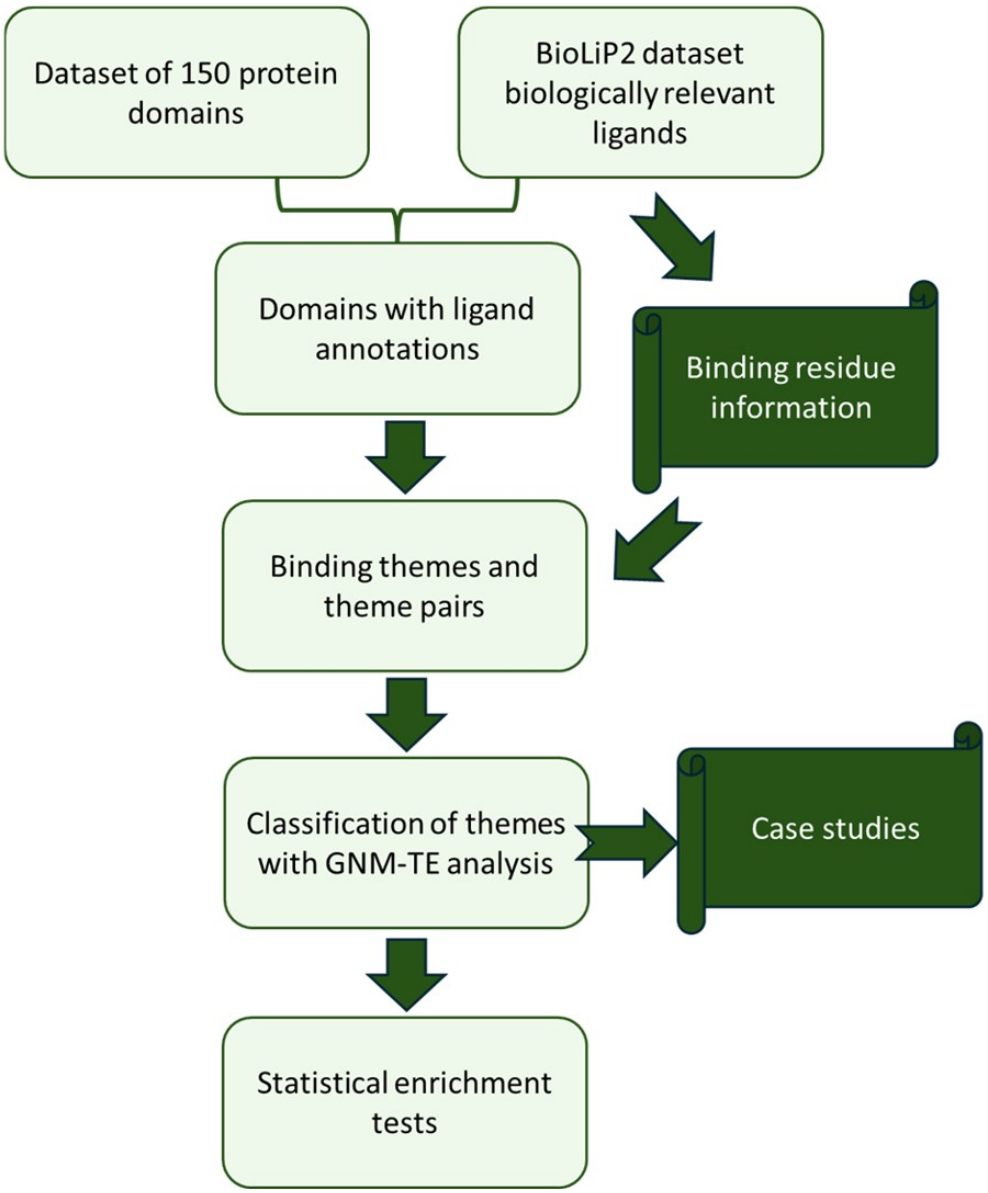


Figure 4.55. Schematic representation of the methodological workflow used for the systematic identification and dynamic classification of binding themes based on GNM-TE communication profiles and enrichment analysis

In order to avoid redundancy, themes participating in binding theme pairs were not subsequently considered as single binding themes, even if their individual overlap exceeded the single-theme threshold. This procedure prioritized theme pairs over individual themes when both options were available, ensuring a more complete and biologically meaningful representation of distributed binding regions.

Applying these criteria, we identified that 64 out of 150 protein domains contained at least one binding theme. In total, these domains include 1121 themes, of which 490 were involved in binding—either individually or as part of a binding theme pair. Across these domains, we identified 1435 binding themes or theme pairs (Table A1). Detailed results for each protein domain, including theme–binding site overlaps and ligand information, are provided as Excel files in the repository [146].

As an example, for the protein domain e2xyiA1, the BioLiP2 database provides annotated binding residues corresponding to a peptide ligand (histone 4). The binding residues listed are 27, 28, 31, 35, 362, 365, 366, 370 and 373.

First, we gathered the list of binding residues and obtained the information about themes identified in domain e2xyiA1. For each theme, we calculated the number and percentage of binding residues it covered (Table 4.23).

Table 4.23. Overlap of identified themes with peptide-binding residues in domain e2xyiA1.

Theme ID	**Start Residue**	**End Residue**	**Binding Residues**	**Overlap Percentage**
14940-1	14	53	4	44.44
14945	16	120	4	44.44
14940-2	21	55	4	44.44

Table 4.23. Overlap of themes and peptide-binding residues in domain e2xyiA1. (cont.)

Theme ID	**Start Residue**	**End Residue**	**Binding Residues**	**Overlap Percentage**
14940-3	21	59	4	44.44
14952	47	120	0	0.00
14946-1	55	120	0	0.00
14946-2	58	91	0	0.00
14942-1	126	162	0	0.00
14942-2	126	170	0	0.00
14953	126	181	0	0.00
14944	126	213	0	0.00
14812	176	213	0	0.00
14954	176	220	0	0.00
14813-1	176	231	0	0.00
14813-2	176	240	0	0.00
14937	227	265	0	0.00
14938	227	287	0	0.00
14939	227	309	0	0.00
15218	230	273	0	0.00
15220	230	297	0	0.00
15221	230	330	0	0.00
14949	243	297	0	0.00
14815-1	244	309	0	0.00
14816	271	309	0	0.00

Table 4.23. Overlap of themes and peptide-binding residues in domain e2xyiA1. (cont.)

Theme ID	**Start Residue**	**End Residue**	**Binding Residues**	**Overlap Percentage**
15219	281	315	0	0.00
14955	283	341	0	0.00
14815-2	289	353	0	0.00
14815-3	299	353	0	0.00
14861-1	314	353	0	0.00
14950	314	408	5	55.56
14956	315	381	5	55.56
15222	336	407	5	55.56
14815-4	343	406	5	55.56
14951	356	415	5	55.56
14861-2	368	408	2	22.22
14941	376	411	0	0.00

Following the identification of overlaps, we systematically checked whether there were any two non-overlapping themes, each covering at least a quarter ($\geq$25%) of the binding site residues. Additionally, we checked for the presence of any single themes that covered at least half ($\geq$50%) of the binding site residues but were not already participating in a theme pair. The final list of binding theme pairs identified for e2xyiA1 is presented in Table 4.24.

As seen in Table 4.24, ten different binding theme pairs were identified for peptide binding in domain e2xyiA1. The binding residues located at the N-terminal region of the sequence are primarily covered by two themes, 14940 and 14945, while the binding

residues at the C-terminal region are covered by multiple themes of varying lengths (14950, 14956, 15222, 14815, and 14951). When considering residue specificity, the theme pair 14940–14951 appears to provide the most compact and precise coverage of the ligand-binding residues, representing both the N-terminal and C-terminal binding sites with minimal sequence distance.

Table 4.24. Binding theme pairs identified for peptide binding in domain e2xyiA1.

Domain ID	**Ligand**	**Theme 1**	**Theme 2**
e2xyiA1	peptide	14940	14950
e2xyiA1	peptide	14940	14956
e2xyiA1	peptide	14940	15222
e2xyiA1	peptide	14940	14815
e2xyiA1	peptide	14940	14951
e2xyiA1	peptide	14945	14950
e2xyiA1	peptide	14945	14956
e2xyiA1	peptide	14945	15222
e2xyiA1	peptide	14945	14815
e2xyiA1	peptide	14945	14951

As seen in Table 4.24, ten different binding theme pairs were identified for peptide binding in domain e2xyiA1. The binding residues located at the N-terminal region of the sequence are primarily covered by two themes, 14940 and 14945, while the binding residues at the C-terminal region are covered by multiple themes of varying lengths (14950, 14956, 15222, 14815, and 14951). When considering residue specificity, the theme pair 14940–14951 appears to provide the most compact and precise coverage of

the ligand-binding residues, representing both the N-terminal and C-terminal binding sites with minimal sequence distance.

4.3.2. Theme Communication Analysis Using GNM-TE

To investigate the dynamic communication between identified themes, we applied the Gaussian Network Model-based Transfer Entropy (GNM-TE) method. GNM-TE combines the principles of elastic network models and information theory to evaluate directional information flow between different regions of a protein based on their collective dynamics.

Unlike standard cross-correlation approaches, which capture only symmetric relationships, GNM-TE provides asymmetric measures of causality — indicating whether fluctuations in one region statistically precede and influence those in another. This feature makes GNM-TE particularly well-suited for identifying potential allosteric pathways and communication hubs within the protein structure (see Materials and Methods section for details). In this study, we adapted the GNM-TE method to operate at the theme level, treating each sequence-defined theme as a dynamic unit of information exchange.

For each protein domain containing binding themes or theme pairs, we constructed a communication network where each node corresponded to a theme. Transfer entropy values were computed between all possible theme pairs, resulting in a directed information flow matrix. Here in this part, as a subset of modes used in GNM-TE method, we decided to include all the low frequency modes, and all the calculations made with 1-10 modes subset.

For each theme, the net transfer entropy (netTE) was calculated as the difference between the total outgoing and incoming information. Themes with positive netTE values acted as net information sources (senders), while themes with negative netTE values functioned as net information sinks (receivers). In addition to netTE analysis,

we analyzed the distribution of global information sources and global sinks within each theme. Global information source and sink residues identified based on TEcol scores which combines residue collectivities with GNM based TE analysis (see Materials and Methods section for details). Themes enriched in global sources or sinks were further characterized in terms of their communication roles.

Themes were classified into communication categories based on the Z-score of their net transfer entropy (netTE) and the proportions of global peak (source-like) and sink (receiver-like) residues they contained. The classification system prioritizes biologically interpretable features — such as directionality of information flow and enrichment of global signal transduction residues — while also accounting for relative dynamic prominence within each domain through netTE standardization. The classification criteria, together with the associated structural and dynamic conditions, are summarized in Table 4.25 along with their potential biological interpretations.

Table 4.25. The classification criteria with structural and dynamic conditions and potential biological interpretations.

Condition	**Class Name**	**Possible Biological Relevance**
Source% $\geq$ 50% and Sink% $\geq$ 50%	Strong sender-receiver	Themes integrating bidirectional dynamics; potential allosteric hubs
Source% $\geq$ 50% or (Z $\geq$ +1 and Source% $\geq$ 25%)	Strong sender	Global information sources with strong structural involvement
Sink% $\geq$ 50% or (Z $\leq$ –1 and Sink% $\geq$ 25%)	Strong receiver	Global information sinks with strong structural involvement

Table 4.25 Classification criteria with structural and dynamic conditions. (cont.)

Condition	**Class Name**	**Possible Biological Relevance**
Z between –1 and +1 and Source% ≥ 25% and Sink% ≥ 25%	Balanced sender-receiver	Themes that may relay or modulate information bi-directionally
Z ≥ +1 or (netTE ¿ 0 and Source% ≥ 25% and Sink% <25%)	Sender	Dynamic information source with moderate local support
Z ≤ –1 or (netTE <0 and Sink% ≥ 25% and Peak% <25%)	Receiver	Dynamic information sink with moderate local support
netTE >0 and Sink% ≥ 25% and Peak% <25%	Sender with local receiver features	Theme acts globally as sender, but locally enriched in sinks
netTE <0 and Peak% ≥ 25% and Sink% <25%	Receiver with local sender features	Theme acts globally as receiver, but locally enriched in sources
netTE Z between –1 and +1, Peak% <25%, Sink% <25%	Weak	Likely passive in signal transduction or purely structural roles

This classification scheme enabled the systematic identification of allosteric communication elements across the protein domains through themes, providing insights into the dynamic roles of themes and binding themes within broader signaling networks. The classification results for the themes in each protein domain are presented in Appendix B. Detailed data for each domain, including classification assignments and information transfer partners, are provided in separate files in the accompanying data

repository [146].

While this classification scheme provided a dynamic perspective on information flow across protein structures, we next sought to determine whether specific communication classes—such as Strong receivers or Senders—were preferentially associated with biologically validated binding events. To this end, we performed a statistical enrichment analysis to evaluate whether certain theme classes were overrepresented among themes identified as binding-related. This allowed us to quantitatively assess the functional significance of communication categories in the context of ligand interaction.

To assess whether specific communication classes were statistically enriched among biologically defined binding themes, we performed Fisher's exact test independently for each class. For each test, a 2 by 2 contingency table was constructed comparing the number of binding and non-binding themes that fell into a given communication category.

The resulting raw p-values were corrected for multiple hypothesis testing using two standard procedures. Bonferroni correction, which controls the family-wise error rate by multiplying each p-value by the total number of tests. While conservative, this method reduces the likelihood of Type I errors. Benjamini–Hochberg false discovery rate (FDR) correction, which controls the expected proportion of false positives among the declared significant results. This method offers improved sensitivity in exploratory settings, particularly when testing multiple interrelated hypotheses.

A communication class was considered significantly enriched if the corrected p-value (either Bonferroni- or FDR-adjusted) was below 0.05. The results of Fisher's exact test, along with corrected p-values using both Bonferroni and Benjamini–Hochberg false discovery rate (FDR) procedures, are summarized in Table 4.26 and class-level enrichment trends are visualized in Figure 4.56.

Table 4.26. Enrichment of communication classes among biologically defined binding themes. For each class, the number of binding themes, total themes, and binding theme fraction are listed. Statistical significance was tested using Fisher's exact test with Bonferroni and FDR corrections. Bold values indicate significance at $p < 0.05$ or $q < 0.05$.

Class	**Binding Themes**	**Total Themes**	**Binding Theme Fraction**	**p-value (Fisher)**	**Bonferroni p-value**	**FDR q-value**
Strong sender receiver	28	46	0.61	**0.0220**	0.1977	**0.0432**
Strong sender	43	74	0.58	**0.0108**	0.0976	**0.0325**
Strong receiver	103	180	0.57	**0.0001**	**0.0007**	**0.0007**
Sender	68	188	0.36	**0.0240**	0.2162	**0.0432**
Receiver	71	213	0.33	**0.0007**	**0.0065**	**0.0032**
Balanced sender receiver	19	51	0.37	0.3874	1.0000	0.3874
Sender with LRF	12	19	0.63	0.1033	0.9299	0.1550
Receiver with LSF	17	31	0.55	0.2703	1.0000	0.3041
Weak	129	319	0.40	0.1820	1.0000	0.2340

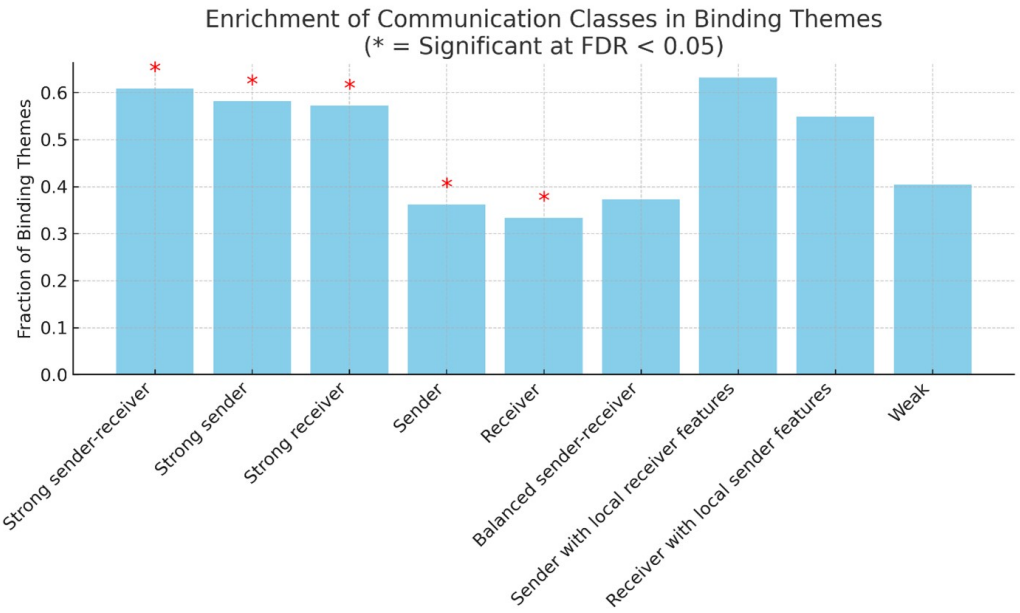


Figure 4.56. Enrichment of communication classes among binding themes. The y-axis shows the fraction of themes in each class involved in binding. Red asterisks (*) mark classes with significant enrichment at FDR q <0.05 (Benjamini–Hochberg corrected Fisher's test). Strong receiver, Receiver, Sender, Strong sender, and Strong sender-receiver classes show notable enrichment among binding themes.

Among all communication classes, Strong receiver and Receiver themes were significantly enriched among biologically validated binding themes, even after correction using the conservative Bonferroni method ($p < 0.001$ and $p < 0.01$, respectively). This finding suggests that signal-absorbing themes play a central role in mediating ligand recognition or regulatory function across diverse protein domains. When using the Benjamini–Hochberg procedure to control the false discovery rate, five out of nine communication classes showed statistically significant enrichment ($q < 0.05$), including Strong sender-receiver, Strong sender, and Sender classes, which were not identified under Bonferroni correction. These results underscore the increased sensitivity of the FDR approach in detecting biologically meaningful, though potentially more subtle, patterns of functional communication.

These findings suggest that communication themes functioning as global receivers or sender-receiver hubs are frequently positioned at, or proximal to, biologically relevant binding sites. This supports the interpretation that such themes are not limited to local structural stabilization but may also serve as key mediators of long-range information transduction initiated by ligand binding. The enrichment of Strong receiver themes among binding sites reinforces the notion that allosteric signal flow often converges at these locations. Similarly, the presence of Strong sender and Sender-receiver themes within binding regions implies a dual functional role—both in sensing external perturbations and broadcasting conformational changes to distal regions. Overall, these results support the view that binding themes are embedded within global information transfer pathways and may serve as structural and dynamic hubs in allosteric regulation.

To further illustrate these functional roles, we selected representative protein domains where binding themes also exhibited prominent sender or receiver behavior based on their transfer entropy profiles. These case studies highlight the structural localization of communication hubs and their integration within broader information flow pathways.

4.3.3. Case Studies of Allosteric Communication Through Binding Themes

As our first case study, we focused on the protein domain e2xyiA1 propeller (PDB ID: 2XYI), one of the main systems analyzed in our previous work [59]. This domain binds to histone 4 (H4), and ten different theme pairs were previously identified as covering the H4-binding site. Here, we highlight one such pair: theme 14940 and theme 14950.

We performed GNM-TE analysis on e2xyiA1, identifying residue-level entropy transfer and locating global source and sink residues across the domain. Using our classification scheme defined in the previous section, theme 14940 was labeled a Sender, while theme 14950 was classified as a Strong Receiver.

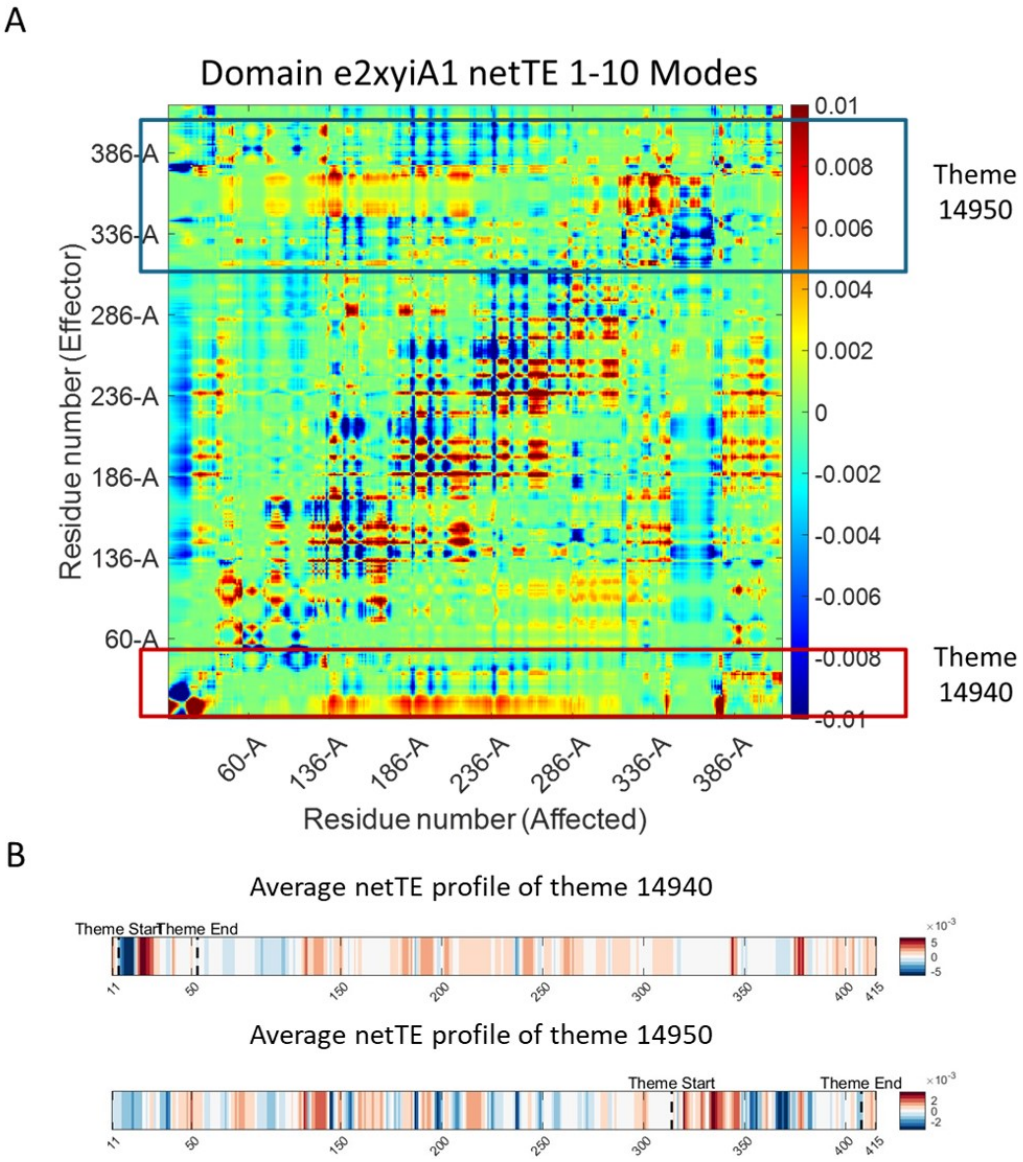


Figure 4.57. GNM-TE analysis of domain e2xyiA1 highlighting themes 14940 and 14950. Top: 2D netTE map with theme positions marked. Bottom: Per residue netTE bar plots for each theme. Theme 14940 (Sender) shows positive netTE; theme 14950 (Strong Receiver) displays negative netTE. Colors follow a red–white–blue scale (sender to receiver).

Theme 14940 exhibits the highest per-residue netTE across the domain, indicating a strong tendency to initiate information flow. It also contains 2 of the 15 global source residues, reinforcing its role as a dynamic sender. In contrast, theme 14950

shows a net negative per-residue netTE and houses 4 of the 7 global sink residues, suggesting it acts as a terminal receiver for allosteric signals. Figure 4.57 presents the residue-level netTE distribution across e2xyiA1 (2D netTE map), together with the netTE profiles of themes 14940 and 14950.

To further investigate their roles, we examined the transfer entropy flow paths originating from the global source residues within the theme 14940 (Figure 4.58) and terminating at the global sink residues in theme 14950 (Figure 4.59).

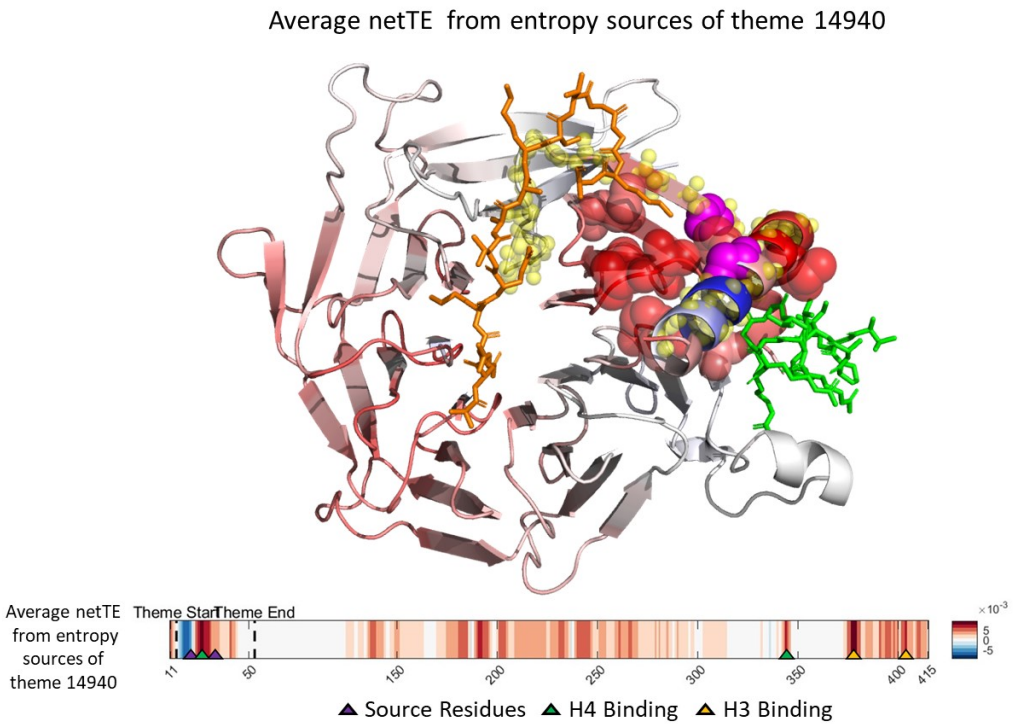


Figure 4.58. Information flow from global source residues in theme 14940 in domain e2xyiA1. Top: 3D structure colored by netTE from source residues (magenta spheres) within the theme (yellow). Transparent spheres indicate major receivers. H4 and H3 ligands shown as green and orange ribbons. Bottom: Bar plot of netTE flow; key residues marked by colored triangles. Red bias confirms theme 14940 as a sender.

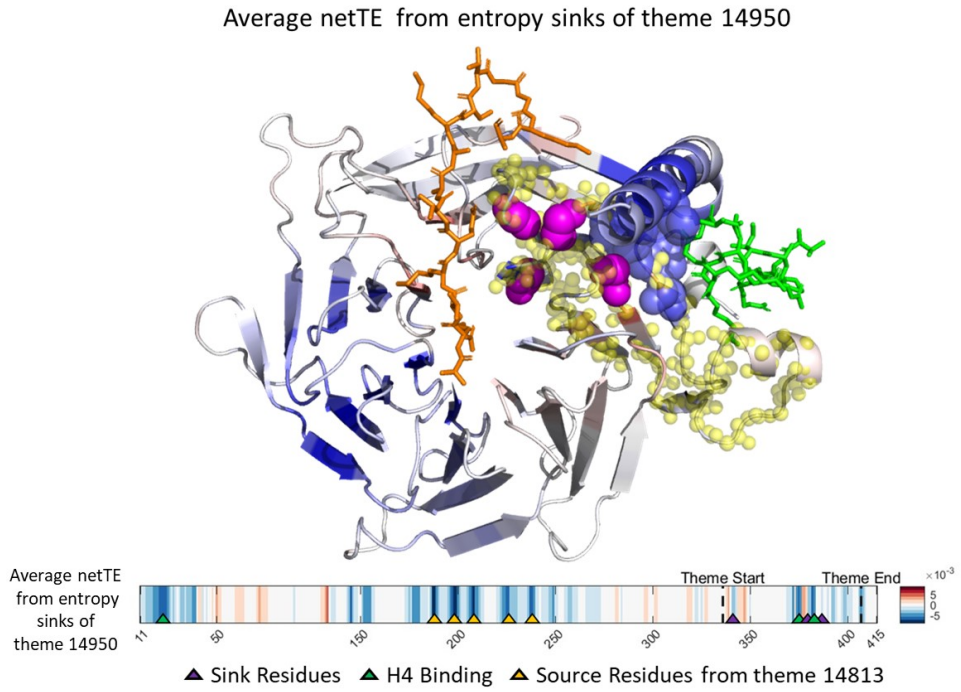


Figure 4.59. Information flow to global sink residues in theme 14950 in domain e2xyiA1. Top: 3D view colored by entropy flow into sink residues (magenta) in theme 14950 (yellow). Transparent spheres indicate major senders. Bottom: Bar plot of netTE to sinks; color-coded triangles mark source and binding residues. Predominant blue supports receiver classification.

Notably, source residues in theme 14940 transfer information to multiple H4-binding residues, not only within their own theme but also toward theme 14950. Interestingly, this includes residues involved in histone 3 (H3) binding — a second interaction not captured in BioLiP2 but reported in our earlier analysis [59]. Conversely, sink residues in theme 14950 receive substantial entropy inflow from source residues of theme 14940, as well as from theme 14813 (residues 170–240), which is also classified as a Sender. This suggests a multi-theme allosteric network, where sender themes are located both proximal and distal to the binding site coordinate information flow toward key functional sinks. Such an arrangement implies that the H4-binding interface may not act solely as a receiver but as a communication endpoint influenced by multiple regulatory hotspots across the structure.

As a second representative example, we analyzed the Rossmann-related domain e2ae2A1 (tropinone reductase-II, PDB ID: 2AE2). This case is particularly informative because the BioLiP2 database lists two biologically relevant ligands for this domain: NAP (nicotinamide adenine dinucleotide phosphate) and PTD (tropinone). The analysis revealed that theme 2962 encompasses all of the PTD-binding residues, while the theme pair 2939–2962 together capture the majority of the NAP-binding interface.

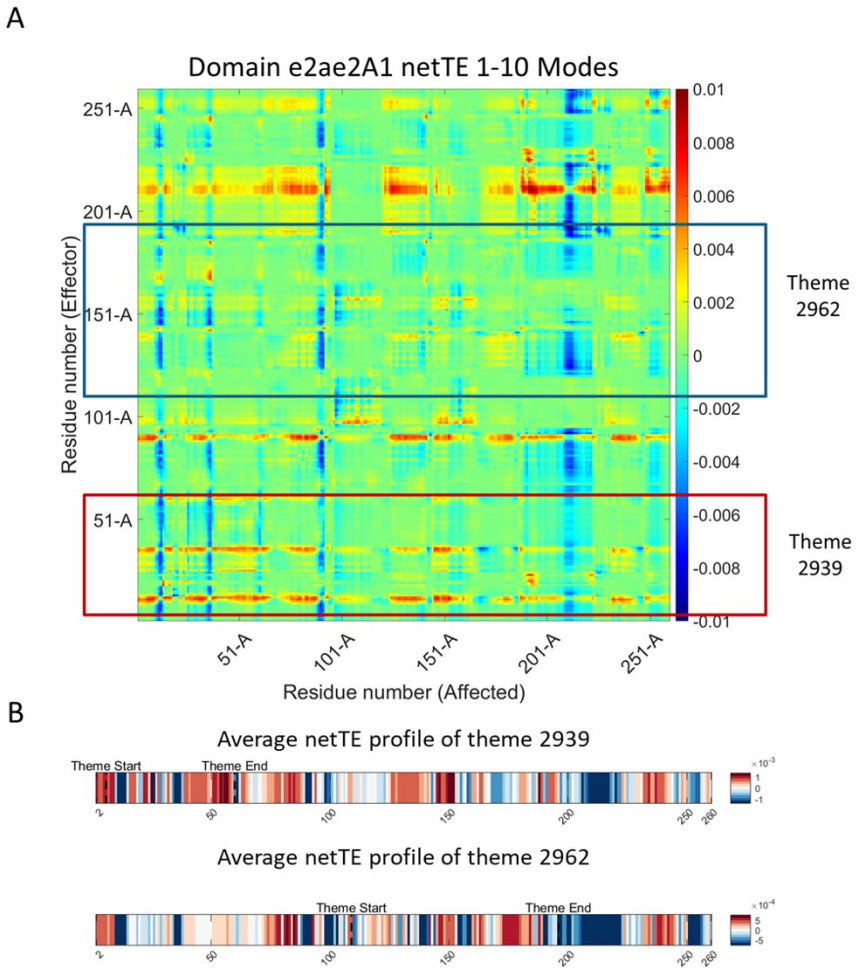


Figure 4.60. GNM-TE analysis of e2ae2A1 (PDB ID: 2AE2) highlighting themes 2939 and 2962. (a) 2D netTE map with theme locations marked. (b) Per residue netTE bar plots for each theme. Theme 2939 (Sender) shows positive netTE; theme 2962 (Strong Receiver) shows negative netTE. Bars are color-coded red–white–blue from sender to receiver.

Interestingly, these themes display contrasting communication profiles in our classification scheme, similar to the case of e2xyiA1. Theme 2939 is categorized as a Sender theme, with a high per-residue netTE score and 2 out of 4 global source residues. Theme 2962 is classified as a Strong Receiver, containing 2 of the 3 global sink residues and exhibiting strongly negative per-residue netTE. The residue-level netTE distribution across the domain and together with the netTE profiles of themes 2939 and 2962 are visualized in Figure 4.60.

To further dissect their communication roles, we traced the transfer entropy flow paths originating from the global source residues in theme 2939 (Figure 4.61) and terminating at the sink residues in theme 2962 (Figure 4.62).

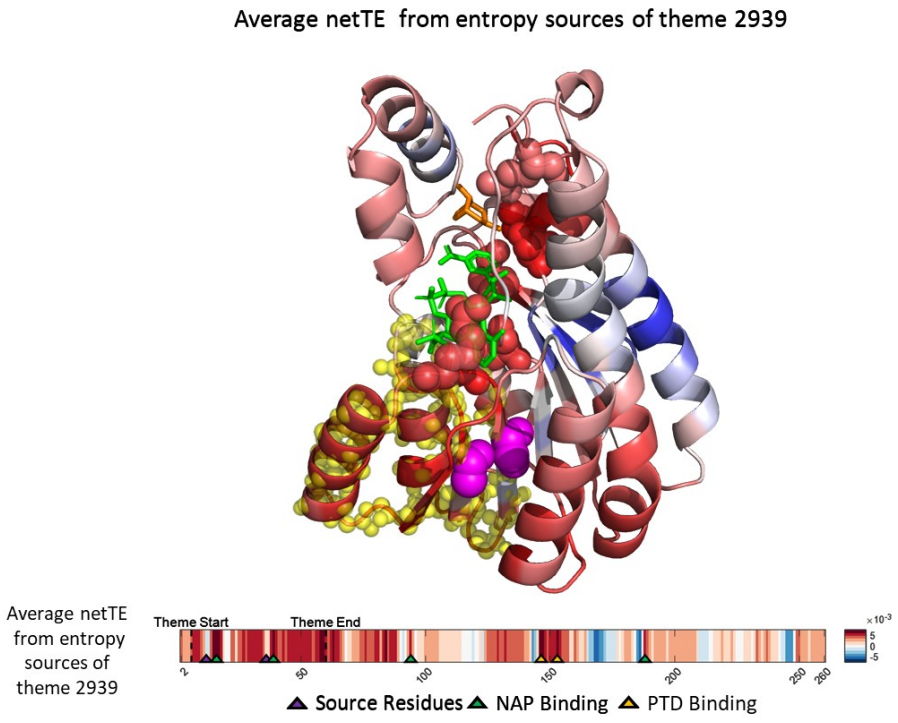


Figure 4.61. Information flow from source residues in theme 2939 of e2ae2A1. Top: Structure colored by netTE from source residues (magenta) in theme 2939 (yellow). Ligands NAP/PTD shown as green/orange ribbons. Bottom: Bar plot of transfer to other residues; source residues and binding residues marked by colored triangles.

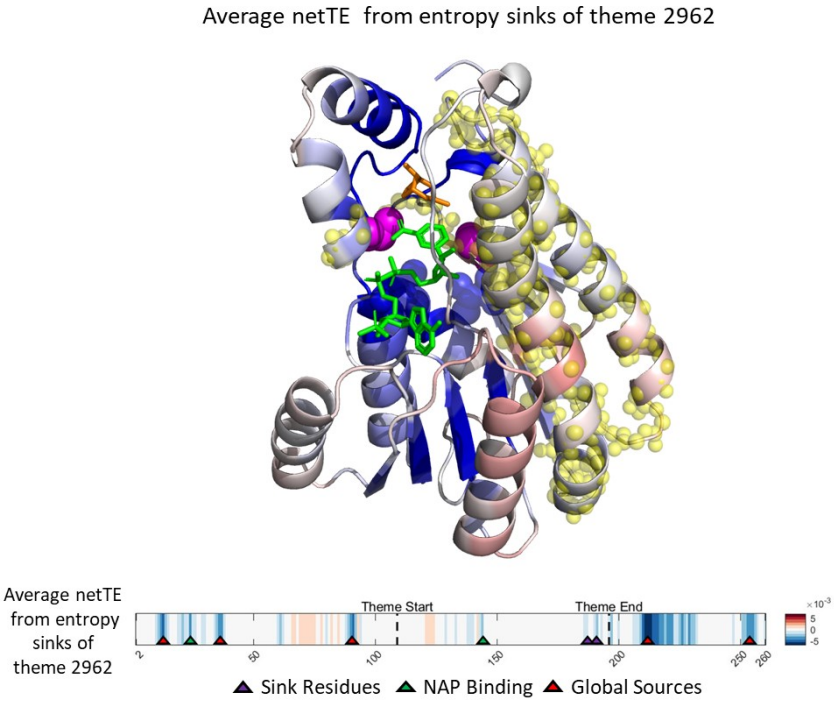


Figure 4.62. Information flow to sink residues in theme 2962 of e2ae2A1. Top: Structure colored by entropy transfer to sink residues (magenta) in theme 2962 (yellow). Key sender residues shown as transparent spheres. Bottom: Bar plot of transfer from other residues; sink residues, binding residues and global sources marked by colored triangles.

Source residues in theme 2939 were found to transmit entropy toward several NAP-binding residues, both within the theme and in adjacent regions. These residues also sent information directly to PTD-binding residues located in theme 2962, suggesting a potential coupling between the ligand-binding sites. In turn, sink residues in theme 2962 received substantial entropy inflow from global sources in theme 2939, as well as from residue 212 (a source outside both themes). Additional input was observed from NAP-binding residue 144 (within theme 2962) and residue 21 (from theme 2939), reinforcing the bidirectional communication between these ligand-binding sites. This example illustrates how ligand-binding themes from distinct regions of the protein can engage in long-range allosteric signaling, and how their dynamic roles—as senders or receivers—can be resolved using GNM-TE analysis.

4.3.4. Conclusion

In this section, we systematically characterized the dynamic roles of sequence-defined protein themes across 150 structurally diverse domains, with a particular focus on those involved in biologically relevant ligand binding. By integrating ligand-binding annotations from the BioLiP2 database with GNM-based transfer entropy (GNM-TE) analysis, we identified binding themes and classified them into communication categories reflecting their roles in information propagation. Statistical enrichment tests revealed that Strong receiver and Receiver themes were significantly overrepresented among binding themes, suggesting that ligand-binding sites often function as entropy-absorbing endpoints in allosteric signaling pathways. Other classes such as Strong sender, Sender and Sender-receiver also showed enrichment under FDR correction, indicating that bidirectional communication may be a common feature of binding regions. Detailed case studies further illustrated how theme-level communication patterns align with ligand interactions and long-range allosteric communication. Together, these findings support the view that binding themes are not merely structural motifs but act as dynamically integrated elements within global signaling networks.

4.4. Additional Contributions

4.4.1. Extension of GNM Perturbation Framework

This work builds upon the foundation of my Master's thesis and led to the publication of the article titled "Global Dynamics Renders Protein Sites with High Functional Response" in The Journal of Physical Chemistry B [131]. The study explores the extent to which protein structural dynamics, modeled using the Gaussian Network Model (GNM), can account for the sensitivity of specific sites to mutations observed in deep mutational scanning (DMS) experiments. It demonstrates that mutation-sensitive residues often act as key modulators of global dynamics and can alter protein function via entropy-driven allosteric mechanisms—even when located distally from canonical functional sites.

As part of this effort, we employed a minimalist GNM-based perturbation–response scanning strategy across a dataset of seven DMS-characterized proteins. The analysis revealed that residues most sensitive to mutations tend to exhibit high allosteric control, decreasing the fluctuation of other residues upon perturbation and thus modulating functional dynamics. The results highlighted a previously underappreciated role of such residues in entropy-based allosteric control, offering new insights into their functional relevance.

To further develop and disseminate this approach, I presented a poster titled "Global motions reveal functionally sensitive mutational sites with allosteric response" at the EMBO Workshop on Computational Structural Biology in Heidelberg, Germany (December 2023). This presentation summarized ongoing improvements and expanded analyses.

4.4.1.1. Poster Summary and Methodological Enhancements. The GNM-based perturbation–response scanning framework simulates point mutations by applying perturbations at individual residues and monitoring changes in residue-level fluctuations, thus reflecting entropy-driven alterations in global motions. In the original study, the method showed strong correlations with experimental mutational fitness data across seven proteins [131].

To broaden the applicability and robustness of the method, we extended the dataset to 20 proteins and implemented several refinements, including Algorithmic improvements, such as integration of peak prominence detection to better identify functionally significant minima positions on Cumulative Residue Fluctuation (CRF) difference graphs and validation via statistical significance tests, where we assessed both residue-wise matches and spatial distances between predicted allosteric sites and mutation-sensitive residues from DMS studies.

These enhancements yielded statistically significant results. In 17 out of 20 proteins, local/global minima positions on delta CRF significantly overlapped with DMS-

sensitive residues ($p < 0.05$). In 19 out of 20 proteins, the predicted residues were in close spatial proximity to mutation-sensitive sites, further supporting the method's reliability.

4.4.1.2. Case Study: ABC Transporter ErfCD. In the ABC transporter ErfCD, our refined GNM-based perturbation–response scanning method successfully identified functionally and allosterically critical residues. The corresponding DMS study focused on 65 residues within the Transmembrane Domain (TMD) and 14 highly conserved residues in the Nucleotide-Binding Domain (NBD). Notably, 13 mutation-sensitive residues overlapped with local or global CRF minima predicted by our method (Figure 4.63), confirming both functional and allosteric relevance. Furthermore, residues such as 103^{EfrC}, 192^{EfrC}, 197^{EfrC}, 102^{EfrD}, 194^{EfrD}, 210^{EfrD}, and 225^{EfrD} displayed strong perturbation responses, despite lacking available DMS data. We propose that these residues may have allosteric coupling potential with nucleotide-binding regions

4.4.1.3. Outlook: Server Development. Given the promising performance of our refined GNM-PRS approach, we have initiated the development of a dedicated web server to enable wider accessibility to the method. This platform aims to provide a user-friendly interface for simulating entropy-driven allosteric responses in proteins, thereby facilitating its adoption by the broader scientific community for mutation sensitivity prediction, functional annotation, and druggable site identification.

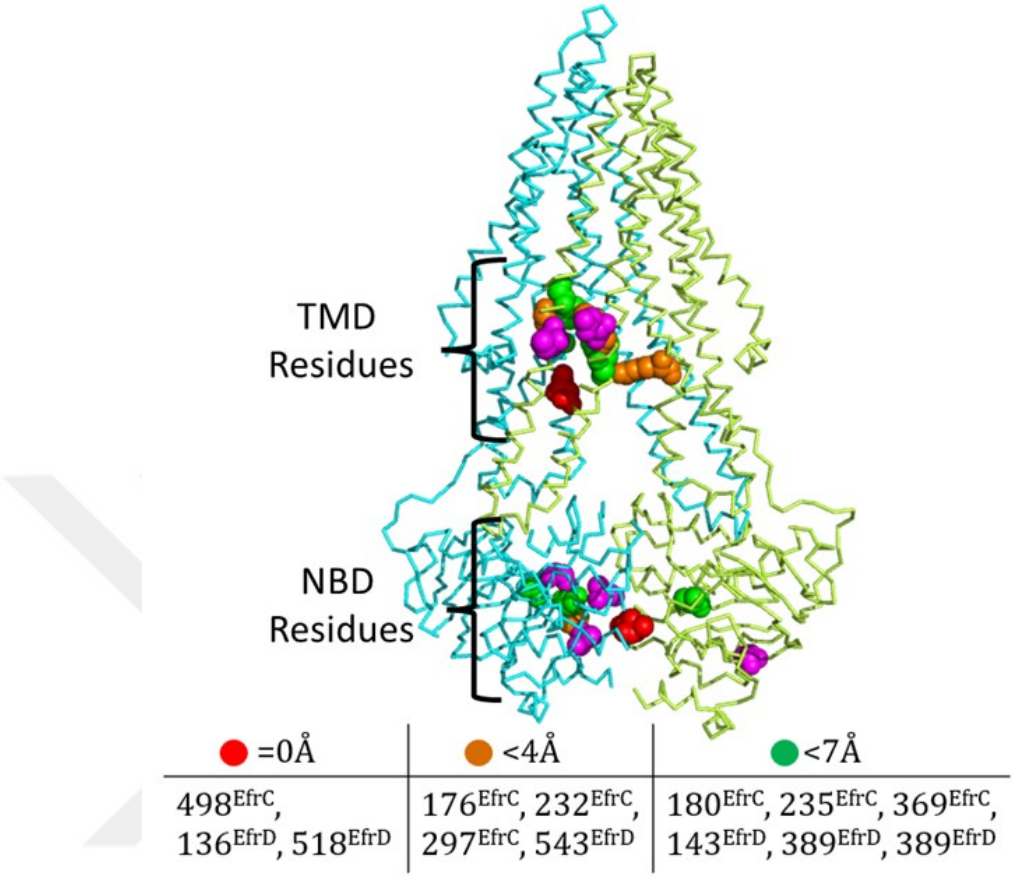


Figure 4.63. Thirteen mutation sensitive positions and related local/global minima positions of delta CRF represented on the structure of the EfrCD as spheres.

4.4.2. Optimization and Extension of the GNM-Based Transfer Entropy Method

The GNM-based transfer entropy (GNM-TE) method quantifies directional information transfer between residues by leveraging time-delayed correlations derived from Gaussian Network Model (GNM) dynamics within a Shannon entropy framework [55, 56]. During my doctoral studies, I contributed extensively to improving the algorithm, focusing on enhancing computational performance, addressing critical bottlenecks, and implementing user-friendly features to support broader application.

To facilitate community-wide accessibility and reproducibility, we are currently preparing a web server implementation of the enhanced GNM-TE method. This platform will enable users to conduct large-scale residue-residue information flow analyses directly from protein structural data. Moreover, this server will support and complement several analyses described in later sections of this thesis, where the GNM-TE method plays a central role.

In addition to protein-only systems, a major advancement involved expanding GNM-TE to handle RNA, DNA, and protein-nucleotide complexes. We developed alternate modules for:

- Nucleotide-only systems (e.g., DNA duplexes)
- Protein-nucleotide complexes, including logic to separately handle protein, nucleotide, and interface components.

This extension represents a novel step toward characterizing information flow in nucleic acids and hybrid systems, an area with limited prior computational investigation. Preliminary analyses on DNA structures have already yielded promising results, suggesting the method's potential to uncover dynamic communication networks beyond proteins. The overarching goal of these enhancements is to establish a generalized dynamic framework applicable to diverse biomolecular systems—ranging from multidomain proteins to nucleic acid-protein assemblies—by exploiting intrinsic fluctuations encoded in available structural data.

Additionally, as part of a collaborative study led by Dr. Hasan, I contributed a detailed GNM and GNM-TE analysis of the KEAP-1 protein to investigate the dynamic underpinnings of DMF binding. This analysis revealed that the allosteric behavior of DMF-binding residues becomes more pronounced in the ambient temperature structure, with specific GNM modes, particularly the third slowest, driving the relevant conformational fluctuations. The GNM-TE results successfully identified effector residues aligned with the DMF binding site, supporting the role of thermal fluctu-

ations in modulating functional dynamics. While this work complements a broader study, it illustrates the effectiveness of GNM-TE in capturing subtle allosteric changes in condition-dependent protein states.

Overall advancements in GNM-TE can be summarized as following.

- Debugged and optimized the original GNM-TE implementation.
- Automated and modularized the workflow for ease of use.
- Developed separate computational strategies for protein, nucleotide, and interface regions.
- Laid the groundwork for a web-based GNM-TE server to promote wider usage and integration.

4.4.3. Enhancements to the ANM-LD Simulation Framework

During my PhD, I contributed significantly to the development and refinement of the ANM-LD simulation method, a hybrid approach that combines Anisotropic Network Model (ANM) dynamics with Langevin Dynamics to model conformational transitions [97, 98]. As part of a collaborative effort toward establishing the ANM-LD Path server, I implemented GPU support to improve computational efficiency and scalability. I also developed a module enabling the extraction and use of selected collective modes during simulations, offering users finer control over the transition pathway generation. To ensure the robustness of the pipeline, I took part in debugging the codebase and contributed to automating the post-simulation analysis workflow, enabling the system to handle premature simulation termination and improve reproducibility. These improvements are currently being integrated into a server interface designed to make the ANM-LD methodology accessible to the broader research community.

Building upon an earlier implementation, we extended the ANM-LD framework to incorporate low-resolution structural data, such as SAXS and cryo-EM, as guides for simulating conformational transitions. This work was presented as a poster ti-

tled "Low-Resolution Structural Data Guides Prediction of Conformational Transition Pathways in Proteins using ANM-Langevin Dynamics Hybrid Methodology" at the EMBO Workshop Physics of Living Systems in Dresden, Germany (June 2023). The case study of Adenylate Kinase (AdK) demonstrated that SAXS-derived open-state profiles could effectively guide ANM-LD simulations starting from closed-state crystal structures, predicting physically plausible intermediate states along the transition pathway. The resulting trajectories exhibited reduced $RMSD$ and χ^2 values and followed low-energy routes consistent with experimentally observed conformers. This implementation highlights the potential of ANM-LD to extract functional dynamics from sparse experimental data and provides a foundation for integrating ensemble-averaged measurements into hybrid modeling approaches. The schematic overview of the SAXS and cryo-EM guided ANM-LD workflow is presented in Figure 4.64, while the application to AdK, including the mapping of simulated intermediates onto the free energy landscape [147] and convergence trends in $RMSD$ and χ^2 values, is illustrated in Figure 4.65.

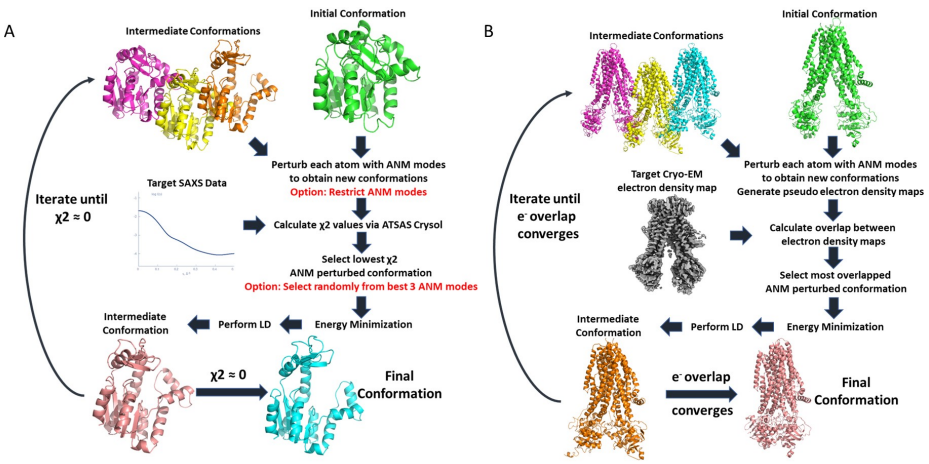


Figure 4.64. Schematic overview of the low-resolution structural data driven ANM-LD hybrid simulation. (a) SAXS data and (b) cryo-EM data

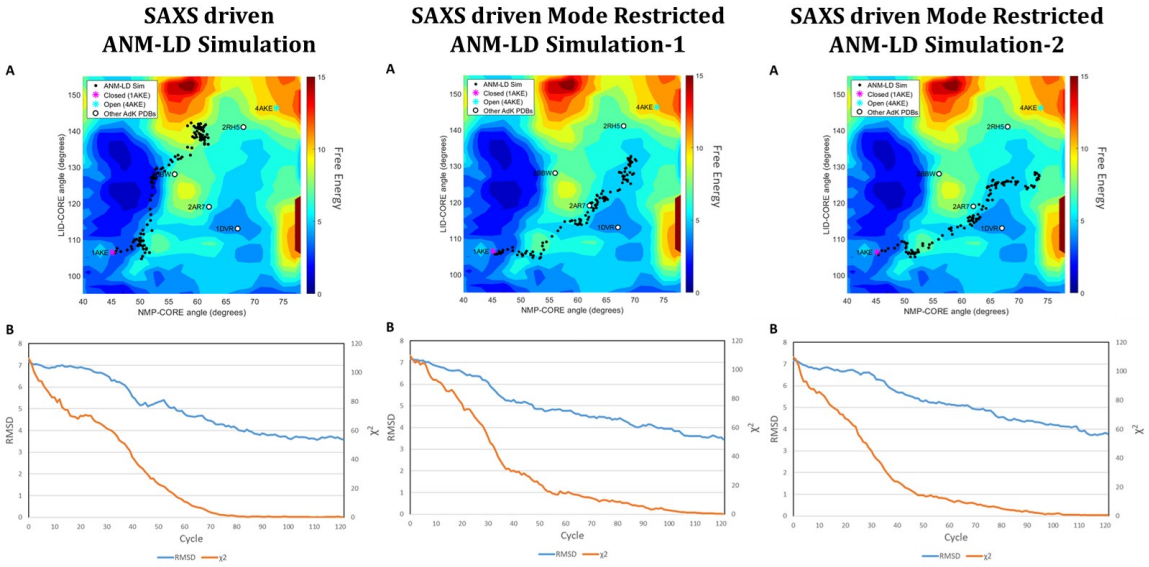


Figure 4.65. Representation of intermediate structures obtained from ANM-LD simulations using SAXS data for the AdK system on the free energy landscape of AdK together with $RMSD$ values to the open form (4AKE) and χ^2 values of each cycle.

4.4.3.1. Preliminary Efforts to Incorporate Forward Flux Sampling into ANM-LD. A limitation of ANM-LD simulations is their inability to directly provide kinetic information, which restricts their use in time-resolved modeling of protein conformational transitions. To address this, we initiated the development of a hybrid approach that aims to integrate Forward Flux Sampling (FFS) into the ANM-LD simulation framework. FFS is a rare-event simulation method that estimates transition rates between defined states by sampling trajectories crossing a series of intermediate order parameters [148]. In this work, we used collective variables extracted from ANM-LD simulations—specifically the LID-Core and NMP-Core angles of adenylate kinase (AdK)(Figure 4.66)—to define successive milestones (e.g., λ_0, λ_1, λ_2,) along the transition pathway. This approach leverages the structural relevance of ANM-LD intermediates while embedding them in a kinetic framework that quantifies transition probabilities.

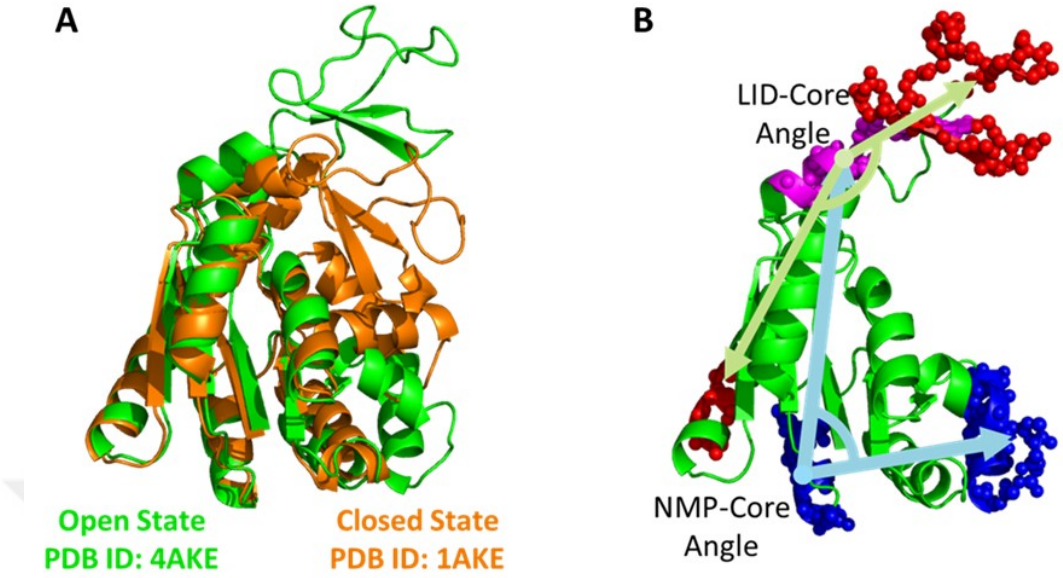


Figure 4.66. Structure of adenylate kinase (AdK). (a) Open-Close States of AdK. (b) Collective Variables of AdK.

In the proof-of-concept application to the AdK system, we first conducted 100 short explicit-solvent MD simulations from the open state to sample initial angle distributions. Structures reaching the 10th percentile of the angle distributions were used to define the first order parameter (λ_0), and corresponding ANM-LD intermediates provided λ_1 and λ_2 (Figure 4.67).

Thousands of short MD runs were then initiated from filtered intermediate structures using PLUMED to monitor collective variable evolution. The success rate between λ_0 and λ_1 was found to be $\approx 40.9\%$, and from λ_1 to λ_2, $\approx 16.7\%$, as shown in Figures 4.68 and 4.69, respectively. This yielded an estimated transition rate of ≈ 0.25 transitions/ns.

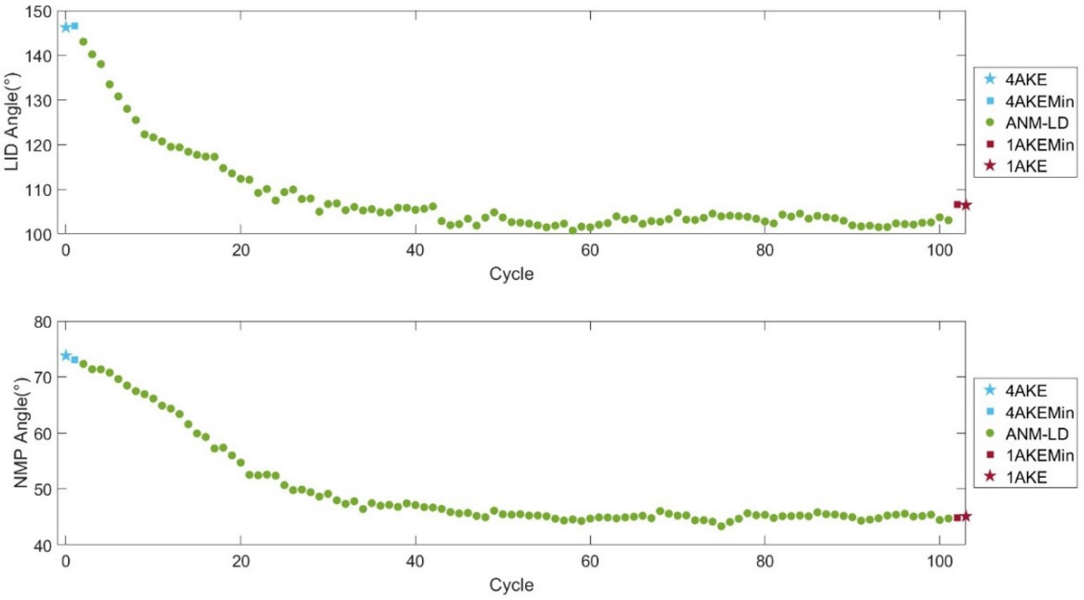


Figure 4.67. Collective variables calculated from intermediate structures obtained at each cycle of the standard ANM-LD simulation for the AdK system.

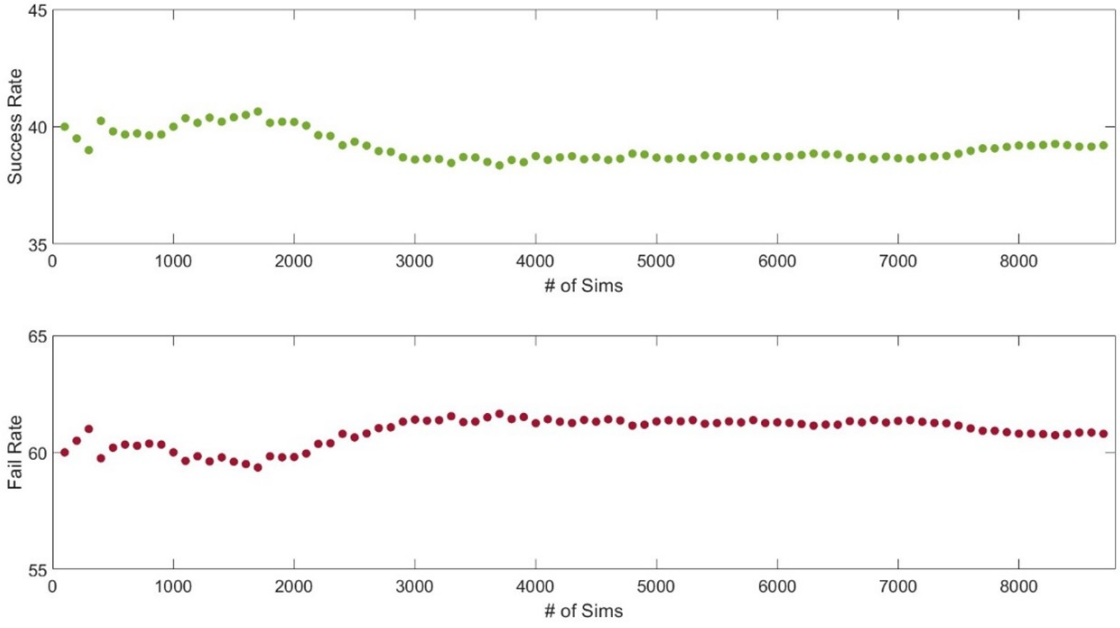


Figure 4.68. Rates of successful and unsuccessful simulations for the first step of FFS.

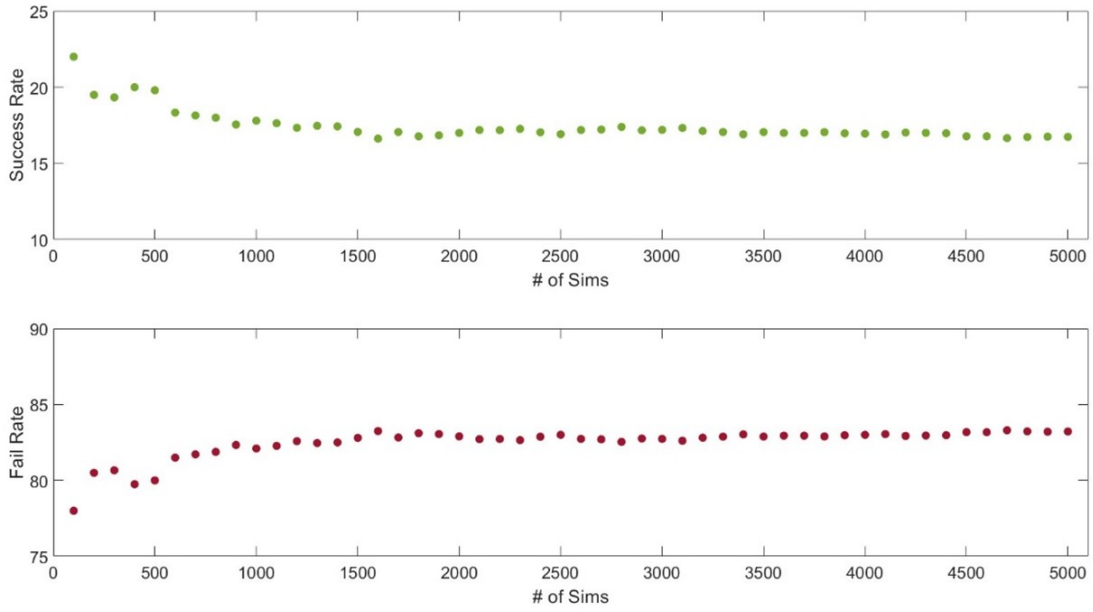


Figure 4.69. Rates of successful and unsuccessful simulations for the second step of FFS.

To support this framework, we implemented custom algorithms for tracking multiple collective variables and automating the success/failure classification. These results show that ANM-LD-derived intermediates can serve as a basis for kinetic modeling, providing both mechanistic and temporal resolution for large-scale conformational changes in proteins.

4.4.4. Comparative Analysis of GNM-Derived Segments and Protein Language Model Representations

In collaboration with Assoc. Prof. Arzucan Özgür from the Computer Engineering Department, I contributed to an ongoing project exploring the relationship between dynamic elements derived from Gaussian Network Model (GNM) analysis and segments predicted by protein language models (pLMs), particularly those generated using Evolutionary Scale Modeling (ESM). These models assign each amino acid to one of 64 predefined clusters, resulting in structural segmentations that differ markedly from

classical motifs in terms of size, continuity, and coverage.

To compare these representations meaningfully, we implemented and evaluated several computational strategies. These included correlation analysis between GNM dynamic domains and pLM segments, distance-based mapping of GNM hinge points to segment boundaries, and mutual information-based comparisons using individual and averaged GNM modes. To overcome challenges in segment extraction from 2D cross-correlation maps, I applied both statistical thresholding methods (e.g., percentile-based change detection) and unsupervised learning approaches, such as k-means clustering on residue-wise correlation profiles. The latter method allowed for consistent segment counts between datasets and enabled preliminary mutual information analyses across seven proteins. These efforts lay the groundwork for broader-scale comparisons in the next phase of the project, with planned integration of functional annotations and binding site information to contextualize the observed correlations.

4.4.5. Summary of Additional Contributions

In addition to the core research presented in this thesis, I undertook several side projects that expanded the methodological scope and computational depth of my doctoral work. These efforts included the refinement and application of GNM-based perturbation analysis to identify mutation-sensitive residues, optimization and extension of the GNM-TE framework to accommodate nucleotide-protein systems, and extensive contributions to the development of the ANM-LD simulation server, including its adaptation for low-resolution data and attempts for integration with kinetic modeling via Forward Flux Sampling. Finally, we initiated exploratory work to bridge protein language model representations with GNM-derived dynamic segments through collaborative analysis of segmentation patterns and mutual information. Collectively, these contributions reflect both the versatility of elastic network-based modeling and a broader commitment to method development and interdisciplinary collaboration.

5. CONCLUSION

This thesis explored how evolutionary reuse of sequence segments, known as themes, relates to the intrinsic dynamics and long-range communication within protein structures. While prior work demonstrated the recurrence of these short segments across proteins with different folds and functions, their dynamic characteristics and allosteric roles remained largely unexamined. Addressing this gap, the present work introduced a unified framework that integrates sequence-based evolutionary signals with structure-based dynamic analyses, thereby providing a novel perspective on the interplay between conservation, motion, and function.

In the first part of the thesis, we performed a large-scale comparative analysis of 150 ECOD domains and investigated the correspondence between themes, defined through statistically significant sequence similarity, and dynamic elements (DEs), extracted from low-frequency collective modes using the Gaussian Network Model (GNM). The results revealed a non-trivial and statistically significant alignment between theme boundaries and DEs across structurally diverse domains. This finding was consistent across detailed in-depth case studies and broader statistical validation, suggesting that reused segments are not only conserved in sequence but also embedded in dynamically coherent regions. These observations imply that evolutionary modularity in protein sequences may be preserved and repurposed through their dynamic roles.

The second part of the thesis focused specifically on adenine-binding themes, leveraging a curated dataset of kinase-like proteins and several motif-defined clusters. By combining GNM with GNM-based Transfer Entropy (GNM-TE) analysis, we showed that residues within these ligand-interacting motifs frequently act as sources or sinks of dynamic information, mediating allosteric signaling across the protein structure. Importantly, not all residues involved in this communication were directly in contact with the ligand, indicating latent allosteric capacities. These results emphasize that adenine-binding motifs function not only as structural binding sites but also as

dynamically active communication hubs.

In the third part, we extended our analysis to a broader classification of biologically defined binding themes across the entire dataset. By integrating ligand annotations with GNM-TE-based communication metrics, we classified themes into communication categories (e.g., Strong sender, Receiver, Sender-receiver). Statistical enrichment analyses using Fisher's exact test and multiple hypothesis correction methods revealed that ligand-binding themes are disproportionately found among dynamic receivers or hubs in allosteric networks. These results establish a strong link between structural motifs involved in binding and their role in system-wide information flow.

Lastly, the thesis includes additional contributions made during the PhD, including the development and application of GNM-TE to diverse systems, enhancements to the ANM-LD simulation framework, and collaborative work on the allosteric architecture of proteins. These contributions further demonstrate the general applicability of the methods developed and their relevance to diverse protein systems.

In summary, this work presents the first systematic framework for linking evolutionarily conserved sequence motifs with structure-based dynamic features and information flow in proteins. By introducing the concept of dynamic elements and combining it with mutual information, transfer entropy, and rigorous statistical validation, this thesis uncovers a fundamental layer of functional organization that had previously been overlooked: that many reused protein segments are not only conserved in sequence and structure but also participate in coherent dynamic behavior and long-range communication. This observation bridges the gap between two major paradigms in protein science, modular evolution and allostery, and suggests that nature may have selected not only for foldability and function, but also for dynamic compatibility across reused elements.

The novelty of this work lies not only in its integrative methodology, but also in its conceptual contribution: it reframes protein themes not merely as evolutionary

or structural units, but as dynamic and communicative modules that contribute to the allosteric landscape of proteins. The statistical rigor applied throughout through information-theoretic analyses, null models, and enrichment testing, ensures that these conclusions are not anecdotal but broadly generalizable. The dynamic characterization of adenine-binding themes and the classification of communication-capable motifs open new directions for identifying latent allosteric sites, understanding functional divergence, and engineering modular proteins with desired dynamic behaviors.

Taken together, the studies in this thesis reveal that evolutionary reuse may extend beyond structural compatibility to include a preserved dynamic logic embedded in the protein scaffold. This insight offers a powerful new lens for studying protein evolution, function, and design, and establishes a foundation for future efforts to uncover dynamic codes of allostery conserved across folds, functions, and time.

5.1. Recommendations

Building on the findings of this thesis, future studies could expand the dynamic analysis to include other classes of conserved motifs—such as metal-binding sites or protein–protein interaction segments—to evaluate whether the established correspondence between reused sequence elements and dynamic modules generalizes across broader functional contexts. Additionally, exploring the relationship between dynamics-based segmentation and structural annotations derived from deep learning models, such as protein language models, may uncover hidden patterns of functional modularity encoded in sequence. Integrating this framework with evolutionary reconstructions could also provide insights into how dynamic properties themselves are conserved or repurposed over time. Finally, the observed link between theme reuse and allosteric communication suggests promising applications in modular protein design, where evolutionarily and dynamically informed segments could be leveraged to engineer proteins with tailored conformational behaviors or regulatory features.

REFERENCES

1. Yang, L., G. Song and R. L. Jernigan, "How Well Can We Understand Large-Scale Protein Motions Using Normal Modes of Elastic Network Models?", *Biophysical Journal*, Vol. 93, No. 3, pp. 920–929, 2007.

2. Haliloglu, T. and I. Bahar, "Adaptability of Protein Structures to Enable Functional Interactions and Evolutionary Implications", *Current Opinion in Structural Biology*, Vol. 35, pp. 17–23, 2015.

3. Kaynak, B. T., J. M. Krieger, B. Dudas, Z. L. Dahmani, M. G. S. Costa, E. Balog, A. L. Scott, P. Doruker, D. Perahia and I. Bahar, "Sampling of Protein Conformational Space Using Hybrid Simulations: A Critical Assessment of Recent Methods", *Frontiers in Molecular Biosciences*, Vol. 9, 2022.

4. Shakhnovich, E., V. Abkevich and O. Ptitsyn, "Conserved Residues and the Mechanism of Protein Folding", *Nature*, Vol. 379, No. 6560, pp. 96–98, 1996.

5. Holm, L. and C. Sander, "Protein Folds and Families: Sequence and Structure Alignments", *Nucleic Acids Research*, Vol. 27, No. 1, pp. 244–247, 1999.

6. Chothia, C., J. Gough, C. Vogel and S. A. Teichmann, "Evolution of the Protein Repertoire", *Science*, Vol. 300, No. 5626, pp. 1701–1703, 2003.

7. Socolich, M., S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner and R. Ranganathan, "Evolutionary Information for Specifying a Protein Fold", *Nature*, Vol. 437, No. 7058, pp. 512–518, 2005.

8. Lee, D., O. Redfern and C. Orengo, "Predicting Protein Function from Sequence and Structure", *Nature Reviews Molecular Cell Biology*, Vol. 8, No. 12, pp. 995–1005, 2007.

9. Liberles, D. A., S. A. Teichmann, I. Bahar, U. Bastolla, J. Bloom, E. Bornberg-Bauer, L. J. Colwell, A. P. J. De Koning, N. V. Dokholyan, J. Echave, A. Elofsson, D. L. Gerloff, R. A. Goldstein, J. A. Grahnen, M. T. Holder, C. Lakner, N. Lartillot, S. C. Lovell, G. Naylor, T. Perica, D. D. Pollock, T. Pupko, L. Regan, A. Roger, N. Rubinstein, E. Shakhnovich, K. Sjölander, S. Sunyaev, A. I. Teufel, J. L. Thorne, J. W. Thornton, D. M. Weinreich and S. Whelan, "The Interface of Protein Structure, Protein Biophysics, and Molecular Evolution", *Protein Science*, Vol. 21, No. 6, pp. 769–785, 2012.

10. Nepomnyachiy, S., N. Ben-Tal and R. Kolodny, "Global View of the Protein Universe", *Proceedings of the National Academy of Sciences*, Vol. 111, No. 32, pp. 11691–11696, 2014.

11. Edwards, H. and C. M. Deane, "Structural Bridges through Fold Space", *Plos Computational Biology*, Vol. 11, No. 9, p. E1004466, 2015.

12. Alva, V. and A. N. Lupas, "From Ancestral Peptides to Designed Proteins", *Current Opinion in Structural Biology*, Vol. 48, pp. 103–109, 2018.

13. Cheng, H., R. D. Schaeffer, Y. Liao, L. N. Kinch, J. Pei, S. Shi, B.-H. Kim and N. V. Grishin, "ECOD: An Evolutionary Classification of Protein Domains", *Plos Computational Biology*, Vol. 10, No. 12, p. E1003926, 2014.

14. Andreeva, A., D. Howorth, C. Chothia, E. Kulesha and A. G. Murzin, "SCOP2 Prototype: A New Approach to Protein Structure Mining", *Nucleic Acids Research*, Vol. 42, No. D1, pp. D310–D314, 2013.

15. Andreeva, A., E. Kulesha, J. Gough and A. G. Murzin, "The SCOP Database in 2020: Expanded Classification of Representative Family and Superfamily Domains of Known Protein Structures", *Nucleic Acids Research*, Vol. 48, No. D1, pp. D376–D382, 2019.

16. Sillitoe, I., N. Dawson, T. E. Lewis, S. Das, J. G. Lees, P. Ashford, A. Tolulope, H. M. Scholes, I. Senatorov, A. Bujan, F. Ceballos Rodriguez-Conde, B. Dowling, J. Thornton and C. A. Orengo, "CATH: Expanding the Horizons of Structure-Based Functional Annotations for Genome Sequences", *Nucleic Acids Research*, Vol. 47, No. D1, pp. D280–D284, 2018.

17. Sillitoe, I., N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, C. S. M. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S. D. Lam, K. Berka, I. H. Varekova, R. Svobodova, J. Lees and C. A. Orengo, "CATH: Increased Structural Coverage of Functional Space", *Nucleic Acids Research*, Vol. 49, No. D1, pp. D266–D273, 2020.

18. Schaeffer, R. D., K. E. Medvedev, A. Andreeva, S. R. Chuguransky, B. L. Pinto, J. Zhang, Q. Cong, A. Bateman and N. V. Grishin, "ECOD: Integrating Classifications of Protein Domains from Experimental and Predicted Structures", *Nucleic Acids Research*, Vol. 53, No. D1, pp. D411–D418, 2024.

19. Eck, R. V. and M. O. Dayhoff, "Evolution of the Structure of Ferredoxin Based on Living Relics of Primitive Amino Acid Sequences", *Science*, Vol. 152, No. 3720, pp. 363–366, 1966.

20. Berezovsky, I. N., A. Y. Grosberg and E. N. Trifonov, "Closed Loops of Nearly Standard Size: Common Basic Element of Protein Structure", *FEBS Letters*, Vol. 466, No. 2–3, pp. 283–286, 2000.

21. Trifonov, E. N., A. Kirzhner, V. M. Kirzhner and I. N. Berezovsky, "Distinct Stages of Protein Evolution as Suggested by Protein Sequence Analysis", *Journal of Molecular Evolution*, Vol. 53, No. 4–5, pp. 394–401, 2001.

22. Lupas, A. N., C. P. Ponting and R. B. Russell, "On the Evolution of Protein Folds: Are Similar Motifs in Different Protein Folds the Result of Convergence, Insertion, or Relics of an Ancient Peptide World?", *Journal of Structural Biology*,

Vol. 134, No. 2–3, pp. 191–203, 2001.

23. Dokholyan, N. V., B. Shakhnovich and E. I. Shakhnovich, "Expanding Protein Universe and Its Origin from the Biological Big Bang", *Proceedings of the National Academy of Sciences*, Vol. 99, No. 22, pp. 14132–14136, 2002.

24. Alva, V., M. Remmert, A. Biegert, A. N. Lupas and J. Söding, "A Galaxy of Folds", *Protein Science*, Vol. 19, No. 1, pp. 124–130, 2009.

25. Alva, V., J. Söding and A. N. Lupas, "A Vocabulary of Ancient Peptides at the Origin of Folded Proteins", *Elife*, Vol. 4, 2015.

26. Raanan, H., D. H. Pike, E. K. Moore, P. G. Falkowski and V. Nanda, "Modular Origins of Biological Electron Transfer Chains", *Proceedings of the National Academy of Sciences*, Vol. 115, No. 6, pp. 1280–1285, 2018.

27. Raanan, H., S. Poudel, D. H. Pike, V. Nanda and P. G. Falkowski, "Small Protein Folds at the Root of an Ancient Metabolic Network", *Proceedings of the National Academy of Sciences*, Vol. 117, No. 13, pp. 7193–7199, 2020.

28. Longo, L. M., J. Jabłońska, P. Vyas, M. Kanade, R. Kolodny, N. Ben-Tal and D. S. Tawfik, "On the Emergence of P-Loop NTPase and Rossmann Enzymes from a Beta-Alpha-Beta Ancestral Fragment", *eLife*, Vol. 9, 2020.

29. Kolodny, R., S. Nepomnyachiy, D. S. Tawfik and N. Ben-Tal, "Bridging Themes: Short Protein Segments Found in Different Architectures", *Molecular Biology and Evolution*, Vol. 38, No. 6, pp. 2191–2208, 2021.

30. Ben-Tal, N. and A. N. Lupas, "Editorial Overview: Sequences and Topology: Paths from Sequence to Structure", *Current Opinion in Structural Biology*, Vol. 68, pp. Vi–Viii, 2021.

31. Qiu, K., N. Ben-Tal and R. Kolodny, "Similar Protein Segments Shared between

Domains of Different Evolutionary Lineages", *Protein Science*, Vol. 31, No. 9, 2022.

32. Farías-Rico, J. A., S. Schmidt and B. Höcker, "Evolutionary Relationship of Two Ancient Protein Superfolds", *Nature Chemical Biology*, Vol. 10, No. 9, pp. 710–715, 2014.

33. Nepomnyachiy, S., N. Ben-Tal and R. Kolodny, "Complex Evolutionary Footprints Revealed in an Analysis of Reused Protein Segments of Diverse Lengths", *Proceedings of the National Academy of Sciences*, Vol. 114, No. 44, pp. 11703–11708, 2017.

34. Romero-Romero, S., S. Kordes, F. Michel and B. Höcker, "Evolution, Folding, and Design of TIM Barrels and Related Proteins", *Current Opinion in Structural Biology*, Vol. 68, pp. 94–104, 2021.

35. Söding, J., "Protein Homology Detection by HMM–HMM Comparison", *Bioinformatics*, Vol. 21, No. 7, pp. 951–960, 2005.

36. Narunsky, A., A. Kessel, R. Solan, V. Alva, R. Kolodny and N. Ben-Tal, "On the Evolution of Protein–Adenine Binding", *Proceedings of the National Academy of Sciences*, Vol. 117, No. 9, pp. 4701–4709, 2020.

37. Bromberg, Y., A. A. Aptekmann, Y. Mahlich, L. Cook, S. Senn, M. Miller, V. Nanda, D. U. Ferreiro and P. G. Falkowski, "Quantifying Structural Relationships of Metal-Binding Sites Suggests Origins of Biological Electron Transfer", *Science Advances*, Vol. 8, No. 2, 2022.

38. Zheng, Z., A. Goncearenco and I. N. Berezovsky, "Back in Time to the Gly-Rich Prototype of the Phosphate Binding Elementary Function", *Current Research in Structural Biology*, Vol. 7, p. 100142, 2024.

39. Alexandrov, V., U. Lehnert, N. Echols, D. Milburn, D. Engelman and M. Gerstein,

"Normal Modes for Predicting Protein Motions: A Comprehensive Database Assessment and Associated Web Tool", *Protein Science*, Vol. 14, No. 3, pp. 633–643, 2005.

40. Bahar, I. and A. Rader, "Coarse-Grained Normal Mode Analysis in Structural Biology", *Current Opinion in Structural Biology*, Vol. 15, No. 5, pp. 586–592, 2005.

41. Fuglebakk, E., S. P. Tiwari and N. Reuter, "Comparing the Intrinsic Dynamics of Multiple Protein Structures Using Elastic Network Models", *Biochimica et Biophysica Acta (BBA)-General Subjects*, Vol. 1850, No. 5, pp. 911–922, 2015.

42. Grudinin, S., E. Laine and A. Hoffmann, "Predicting Protein Functional Motions: An Old Recipe with a New Twist", *Biophysical Journal*, Vol. 118, No. 10, pp. 2513–2525, 2020.

43. Henzler-Wildman, K. A., M. Lei, V. Thai, S. J. Kerns, M. Karplus and D. Kern, "A Hierarchy of Timescales in Protein Dynamics Is Linked to Enzyme Catalysis", *Nature*, Vol. 450, No. 7171, pp. 913–916, 2007.

44. Mcclendon, C. L., A. P. Kornev, M. K. Gilson and S. S. Taylor, "Dynamic Architecture of a Protein Kinase", *Proceedings of the National Academy of Sciences*, Vol. 111, No. 43, pp. E4623–E4631, 2014.

45. Chopra, N., T. E. Wales, R. E. Joseph, S. E. Boyken, J. R. Engen, R. L. Jernigan and A. H. Andreotti, "Dynamic Allostery Mediated by a Conserved Tryptophan in the Tec Family Kinases", *Plos Computational Biology*, Vol. 12, No. 3, p. E1004826, 2016.

46. Tiwari, S. P. and N. Reuter, "Conservation of Intrinsic Dynamics in Proteins—What Have Computational Models Taught Us?", *Current Opinion in Structural Biology*, Vol. 50, pp. 75–81, 2018.

47. Süel, G. M., S. W. Lockless, M. A. Wall and R. Ranganathan, "Evolutionarily Conserved Networks of Residues Mediate Allosteric Communication in Proteins", *Nature Structural Biology*, Vol. 10, No. 1, pp. 59–69, 2003.

48. Granata, D., L. Ponzoni, C. Micheletti and V. Carnevale, "Patterns of Coevolving Amino Acids Unveil Structural and Dynamical Domains", *Proceedings of the National Academy of Sciences*, Vol. 114, No. 50, pp. E10612–E10621, 2017.

49. Maguid, S., S. Fernandez-Alberti, L. Ferrelli and J. Echave, "Exploring the Common Dynamics of Homologous Proteins. Application to the Globin Family", *Biophysical Journal*, Vol. 89, No. 1, pp. 3–13, 2005.

50. Bastolla, U., Y. Dehouck and J. Echave, "What Evolution Tells Us About Protein Physics, and Protein Physics Tells Us About Evolution", *Current Opinion in Structural Biology*, Vol. 42, pp. 59–66, 2017.

51. Campitelli, P., T. Modi, S. Kumar and S. B. Ozkan, "The Role of Conformational Dynamics and Allostery in Modulating Protein Evolution", *Annual Review of Biophysics*, Vol. 49, No. 1, pp. 267–288, 2020.

52. Bahar, I., A. R. Atilgan and B. Erman, "Direct Evaluation of Thermal Fluctuations in Proteins Using A Single-Parameter Harmonic Potential", *Folding and Design*, Vol. 2, No. 3, pp. 173–181, 1997.

53. Haliloglu, T., I. Bahar and B. Erman, "Gaussian Dynamics of Folded Proteins", *Physical Review Letters*, Vol. 79, No. 16, p. 3090, 1997.

54. Atilgan, A., S. Durell, R. Jernigan, M. Demirel, O. Keskin and I. Bahar, "Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model", *Biophysical Journal*, Vol. 80, No. 1, pp. 505–515, 2001.

55. Hacisuleyman, A. and B. Erman, "Causality, Transfer Entropy, and Allosteric Communication Landscapes in Proteins with Harmonic Interactions", *Proteins:*

Structure, Function, and Bioinformatics, Vol. 85, No. 6, pp. 1056–1064, 2017.

56. Altintel, B., B. Acar, B. Erman and T. Haliloglu, "Subsets of Slow Dynamic Modes Reveal Global Information Sources as Allosteric Sites", *Journal of Molecular Biology*, Vol. 434, No. 17, p. 167644, 2022.

57. Ersoy, A., B. Altintel, N. Livnat Levanon, N. Ben-Tal, T. Haliloglu and O. Lewinson, "Computational Analysis of Long-Range Allosteric Communications in CFTR", *Elife*, Vol. 12, p. Rp88659, 2023.

58. Duman, O., A. Kuznetsova, N. L. Levanon, M. Grupper, A. A. Ersoy, B. Acar, A. Kessel, N. Ben-Tal, O. Lewinson and T. Haliloglu, "Computational and Experimental Mapping of the Allosteric Network of Two Manganese ABC Transporters", *Protein Science*, Vol. 34, No. 2, p. E70039, 2025.

59. Kutlu, Y., G. Axel, R. Kolodny, N. Ben-Tal and T. Haliloglu, "Reused Protein Segments Linked to Functional Dynamics", *Molecular Biology and Evolution*, Vol. 41, No. 9, 2024.

60. Boehr, D. D., R. Nussinov and P. E. Wright, "The Role of Dynamic Conformational Ensembles in Biomolecular Recognition", *Nature Chemical Biology*, Vol. 5, No. 11, pp. 789–796, 2009.

61. Nussinov, R. and C.-J. Tsai, "Allostery in Disease and in Drug Discovery", *Cell*, Vol. 153, No. 2, pp. 293–305, 2013.

62. Bahar, I., T. R. Lezon, L.-W. Yang and E. Eyal, "Global Dynamics of Proteins: Bridging between Structure and Function", *Annual Review of Biophysics*, Vol. 39, pp. 23–42, 2010.

63. Cramer, P., "AlphaFold2 and the Future of Structural Biology", *Nature Structural and Molecular Biology*, Vol. 28, No. 9, pp. 704–705, 2021.

64. Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, "Highly Accurate Protein Structure Prediction with AlphaFold", *Nature*, Vol. 596, No. 7873, pp. 583–589, 2021.

65. Abramson, J., J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis and J. M. Jumper, "Accurate Structure Prediction of Biomolecular Interactions with AlphaFold3", *Nature*, Vol. 630, No. 8016, pp. 493–500, 2024.

66. Li, H., S. Sakuraba, A. Chandrasekaran and L.-W. Yang, "Molecular Binding Sites Are Located Near the Interface of Intrinsic Dynamics Domains (IDDs)", *Journal of Chemical Information and Modeling*, Vol. 54, No. 8, pp. 2275–2285, 2014.

67. Chandrasekaran, A., J. Chan, C. Lim and L.-W. Yang, "Protein Dynamics and Contact Topology Reveal Protein–Dna Binding Orientation", *Journal of Chemical Theory and Computation*, Vol. 12, No. 11, pp. 5269–5277, 2016.

68. Guclu, T. F., A. R. Atilgan and C. Atilgan, "Dynamic Community Composition Unravels Allosteric Communication in PDZ3", *The Journal of Physical Chemistry B*, Vol. 125, No. 9, pp. 2266–2276, 2021.

69. Zhang, Y., P. Doruker, B. Kaynak, S. Zhang, J. Krieger, H. Li and I. Bahar, "Intrinsic Dynamics Is Evolutionarily Optimized to Enable Allosteric Behavior", *Current Opinion in Structural Biology*, Vol. 62, pp. 14–21, 2020.

70. Hatley, M. E., S. W. Lockless, S. K. Gibson, A. G. Gilman and R. Ranganathan, "Allosteric Determinants in Guanine Nucleotide-Binding Proteins", *Proceedings of the National Academy of Sciences*, Vol. 100, No. 24, pp. 14445–14450, 2003.

71. Modi, T., P. Campitelli, I. C. Kazan and S. B. Ozkan, "Protein Folding Stability and Binding Interactions through the Lens of Evolution: A Dynamical Perspective", *Current Opinion in Structural Biology*, Vol. 66, pp. 207–215, 2021.

72. Emekli, U., D. Schneidman-Duhovny, H. J. Wolfson, R. Nussinov and T. Haliloglu, "Hingeprot: Automated Prediction of Hinges in Protein Structures", *Proteins: Structure, Function, and Bioinformatics*, Vol. 70, No. 4, pp. 1219–1227, 2008.

73. Butler, B. M., I. C. Kazan, A. Kumar and S. B. Ozkan, "Coevolving Residues Inform Protein Dynamics Profiles and Disease Susceptibility of nSNVs", *Plos Computational Biology*, Vol. 14, No. 11, p. E1006626, 2018.

74. Zhang, H. and L. Kurgan, "Sequence-Based Gaussian Network Model for Protein Dynamics", *Bioinformatics*, Vol. 30, No. 4, pp. 497–505, 2014.

75. Ose, N. J., B. M. Butler, A. Kumar, I. C. Kazan, M. Sanderford, S. Kumar and S. B. Ozkan, "Dynamic Coupling of Residues within Proteins as a Mechanistic Foundation of Many Enigmatic Pathogenic Missense Variants", *Plos Computational Biology*, Vol. 18, No. 4, p. E1010006, 2022.

76. Zen, A., C. Micheletti, O. Keskin and R. Nussinov, "Comparing Interfacial Dynamics in Protein-Protein Complexes: An Elastic Network Approach", *Bmc Structural Biology*, Vol. 10, pp. 1–13, 2010.

77. Demirtas, K., B. Erman and T. Haliloglu, "Dynamic Correlations: Exact and

Approximate Methods for Mutual Information", *Bioinformatics*, Vol. 40, No. 2, p. Btae076, 2024.

78. Chong, W. L., V. Vao-Soongnern, P. Nimmanpipug, C. Tayapiwatana, J.-H. Lin, Y.-L. Lin, H. Y. Chee, S. M. Zain, N. Abd Rahman and V. S. Lee, "Molecular Dynamics Simulations and Gaussian Network Model for Designing Antibody Mimicking Protein Towards Dengue Envelope Protein", *Journal of Molecular Liquids*, Vol. 346, p. 118086, 2022.

79. Feng, H., J. Y. Zhao and G.-W. Wei, "Multiscale Differential Geometry Learning for Protein Flexibility Analysis", *Journal of Computational Chemistry*, Vol. 46, No. 7, p. E70073, 2025.

80. Banerjee, A., S. Zhang and I. Bahar, "Genome Structural Dynamics: Insights from Gaussian Network Analysis of Hi-C Data", *Briefings in Functional Genomics*, Vol. 23, No. 5, pp. 525–537, 2024.

81. Doruker, P., A. R. Atilgan and I. Bahar, "Dynamics of Proteins Predicted by Molecular Dynamics Simulations and Analytical Approaches: Application to α-Amylase Inhibitor", *Proteins: Structure, Function, and Bioinformatics*, Vol. 40, No. 3, pp. 512–524, 2000.

82. Eyal, E., G. Lum and I. Bahar, "The Anisotropic Network Model Web Server at 2015 (ANM 2.0)", *Bioinformatics*, Vol. 31, No. 9, pp. 1487–1489, 2015.

83. Lam, J. H., A. Nakano and V. Katritch, "Scalable Computation of Anisotropic Vibrations for Large Macromolecular Assemblies", *Nature Communications*, Vol. 15, No. 1, 2024.

84. Wolff, A. M., E. Nango, I. D. Young, A. S. Brewster, M. Kubo, T. Nomura, M. Sugahara, S. Owada, B. A. Barad, K. Ito, A. Bhowmick, S. Carbajo, T. Hino, J. M. Holton, D. Im, L. J. O'Riordan, T. Tanaka, R. Tanaka, R. G. Sierra,

F. Yumoto, K. Tono, S. Iwata, N. K. Sauter, J. S. Fraser and M. C. Thompson, "Mapping Protein Dynamics at High Spatial Resolution with Temperature-Jump X-Ray Crystallography", *Nature Chemistry*, Vol. 15, No. 11, pp. 1549–1558, 2023.

85. Ventura, C., A. Banerjee, M. Zacharopoulou, L. S. Itzhaki and I. Bahar, "Tandem-Repeat Proteins Conformational Mechanics Are Optimized to Facilitate Functional Interactions and Complexations", *Current Opinion in Structural Biology*, Vol. 84, p. 102744, 2024.

86. Karplus, M. and G. A. Petsko, "Molecular Dynamics Simulations in Biology", *Nature*, Vol. 347, No. 6294, pp. 631–639, 1990.

87. Adcock, S. A. and J. A. McCammon, "Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins", *Chemical Reviews*, Vol. 106, No. 5, pp. 1589–1615, 2006.

88. McCammon, J. A., B. R. Gelin and M. Karplus, "Dynamics of Folded Proteins", *Nature*, Vol. 267, No. 5612, pp. 585–590, 1977.

89. Vanommeslaeghe, K., O. Guvench and A. D. Mackerell, "Molecular Mechanics", *Current Pharmaceutical Design*, Vol. 20, No. 20, pp. 3281–3292, 2014.

90. Bernardi, R. C., M. C. Melo and K. Schulten, "Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems", *Biochimica et Biophysica Acta (BBA) - General Subjects*, Vol. 1850, No. 5, pp. 872–877, 2015.

91. Shrimpton-Phoenix, E., E. Notari, T. Kluonis and C. W. Wood, "DrMD: Molecular Dynamics for Experimentalists", *Journal of Molecular Biology*, Vol. 436, No. 24, p. 168918, 2024.

92. Roy, A., E. Ward, I. Choi, M. Cosi, T. Edgin, T. S. Hughes, M. S. Islam, A. M. Khan, A. Kolekar, M. Rayl, I. Robinson, P. Sarando, E. Skidmore, T. L. Swetnam, M. Wall, Z. Xu, M. L. Yung, N. Merchant and T. J. Wheeler, "MDRepo —

An Open Data Warehouse for Community-Contributed Molecular Dynamics Simulations of Proteins", *Nucleic Acids Research*, Vol. 53, No. D1, pp. D477–D486, 2024.

93. Persson, L. J., C. Sahin, M. Landreh and E. G. Marklund, "High-Performance Molecular Dynamics Simulations for Native Mass Spectrometry of Large Protein Complexes with the Fast Multipole Method", *Analytical Chemistry*, Vol. 96, No. 37, pp. 15023–15030, 2024.

94. Wang, T., X. He, M. Li, Y. Li, R. Bi, Y. Wang, C. Cheng, X. Shen, J. Meng, H. Zhang, H. Liu, Z. Wang, S. Li, B. Shao and T.-Y. Liu, "Ab Initio Characterization of Protein Molecular Dynamics with AI^{2}BMD", *Nature*, Vol. 635, No. 8040, pp. 1019–1027, 2024.

95. Roe, D. R., C. Bergonzo and T. E. Cheatham, "Evaluation of Enhanced Sampling Provided by Accelerated Molecular Dynamics with Hamiltonian Replica Exchange Methods", *The Journal of Physical Chemistry B*, Vol. 118, No. 13, pp. 3543–3552, 2014.

96. Marsili, S., G. F. Signorini, R. Chelli, M. Marchi and P. Procacci, "ORAC: A Molecular Dynamics Simulation Program to Explore Free Energy Surfaces in Biomolecular Systems at the Atomistic Level", *Journal of Computational Chemistry*, Vol. 31, No. 5, pp. 1106–1116, 2009.

97. Yang, M., N. Livnat Levanon, B. Acar, B. Aykac Fas, G. Masrati, J. Rose, N. Ben-Tal, T. Haliloglu, Y. Zhao and O. Lewinson, "Single-Molecule Probing of the Conformational Homogeneity of the ABC Transporter BtuCD", *Nature Chemical Biology*, Vol. 14, No. 7, pp. 715–722, 2018.

98. Acar, B., J. Rose, B. A. Fas, N. Ben-Tal, O. Lewinson and T. Haliloglu, "Distinct Allosteric Networks Underlie Mechanistic Speciation of ABC Transporters", *Structure*, Vol. 28, No. 6, pp. 651–663, 2020.

99. Bonati, L., G. Piccini and M. Parrinello, "Deep Learning the Slow Modes for Rare Events Sampling", *Proceedings of the National Academy of Sciences*, Vol. 118, No. 44, 2021.

100. Mcgibbon, R. T., B. E. Husic and V. S. Pande, "Identification of Simple Reaction Coordinates from Complex Dynamics", *The Journal of Chemical Physics*, Vol. 146, No. 4, 2017.

101. Glielmo, A., B. E. Husic, A. Rodriguez, C. Clementi, F. Noé and A. Laio, "Unsupervised Learning Methods for Molecular Simulation Data", *Chemical Reviews*, Vol. 121, No. 16, pp. 9722–9758, 2021.

102. Bonati, L., Y.-Y. Zhang and M. Parrinello, "Neural Networks-Based Variationally Enhanced Sampling", *Proceedings of the National Academy of Sciences*, Vol. 116, No. 36, pp. 17641–17647, 2019.

103. Bussi, G. and G. A. Tribello, "Analyzing and Biasing Simulations with PLUMED", *Biomolecular Simulations: Methods and Protocols*, Vol. 2, pp. 529–578, 2019.

104. Bussi, G. and A. Laio, "Using Metadynamics to Explore Complex Free-Energy Landscapes", *Nature Reviews Physics*, Vol. 2, No. 4, pp. 200–212, 2020.

105. Gkeka, P., G. Stoltz, A. Barati Farimani, Z. Belkacemi, M. Ceriotti, J. D. Chodera, A. R. Dinner, A. L. Ferguson, J.-B. Maillet, H. Minoux, C. Peter, F. Pietrucci, A. Silveira, A. Tkatchenko, Z. Trstanova, R. Wiewiora and T. Lelièvre, "Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems", *Journal of Chemical Theory and Computation*, Vol. 16, No. 8, pp. 4757–4775, 2020.

106. Akdel, M., D. E. V. Pires, E. P. Pardo, J. Jänes, A. O. Zalevsky, B. Mészáros, P. Bryant, L. L. Good, R. A. Laskowski, G. Pozzati, A. Shenoy, W. Zhu, P. Kun-

drotas, V. R. Serra, C. H. M. Rodrigues, A. S. Dunham, D. Burke, N. Borkakoti, S. Velankar, A. Frost, J. Basquin, K. Lindorff-Larsen, A. Bateman, A. V. Kajava, A. Valencia, S. Ovchinnikov, J. Durairaj, D. B. Ascher, J. M. Thornton, N. E. Davey, A. Stein, A. Elofsson, T. I. Croll and P. Beltrao, "A Structural Biology Community Assessment of AlphaFold2 Applications", *Nature Structural and Molecular Biology*, Vol. 29, No. 11, pp. 1056–1067, 2022.

107. Yang, Z., X. Zeng, Y. Zhao and R. Chen, "AlphaFold2 and Its Applications in the Fields of Biology and Medicine", *Signal Transduction and Targeted Therapy*, Vol. 8, No. 1, 2023.

108. Homma, F., J. Huang and R. A. L. Van Der Hoorn, "AlphaFold-Multimer Predicts Cross-Kingdom Interactions at the Plant-Pathogen Interface", *Nature Communications*, Vol. 14, No. 1, 2023.

109. Söding, J. and A. N. Lupas, "More than the Sum of Their Parts: On the Evolution of Proteins from Peptides", *Bioessays*, Vol. 25, No. 9, pp. 837–846, 2003.

110. Ferruz, N., F. Lobos, D. Lemm, S. Toledo-Patino, J. A. Farías-Rico, S. Schmidt and B. Höcker, "Identification and Analysis of Natural Building Blocks for Evolution-Guided Fragment-Based Protein Design", *Journal of Molecular Biology*, Vol. 432, No. 13, pp. 3898–3914, 2020.

111. Vogel, C., M. Bashton, N. D. Kerrison, C. Chothia and S. A. Teichmann, "Structure, Function and Evolution of Multidomain Proteins", *Current Opinion in Structural Biology*, Vol. 14, No. 2, pp. 208–216, 2004.

112. Ezerzer, Y., M. Frenkel-Pinter, R. Kolodny and N. Ben-Tal, "A Building Blocks Perspective on Protein Emergence and Evolution", *Current Opinion in Structural Biology*, Vol. 91, p. 102996, 2025.

113. Longo, L. M., R. Kolodny and S. E. Mcglynn, "Evidence for the Emergence of β-

Trefoils by 'Peptide Budding' from an Igg-Like β-Sandwich", *Plos Computational Biology*, Vol. 18, No. 2, p. E1009833, 2022.

114. Alvarez-Carreño, C., R. J. Gupta, A. S. Petrov and L. D. Williams, "Creative Destruction: New Protein Folds from Old", *Proceedings of the National Academy of Sciences*, Vol. 119, No. 52, 2022.

115. Kolodny, R., "The Themes Database of Reuse in Proteins", 2017, `https://trachel-srv.cs.haifa.ac.il/rachel/ppi/themes/`, accessed on May 30, 2022.

116. Axel, G. and Y. Kutlu, "Reused Protein Segments Linked to Functional Dynamics - Supplementary Data", 2024, `https://gabiaxel.github.io/themes-dynamics/`, accessed on November 15, 2024.

117. Brüschweiler, R., "Collective Protein Dynamics and Nuclear Spin Relaxation", *The Journal of Chemical Physics*, Vol. 102, No. 8, pp. 3396–3403, 1995.

118. Vinh, N. X., J. Epps and J. Bailey, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance.", *Journal of Machine Learning Research*, Vol. 11, No. 2, pp. 2837–2854, 2010.

119. Romano, S., J. Bailey, V. Nguyen and K. Verspoor, "Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance", *International Conference on Machine Learning*, Vol. 32, pp. 1143–1151, Pmlr, Bejing, China, 2014.

120. Fisher, R. A., "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P", *Journal of the Royal Statistical Society*, Vol. 85, No. 1, p. 87, 1922.

121. Dunn, O. J., "Multiple Comparisons Among Means", *Journal of the American Statistical Association*, Vol. 56, No. 293, pp. 52–64, 1961.

122. Benjamini, Y. and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Vol. 57, No. 1, pp. 289–300, 1995.

123. Amadei, A., A. B. M. Linssen and H. J. C. Berendsen, "Essential Dynamics of Proteins", *Proteins: Structure, Function, and Bioinformatics*, Vol. 17, No. 4, pp. 412–425, 1993.

124. Chu, N., T. Viennet, H. Bae, A. Salguero, A. Boeszoermenyi, H. Arthanari and P. A. Cole, "The Structural Determinants of PH Domain-Mediated Regulation of Akt Revealed by Segmental Labeling", *eLife*, Vol. 9, 2020.

125. Nowak, A. J., C. Alfieri, C. U. Stirnimann, V. Rybin, F. Baudin, N. Ly-Hartig, D. Lindner and C. W. MÜLler, "Chromatin-Modifying Complex Component Nurf55/P55 Associates with Histones H3 and H4 and Polycomb Repressive Complex 2 Subunit Su(Z)12 through Partially Overlapping Binding Sites", *Journal of Biological Chemistry*, Vol. 286, No. 26, pp. 23388–23396, 2011.

126. Ma, B., C.-J. Tsai, T. Haliloğlu and R. Nussinov, "Dynamic Allostery: Linkers Are Not Merely Flexible", *Structure*, Vol. 19, No. 7, pp. 907–917, 2011.

127. Liu, Y. and I. Bahar, "Sequence Evolution Correlates with Structural Dynamics", *Molecular Biology and Evolution*, Vol. 29, No. 9, pp. 2253–2263, 2012.

128. Berezovsky, I. N., E. Guarnera, Z. Zheng, B. Eisenhaber and F. Eisenhaber, "Protein Function Machinery: From Basic Structural Units to Modulation of Activity", *Current Opinion in Structural Biology*, Vol. 42, pp. 67–74, 2017.

129. Mishra, S. K. and R. L. Jernigan, "Protein Dynamic Communities from Elastic Network Models Align Closely to the Communities Defined by Molecular Dynam-

ics", *Plos One*, Vol. 13, No. 6, p. E0199225, 2018.

130. Saavedra, H. G., J. O. Wrabl, J. A. Anderson, J. Li and V. J. Hilser, "Dynamic Allostery Can Drive Cold Adaptation in Enzymes", *Nature*, Vol. 558, No. 7709, pp. 324–328, 2018.

131. Kutlu, Y., N. Ben-Tal and T. Haliloglu, "Global Dynamics Renders Protein Sites with High Functional Response", *The Journal of Physical Chemistry B*, Vol. 125, No. 18, pp. 4734–4745, 2021.

132. Tang, Q.-Y. and K. Kaneko, "Dynamics-Evolution Correspondence in Protein Structures", *Physical Review Letters*, Vol. 127, No. 9, 2021.

133. Micheletti, C., P. Carloni and A. Maritan, "Accurate and Efficient Description of Protein Vibrational Dynamics: Comparing Molecular Dynamics and Gaussian Models", *Proteins: Structure, Function, and Bioinformatics*, Vol. 55, No. 3, pp. 635–645, 2004.

134. Wang, Y., A. Rader, I. Bahar and R. L. Jernigan, "Global Ribosome Motions Revealed with Elastic Network Model", *Journal of Structural Biology*, Vol. 147, No. 3, pp. 302–314, 2004.

135. Togashi, Y. and H. Flechsig, "Coarse-Grained Protein Dynamics Studies Using Elastic Network Models", *International Journal of Molecular Sciences*, Vol. 19, No. 12, p. 3899, 2018.

136. Rodríguez-Espigares, I., M. Torrens-Fontanals, J. K. S. Tiemann, D. Aranda-García, J. M. Ramírez-Anguita, T. M. Stepniewski, N. Worp, A. Varela-Rial, A. Morales-Pastor, B. Medel-Lacruz, G. Pándy-Szekeres, E. Mayol, T. Giorgino, J. Carlsson, X. Deupi, S. Filipek, M. Filizola, J. C. Gómez-Tamayo, A. Gonzalez, H. Gutiérrez-De-Terán, M. Jiménez-Rosés, W. Jespers, J. Kapla, G. Khelashvili, P. Kolb, D. Latek, M. Marti-Solano, P. Matricon, M.-T. Matsoukas,

P. Miszta, M. Olivella, L. Perez-Benito, D. Provasi, S. Ríos, I. R. Torrecillas, J. Sallander, A. Sztyler, S. Vasile, H. Weinstein, U. Zachariae, P. W. Hildebrand, G. De Fabritiis, F. Sanz, D. E. Gloriam, A. Cordomi, R. Guixà-González and J. Selent, "GPCRmd Uncovers the Dynamics of the 3D-GPCRome", *Nature Methods*, Vol. 17, No. 8, pp. 777–787, 2020.

137. Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate and M. Punta, "Pfam: The Protein Families Database", *Nucleic Acids Research*, Vol. 42, No. D1, pp. D222–D230, 2013.

138. Höcker, B., "Design of Proteins from Smaller Fragments—Learning from Evolution", *Current Opinion in Structural Biology*, Vol. 27, pp. 56–62, 2014.

139. Jacobs, T. M., B. Williams, T. Williams, X. Xu, A. Eletsky, J. F. Federizon, T. Szyperski and B. Kuhlman, "Design of Structurally Distinct Proteins Using Strategies Inspired by Evolution", *Science*, Vol. 352, No. 6286, pp. 687–690, 2016.

140. Laronde-Leblanc, N., T. Guszczynski, T. Copeland and A. Wlodawer, "Autophosphorylation of Archaeoglobus Fulgidus Rio2 and Crystal Structures of Its Nucleotide–Metal Ion Complexes", *The FEBS Journal*, Vol. 272, No. 11, pp. 2800–2810, 2005.

141. Ferreira-Cerca, S., V. Sagar, T. Schäfer, M. Diop, A.-M. Wesseling, H. Lu, E. Chai, E. Hurt and N. Laronde-Leblanc, "ATPase-Dependent Role of the Atypical Kinase Rio2 on the Evolving Pre-40S Ribosomal Subunit", *Nature Structural and Molecular Biology*, Vol. 19, No. 12, pp. 1316–1323, 2012.

142. Maurice, F., N. Pérébaskine, S. Thore and S. Fribourg, "In Vitro Dimerization of Human Rio2 Kinase", *RNA Biology*, Vol. 16, No. 11, pp. 1633–1642, 2019.

143. Mao, D. Y., D. Neculai, M. Downey, S. Orlicky, Y. Z. Haffani, D. F. Ceccarelli,

J. S. Ho, R. K. Szilard, W. Zhang, C. S. Ho, L. Wan, C. Fares, S. Rumpel, I. Kurinov, C. H. Arrowsmith, D. Durocher and F. Sicheri, "Atomic Structure of the Keops Complex: An Ancient Protein Kinase-Containing Molecular Machine", *Molecular Cell*, Vol. 32, No. 2, pp. 259–275, 2008.

144. Zhang, W., B. Collinet, M. Graille, M.-C. Daugeron, N. Lazar, D. Libri, D. Durand and H. Van Tilbeurgh, "Crystal Structures of the Gon7/Pcc1 and Bud32/Cgi121 Complexes Provide A Model for the Complete Yeast KEOPS Complex", *Nucleic Acids Research*, Vol. 43, No. 6, pp. 3358–3372, 2015.

145. Zhang, C., X. Zhang, L. Freddolino and Y. Zhang, "BioLiP2: An Updated Structure Database for Biologically Relevant Ligand–Protein Interactions", *Nucleic Acids Research*, Vol. 52, No. D1, pp. D404–D412, 2023.

146. Kutlu, Y., "PhD Thesis Data", 2025, `https://github.com/yigitkutlu/PhD-Thesis-Data`, accessed on May 15, 2025.

147. Beckstein, O., E. J. Denning, J. R. Perilla and T. B. Woolf, "Zipping and Unzipping of Adenylate Kinase: Atomistic Insights Into the Ensemble of Open $\Leftrightarrow$ Closed Transitions", *Journal of Molecular Biology*, Vol. 394, No. 1, pp. 160–176, 2009.

148. Allen, R. J., D. Frenkel and P. R. Ten Wolde, "Forward Flux Sampling-Type Schemes for Simulating Rare Events: Efficiency Analysis", *The Journal of Chemical Physics*, Vol. 124, No. 19, 2006.

APPENDIX A: Complete List of Binding Themes

Table A.1. Complete list of binding themes.

PDB ID	Ligand	Pair Theme 1	Pair Theme 2	PDB ID	Ligand	Pair Theme 1	Pair Theme 2
e2xyiA1	peptide	14940	14950	e4c1hA1	ZN	2826	2820
e2xyiA1	peptide	14940	14956	e4c1hA1	ZN	2826	2836
e2xyiA1	peptide	14940	15222	e4c1hA1	ZN	2826	2862
e2xyiA1	peptide	14940	14815	e4c1hA1	ZN	2826	2821
e2xyiA1	peptide	14940	14951	e4c1hA1	ZN	2826	2822
e2xyiA1	peptide	14945	14950	e4c1hA1	ZN	2826	2823
e2xyiA1	peptide	14945	14956	e4c1hA1	ZN	2826	2838
e2xyiA1	peptide	14945	15222	e4c1hA1	ZN	2828	2823
e2xyiA1	peptide	14945	14815	e4c1hA1	ZN	2828	2838
e2xyiA1	peptide	14945	14951	e4c1hA1	ZN	2895	2823
e3emhA1	peptide	14813	14986	e4c1hA1	ZN	2895	2838
e3emhA1	peptide	14813	14879	e4c1hA1	ZN	2820	2823
e3emhA1	peptide	14813	14957	e4c1hA1	ZN	2820	2838
e3emhA1	peptide	14813	14813	e4c1hA1	ZN	2828	single
e3emhA1	peptide	14813	14958	e4c1hA1	ZN	2921	2895
e3emhA1	peptide	14813	14815	e4c1hA1	ZN	2921	2820
e3emhA1	peptide	14813	14883	e4c1hA1	ZN	2921	2836
e3emhA1	peptide	14813	14884	e4c1hA1	ZN	2921	2862

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e3emhA1	peptide	14815	14957	e4c1hA1	ZN	2921	2821
e3emhA1	peptide	14815	14813	e4c1hA1	ZN	2921	2822
e3emhA1	peptide	14815	14958	e4c1hA1	ZN	2893	2895
e3emhA1	peptide	14815	14815	e4c1hA1	ZN	2893	2820
e3emhA1	peptide	14815	14883	e4c1hA1	ZN	2893	2836
e3emhA1	peptide	14815	14884	e4c1hA1	ZN	2893	2862
e3emhA1	peptide	14862	14957	e4c1hA1	ZN	2893	2821
e3emhA1	peptide	14862	14813	e4c1hA1	ZN	2893	2822
e3emhA1	peptide	14862	14958	e4c1hA1	ZN	2894	2895
e3emhA1	peptide	14862	14815	e4c1hA1	ZN	2894	2820
e3emhA1	peptide	14862	14883	e4c1hA1	ZN	2894	2836
e3emhA1	peptide	14862	14884	e4c1hA1	ZN	2894	2862
e3emhA1	peptide	14986	14815	e4c1hA1	ZN	2894	2821
e3emhA1	peptide	14986	14883	e4c1hA1	ZN	2894	2822
e3emhA1	peptide	14986	14884	e4c1hA1	ZN	2839	2895
e3emhA1	peptide	14879	14957	e4c1hA1	ZN	2839	2820
e3emhA1	peptide	14879	14813	e4c1hA1	ZN	2839	2836
e3emhA1	peptide	14879	14958	e4c1hA1	ZN	2839	2862
e3emhA1	peptide	14879	14815	e4c1hA1	ZN	2839	2821
e3emhA1	peptide	14879	14883	e4c1hA1	ZN	2839	2822
e3emhA1	peptide	14879	14884	e4c1hA1	ZN	2826	2895

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e3ignA1	C2E	10435	single	e4c1hA1	ZN	2826	2820
e3ignA1	C2E	10456	10447	e4c1hA1	ZN	2826	2836
e3ignA1	C2E	10445	10436	e4c1hA1	ZN	2826	2862
e3ignA1	C2E	10445	10447	e4c1hA1	ZN	2826	2821
e3ignA1	C2E	10456	single	e4c1hA1	ZN	2826	2822
e3ignA1	C2E	10439	single	e1m2xA1	ZN	2828	single
e3ignA1	C2E	10460	single	e1m2xA1	ZN	2893	2842
e3ignA1	C2E	10435	single	e1m2xA1	ZN	2893	2895
e3ignA1	C2E	10445	10436	e1m2xA1	ZN	2893	2820
e4urgA1	C2E	10435	single	e1m2xA1	ZN	2893	2836
e4urgA1	C2E	10456	10447	e1m2xA1	ZN	2893	2862
e4urgA1	C2E	10445	10436	e1m2xA1	ZN	2893	2821
e4urgA1	C2E	10445	10447	e1m2xA1	ZN	2893	2822
e4urgA1	C2E	10456	single	e1m2xA1	ZN	2894	2895
e4urgA1	C2E	10439	single	e1m2xA1	ZN	2894	2820
e4urgA1	C2E	10460	single	e1m2xA1	ZN	2894	2836
e4urgA1	C2E	10435	single	e1m2xA1	ZN	2894	2862
e4urgA1	C2E	10445	10436	e1m2xA1	ZN	2894	2821
e1o04A1	NAD	7084	7041	e1m2xA1	ZN	2894	2822
e1o04A1	NAD	7084	7152	e1m2xA1	ZN	2839	2820
e1o04A1	NAD	7084	7153	e1m2xA1	ZN	2839	2836

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1o04A1	NAD	7084	7021	e1m2xA1	ZN	2839	2862
e1o04A1	NAD	7084	7100	e1m2xA1	ZN	2839	2821
e1o04A1	NAD	7084	7099	e1m2xA1	ZN	2839	2822
e1o04A1	NAD	7084	7101	e1m2xA1	ZN	2826	2820
e1o04A1	NAD	7084	7045	e1m2xA1	ZN	2826	2836
e1o04A1	NAD	7084	7022	e1m2xA1	ZN	2826	2862
e1o04A1	NAD	7084	7137	e1m2xA1	ZN	2826	2821
e1o04A1	NAD	7084	7110	e1m2xA1	ZN	2826	2822
e1o04A1	NAD	7084	7111	e1m2xA1	ZN	2826	2895
e1o04A1	NAD	7084	7091	e1m2xA1	ZN	2842	2820
e1o04A1	NAD	7084	7092	e1m2xA1	ZN	2842	2836
e1o04A1	NAD	7085	7153	e1m2xA1	ZN	2842	2862
e1o04A1	NAD	7085	7100	e1m2xA1	ZN	2842	2821
e1o04A1	NAD	7085	7099	e1m2xA1	ZN	2842	2822
e1o04A1	NAD	7085	7101	e1m2xA1	ZN	2893	2842
e1o04A1	NAD	7085	7045	e1m2xA1	ZN	2893	2895
e1o04A1	NAD	7085	7022	e1m2xA1	ZN	2893	2820
e1o04A1	NAD	7085	7137	e1m2xA1	ZN	2893	2836
e1o04A1	NAD	7085	7110	e1m2xA1	ZN	2893	2862
e1o04A1	NAD	7085	7111	e1m2xA1	ZN	2893	2821
e1o04A1	NAD	7085	7091	e1m2xA1	ZN	2893	2822

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1o04A1	NAD	7085	7092	e1m2xA1	ZN	2893	2838
e1o04A1	NAD	7108	7153	e1m2xA1	ZN	2893	2823
e1o04A1	NAD	7108	7100	e1m2xA1	ZN	2894	2895
e1o04A1	NAD	7108	7099	e1m2xA1	ZN	2894	2820
e1o04A1	NAD	7108	7101	e1m2xA1	ZN	2894	2836
e1o04A1	NAD	7108	7045	e1m2xA1	ZN	2894	2862
e1o04A1	NAD	7108	7022	e1m2xA1	ZN	2894	2821
e1o04A1	NAD	7108	7137	e1m2xA1	ZN	2894	2822
e1o04A1	NAD	7108	7110	e1m2xA1	ZN	2894	2838
e1o04A1	NAD	7108	7111	e1m2xA1	ZN	2894	2823
e1o04A1	NAD	7108	7091	e1m2xA1	ZN	2839	2820
e1o04A1	NAD	7108	7092	e1m2xA1	ZN	2839	2836
e1o04A1	NAD	7088	7153	e1m2xA1	ZN	2839	2862
e1o04A1	NAD	7088	7021	e1m2xA1	ZN	2839	2821
e1o04A1	NAD	7088	7100	e1m2xA1	ZN	2839	2822
e1o04A1	NAD	7088	7099	e1m2xA1	ZN	2839	2838
e1o04A1	NAD	7088	7101	e1m2xA1	ZN	2839	2823
e1o04A1	NAD	7088	7045	e1m2xA1	ZN	2826	2820
e1o04A1	NAD	7088	7022	e1m2xA1	ZN	2826	2836
e1o04A1	NAD	7088	7137	e1m2xA1	ZN	2826	2862
e1o04A1	NAD	7088	7110	e1m2xA1	ZN	2826	2821

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1o04A1	NAD	7088	7111	e1m2xA1	ZN	2826	2822
e1o04A1	NAD	7088	7091	e1m2xA1	ZN	2826	2838
e1o04A1	NAD	7088	7092	e1m2xA1	ZN	2826	2823
e1o04A1	NAD	7089	7045	e1m2xA1	ZN	2828	2838
e1o04A1	NAD	7089	7022	e1m2xA1	ZN	2828	2823
e1o04A1	NAD	7089	7137	e1m2xA1	ZN	2826	2895
e1o04A1	NAD	7089	7110	e1m2xA1	ZN	2842	2820
e1o04A1	NAD	7089	7111	e1m2xA1	ZN	2842	2836
e1o04A1	NAD	7089	7091	e1m2xA1	ZN	2842	2862
e1o04A1	NAD	7089	7092	e1m2xA1	ZN	2842	2821
e1o04A1	NAD	7040	7153	e1m2xA1	ZN	2842	2822
e1o04A1	NAD	7040	7021	e1m2xA1	ZN	2842	2838
e1o04A1	NAD	7040	7100	e1m2xA1	ZN	2842	2823
e1o04A1	NAD	7040	7099	e1m2xA1	ZN	2895	2838
e1o04A1	NAD	7040	7101	e1m2xA1	ZN	2895	2823
e1o04A1	NAD	7040	7045	e1m2xA1	ZN	2820	2838
e1o04A1	NAD	7040	7022	e1m2xA1	ZN	2820	2823
e1o04A1	NAD	7040	7137	e1m2xA1	ZN	2862	single
e1o04A1	NAD	7040	7110	e1m2xA1	ZN	2822	single
e1o04A1	NAD	7040	7111	e1m2xA1	ZN	2823	single
e1o04A1	NAD	7040	7091	e1m2xA1	MCO	2828	single

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1o04A1	NAD	7040	7092	e1m2xA1	MCO	2893	2842
e1o04A1	NAD	7020	7100	e1m2xA1	MCO	2893	2895
e1o04A1	NAD	7020	7099	e1m2xA1	MCO	2893	2836
e1o04A1	NAD	7020	7101	e1m2xA1	MCO	2893	2862
e1o04A1	NAD	7020	7045	e1m2xA1	MCO	2893	2821
e1o04A1	NAD	7020	7022	e1m2xA1	MCO	2893	2822
e1o04A1	NAD	7020	7137	e1m2xA1	MCO	2894	2895
e1o04A1	NAD	7020	7110	e1m2xA1	MCO	2894	2836
e1o04A1	NAD	7020	7111	e1m2xA1	MCO	2894	2862
e1o04A1	NAD	7020	7091	e1m2xA1	MCO	2894	2821
e1o04A1	NAD	7020	7092	e1m2xA1	MCO	2894	2822
e1o04A1	NAD	7019	7152	e1m2xA1	MCO	2839	2836
e1o04A1	NAD	7019	7153	e1m2xA1	MCO	2839	2862
e1o04A1	NAD	7019	7021	e1m2xA1	MCO	2839	2821
e1o04A1	NAD	7019	7100	e1m2xA1	MCO	2839	2822
e1o04A1	NAD	7019	7099	e1m2xA1	MCO	2826	2836
e1o04A1	NAD	7019	7101	e1m2xA1	MCO	2826	2862
e1o04A1	NAD	7019	7045	e1m2xA1	MCO	2826	2821
e1o04A1	NAD	7019	7022	e1m2xA1	MCO	2826	2822
e1o04A1	NAD	7019	7137	e1m2xA1	MCO	2826	2895
e1o04A1	NAD	7019	7110	e1m2xA1	MCO	2842	2836

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1o04A1	NAD	7019	7111	e1m2xA1	MCO	2842	2862
e1o04A1	NAD	7019	7091	e1m2xA1	MCO	2842	2821
e1o04A1	NAD	7019	7092	e1m2xA1	MCO	2842	2822
e1o04A1	NAD	7041	7153	e1a7tA1	ZN	2828	single
e1o04A1	NAD	7041	7021	e1a7tA1	ZN	2921	2895
e1o04A1	NAD	7041	7100	e1a7tA1	ZN	2921	2820
e1o04A1	NAD	7041	7099	e1a7tA1	ZN	2921	2836
e1o04A1	NAD	7041	7101	e1a7tA1	ZN	2921	2862
e1o04A1	NAD	7041	7045	e1a7tA1	ZN	2921	2821
e1o04A1	NAD	7041	7022	e1a7tA1	ZN	2921	2822
e1o04A1	NAD	7041	7137	e1a7tA1	ZN	2894	2895
e1o04A1	NAD	7041	7110	e1a7tA1	ZN	2894	2820
e1o04A1	NAD	7041	7111	e1a7tA1	ZN	2894	2836
e1o04A1	NAD	7041	7091	e1a7tA1	ZN	2894	2862
e1o04A1	NAD	7041	7092	e1a7tA1	ZN	2894	2821
e1o04A1	NAD	7151	7045	e1a7tA1	ZN	2894	2822
e1o04A1	NAD	7151	7022	e1a7tA1	ZN	2839	2895
e1o04A1	NAD	7151	7137	e1a7tA1	ZN	2839	2820
e1o04A1	NAD	7151	7110	e1a7tA1	ZN	2839	2836
e1o04A1	NAD	7151	7111	e1a7tA1	ZN	2839	2862
e1o04A1	NAD	7151	7091	e1a7tA1	ZN	2839	2821

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1o04A1	NAD	7151	7092	e1a7tA1	ZN	2839	2822
e1o04A1	NAD	7152	7022	e1a7tA1	ZN	2893	2895
e1o04A1	NAD	7152	7092	e1a7tA1	ZN	2893	2820
e1o04A1	NAD	7021	7022	e1a7tA1	ZN	2893	2836
e1o04A1	NAD	7021	7091	e1a7tA1	ZN	2893	2862
e1o04A1	NAD	7021	7092	e1a7tA1	ZN	2893	2821
e1o04A1	NAD	7100	7022	e1a7tA1	ZN	2893	2822
e1o04A1	NAD	7100	7091	e1a7tA1	ZN	2826	2895
e1o04A1	NAD	7100	7092	e1a7tA1	ZN	2826	2820
e4i3vA1	NAD	7088	7021	e1a7tA1	ZN	2826	2836
e4i3vA1	NAD	7088	7100	e1a7tA1	ZN	2826	2862
e4i3vA1	NAD	7088	7099	e1a7tA1	ZN	2826	2821
e4i3vA1	NAD	7088	7101	e1a7tA1	ZN	2826	2822
e4i3vA1	NAD	7088	7022	e1a7tA1	ZN	2842	2820
e4i3vA1	NAD	7019	7021	e1a7tA1	ZN	2842	2836
e4i3vA1	NAD	7019	7100	e1a7tA1	ZN	2842	2862
e4i3vA1	NAD	7019	7099	e1a7tA1	ZN	2842	2821
e4i3vA1	NAD	7019	7101	e1a7tA1	ZN	2842	2822
e4i3vA1	NAD	7019	7022	e1a7tA1	ZN	2921	2895
e4i3vA1	NAD	7020	7100	e1a7tA1	ZN	2921	2820
e4i3vA1	NAD	7020	7099	e1a7tA1	ZN	2921	2836

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e4i3vA1	NAD	7020	7101	e1a7tA1	ZN	2921	2862
e4i3vA1	NAD	7020	7022	e1a7tA1	ZN	2921	2821
e4i3vA1	POA	7088	7022	e1a7tA1	ZN	2921	2822
e4i3vA1	POA	7088	7024	e1a7tA1	ZN	2921	2823
e4i3vA1	POA	7088	7025	e1a7tA1	ZN	2921	2838
e4i3vA1	POA	7088	7102	e1a7tA1	ZN	2894	2895
e4i3vA1	POA	7088	7103	e1a7tA1	ZN	2894	2820
e4i3vA1	POA	7088	7027	e1a7tA1	ZN	2894	2836
e4i3vA1	POA	7088	7026	e1a7tA1	ZN	2894	2862
e4i3vA1	POA	7019	7022	e1a7tA1	ZN	2894	2821
e4i3vA1	POA	7019	7024	e1a7tA1	ZN	2894	2822
e4i3vA1	POA	7019	7025	e1a7tA1	ZN	2894	2823
e4i3vA1	POA	7019	7102	e1a7tA1	ZN	2894	2838
e4i3vA1	POA	7019	7103	e1a7tA1	ZN	2839	2895
e4i3vA1	POA	7019	7027	e1a7tA1	ZN	2839	2820
e4i3vA1	POA	7019	7026	e1a7tA1	ZN	2839	2836
e4i3vA1	POA	7020	7022	e1a7tA1	ZN	2839	2862
e4i3vA1	POA	7020	7024	e1a7tA1	ZN	2839	2821
e4i3vA1	POA	7020	7025	e1a7tA1	ZN	2839	2822
e4i3vA1	POA	7020	7102	e1a7tA1	ZN	2839	2823
e4i3vA1	POA	7020	7103	e1a7tA1	ZN	2839	2838

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e4i3vA1	POA	7020	7027	e1a7tA1	ZN	2893	2895
e4i3vA1	POA	7020	7026	e1a7tA1	ZN	2893	2820
e1uxtA1	G1P	7107	7149	e1a7tA1	ZN	2893	2836
e1uxtA1	G1P	7107	7150	e1a7tA1	ZN	2893	2862
e1uxtA1	G1P	7107	7088	e1a7tA1	ZN	2893	2821
e1uxtA1	G1P	7107	7019	e1a7tA1	ZN	2893	2822
e1uxtA1	G1P	7107	7020	e1a7tA1	ZN	2893	2823
e1uxtA1	G1P	7106	7069	e1a7tA1	ZN	2893	2838
e1uxtA1	G1P	7106	7149	e1a7tA1	ZN	2826	2895
e1uxtA1	G1P	7106	7150	e1a7tA1	ZN	2826	2820
e1uxtA1	G1P	7106	7088	e1a7tA1	ZN	2826	2836
e1uxtA1	G1P	7106	7019	e1a7tA1	ZN	2826	2862
e1uxtA1	G1P	7106	7020	e1a7tA1	ZN	2826	2821
e1uxtA1	G1P	7069	7149	e1a7tA1	ZN	2826	2822
e1uxtA1	G1P	7069	7150	e1a7tA1	ZN	2826	2823
e1uxtA1	G1P	7069	7088	e1a7tA1	ZN	2826	2838
e1uxtA1	G1P	7069	7019	e1a7tA1	ZN	2828	2823
e1uxtA1	G1P	7069	7020	e1a7tA1	ZN	2828	2838
e1uxtA1	NAD	7150	7153	e1a7tA1	ZN	2842	2820
e1uxtA1	NAD	7150	7100	e1a7tA1	ZN	2842	2836
e1uxtA1	NAD	7150	7099	e1a7tA1	ZN	2842	2862

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1uxtA1	NAD	7150	7101	e1a7tA1	ZN	2842	2821
e1uxtA1	NAD	7150	7058	e1a7tA1	ZN	2842	2822
e1uxtA1	NAD	7150	7059	e1a7tA1	ZN	2842	2823
e1uxtA1	NAD	7150	7022	e1a7tA1	ZN	2842	2838
e1uxtA1	NAD	7150	7111	e1a7tA1	ZN	2895	2823
e1uxtA1	NAD	7150	7091	e1a7tA1	ZN	2895	2838
e1uxtA1	NAD	7088	7153	e1a7tA1	ZN	2820	2823
e1uxtA1	NAD	7088	7100	e1a7tA1	ZN	2820	2838
e1uxtA1	NAD	7088	7099	e4c1dA1	X8Z	2921	2862
e1uxtA1	NAD	7088	7101	e4c1dA1	X8Z	2921	2821
e1uxtA1	NAD	7088	7058	e4c1dA1	X8Z	2921	2822
e1uxtA1	NAD	7088	7059	e4c1dA1	X8Z	2825	2826
e1uxtA1	NAD	7088	7022	e4c1dA1	X8Z	2825	2862
e1uxtA1	NAD	7088	7111	e4c1dA1	X8Z	2825	2821
e1uxtA1	NAD	7088	7091	e4c1dA1	X8Z	2825	2822
e1uxtA1	NAD	7019	7153	e4c1dA1	X8Z	2893	2862
e1uxtA1	NAD	7019	7100	e4c1dA1	X8Z	2893	2821
e1uxtA1	NAD	7019	7099	e4c1dA1	X8Z	2893	2822
e1uxtA1	NAD	7019	7101	e4c1dA1	X8Z	2894	2862
e1uxtA1	NAD	7019	7058	e4c1dA1	X8Z	2894	2821
e1uxtA1	NAD	7019	7059	e4c1dA1	X8Z	2894	2822

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1uxtA1	NAD	7019	7022	e4c1dA1	X8Z	2839	2862
e1uxtA1	NAD	7019	7111	e4c1dA1	X8Z	2839	2821
e1uxtA1	NAD	7019	7091	e4c1dA1	X8Z	2839	2822
e1uxtA1	NAD	7020	7100	e4c1dA1	X8Z	2826	2862
e1uxtA1	NAD	7020	7099	e4c1dA1	X8Z	2826	2821
e1uxtA1	NAD	7020	7101	e4c1dA1	X8Z	2826	2822
e1uxtA1	NAD	7020	7058	e4c1dA1	ZN	2921	2862
e1uxtA1	NAD	7020	7059	e4c1dA1	ZN	2921	2820
e1uxtA1	NAD	7020	7022	e4c1dA1	ZN	2921	2821
e1uxtA1	NAD	7020	7111	e4c1dA1	ZN	2921	2822
e1uxtA1	NAD	7020	7091	e4c1dA1	ZN	2893	2862
e1uxtA1	NAD	7041	7153	e4c1dA1	ZN	2893	2820
e1uxtA1	NAD	7041	7100	e4c1dA1	ZN	2893	2821
e1uxtA1	NAD	7041	7099	e4c1dA1	ZN	2893	2822
e1uxtA1	NAD	7041	7101	e4c1dA1	ZN	2894	2862
e1uxtA1	NAD	7041	7058	e4c1dA1	ZN	2894	2820
e1uxtA1	NAD	7041	7059	e4c1dA1	ZN	2894	2821
e1uxtA1	NAD	7041	7022	e4c1dA1	ZN	2894	2822
e1uxtA1	NAD	7041	7111	e4c1dA1	ZN	2839	2862
e1uxtA1	NAD	7041	7091	e4c1dA1	ZN	2839	2820
e3v30A2	peptide	13330	13333	e4c1dA1	ZN	2839	2821

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e3v30A2	peptide	13330	13334	e4c1dA1	ZN	2839	2822
e3v30A2	peptide	13330	13328	e4c1dA1	ZN	2826	2862
e3v30A2	peptide	13330	13327	e4c1dA1	ZN	2826	2820
e3v30A2	peptide	13330	13336	e4c1dA1	ZN	2826	2821
e3v30A2	peptide	13330	13323	e4c1dA1	ZN	2826	2822
e3v30A2	peptide	13330	13357	e4c1dA1	ZN	2921	2862
e3v30A2	peptide	13330	13324	e4c1dA1	ZN	2921	2820
e3v30A2	peptide	13344	13333	e4c1dA1	ZN	2921	2821
e3v30A2	peptide	13344	13334	e4c1dA1	ZN	2921	2822
e3v30A2	peptide	13344	13328	e4c1dA1	ZN	2921	2823
e3v30A2	peptide	13344	13327	e4c1dA1	ZN	2893	2862
e3v30A2	peptide	13344	13336	e4c1dA1	ZN	2893	2820
e3v30A2	peptide	13344	13323	e4c1dA1	ZN	2893	2821
e3v30A2	peptide	13344	13357	e4c1dA1	ZN	2893	2822
e3v30A2	peptide	13344	13324	e4c1dA1	ZN	2893	2823
e3v30A2	peptide	13356	13333	e4c1dA1	ZN	2894	2862
e3v30A2	peptide	13356	13336	e4c1dA1	ZN	2894	2820
e3v30A2	peptide	13356	13323	e4c1dA1	ZN	2894	2821
e3v30A2	peptide	13356	13357	e4c1dA1	ZN	2894	2822
e3v30A2	peptide	13356	13324	e4c1dA1	ZN	2894	2823
e3v30A2	peptide	13368	13333	e4c1dA1	ZN	2839	2862

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e3v30A2	peptide	13368	13334	e4c1dA1	ZN	2839	2820
e3v30A2	peptide	13368	13328	e4c1dA1	ZN	2839	2821
e3v30A2	peptide	13368	13327	e4c1dA1	ZN	2839	2822
e3v30A2	peptide	13368	13336	e4c1dA1	ZN	2839	2823
e3v30A2	peptide	13368	13323	e4c1dA1	ZN	2826	2862
e3v30A2	peptide	13368	13357	e4c1dA1	ZN	2826	2820
e3v30A2	peptide	13368	13324	e4c1dA1	ZN	2826	2821
e3v30A2	peptide	13340	13336	e4c1dA1	ZN	2826	2822
e3v30A2	peptide	13340	13323	e4c1dA1	ZN	2826	2823
e3v30A2	peptide	13340	13357	e4c1dA1	ZN	2820	2823
e3v30A2	peptide	13340	13324	e2bcqA3	dna	9365	single
e3v30A2	peptide	13334	13323	e2bcqA3	dna	9362	single
e3v30A2	peptide	13334	13357	e2bcqA3	dna	9363	single
e3v30A2	peptide	13334	13324	e2bcqA3	dna	9357	single
e3v30A2	peptide	13328	13357	e2bcqA3	dna	9370	single
e3v30A2	peptide	13328	13324	e2bcqA3	dna	9365	single
e1elrA1	peptide	13386	single	e2bcqA3	dna	9361	9362
e1elrA1	peptide	13394	13394	e2bcqA3	dna	9361	9363
e3ro3A1	peptide	13386	single	e2bcqA3	dna	9361	9357
e3ro3A1	peptide	13390	single	e2bcqA3	dna	9360	9357
e3ro3A1	peptide	13643	13394	e2bcqA3	MG	9358	single

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e3ro3A1	peptide	13643	13396	e2bcqA3	MG	9371	single
e3ro3A1	peptide	13643	13644	e4kliA3	dna	9365	single
e3ro3A1	peptide	13689	13394	e4kliA3	dna	9362	single
e3ro3A1	peptide	13689	13396	e4kliA3	dna	9361	9363
e3ro3A1	peptide	13689	13644	e4kliA3	dna	9361	9370
e3ro3A1	peptide	13396	13644	e4kliA3	dna	9361	9357
e3ro3A1	peptide	13396	13394	e4kliA3	dna	9360	9363
e3ro3A1	peptide	13396	13396	e4kliA3	dna	9360	9370
e1elwA1	peptide	13390	single	e4kliA3	dna	9360	9357
e1elwA1	peptide	13386	13440	e4kliA3	dna	9365	single
e1elwA1	peptide	13386	13391	e4kliA3	dna	9362	single
e1elwA1	peptide	13386	13702	e4kliA3	MG	9358	single
e1elwA1	peptide	13386	13396	e4kliA3	MG	9371	single
e1elwA1	peptide	13394	13391	e4kliA3	PPV	9358	single
e1elwA1	peptide	13394	13396	e4kliA3	PPV	9371	single
e4psxB1	peptide	14945	15221	e3q25A1	GLC	6533	6535
e3j80g1	rna	14987	15011	e3q25A1	GLC	6533	6824
e3j80g1	rna	14987	15167	e3q25A1	GLC	6533	6673
e3j80g1	rna	14987	14815	e3q25A1	GLC	6533	6878
e3j80g1	rna	14987	15090	e3q25A1	GLC	6875	6535
e3j80g1	rna	14987	15010	e3q25A1	GLC	6875	6824

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e3j80g1	rna	14987	14991	e3q25A1	GLC	6875	6673
e3j80g1	rna	14990	15090	e3q25A1	GLC	6875	6878
e3j80g1	rna	14990	15010	e3q25A1	GLC	6637	6535
e3j80g1	rna	14990	14991	e3q25A1	GLC	6637	6824
e3j80g1	rna	14812	15090	e3q25A1	GLC	6637	6673
e3j80g1	rna	14812	15010	e3q25A1	GLC	6637	6878
e3j80g1	rna	14812	14991	e3q25A1	GLC	6533	6824
e3j80g1	rna	15011	15090	e3q25A1	GLC	6533	6871
e3j80g1	rna	15011	15010	e3q25A1	GLC	6533	6530
e3j80g1	rna	15011	14991	e3q25A1	GLC	6533	6638
e3j80g1	rna	15167	15090	e3q25A1	GLC	6533	6639
e3j80g1	rna	15167	15010	e3q25A1	GLC	6533	6872
e3j80g1	rna	15167	14991	e3q25A1	GLC	6875	6824
e3j80g1	rna	14815	15090	e3q25A1	GLC	6875	6871
e3j80g1	rna	14815	15010	e3q25A1	GLC	6875	6530
e3j80g1	rna	14815	14991	e3q25A1	GLC	6875	6638
e1p22A1	peptide	14917	14879	e3q25A1	GLC	6875	6639
e1p22A1	peptide	14917	14860	e3q25A1	GLC	6875	6872
e1p22A1	peptide	14917	14918	e3e13X1	FE	6616	6613
e1p22A1	peptide	14917	15255	e3e13X1	FE	6616	6614
e1p22A1	peptide	14917	14884	e3e13X1	FE	6616	6718

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1p22A1	peptide	14926	14860	e3e13X1	FE	6616	6615
e1p22A1	peptide	14926	14918	e3e13X1	FE	6616	6717
e1p22A1	peptide	14926	15255	e3e13X1	FE	6616	6617
e1p22A1	peptide	14926	14884	e3e13X1	FE	6616	6722
e1p22A1	peptide	14862	14860	e3e13X1	FE	6616	6721
e1p22A1	peptide	14862	14918	e3f11A1	FE	6616	6613
e1p22A1	peptide	14862	15255	e3f11A1	FE	6616	6614
e1p22A1	peptide	14862	14884	e3f11A1	FE	6616	6718
e1p22A1	peptide	14879	14860	e3f11A1	FE	6616	6615
e1p22A1	peptide	14879	14918	e3f11A1	FE	6616	6717
e1p22A1	peptide	14879	15255	e3f11A1	FE	6616	6617
e1p22A1	peptide	14879	14884	e3f11A1	FE	6616	6721
e2r25B1	MG	5391	single	e3f11A1	FE	6616	6722
e2r25B1	MG	5408	single	e3f11A1	FE	6613	6721
e2r25B1	MG	5407	single	e3f11A1	FE	6614	6721
e2r25B1	MG	5397	5451	e3f11A1	FE	6615	6721
e2r25B1	BEF	5407	single	e3osrA1	GLC	6533	6824
e2r25B1	BEF	5391	5488	e3osrA1	GLC	6533	6530
e2r25B1	BEF	5391	5400	e3osrA1	GLC	6533	6639
e2r25B1	BEF	5408	5488	e3osrA1	GLC	6533	6872
e2r25B1	BEF	5408	5400	e3osrA1	GLC	6875	6824

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e2r25B1	BEF	5451	5400	e3osrA1	GLC	6875	6530
e3ozyA4	MG	1217	single	e3osrA1	GLC	6875	6639
e3ozyA4	MG	9624	single	e3osrA1	GLC	6875	6872
e3ozyA4	MG	9614	single	e3osrA1	GLC	6637	6824
e3ozyA4	MG	1216	1219	e3osrA1	GLC	6637	6530
e2j13A1	ZN	1673	single	e3osrA1	GLC	6637	6639
e2j13A1	ZN	1668	single	e3osrA1	GLC	6637	6872
e2j13A1	ZN	1685	single	e3osrA1	GLC	6533	6824
e2j13A1	ZN	1675	single	e3osrA1	GLC	6533	6871
e2j13A1	ZN	1674	single	e3osrA1	GLC	6533	6530
e2j13A1	ZN	1671	1669	e3osrA1	GLC	6533	6638
e2j13A1	ZN	1671	1670	e3osrA1	GLC	6533	6639
e2j13A1	ZN	1673	1669	e3osrA1	GLC	6533	6872
e2j13A1	ZN	1673	1670	e3osrA1	GLC	6875	6824
e2j13A1	ZN	1668	1669	e3osrA1	GLC	6875	6871
e2j13A1	ZN	1668	1670	e3osrA1	GLC	6875	6530
e1y2kA1	ZN	12905	single	e3osrA1	GLC	6875	6638
e1y2kA1	ZN	12906	single	e3osrA1	GLC	6875	6639
e1y2kA1	ZN	12932	single	e3osrA1	GLC	6875	6872
e1y2kA1	ZN	12916	12881	e3osrA1	GLC	6642	6824
e1y2kA1	ZN	12916	12882	e3osrA1	GLC	6642	6871

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1y2kA1	ZN	12916	12883	e3osrA1	GLC	6642	6530
e1y2kA1	ZN	12916	12884	e3osrA1	GLC	6642	6638
e1y2kA1	7DE	12904	single	e3osrA1	GLC	6642	6639
e1y2kA1	7DE	12908	single	e3osrA1	GLC	6642	6872
e1y2kA1	7DE	12891	single	e3g7wA1	GLC	6533	6824
e1y2kA1	7DE	12915	single	e3g7wA1	GLC	6533	6530
e1y2kA1	7DE	12893	single	e3g7wA1	GLC	6533	6872
e4qgeA1	35O	12908	single	e3g7wA1	GLC	6875	6824
e4qgeA1	35O	12914	12929	e3g7wA1	GLC	6875	6530
e4qgeA1	35O	12914	12910	e3g7wA1	GLC	6875	6872
e4qgeA1	35O	12914	12894	e3g7wA1	GLC	6637	6824
e4qgeA1	35O	12954	12894	e3g7wA1	GLC	6637	6530
e4qgeA1	35O	12891	12910	e3g7wA1	GLC	6637	6872
e4qgeA1	35O	12891	12894	e3g7wA1	GLC	6533	6824
e4qgeA1	35O	12915	12910	e3g7wA1	GLC	6533	6530
e4qgeA1	35O	12915	12894	e3g7wA1	GLC	6533	6871
e4qgeA1	35O	12909	12894	e3g7wA1	GLC	6533	6638
e4qgeA1	ZN	12906	single	e3g7wA1	GLC	6533	6639
e4qgeA1	ZN	12932	single	e3g7wA1	GLC	6533	6872
e4qgeA1	ZN	12916	12917	e3g7wA1	GLC	6875	6824
e1zklA1	ZN	12905	single	e3g7wA1	GLC	6875	6530

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1zklA1	ZN	12906	single	e3g7wA1	GLC	6875	6871
e1zklA1	ZN	12916	12881	e3g7wA1	GLC	6875	6638
e1zklA1	ZN	12916	12882	e3g7wA1	GLC	6875	6639
e1zklA1	ZN	12916	12883	e3g7wA1	GLC	6875	6872
e1zklA1	ZN	12916	12884	e3g7wA1	GLC	6533	6824
e1zklA1	IBM	12916	12908	e3g7wA1	GLC	6533	6530
e1zklA1	IBM	12916	12892	e3g7wA1	GLC	6533	6871
e1zklA1	IBM	12916	12893	e3g7wA1	GLC	6533	6638
e1zklA1	IBM	12916	12910	e3g7wA1	GLC	6533	6639
e1zklA1	IBM	12916	12894	e3g7wA1	GLC	6533	6872
e1zklA1	IBM	12905	12908	e3g7wA1	GLC	6533	6532
e1zklA1	IBM	12905	12892	e3g7wA1	GLC	6533	12238
e1zklA1	IBM	12905	12893	e3g7wA1	GLC	6533	6877
e1zklA1	IBM	12905	12910	e3g7wA1	GLC	6875	6824
e1zklA1	IBM	12905	12894	e3g7wA1	GLC	6875	6530
e1zklA1	IBM	12906	12908	e3g7wA1	GLC	6875	6871
e1zklA1	IBM	12906	12892	e3g7wA1	GLC	6875	6638
e1zklA1	IBM	12906	12893	e3g7wA1	GLC	6875	6639
e1zklA1	IBM	12906	12910	e3g7wA1	GLC	6875	6872
e1zklA1	IBM	12906	12894	e3g7wA1	GLC	6875	6532
e1zklA1	IBM	12904	12910	e3g7wA1	GLC	6875	12238

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1zklA1	IBM	12904	12894	e3g7wA1	GLC	6875	6877
e1zklA1	IBM	12914	12910	e3g7wA1	GLC	6824	6532
e1zklA1	IBM	12914	12894	e3g7wA1	GLC	6824	12238
e1zklA1	IBM	12891	12910	e3g7wA1	GLC	6824	6877
e1zklA1	IBM	12891	12894	e3g7wA1	GLC	6530	6532
e1zklA1	IBM	12915	12910	e3g7wA1	GLC	6530	12238
e1zklA1	IBM	12915	12894	e3g7wA1	GLC	6530	6877
e1zklA1	IBM	12909	12894	e3g7wA1	GLC	6871	6532
e1zklA1	IBM	12892	12894	e3g7wA1	GLC	6871	12238
e3ecmA1	ZN	12905	single	e3g7wA1	GLC	6871	6877
e3ecmA1	ZN	12906	single	e3g7wA1	GLC	6638	6532
e3ecmA1	ZN	12916	12881	e3g7wA1	GLC	6638	12238
e3ecmA1	ZN	12916	12882	e3g7wA1	GLC	6638	6877
e3ecmA1	ZN	12916	12883	e3g7wA1	GLC	6639	6532
e3ecmA1	ZN	12916	12884	e3g7wA1	GLC	6639	12238
e3ecmA1	ZN	12897	12881	e3g7wA1	GLC	6639	6877
e3ecmA1	ZN	12897	12882	e3g7wA1	GLC	6872	6532
e3ecmA1	ZN	12897	12883	e3g7wA1	GLC	6872	12238
e3ecmA1	ZN	12897	12884	e3g7wA1	GLC	6872	6877
e1so2A1	MG	12905	single	e2vd3A1	MG	6656	single
e1so2A1	MG	12906	single	e2vd3A1	MG	6545	single

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1so2A1	MG	12897	12917	e2vd3A1	MG	6546	single
e1so2A1	MG	12897	12881	e2vd3A1	MG	6592	single
e1so2A1	MG	12897	12882	e2vd3A1	MG	6536	single
e1so2A1	MG	12897	12883	e2vd3A1	MG	6651	single
e1so2A1	MG	12897	12884	e2vd3A1	MG	6714	single
e1so2A1	HG9	12908	single	e2vd3A1	MG	6537	single
e1so2A1	HG9	12904	12910	e2vd3A1	MG	6540	single
e1so2A1	HG9	12914	12893	e2vd3A1	MG	6541	single
e1so2A1	HG9	12914	12910	e4qvhB1	GLC	6875	6871
e1so2A1	HG9	12891	12910	e4qvhB1	GLC	6533	6871
e1so2A1	HG9	12915	12910	e1nh8A2	AMP	6546	single
e1so2A1	HG9	12892	12910	e1nh8A2	AMP	6714	single
e1so2A1	666	12908	single	e1nh8A2	AMP	6537	single
e1so2A1	666	12904	12910	e1nh8A2	AMP	6536	6538
e1so2A1	666	12904	12894	e1h3dA2	AMP	6546	single
e1so2A1	666	12914	12893	e1h3dA2	AMP	6651	single
e1so2A1	666	12914	12910	e1h3dA2	AMP	6536	single
e1so2A1	666	12914	12894	e1h3dA2	AMP	6714	single
e1so2A1	666	12891	12910	e1h3dA2	AMP	6537	single
e1so2A1	666	12891	12894	e1h3dA2	AMP	6536	6538
e1so2A1	666	12915	12910	e4pf4A1	MG	1829	1826

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1so2A1	666	12915	12894	e4pf4A1	MG	1829	1916
e1so2A1	666	12892	12910	e4pf4A1	MG	1829	1973
e1so2A1	666	12892	12894	e4pf4A1	MG	1837	1826
e4k5uB1	GAL	15301	15301	e4pf4A1	MG	1837	1916
e4k5uB1	GAL	15301	15330	e4pf4A1	MG	1837	1973
e4k5uB1	GAL	15301	15881	e3m7nH2	rna	9545	9526
e4k5uB1	GAL	15301	15809	e3m7nH2	rna	9545	9528
e4k5uB1	GAL	15301	15347	e3m7nH2	rna	9530	9536
e4k5uB1	GAL	15301	15875	e3m7nH2	rna	9530	9547
e4k5uB1	GAL	15301	15471	e3m7nH2	rna	9530	9537
e4k5uB1	GAL	15301	15890	e3m7nH2	rna	9530	9526
e4k5uB1	GAL	15301	15420	e3m7nH2	rna	9530	9528
e4k5uB1	GAL	15914	15301	e3m7nH2	rna	9535	9526
e4k5uB1	GAL	15914	15330	e3m7nH2	rna	9535	9528
e4k5uB1	GAL	15914	15881	e3m7nH2	rna	9536	9526
e4k5uB1	GAL	15914	15809	e3m7nH2	rna	9536	9528
e4k5uB1	GAL	15914	15347	e3m7nH2	rna	9547	9528
e4k5uB1	GAL	15914	15875	e3bwhA1	MAN	10077	single
e4k5uB1	GAL	15914	15471	e3bwhA1	MAN	10091	single
e4k5uB1	GAL	15914	15890	e3bwhA1	MAN	10092	single
e4k5uB1	GAL	15914	15420	e2ae2A1	NAP	2939	2962

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e4k5uB1	GAL	15680	15471	e2ae2A1	NAP	2939	2963
e4k5uB1	GAL	15680	15890	e2ae2A1	NAP	2939	2964
e4k5uB1	GAL	15680	15301	e2ae2A1	NAP	2939	2965
e4k5uB1	GAL	15680	15420	e2ae2A1	PTO	2962	single
e4k5uB1	GAL	15723	15471	e2ae2A1	PTO	2969	2963
e4k5uB1	GAL	15723	15890	e2ae2A1	PTO	2969	2964
e4k5uB1	GAL	15723	15420	e2ae2A1	PTO	2969	2965
e4k5uB1	GAL	15723	15301	e2ae2A1	PTO	2960	2963
e4k5uB1	GAL	15586	15301	e2ae2A1	PTO	2960	2964
e4k5uB1	GAL	15586	15330	e2ae2A1	PTO	2960	2965
e4k5uB1	GAL	15586	15881	e2ae2A1	PTO	2962	2965
e4k5uB1	GAL	15586	15809	e2wsbA1	NAD	2939	2962
e4k5uB1	GAL	15586	15347	e2wsbA1	NAD	2939	2963
e4k5uB1	GAL	15586	15875	e2wsbA1	NAD	2939	3395
e4k5uB1	GAL	15586	15471	e2wsbA1	NAD	2939	3313
e4k5uB1	GAL	15586	15890	e2wsbA1	NAD	2939	2964
e4k5uB1	GAL	15586	15420	e2wsbA1	NAD	2939	2965
e4k5uB1	GAL	15351	15471	e2wsbA1	NAD	2939	2972
e4k5uB1	GAL	15351	15301	e2wsbA1	NAD	2962	2965
e4k5uB1	GAL	15351	15890	e2wsbA1	NAD	3395	2965
e4k5uB1	GAL	15351	15420	e2wsbA1	NAD	3313	2965

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e4k5uB1	GAL	15434	15301	e2wsbA1	POL	2962	single
e4k5uB1	GAL	15434	15471	e2wsbA1	POL	2963	single
e4k5uB1	GAL	15434	15890	e2wsbA1	POL	3395	single
e4k5uB1	GAL	15434	15420	e2wsbA1	POL	3313	single
e3rg1E1	MAN	15518	single	e2wsbA1	POL	3396	single
e3rg1E1	MAN	15509	15593	e2wsbA1	POL	2964	single
e3rg1E1	MAN	15509	15519	e2wsbA1	POL	2961	2965
e3rg1E1	MAN	15595	15519	e2wsbA1	POL	2961	2964
e3rg1A6	MAN	15518	single	e2wsbA1	POL	2961	2972
e3rg1A6	MAN	15509	15593	e2wsbA1	POL	2962	single
e3rg1A6	MAN	15509	15519	e2wsbA1	POL	2963	single
e3rg1A6	MAN	15595	15519	e2wsbA1	POL	2974	2963
e4lscA1	FUC	15609	single	e2wsbA1	POL	2974	3395
e4lscA1	FUC	15631	single	e2wsbA1	POL	2974	3313
e4lscA1	FUC	15663	single	e2wsbA1	POL	2974	3396
e4lscA1	FUC	15818	single	e2wsbA1	POL	2974	2964
e4mn8B1	peptide	15325	single	e2wsbA1	POL	2974	2965
e3b2dA4	MAN	15518	single	e2wsbA1	POL	2974	2972
e3b2dA4	MAN	15509	15593	e2wsbA1	POL	2960	3395
e3b2dA4	MAN	15509	15519	e2wsbA1	POL	2960	3313
e3b2dA4	MAN	15595	15519	e2wsbA1	POL	2960	3396

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e1dzkA1	PRZ	14284	14280	e2wsbA1	POL	2960	2964
e1dzkA1	PRZ	14281	14313	e2wsbA1	POL	2960	2965
e1dzkA1	PRZ	14281	14285	e2wsbA1	POL	2960	2972
e1dzkA1	PRZ	14281	14280	e2wsbA1	POL	2961	2965
e1dzkA1	PRZ	14282	14313	e2wsbA1	POL	2961	2964
e1dzkA1	PRZ	14282	14285	e2wsbA1	POL	2961	2972
e1dzkA1	PRZ	14282	14280	e4qedA1	NAP	2939	2968
e1dzkA1	PRZ	14283	14313	e4qedA1	NAP	2939	2967
e1dzkA1	PRZ	14283	14285	e4qedA1	NAP	2939	3380
e1dzkA1	PRZ	14283	14280	e4qedA1	NAP	2939	2962
e1dzkA1	PRZ	14313	14280	e4qedA1	NAP	2939	3485
e3kffA1	XBT	14284	single	e4qedA1	NAP	2968	2962
e3kffA1	XBT	14286	single	e4qedA1	NAP	2968	3485
e3kffA1	XBT	14283	14313	e4qedA1	NAP	2967	2962
e3kffA1	XBT	14283	14285	e4qedA1	NAP	2967	3485
e3kffA1	ZBT	14283	single	e4qedA1	NAP	3380	3485
e3kffA1	ZBT	14284	single	e4qedA1	NAP	3380	2962
e3kffA1	ZBT	14283	14313	e2o23A1	NAD	2939	2962
e3kffA1	ZBT	14282	14286	e2o23A1	NAD	2939	2963
e3kffA1	ZBT	14282	14313	e2o23A1	NAD	2939	2964
e2yntB1	ZN	2828	single	e2o23A1	NAD	2939	2965

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e2yntB1	ZN	2921	2820	e2o23A1	NAD	2967	2962
e2yntB1	ZN	2921	2836	e2o23A1	NAD	2967	2963
e2yntB1	ZN	2921	2862	e2o23A1	NAD	2967	2964
e2yntB1	ZN	2921	2821	e2o23A1	NAD	2967	2965
e2yntB1	ZN	2921	2822	e2o23A1	NAD	2968	2962
e2yntB1	ZN	2893	2842	e2o23A1	NAD	2968	2963
e2yntB1	ZN	2893	2895	e2o23A1	NAD	2968	2964
e2yntB1	ZN	2893	2820	e2o23A1	NAD	2968	2965
e2yntB1	ZN	2893	2836	e4fn4A1	NAD	2939	2967
e2yntB1	ZN	2893	2862	e4fn4A1	NAD	2939	2968
e2yntB1	ZN	2893	2821	e4fn4A1	NAD	2939	2962
e2yntB1	ZN	2893	2822	e4fn4A1	NAD	2939	2964
e2yntB1	ZN	2894	2895	e4fn4A1	NAD	2967	2962
e2yntB1	ZN	2894	2820	e4fn4A1	NAD	2967	2964
e2yntB1	ZN	2894	2836	e4fn4A1	NAD	2968	2962
e2yntB1	ZN	2894	2862	e4fn4A1	NAD	2968	2964
e2yntB1	ZN	2894	2821	e1fmcA1	CHO	2962	single
e2yntB1	ZN	2894	2822	e1fmcA1	CHO	2969	2963
e2yntB1	ZN	2839	2820	e1fmcA1	CHO	2969	2965
e2yntB1	ZN	2839	2836	e1fmcA1	CHO	2961	2965
e2yntB1	ZN	2839	2862	e1fmcA1	NAI	2939	2968

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e2yntB1	ZN	2839	2821	e1fmcA1	NAI	2939	2967
e2yntB1	ZN	2839	2822	e1fmcA1	NAI	2939	2962
e2yntB1	ZN	2826	2820	e1fmcA1	NAI	2939	2963
e2yntB1	ZN	2826	2836	e1fmcA1	NAI	2939	2964
e2yntB1	ZN	2826	2862	e1fmcA1	NAI	2939	2965
e2yntB1	ZN	2826	2821	e1fmcA1	NAI	2968	2962
e2yntB1	ZN	2826	2822	e1fmcA1	NAI	2968	2963
e2yntB1	ZN	2842	2820	e1fmcA1	NAI	2968	2964
e2yntB1	ZN	2842	2836	e1fmcA1	NAI	2968	2965
e2yntB1	ZN	2842	2862	e1fmcA1	NAI	2967	2962
e2yntB1	ZN	2842	2821	e1fmcA1	NAI	2967	2963
e2yntB1	ZN	2842	2822	e1fmcA1	NAI	2967	2964
e2yntB1	ZN	2921	2820	e1fmcA1	NAI	2967	2965
e2yntB1	ZN	2921	2836	e1h5qA1	NAP	2939	2967
e2yntB1	ZN	2921	2862	e1h5qA1	NAP	2939	2968
e2yntB1	ZN	2921	2821	e1h5qA1	NAP	2939	2962
e2yntB1	ZN	2921	2822	e1h5qA1	NAP	2939	2963
e2yntB1	ZN	2921	2823	e1h5qA1	NAP	2967	2962
e2yntB1	ZN	2921	2838	e1h5qA1	NAP	2967	2963
e2yntB1	ZN	2893	2842	e1h5qA1	NAP	2968	2962
e2yntB1	ZN	2893	2895	e1h5qA1	NAP	2968	2963

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e2yntB1	ZN	2893	2820	e1h5qA1	NAP	2969	2963
e2yntB1	ZN	2893	2836	e1h5qA1	NAP	2974	2963
e2yntB1	ZN	2893	2862	e2zatA1	NAP	2939	2962
e2yntB1	ZN	2893	2821	e2zatA1	NAP	2939	2963
e2yntB1	ZN	2893	2822	e2zatA1	NAP	2939	2964
e2yntB1	ZN	2893	2823	e2zatA1	NAP	2939	2965
e2yntB1	ZN	2893	2838	e2zatA1	NAP	2962	2965
e2yntB1	ZN	2894	2895	e2ag5A1	SO4	3485	single
e2yntB1	ZN	2894	2820	e2ag5A1	SO4	2961	2965
e2yntB1	ZN	2894	2836	e2ag5A1	SO4	2961	2966
e2yntB1	ZN	2894	2862	e2ag5A1	SO4	2962	2965
e2yntB1	ZN	2894	2821	e2ag5A1	SO4	2962	2966
e2yntB1	ZN	2894	2822	e2ag5A1	SO4	2972	2966
e2yntB1	ZN	2894	2823	e2ag5A1	NAD	2939	2968
e2yntB1	ZN	2894	2838	e2ag5A1	NAD	2939	2962
e2yntB1	ZN	2839	2820	e2ag5A1	NAD	2939	3485
e2yntB1	ZN	2839	2836	e2ag5A1	NAD	2939	2964
e2yntB1	ZN	2839	2862	e2ag5A1	NAD	2939	2972
e2yntB1	ZN	2839	2821	e2ag5A1	NAD	2967	2962
e2yntB1	ZN	2839	2822	e2ag5A1	NAD	2967	3485
e2yntB1	ZN	2839	2823	e2ag5A1	NAD	2967	2964

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e2yntB1	ZN	2839	2838	e2ag5A1	NAD	2967	2972
e2yntB1	ZN	2826	2820	e2ag5A1	NAD	2968	2962
e2yntB1	ZN	2826	2836	e2ag5A1	NAD	2968	3485
e2yntB1	ZN	2826	2862	e2ag5A1	NAD	2968	2964
e2yntB1	ZN	2826	2821	e2ag5A1	NAD	2968	2972
e2yntB1	ZN	2826	2822	e4wecA1	NAD	2939	2962
e2yntB1	ZN	2826	2823	e4wecA1	NAD	2939	3471
e2yntB1	ZN	2826	2838	e4wecA1	NAD	2939	2963
e2yntB1	ZN	2828	2823	e4wecA1	NAD	2939	3395
e2yntB1	ZN	2828	2838	e4wecA1	NAD	2939	3396
e2yntB1	ZN	2842	2820	e4wecA1	NAD	2939	2964
e2yntB1	ZN	2842	2836	e4wecA1	NAD	2939	2965
e2yntB1	ZN	2842	2862	e4wecA1	NAD	2939	2972
e2yntB1	ZN	2842	2821	e1zk4A1	NAP	2939	2968
e2yntB1	ZN	2842	2822	e1zk4A1	NAP	2939	2967
e2yntB1	ZN	2842	2823	e1zk4A1	NAP	2939	2962
e2yntB1	ZN	2842	2838	e1zk4A1	NAP	2939	3471
e2yntB1	ZN	2895	2823	e1zk4A1	NAP	2939	3485
e2yntB1	ZN	2895	2838	e1zk4A1	NAP	2939	2964
e2yntB1	ZN	2820	2823	e1zk4A1	NAP	2939	2965
e2yntB1	ZN	2820	2838	e1zk4A1	NAP	2968	2962

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e4c1hA1	X8Z	2921	2836	e1zk4A1	NAP	2968	3471
e4c1hA1	X8Z	2921	2862	e1zk4A1	NAP	2968	3485
e4c1hA1	X8Z	2921	2821	e1zk4A1	NAP	2968	2964
e4c1hA1	X8Z	2921	2822	e1zk4A1	NAP	2968	2965
e4c1hA1	X8Z	2921	2823	e1zk4A1	NAP	2967	2962
e4c1hA1	X8Z	2921	2838	e1zk4A1	NAP	2967	3471
e4c1hA1	X8Z	2893	2836	e1zk4A1	NAP	2967	3485
e4c1hA1	X8Z	2893	2862	e1zk4A1	NAP	2967	2964
e4c1hA1	X8Z	2893	2821	e1zk4A1	NAP	2967	2965
e4c1hA1	X8Z	2893	2822	e1zk4A1	AC0	2968	2962
e4c1hA1	X8Z	2893	2823	e1zk4A1	AC0	2968	2961
e4c1hA1	X8Z	2893	2838	e1zk4A1	AC0	2968	3471
e4c1hA1	X8Z	2894	2836	e1zk4A1	AC0	2968	3472
e4c1hA1	X8Z	2894	2862	e1zk4A1	AC0	2968	3485
e4c1hA1	X8Z	2894	2821	e1zk4A1	AC0	2968	3313
e4c1hA1	X8Z	2894	2822	e1zk4A1	AC0	2968	2964
e4c1hA1	X8Z	2894	2823	e1zk4A1	AC0	2968	2965
e4c1hA1	X8Z	2894	2838	e1zk4A1	AC0	2967	2962
e4c1hA1	X8Z	2839	2836	e1zk4A1	AC0	2967	2961
e4c1hA1	X8Z	2839	2862	e1zk4A1	AC0	2967	3471
e4c1hA1	X8Z	2839	2821	e1zk4A1	AC0	2967	3472

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e4c1hA1	X8Z	2839	2822	e1zk4A1	AC0	2967	3485
e4c1hA1	X8Z	2839	2823	e1zk4A1	AC0	2967	3313
e4c1hA1	X8Z	2839	2838	e1zk4A1	AC0	2967	2964
e4c1hA1	X8Z	2826	2836	e1zk4A1	AC0	2967	2965
e4c1hA1	X8Z	2826	2862	e1zk4A1	AC0	2962	2965
e4c1hA1	X8Z	2826	2821	e1zk4A1	AC0	2961	2964
e4c1hA1	X8Z	2826	2822	e1zk4A1	AC0	2961	2965
e4c1hA1	X8Z	2826	2823	e1zk4A1	AC0	3472	2965
e4c1hA1	X8Z	2826	2838	e1zk4A1	AC0	3313	2965
e4c1hA1	X8Z	2828	2823	e2dtxA1	BMA	2969	3241
e4c1hA1	X8Z	2828	2838	e2dtxA1	BMA	2969	2963
e4c1hA1	X8Z	2895	2823	e2dtxA1	BMA	2969	3485
e4c1hA1	X8Z	2895	2838	e2dtxA1	BMA	2969	3313
e4c1hA1	X8Z	2820	2823	e2dtxA1	BMA	2969	2964
e4c1hA1	X8Z	2820	2838	e2dtxA1	BMA	2969	2972
e4c1hA1	ZN	2921	2895	e2dtxA1	BMA	2969	2965
e4c1hA1	ZN	2921	2820	e2dtxA1	BMA	2960	2963
e4c1hA1	ZN	2921	2836	e2dtxA1	BMA	2960	3485
e4c1hA1	ZN	2921	2862	e2dtxA1	BMA	2960	3313
e4c1hA1	ZN	2921	2821	e2dtxA1	BMA	2960	2964
e4c1hA1	ZN	2921	2822	e2dtxA1	BMA	2960	2972

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e4c1hA1	ZN	2921	2823	e2dtxA1	BMA	2960	2965
e4c1hA1	ZN	2921	2838	e2dtxA1	BMA	2962	2965
e4c1hA1	ZN	2893	2895	e2dtxA1	BMA	2961	2972
e4c1hA1	ZN	2893	2820	e2dtxA1	BMA	2961	2965
e4c1hA1	ZN	2893	2836	e2dtxA1	BMA	3313	2965
e4c1hA1	ZN	2893	2862	e1zemA1	NAD	2939	2967
e4c1hA1	ZN	2893	2821	e1zemA1	NAD	2939	2968
e4c1hA1	ZN	2893	2822	e1zemA1	NAD	2939	2969
e4c1hA1	ZN	2893	2823	e1zemA1	NAD	2939	2962
e4c1hA1	ZN	2893	2838	e1zemA1	NAD	2939	3241
e4c1hA1	ZN	2894	2895	e1zemA1	NAD	2939	2963
e4c1hA1	ZN	2894	2820	e1zemA1	NAD	2939	3313
e4c1hA1	ZN	2894	2836	e1zemA1	NAD	2967	2962
e4c1hA1	ZN	2894	2862	e1zemA1	NAD	2967	3241
e4c1hA1	ZN	2894	2821	e1zemA1	NAD	2967	2963
e4c1hA1	ZN	2894	2822	e1zemA1	NAD	2967	3313
e4c1hA1	ZN	2894	2823	e1zemA1	NAD	2968	2962
e4c1hA1	ZN	2894	2838	e1zemA1	NAD	2968	3241
e4c1hA1	ZN	2839	2895	e1zemA1	NAD	2968	2963
e4c1hA1	ZN	2839	2820	e1zemA1	NAD	2968	3313
e4c1hA1	ZN	2839	2836	e1zemA1	NAD	2969	3241

Table A.1 Complete list of binding themes. (cont.)

PDB ID	**Ligand**	**Pair Theme 1**	**Pair Theme 2**	**PDB ID**	**Ligand**	**Pair Theme 1**	**Pair Theme 2**
e4c1hA1	ZN	2839	2862	e1zemA1	NAD	2969	2963
e4c1hA1	ZN	2839	2821	e1zemA1	NAD	2969	3313
e4c1hA1	ZN	2839	2822	e3pk0A1	CA	2939	single
e4c1hA1	ZN	2839	2823	e3pk0A1	CA	2939	2967
e4c1hA1	ZN	2839	2838	e4h4rA3	FAD	2939	3519
e4c1hA1	ZN	2826	2895	e4h4rA3	FAD	2939	3499

APPENDIX B: Theme Classification Results for All Domains

Table B.1. Classification results for e1a7tA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2825	Weak	-0.0108	22.22	0	2826	2823
2921	Strong receiver	-0.0154	22.22	100	nan	2823
2824	Weak	0.0084	22.22	0	2828	2823
2894	Strong receiver	-0.014	22.22	100	nan	2823
2839	Strong receiver	-0.0116	22.22	100	nan	2823
2893	Weak	-0.005	22.22	0	2836	2823
2826	Strong receiver	-0.0258	0	100	nan	2823
2828	Strong receiver	-0.0157	11.11	100	nan	2823
2842	Strong receiver	-0.0259	0	100	nan	2823
2895	Weak	-0.008	11.11	0	nan	2823
2820	Weak	-0.0082	11.11	0	2842	2823

Table B.1 Classification results for e1a7tA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2836	Receiver with local sender features	-0.0089	44.44	0	2842	2824
2862	Strong sender	0.0148	77.78	0	2826	nan
2821	Receiver with local sender features	-0.0064	44.44	0	2826	2824
2822	Strong sender	-0.0011	55.56	0	2921	2824
2823	Strong sender	0.0446	66.67	0	2825	nan
2838	Receiver with local sender features	-0.0019	33.33	0	2826	2824

Table B.2. Classification results for e1dzkA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
14284	Strong sender-receiver	0.0285	100	50	14280	nan
14281	Receiver with local sender features	-0.0209	25	16.67	14280	14286
14282	Strong sender	0.008	50	16.67	14280	14286
14283	Strong sender	0.0252	75	33.33	14280	14313
14286	Balanced sender-receiver	0.04	37.5	33.33	14280	nan
14313	Sender	0.0386	25	16.67	14280	nan
14285	Strong receiver	-0.0293	25	66.67	nan	14283
14280	Strong receiver	-0.0508	12.5	50	nan	14286

Table B.3. Classification results for e1elrA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
13394	Sender	0.0606	33.33	0	13396	nan
13386	Strong receiver	0.0235	33.33	50	13396	nan
13396	Strong sender-receiver	-0.0423	50	50	nan	13394

Table B.4. Classification results for e1elwA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
13394	Strong sender	0.0552	60	0	13391	nan
13386	Strong sender-receiver	0.0325	60	66.67	13391	nan
13390	Strong sender-receiver	0.0033	60	100	nan	13394
13389	Strong receiver	0.0245	20	66.67	13391	nan

Table B.4 Classification results for e1elwA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
13469	Strong receiver	0.0261	20	66.67	13396	nan
13440	Balanced sender-receiver	-0.0134	40	33.33	nan	13394
13391	Receiver	-0.0342	40	33.33	nan	13394
13702	Balanced sender-receiver	-0.0291	40	33.33	nan	13394
13396	Receiver	-0.0342	40	33.33	nan	13394

Table B.5. Classification results for e1fmcA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Sender	0.0115	22.22	0	2968	nan
2968	Weak	-0.0056	22.22	0	nan	2939
2967	Weak	-0.0066	22.22	0	nan	2939
2969	Receiver with local sender features	-0.0025	33.33	0	nan	2939

Table B.5 Classification results for e1fmcA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2960	Receiver	-0.0101	11.11	0	nan	2939
2962	Strong receiver	-0.0059	22.22	100	nan	nan
2961	Receiver	-0.0077	11.11	0	nan	2939
2963	Strong receiver	0.0009	44.44	100	nan	nan
2964	Strong receiver	-0.0006	11.11	100	nan	nan
2965	Strong receiver	0.0033	33.33	100	nan	nan
2966	Sender	0.0052	22.22	0	2967	nan

Table B.6. Classification results for e1h3dA2 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6540	Receiver	-0.2792	12.5	0	nan	6537
6541	Receiver	-0.2839	12.5	33.33	nan	6537
6590	Receiver	-0.2217	0	33.33	6540	6537
6654	Strong receiver	-0.1911	12.5	66.67	6540	6537

Table B.6 Classification results for e1h3dA2 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6656	Strong sender-receiver	0.3031	75	66.67	6541	6537
6543	Strong receiver	0.2034	37.5	66.67	6541	6537
6544	Strong sender-receiver	0.3271	62.5	66.67	6540	6537
6545	Strong sender-receiver	0.4262	75	66.67	6540	nan
6546	Strong sender-receiver	0.2056	87.5	66.67	6541	nan
6536	Strong sender	0.4958	62.5	33.33	6540	nan
6651	Strong sender	0.248	75	33.33	6540	nan
6714	Strong sender	0.5239	75	33.33	6541	nan
6537	Sender	0.6824	37.5	0	6541	nan
6538	Sender	0.0805	25	0	6541	6536
6653	Receiver	-0.1431	12.5	0	6541	6537

Table B.7. Classification results for e1h5qA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Receiver	-0.0136	22.22	0	nan	2962
2967	Strong receiver	-0.013	11.11	50	nan	2962
2968	Strong receiver	-0.0106	11.11	50	nan	2963
2969	Strong receiver	-0.009	22.22	100	nan	2965
2974	Strong receiver	-0.0038	22.22	50	nan	nan
2960	Strong receiver	-0.0132	11.11	50	nan	2965
2962	Strong receiver	0.0014	44.44	50	2939	2966
2963	Strong sender-receiver	0.0087	66.67	50	2939	nan
2964	Weak	-0.0029	11.11	0	2939	2966
2972	Weak	-0.0008	11.11	0	2939	2966
2965	Sender	0.0101	33.33	0	2939	nan
2966	Sender	0.0184	22.22	0	2939	nan

Table B.8. Classification results for e1m2xA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2825	Strong receiver	-0.0283	10	66.67	2842	2823
2840	Strong receiver	-0.0145	10	66.67	2826	2823
2893	Receiver	-0.0234	20	33.33	2842	2823
2894	Strong receiver	-0.0321	20	66.67	nan	2823
2839	Strong receiver	-0.0292	20	66.67	nan	2823
2826	Receiver	-0.0313	10	33.33	nan	2823
2828	Balanced sender-receiver	-0.0135	30	33.33	nan	2823
2842	Receiver	-0.0293	0	33.33	nan	2823
2895	Weak	-0.0021	20	0	2893	2838
2820	Weak	0.0049	20	0	2825	2823
2836	Strong sender	0.0106	60	0	2825	nan
2862	Strong sender	0.029	80	0	2825	nan
2821	Strong sender	0.0154	60	0	2825	nan

Table B.8 Classification results for e1m2xA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2822	Strong sender	0.0132	60	0	2825	nan
2838	Sender	0.0172	40	0	2825	nan
2823	Strong sender	0.0563	60	0	2894	nan

Table B.9. Classification results for e1nh8A2 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6540	Weak	-0.1092	20	0	6538	6545
6541	Receiver	-0.1571	30	0	6538	6545
6590	Receiver	-0.2564	10	0	6538	6545
6543	Sender	0.2395	30	0	6538	6537
6544	Strong sender	0.3308	50	0	6538	6537
6545	Strong sender	0.4904	60	0	6538	nan
6546	Strong sender-receiver	0.1667	70	100	6590	nan

Table B.9 Classification results for e1nh8A2 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6536	Strong sender	0.4493	50	0	6538	nan
6714	Strong sender-receiver	0.315	60	100	6590	nan
6537	Sender	0.4579	30	0	6590	nan
6538	Strong receiver	-0.3368	10	100	nan	6545

Table B.10. Classification results for e1o04A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
7042	Receiver	-0.3554	0	20	7110	7026
7044	Weak	-0.2548	0	0	7032	7026
7035	Receiver	-0.3956	0	20	7052	7026
7104	Weak	-0.1933	0	0	7110	7026
7095	Receiver	-0.6162	0	20	7110	7026
7032	Receiver	-0.683	0	20	nan	7026
7052	Receiver	-0.5105	0	20	nan	7026
7053	Receiver	-0.4137	0	20	nan	7026

Table B.10 Classification results for e1o04A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
7036	Weak	-0.2689	5.88	0	7028	7026
7084	Weak	0.3643	5.88	0	7028	7026
7085	Weak	0.1706	11.76	0	7028	7026
7108	Weak	0.2829	11.76	0	7028	7026
7088	Sender	0.4379	11.76	0	7028	7026
7089	Weak	0.1746	11.76	0	7028	7026
7040	Weak	0.1858	5.88	0	7032	7026
7020	Weak	0.0875	5.88	0	7032	7026
7019	Weak	0.1446	5.88	0	7095	7026
7041	Weak	0.0587	5.88	0	7032	7026
7043	Weak	-0.0053	5.88	0	7032	7026
7151	Weak	-0.1528	5.88	0	7032	7026
7152	Weak	-0.1988	5.88	0	7032	7026
7109	Weak	-0.3059	0	0	7032	7026
7153	Weak	-0.1535	11.76	20	7032	7026
7021	Weak	-0.2876	0	0	7032	7026
7100	Weak	-0.2721	0	0	7032	7026
7099	Weak	-0.288	0	20	7032	7026
7101	Weak	-0.0244	17.65	20	7032	7026
7045	Weak	-0.1907	5.88	20	7032	7026
7022	Sender	0.208	29.41	20	7032	7026

Table B.10 Classification results for e1o04A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
7137	Balanced sender-receiver	0.2992	41.18	40	7032	7088
7110	Receiver	-0.3551	0	20	7052	7026
7111	Weak	-0.0509	5.88	20	7032	7026
7091	Weak	0.3432	23.53	20	7032	7026
7092	Balanced sender-receiver	0.1743	41.18	40	7095	7084
7023	Weak	-0.0192	5.88	20	7032	7026
7025	Sender	0.6577	41.18	20	7050	nan
7046	Sender	0.6216	29.41	20	7050	7047
7093	Sender	0.8814	41.18	20	7050	nan
7026	Sender	1.096	35.29	20	7050	nan
7047	Sender	0.9565	23.53	20	7050	nan
7138	Sender	0.5281	17.65	20	7095	7046
7113	Weak	0.076	5.88	20	7095	7084
7027	Weak	-0.1547	5.88	20	7095	7084
7115	Weak	-0.1996	17.65	20	7095	7026
7028	Receiver	-0.4931	0	0	7032	7026

Table B.10 Classification results for e1o04A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
7141	Receiver with local sender features	-0.1283	35.29	20	7032	7026
7048	Weak	0.3481	23.53	0	7095	7026
7029	Sender	0.1443	35.29	20	7095	7026
7049	Sender	0.0501	35.29	20	7095	7026
7030	Balanced sender-receiver	0.0675	47.06	40	7095	7026
7142	Sender	0.2143	29.41	20	7095	7026
7143	Weak	-0.0375	23.53	20	7032	7026
7154	Weak	-0.096	23.53	20	7032	7026
7117	Balanced sender-receiver	-0.0617	29.41	40	7095	7026
7094	Weak	-0.1848	17.65	20	7032	7026
7144	Balanced sender-receiver	-0.0664	29.41	40	7095	7026
7050	Weak	-0.287	11.76	20	7032	7026
7031	Weak	0.0087	11.76	20	7028	7026

Table B.10 Classification results for e1o04A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
7051	Weak	0.0614	5.88	20	7028	7026

Table B.11. Classification results for e1p22A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
14813	Strong receiver	-0.0792	8.33	55.56	nan	14919
14919	Sender	0.1189	25	0	14921	nan
14916	Sender	0.0804	25	0	14921	14956
14917	Sender	0.0843	41.67	11.11	14921	nan
15261	Sender	0.1177	25	0	14921	nan
14926	Strong sender	0.0645	50	11.11	14816	nan
14863	Sender	0.0501	33.33	11.11	14816	14861
14862	Strong sender	0.0328	50	11.11	14816	14919
14812	Sender	0.0359	25	11.11	14816	14919
14997	Sender	0.05	41.67	11.11	14816	15261
14879	Strong sender	0.0191	50	11.11	14955	14919

Table B.11 Classification results for e1p22A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
14993	Weak	-0.0207	16.67	11.11	14816	15261
14815	Strong receiver	-0.0773	8.33	66.67	nan	15261
15263	Sender	0.0345	25	11.11	14815	15261
14925	Sender	0.0089	33.33	11.11	14815	15261
15265	Weak	-0.0134	16.67	0	14813	14919
14859	Weak	-0.0023	8.33	0	14921	14919
14860	Strong receiver	-0.0561	16.67	77.78	nan	14919
14918	Receiver	-0.0385	16.67	44.44	nan	14919
14921	Receiver	-0.1125	0	44.44	nan	14919
15255	Receiver	-0.051	8.33	33.33	nan	14919
14883	Receiver	-0.0674	0	33.33	nan	14919
14920	Receiver	-0.0478	8.33	44.44	nan	14919
14884	Strong receiver	-0.0687	8.33	77.78	nan	15261
14816	Receiver	-0.108	0	33.33	nan	14919
14955	Strong receiver	-0.0889	8.33	66.67	nan	15261

Table B.11 Classification results for e1p22A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
14956	Sender with local receiver features	0.0037	16.67	33.33	14993	14859
14861	Weak	0.0274	16.67	11.11	14993	nan

Table B.12. Classification results for e1so2A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
12885	Weak	-0.0016	10	11.11	12899	12892
12902	Weak	-0.0048	10	11.11	12899	12892
12905	Strong receiver	-0.0116	30	55.56	12899	12892
12897	Receiver	-0.0018	10	33.33	12917	12892
12906	Balanced sender-receiver	-0.0127	30	44.44	12899	12892
12917	Weak	-0.0177	20	22.22	12899	12892
12881	Weak	0.0024	20	11.11	12899	12892
12882	Weak	-0.0036	20	11.11	12887	12892

Table B.12 Classification results for e1so2A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
12883	Weak	-0.0078	20	11.11	nan	12892
12884	Weak	-0.0069	20	11.11	nan	12892
12899	Receiver	-0.0229	0	0	nan	12915
12900	Weak	-0.0181	0	0	nan	12892
12887	Weak	-0.0207	0	0	nan	12892
12888	Receiver	-0.0222	0	0	nan	12892
12901	Weak	-0.0142	0	0	nan	12885
12903	Weak	-0.0093	0	0	nan	12881
12904	Sender	0.0306	40	22.22	12905	nan
12914	Sender	0.0212	30	11.11	12917	nan
12891	Sender	0.0415	40	22.22	12906	nan
12908	Strong sender	0.0227	60	33.33	12917	nan
12915	Sender	0.0478	30	11.11	12906	nan
12892	Sender	0.0793	40	33.33	12906	nan
12893	Strong sender	0.0363	60	33.33	12917	12915
12910	Weak	0.0084	20	11.11	12905	12891
12894	Weak	-0.0078	20	11.11	nan	12892

Table B.13. Classification results for e1uxtA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
7034	Weak	-0.1796	0	0	7024	7088
7104	Weak	-0.2221	0	0	7024	7088
7107	Weak	-0.2646	0	0	7058	7088
7106	Weak	-0.2423	0	0	7058	7088
7069	Weak	-0.2995	0	0	nan	7088
7105	Weak	-0.2818	0	0	7061	7149
7149	Sender	0.8736	7.69	12.5	7028	nan
7150	Sender	0.6781	15.38	12.5	7028	nan
7088	Sender	0.8538	15.38	0	7028	nan
7019	Weak	0.1504	7.69	0	7058	7061
7020	Weak	0.1343	7.69	0	7069	7029
7041	Weak	0.1919	7.69	0	7058	7028
7153	Weak	-0.1301	7.69	12.5	7028	7088
7100	Weak	-0.1213	0	0	7069	7019
7099	Weak	-0.1431	7.69	0	7028	7088
7101	Weak	-0.1195	7.69	12.5	7028	7088
7058	Weak	-0.2738	0	0	7069	7019
7059	Weak	-0.304	0	0	7069	7088

Table B.13 Classification results for e1uxtA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
7022	Receiver with local sender features	-0.109	30.77	12.5	7028	7088
7111	Weak	-0.1621	7.69	12.5	7028	7149
7091	Weak	-0.045	15.38	12.5	7028	7088
7023	Weak	-0.1622	7.69	12.5	7028	7149
7024	Balanced sender-receiver	0.1388	38.46	25	7028	7088
7025	Balanced sender-receiver	0.0287	38.46	37.5	7028	7088
7093	Balanced sender-receiver	0.0869	38.46	37.5	7028	7088
7060	Sender	0.0574	30.77	12.5	7028	7149
7061	Sender	0.3453	38.46	25	7094	7088
7102	Balanced sender-receiver	0.0853	30.77	37.5	7028	7088
7103	Strong receiver	-0.2901	30.77	62.5	7094	7149

Table B.13 Classification results for e1uxtA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
7026	Sender	0.2108	30.77	25	7028	7088
7027	Strong receiver	-0.3386	30.77	50	7094	7149
7063	Receiver	-0.2721	7.69	37.5	7094	7149
7064	Strong receiver	-0.4324	15.38	50	7106	7061
7028	Receiver	-0.6884	0	25	7106	7061
7115	Receiver	-0.4892	15.38	37.5	7106	7061
7030	Balanced sender-receiver	-0.2813	38.46	25	7028	7061
7029	Receiver with local sender features	-0.2149	30.77	12.5	7028	7061
7049	Balanced sender-receiver	-0.2492	30.77	25	7028	7061
7067	Weak	-0.1567	23.08	0	7028	7061
7154	Weak	-0.3267	23.08	12.5	7028	7061
7094	Weak	-0.4289	15.38	12.5	7028	7061
7031	Weak	-0.0154	15.38	0	7028	7149

Table B.13 Classification results for e1uxtA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
7051	Weak	0.1124	7.69	0	7028	7149

Table B.14. Classification results for e1y2kA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
12885	Weak	-0.0194	10	11.11	12916	12892
12902	Weak	-0.023	20	11.11	12916	12892
12916	Receiver	-0.1033	10	33.33	12887	12892
12905	Balanced sender-receiver	-0.0753	30	44.44	12900	12892
12906	Receiver	-0.0766	20	44.44	12900	12892
12932	Receiver	-0.0646	20	33.33	12903	12892
12881	Weak	-0.0367	20	11.11	12903	12892
12882	Weak	-0.0374	20	11.11	12916	12892
12883	Weak	-0.0372	20	11.11	12916	12892
12884	Weak	-0.0497	20	22.22	12916	12892
12899	Weak	-0.0342	10	11.11	12903	12892
12900	Weak	-0.0588	0	22.22	nan	12892

Table B.14 Classification results for e1y2kA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
12901	Weak	-0.0639	0	22.22	nan	12892
12887	Weak	-0.0511	0	11.11	12903	12893
12888	Weak	-0.0674	0	22.22	nan	12892
12903	Weak	-0.0842	0	11.11	nan	12892
12952	Weak	-0.0692	10	22.22	12916	12892
12904	Sender	0.0353	30	22.22	12916	12910
12908	Strong sender	0.0791	60	22.22	12916	nan
12891	Weak	0.0469	20	22.22	12916	12910
12915	Weak	0.0343	10	11.11	12916	12910
12909	Sender	0.1722	20	0	12906	12894
12892	Sender	0.2399	40	0	12906	nan
12893	Strong sender	0.1848	50	0	12916	nan
12910	Sender	0.1771	40	0	12916	nan
12894	Sender	0.1635	30	0	12916	nan

Table B.15. Classification results for e1zemA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Weak	-0.0021	25	nan	2960	2966
2967	Weak	-0.0368	25	nan	2960	2966
2968	Weak	-0.0075	37.5	nan	2960	2966
2969	Weak	-0.0442	25	nan	2964	2966
2960	Receiver	-0.1183	0	nan	2964	2966
2962	Weak	-0.0978	12.5	nan	nan	2966
2961	Weak	-0.0889	12.5	nan	2964	2966
3241	Weak	0.0407	37.5	nan	2960	nan
2963	Weak	0.051	37.5	nan	2960	nan
3313	Weak	-0.0793	12.5	nan	nan	2966
2964	Receiver	-0.1157	0	nan	nan	2966
2965	Sender	0.0836	25	nan	2967	nan
2966	Sender	0.2075	25	nan	2967	nan

Table B.16. Classification results for e1zk4A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Receiver with local sender features	-0.0413	28.57	0	nan	2966
2968	Weak	-0.0482	14.29	0	2939	2966
2967	Receiver	-0.0734	14.29	0	2960	2966
2962	Receiver	-0.0283	14.29	33.33	2939	2966
2960	Receiver	-0.0801	0	0	3313	2966
2961	Weak	-0.0424	0	0	2939	2966
3471	Strong receiver	0.0356	42.86	100	2967	nan
3472	Weak	-0.0157	0	0	2939	2966
3485	Strong sender-receiver	0.0644	57.14	100	2967	nan
3313	Receiver	-0.0345	0	33.33	2939	2966
2964	Sender with local receiver features	0.0156	14.29	33.33	2939	2966
2965	Strong sender-receiver	0.0812	57.14	100	2967	nan

Table B.16 Classification results for e1zk4A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2966	Strong receiver	0.1421	42.86	66.67	2967	nan

Table B.17. Classification results for e1zklA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
12885	Weak	-0.0755	10	11.11	12916	12892
12902	Weak	-0.0667	10	11.11	12916	12892
12916	Receiver	-0.1523	0	44.44	nan	12892
12905	Strong receiver	-0.1548	20	66.67	nan	12892
12906	Strong receiver	-0.1423	20	66.67	nan	12892
12881	Receiver	-0.13	20	22.22	12885	12892
12882	Weak	-0.0988	20	22.22	12885	12892
12883	Weak	-0.0886	20	22.22	12885	12892
12884	Receiver with local sender features	-0.0701	30	22.22	12885	12892

Table B.17 Classification results for e1zklA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
12899	Weak	0.0169	10	0	12903	12892
12900	Weak	-0.0653	10	11.11	12885	12892
12887	Weak	0.008	10	0	12903	12892
12901	Weak	-0.0679	10	11.11	12885	12892
12903	Receiver	-0.1283	0	11.11	nan	12892
12904	Sender	0.1615	40	11.11	12905	12910
12914	Sender	0.0943	30	11.11	12905	12910
12891	Sender	0.2071	40	11.11	12905	12910
12908	Strong sender	0.1782	60	11.11	12905	nan
12915	Sender	0.1829	30	11.11	12881	12910
12909	Sender	0.3813	40	11.11	12881	nan
12892	Strong sender	0.4856	60	11.11	12881	nan
12893	Strong sender	0.2714	60	11.11	12905	nan
12910	Sender	0.2124	30	0	12905	nan
12894	Weak	0.1137	20	0	12905	12892

Table B.18. Classification results for e2ae2A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Sender	0.0398	25	0	2967	2966
2967	Balanced sender-receiver	-0.0532	25	33.33	2960	2966
2968	Balanced sender-receiver	-0.0224	25	33.33	2960	2966
3193	Balanced sender-receiver	-0.0474	25	33.33	2960	2966
2969	Receiver with local sender features	-0.0103	25	0	2964	2966
2960	Receiver	-0.0581	0	0	nan	2966
2962	Strong receiver	-0.0567	12.5	66.67	nan	2966
2963	Strong receiver	0.0277	37.5	66.67	2967	nan
2964	Strong receiver	-0.0443	12.5	66.67	nan	2966
2965	Strong receiver	0.0487	37.5	66.67	2967	nan

Table B.18 Classification results for e2ae2A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2966	Sender	0.1073	25	0	2967	nan

Table B.19. Classification results for e2ag5A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Sender	0.0119	37.5	0	2961	2972
2967	Sender	0.0065	25	0	2961	2966
2968	Sender	0.0071	25	0	2961	2966
2961	Weak	-0.0379	0	0	nan	2966
2960	Receiver	-0.0592	0	0	nan	2966
2962	Strong receiver	-0.0531	12.5	50	nan	2966
3485	Strong receiver	0.0323	37.5	100	2960	nan
2964	Strong receiver	-0.0611	12.5	50	2939	2966
2972	Strong receiver	0.0524	25	50	2960	2966
2965	Strong receiver	0.0656	37.5	100	2964	nan

Table B.19 Classification results for e2ag5A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2966	Strong receiver	0.1153	25	50	2964	nan

Table B.20. Classification results for e2bcqA3 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
9358	Receiver	-0.3726	0	0	nan	9357
9371	Weak	-0.3282	0	0	nan	9357
9361	Receiver	-0.3481	0	33.33	nan	9357
9360	Receiver	-0.2444	16.67	33.33	9358	9357
9365	Strong sender-receiver	0.0271	50	100	9358	nan
9362	Strong sender-receiver	0.3191	100	66.67	9358	nan
9363	Strong sender-receiver	0.2471	50	66.67	9361	nan

Table B.20 Classification results for e2bcqA3 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
9357	Strong sender-receiver	0.3833	100	66.67	9358	nan
9370	Strong sender	0.146	50	33.33	9358	nan

Table B.21. Classification results for e2dtxA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Receiver	-0.0359	22.22	25	nan	2972
2968	Receiver	-0.0135	22.22	25	2939	3485
2967	Receiver	-0.0192	11.11	25	2939	2972
2969	Balanced sender-receiver	-0.0034	44.44	25	2939	2966
2960	Receiver with local sender features	-0.0089	33.33	0	2967	2966
2962	Strong receiver	-0.0162	33.33	50	2939	2966

Table B.21 Classification results for e2dtxA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2961	Receiver with local sender features	-0.0061	33.33	0	2939	2966
3241	Strong receiver	0.0032	44.44	50	2939	nan
2963	Strong receiver	0.0032	22.22	50	2939	nan
3485	Strong receiver	0.0055	22.22	50	2939	nan
3313	Strong receiver	-0.0153	11.11	50	2939	2966
2964	Strong receiver	-0.0234	0	50	2939	2966
2972	Strong receiver	0.0172	0	50	2939	nan
2965	Strong receiver	0.0048	11.11	50	2939	nan
2966	Sender	0.0252	11.11	0	2939	nan

Table B.22. Classification results for e2j13A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
1671	Weak	-0.0247	10	20	nan	1669
1673	Receiver	-0.0217	10	40	1670	1669
1668	Strong sender-receiver	0.0151	50	80	1670	nan
1685	Receiver	-0.0297	10	40	1670	1669
1675	Receiver	-0.0273	10	40	nan	1669
1674	Strong receiver	0.0042	20	60	1670	nan
1669	Strong sender	0.0186	50	20	1671	nan
1670	Receiver with local sender features	-0.0255	30	20	nan	1668

Table B.23. Classification results for e2o23A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Strong receiver	-0.162	33.33	50	2967	2968
2967	Strong receiver	-0.1706	11.11	100	2960	2964
2968	Strong sender-receiver	0.2793	55.56	50	2960	nan
3315	Weak	-0.0552	22.22	0	2939	2968
2960	Weak	-0.145	11.11	0	2939	2968
2962	Weak	-0.0429	22.22	0	2939	2968
2963	Weak	-0.0068	11.11	0	2960	2968
3313	Weak	-0.0018	11.11	0	2939	2968
3316	Weak	0.0348	11.11	0	2939	2968
2964	Weak	0.0434	11.11	0	2939	2968
2965	Weak	-0.017	0	0	2967	2968

Table B.24. Classification results for e2r25B1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
5397	Sender	0.0089	42.86	0	5488	nan
5391	Sender	0.0063	42.86	0	5488	nan
5408	Strong sender-receiver	0.0022	57.14	75	5400	nan
5407	Strong sender-receiver	-0.0019	57.14	100	nan	nan
5451	Sender	0.0094	14.29	25	5400	nan
5488	Strong receiver	-0.0112	14.29	75	nan	5397
5400	Strong receiver	-0.0046	28.57	50	nan	5397

Table B.25. Classification results for e2vd3A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6540	Receiver	-0.5692	9.09	33.33	nan	6592
6541	Strong receiver	-0.4805	18.18	66.67	nan	6592

Table B.25 Classification results for e2vd3A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6590	Receiver	-0.2652	9.09	0	6540	6592
6713	Receiver	-0.1886	18.18	0	6540	6592
6654	Weak	-0.0805	18.18	0	6540	6536
6656	Strong sender	0.5431	63.64	0	6540	6537
6543	Sender	0.3608	27.27	0	6540	6537
6544	Sender	0.4638	45.45	0	6540	6537
6545	Strong sender	0.6366	54.55	0	6540	nan
6546	Strong sender	0.2964	72.73	33.33	6540	nan
6592	Strong sender	0.7113	54.55	0	6540	nan
6536	Sender	0.7193	45.45	0	6540	nan
6651	Strong sender	0.3066	63.64	33.33	6540	6543
6714	Strong sender	0.6073	54.55	0	6540	nan
6537	Sender	0.6642	27.27	0	6540	nan
6538	Receiver	-0.1725	18.18	33.33	6540	6536
6653	Receiver	-0.0601	18.18	33.33	6540	6536

Table B.26. Classification results for e2wsbA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Sender	0.0041	25	0	2967	3396
2967	Receiver	-0.025	12.5	33.33	3397	2966
2968	Balanced sender-receiver	-0.0023	25	33.33	nan	3493
2974	Sender	0.0041	25	0	2967	2966
3397	Receiver	-0.0395	0	0	nan	3493
2960	Receiver	-0.0473	0	0	nan	2965
2962	Strong receiver	-0.0213	25	66.67	nan	2966
2961	Receiver	-0.0183	12.5	33.33	nan	2965
2963	Strong sender-receiver	0.013	50	66.67	2967	nan
3395	Strong receiver	0.0006	25	66.67	2960	2966
3313	Strong receiver	-0.0094	12.5	66.67	2939	3493
3396	Strong receiver	0.0107	25	66.67	2939	2966
2964	Receiver	-0.0103	12.5	33.33	2939	2966
2965	Sender	0.0332	37.5	33.33	2967	nan
2972	Receiver	-0.0188	12.5	33.33	2967	2966

Table B.26 Classification results for e2wsbA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
3493	Sender	0.0444	25	0	2967	nan
2966	Sender	0.0601	25	0	2967	nan

Table B.27. Classification results for e2xyiA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
14940	Sender	0.1611	13.33	0	14954	14946
14945	Sender	0.1317	13.33	0	15219	nan
14952	Sender	0.0733	0	0	15219	14942
14946	Sender	0.1444	0	0	15219	14942
14942	Weak	0.0133	20	14.29	14954	14815
14953	Receiver with local sender features	-0.0338	26.67	14.29	14954	14815
14944	Receiver with local sender features	-0.0035	46.67	14.29	14941	14940
14812	Weak	0.0352	20	0	14941	14940

Table B.27 Classification results for e2xyiA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
14954	Weak	-0.0553	20	14.29	14941	14940
14813	Sender	0.0238	33.33	14.29	14941	14940
14937	Weak	0.0096	20	14.29	15219	14812
14938	Weak	-0.0189	20	14.29	15219	14813
14939	Weak	-0.0697	20	14.29	14941	14946
15218	Weak	-0.0172	20	14.29	15219	14812
15220	Weak	-0.0393	20	14.29	14955	14946
15221	Weak	-0.0975	20	14.29	14941	14946
14949	Weak	-0.0964	13.33	14.29	14941	14813
14815	Strong receiver	0.0464	13.33	57.14	14861	14937
14816	Weak	-0.0992	0	0	14812	14946
15219	Receiver	-0.1761	0	0	14812	15218
14955	Receiver	-0.1791	0	0	14812	14946
14861	Receiver	-0.1317	0	14.29	15219	14951
14950	Strong receiver	-0.0186	13.33	57.14	14942	14937
14956	Sender with local receiver features	0.0501	13.33	28.57	14954	14941

Table B.27 Classification results for e2xyiA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15222	Strong receiver	0.0239	13.33	57.14	14955	14812
14951	Sender with local receiver features	0.0067	6.67	42.86	14861	14812
14941	Receiver	-0.1774	0	42.86	14861	14812

Table B.28. Classification results for e2yntB1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2921	Receiver	-0.0263	22.22	42.86	nan	2823
2840	Receiver	-0.0006	11.11	28.57	2842	2823
2825	Receiver	-0.0061	11.11	28.57	2842	2823
2824	Receiver	-0.0394	11.11	14.29	2842	2823
2893	Weak	-0.0256	22.22	14.29	2842	2823
2894	Receiver	-0.0375	22.22	28.57	nan	2823
2839	Receiver	-0.0343	22.22	28.57	nan	2823
2826	Receiver	-0.0364	11.11	14.29	nan	2823

Table B.28 Classification results for e2yntB1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2828	Balanced sender-receiver	-0.0133	33.33	42.86	2824	2823
2842	Weak	-0.0348	0	14.29	nan	2823
2895	Receiver	-0.0137	22.22	42.86	2824	2840
2820	Sender with local receiver features	0.0068	22.22	28.57	2824	2840
2836	Strong sender-receiver	0.0078	55.56	57.14	2824	nan
2862	Strong sender-receiver	0.0315	77.78	57.14	2824	nan
2821	Strong sender-receiver	0.0106	55.56	57.14	2824	nan
2822	Strong sender	0.0165	66.67	42.86	2824	nan
2823	Strong sender	0.0473	55.56	28.57	2824	nan
2838	Sender	0.0321	33.33	28.57	2824	nan

Table B.29. Classification results for e2zatA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Sender	0.0329	22.22	0	2967	2964
2967	Strong receiver	-0.0564	11.11	100	3314	2966
3193	Strong receiver	-0.0489	11.11	100	3314	2966
2968	Strong receiver	-0.0242	11.11	100	3314	2966
2969	Strong receiver	-0.0298	11.11	100	nan	2966
2962	Weak	-0.0075	22.22	0	2939	2966
2960	Receiver	-0.0419	0	0	nan	2966
2961	Weak	-0.0073	11.11	0	2967	2966
3314	Receiver	-0.0604	0	0	nan	2966
3312	Weak	-0.0346	0	0	nan	2966
2963	Strong sender	0.0324	55.56	0	2967	nan
3313	Weak	0.0119	11.11	0	2939	2966
2964	Weak	0.024	22.22	0	2939	2966
2965	Sender	0.0382	44.44	0	3193	nan
2966	Sender	0.0686	44.44	0	3193	nan

Table B.30. Classification results for e3b2dA4 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15670	Weak	-0.1639	8.33	0	15510	15594
15509	Sender	0.12	25	0	15510	15594
15512	Weak	-0.0585	8.33	0	15510	15594
15595	Sender	0.2629	33.33	0	15510	15594
15518	Weak	0.3589	16.67	0	15510	15594
15593	Receiver	-0.1516	8.33	28.57	15510	15594
15519	Receiver	-0.4596	8.33	42.86	15510	15594
15481	Receiver	-0.7439	0	28.57	15520	15594
15596	Receiver	-0.5091	0	28.57	15520	15594
15592	Weak	-0.448	0	14.29	nan	15594
15507	Receiver	-0.2638	0	28.57	15592	15518
15508	Receiver	-0.3755	0	42.86	15592	15518
15506	Weak	0.0913	0	0	15594	nan
15520	Receiver	-0.5037	0	42.86	15592	15595
15510	Receiver	-0.6378	0	28.57	15592	15518
15511	Receiver	-0.2993	16.67	42.86	15481	15518
15513	Balanced sender-receiver	-0.0614	25	42.86	15481	15594
15521	Balanced sender-receiver	0.3994	33.33	28.57	15481	15594

Table B.30 Classification results for e3b2dA4 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15514	Sender	0.5682	8.33	14.29	15481	15594
15515	Strong sender	0.7042	50	14.29	15481	15506
15516	Sender	0.7545	41.67	0	15481	15506
15594	Sender	0.9861	33.33	0	15481	15506
15591	Sender	0.7465	41.67	0	15481	15506
15517	Sender	0.471	25	0	15481	15506

Table B.31. Classification results for e3bwhA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
10078	Weak	-0.0131	10	0	10096	10101
10081	Weak	-0.0188	20	0	nan	10101
10080	Weak	-0.016	10	0	nan	10101
10103	Weak	-0.0246	10	0	nan	10092
10079	Weak	-0.0286	10	0	nan	10101
10109	Weak	-0.0359	10	0	nan	10092
10095	Weak	-0.0357	10	0	nan	10092
10104	Weak	-0.0291	10	0	nan	10101

Table B.31 Classification results for e3bwhA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
10105	Weak	-0.0336	10	0	nan	10101
10107	Strong receiver	-0.0329	10	100	nan	10101
10096	Weak	-0.039	0	0	nan	10092
10097	Weak	-0.0404	0	0	nan	10092
10098	Strong receiver	-0.0297	10	100	nan	10092
10090	Strong receiver	-0.0211	10	100	nan	10101
10099	Strong receiver	-0.0215	10	100	10096	10092
10100	Strong receiver	-0.0169	10	100	10109	10092
10075	Strong receiver	-0.0159	10	100	10097	10092
10076	Strong receiver	-0.0103	20	100	10109	nan
10077	Strong sender-receiver	0.0444	80	100	10109	nan
10091	Strong sender-receiver	0.0536	60	100	10109	nan

Table B.31 Classification results for e3bwhA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
10101	Sender	0.0817	30	0	10079	nan
10092	Strong sender	0.0919	50	0	10109	nan

Table B.32. Classification results for e3e13X1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6616	Weak	0.0156	23.08	0	6623	6723
6624	Receiver	-0.0907	23.08	33.33	6622	6723
6771	Strong receiver	-0.088	38.46	66.67	6622	6618
6625	Strong receiver	-0.0894	38.46	66.67	6622	6723
6728	Receiver	-0.1451	15.38	33.33	6622	6615
6729	Balanced sender-receiver	-0.1013	30.77	33.33	6622	6618
6730	Balanced sender-receiver	-0.1163	30.77	33.33	6622	6618

Table B.32 Classification results for e3e13X1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6780	Receiver	-0.1384	15.38	33.33	6622	6618
6611	Weak	-0.0633	15.38	0	6728	6618
6612	Weak	-0.0594	15.38	0	6622	6618
6715	Receiver	-0.0733	15.38	33.33	6622	6618
6613	Sender with local receiver features	0.0397	23.08	33.33	6728	6618
6614	Sender with local receiver features	0.027	7.69	33.33	6728	6618
6716	Receiver	-0.0419	0	33.33	6728	6618
6718	Sender with local receiver features	0.0099	23.08	33.33	6728	6618
6719	Receiver	-0.0041	0	33.33	6728	6618
6615	Sender with local receiver features	0.1358	7.69	33.33	6728	6618
6717	Receiver	-0.0288	15.38	33.33	6728	6618

Table B.32 Classification results for e3e13X1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6617	Receiver	-0.0161	15.38	33.33	6728	6618
6722	Weak	0.07	23.08	0	6728	6723
6721	Weak	-0.0511	15.38	0	6728	6723
6770	Weak	-0.0694	15.38	0	6780	6723
6618	Sender	0.2477	23.08	0	6728	nan
6619	Sender	0.1788	23.08	0	6624	nan
6723	Sender	0.2927	23.08	0	6624	nan
6620	Weak	0.1346	23.08	0	6624	nan
6725	Sender	0.2104	23.08	0	6624	nan
6724	Sender	0.2845	15.38	0	6624	nan
6726	Sender	0.2219	15.38	0	6616	nan
6622	Receiver	-0.1176	0	0	nan	6618
6623	Receiver	-0.117	0	0	nan	6618

Table B.33. Classification results for e3ecmA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
12885	Weak	0.0158	22.22	0	12916	12952
12902	Sender	0.0206	33.33	0	12888	12952
12916	Receiver	-0.1112	11.11	25	12887	12952
12905	Strong receiver	-0.0951	33.33	75	12888	12952
12897	Receiver	-0.1172	11.11	25	12887	12952
12906	Strong receiver	-0.1064	22.22	75	12900	12952
12881	Strong receiver	-0.061	22.22	50	12888	12952
12882	Strong receiver	-0.0689	22.22	50	12887	12952
12883	Strong receiver	-0.0733	22.22	50	12916	12952
12884	Strong receiver	-0.0751	22.22	50	12897	12952
12900	Weak	-0.0628	0	0	nan	12892
12899	Weak	-0.0724	0	0	nan	12952
12901	Weak	-0.0537	0	0	12916	12904
12887	Weak	-0.0781	0	0	nan	12952
12888	Weak	-0.0842	0	0	nan	12902
12903	Weak	-0.0289	0	0	12916	12891

Table B.33 Classification results for e3ecmA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
12952	Sender	0.1628	22.22	0	12916	nan
12904	Sender	0.1116	44.44	25	12916	nan
12914	Sender	0.1258	22.22	25	12916	nan
12908	Sender	0.0802	44.44	25	12916	nan
12891	Sender	0.1141	33.33	25	12916	nan
12915	Receiver	-0.0062	11.11	25	12916	12952
12909	Sender with local receiver features	0.0535	22.22	25	12906	12952
12892	Sender	0.0777	22.22	25	12906	12952
12893	Sender with local receiver features	0.0293	22.22	25	12906	12952
12910	Weak	0.009	0	0	12916	12952
12894	Weak	-0.0055	0	0	12905	12952

Table B.34. Classification results for e3emhA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
14812	Strong receiver	-0.0127	25	66.67	14995	14884
14813	Strong receiver	-0.0072	33.33	66.67	14881	14884
14815	Strong receiver	-0.0004	25	50	14881	14884
14995	Weak	-0.0143	8.33	16.67	nan	14812
14987	Weak	-0.0205	0	16.67	nan	14816
14863	Strong receiver	-0.0134	25	66.67	14881	14884
14862	Strong sender-receiver	-0.0097	50	100	14881	14884
14816	Strong receiver	0.0007	25	50	14988	14884
14997	Strong receiver	-0.0042	41.67	83.33	14881	14884
14986	Strong sender-receiver	-0.0152	50	100	nan	14884
14879	Strong sender-receiver	-0.0109	50	83.33	14987	14884

Table B.34 Classification results for e3emhA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
14955	Strong receiver	0.0087	41.67	66.67	14881	14884
14993	Strong receiver	-0.0262	16.67	50	14881	14884
14880	Balanced sender-receiver	-0.0154	25	33.33	14989	14813
14861	Sender	0.0109	25	16.67	14994	14884
14988	Balanced sender-receiver	-0.0128	25	33.33	14957	14816
14881	Receiver	-0.0282	16.67	33.33	14989	14813
14883	Weak	-0.0141	16.67	16.67	nan	14861
14994	Receiver	-0.0253	8.33	16.67	nan	14955
14957	Weak	-0.0122	8.33	0	14987	14816
14859	Weak	-0.0157	0	0	nan	14861
14989	Weak	-0.0161	0	0	nan	14812
14958	Weak	-0.0081	8.33	0	14987	14816
14884	Sender	0.0209	41.67	0	14993	nan

Table B.35. Classification results for e3f11A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6616	Weak	0.1012	21.43	nan	6780	6723
6624	Weak	-0.0866	21.43	nan	6780	6723
6771	Weak	-0.1291	35.71	nan	6622	6723
6625	Weak	-0.1127	35.71	nan	6622	6723
6728	Weak	-0.1558	14.29	nan	6622	6723
6729	Weak	-0.1424	28.57	nan	6622	6723
6730	Weak	-0.1643	28.57	nan	6622	6618
6780	Receiver	-0.2007	14.29	nan	6622	6618
6611	Weak	-0.1549	14.29	nan	6622	6723
6612	Weak	-0.1586	14.29	nan	6622	6618
6715	Weak	-0.1242	14.29	nan	6622	6618
6613	Weak	-0.0339	21.43	nan	6728	6618
6614	Weak	-0.0454	7.14	nan	6728	6618
6716	Weak	-0.1274	0	nan	6728	6618
6718	Weak	-0.0364	21.43	nan	6780	6723
6719	Weak	-0.0846	0	nan	6728	6618
6615	Weak	0.0656	7.14	nan	6728	6618
6717	Weak	-0.0917	14.29	nan	6728	6618
6617	Weak	-0.0883	14.29	nan	6780	6723
6721	Weak	-0.0991	14.29	nan	6780	6723
6722	Weak	0.0891	21.43	nan	6780	6723

Table B.35 Classification results for e3f11A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6770	Weak	-0.0049	14.29	nan	6780	6723
6618	Sender	0.3897	28.57	nan	6780	nan
6619	Sender	0.2881	28.57	nan	6771	nan
6620	Weak	0.1935	28.57	nan	6624	nan
6723	Sender	0.4781	28.57	nan	6771	nan
6725	Sender	0.3132	28.57	nan	6624	nan
6724	Sender	0.4268	21.43	nan	6624	nan
6726	Sender	0.3261	21.43	nan	6624	nan
6622	Weak	-0.1694	0	nan	nan	6618
6623	Weak	-0.1658	0	nan	nan	6618

Table B.36. Classification results for e3g7wA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6533	Sender	0.0783	25	18.18	6878	6874
6875	Balanced sender-receiver	0.0479	33.33	27.27	6878	6874
6637	Weak	0.0332	16.67	9.09	6878	6874

Table B.36 Classification results for e3g7wA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6535	Receiver	-0.1664	16.67	36.36	6876	6874
6824	Strong receiver	-0.0843	41.67	54.55	6876	6874
6673	Receiver	-0.1768	8.33	27.27	6876	6874
6878	Receiver	-0.2341	8.33	36.36	6876	6874
6530	Receiver with local sender features	-0.0326	25	18.18	6878	6874
6871	Weak	-0.0239	8.33	18.18	6878	6874
6638	Weak	0.0744	8.33	9.09	6535	6874
6639	Weak	0.0578	16.67	9.09	6878	6874
6872	Receiver with local sender features	-0.0053	25	9.09	6878	6874
6873	Weak	-0.1562	16.67	9.09	6876	6874
6676	Weak	-0.1298	8.33	0	6876	6874
6677	Sender	0.336	41.67	9.09	6878	nan
6678	Sender	0.2572	33.33	18.18	6535	nan
6679	Sender	0.7953	33.33	9.09	6878	nan
6783	Sender	0.4441	33.33	18.18	6535	nan

Table B.36 Classification results for e3g7wA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6874	Sender	0.8899	33.33	0	6535	nan
6876	Receiver	-0.219	0	18.18	nan	6874
6532	Weak	0.0168	0	9.09	6673	6874
12238	Weak	0.0705	0	0	6876	6679
6877	Weak	0.0989	0	0	6876	6679

Table B.37. Classification results for e3ignA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
10446	Strong sender	0.0468	50	33.33	10447	nan
10456	Strong sender-receiver	0.0188	50	66.67	10447	nan
10440	Sender	0.0672	37.5	33.33	10447	nan
10439	Strong receiver	0.0236	37.5	66.67	10438	nan
10445	Sender	0.0672	25	33.33	10447	nan

Table B.37 Classification results for e3ignA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
10460	Balanced sender-receiver	0.0456	25	33.33	10447	nan
10435	Strong receiver	-0.0509	0	66.67	nan	10440
10436	Strong receiver	-0.0509	0	50	nan	10440
10447	Strong receiver	-0.0517	12.5	50	nan	10440
10438	Strong sender	-0.0029	50	0	10435	10440

Table B.38. Classification results for e3j80g1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
14812	Sender	0.0297	27.27	15.38	14861	14991
14987	Weak	-0.0367	0	23.08	14861	14816
14990	Balanced sender-receiver	-0.0053	36.36	38.46	14861	14991

Table B.38 Classification results for e3j80g1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15011	Sender	0.021	27.27	23.08	14861	14991
15167	Balanced sender-receiver	0.0071	45.45	30.77	14861	14991
14815	Sender	0.0008	36.36	23.08	14861	14991
15004	Weak	-0.0075	18.18	15.38	14883	14991
15005	Weak	-0.0119	9.09	7.69	14883	14991
14816	Sender	0.0446	18.18	0	14987	14991
14988	Weak	0.0262	9.09	7.69	14987	15005
14883	Weak	-0.014	9.09	15.38	14987	14812
14861	Receiver	-0.064	0	7.69	nan	14812
14989	Receiver	-0.0624	9.09	15.38	14816	14812
15090	Balanced sender-receiver	0.0059	27.27	30.77	15005	14812
15010	Balanced sender-receiver	0.0276	27.27	30.77	15004	14812
14991	Sender	0.069	27.27	15.38	15004	nan

Table B.39. Classification results for e3kffA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
14282	Receiver	-0.0097	11.11	33.33	14280	14286
14283	Strong sender-receiver	0.0084	55.56	50	14280	14313
14284	Strong sender-receiver	0.0105	77.78	66.67	14280	nan
14281	Weak	-0.0098	11.11	16.67	14280	14286
14286	Strong sender	0.0224	55.56	33.33	14280	nan
14313	Sender	0.0177	33.33	16.67	14280	nan
14285	Strong receiver	-0.0104	44.44	50	nan	14283
14280	Receiver	-0.0216	22.22	16.67	nan	14286

Table B.40. Classification results for e3m7nH2 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
9527	Sender	1.8673	33.33	nan	9526	nan
9545	Strong sender	0.4111	66.67	nan	9526	nan
9533	Weak	0.5815	33.33	nan	9547	nan
9530	Weak	0.2884	33.33	nan	9547	nan
9535	Weak	-0.4391	33.33	nan	nan	9527
9536	Weak	-0.5903	33.33	nan	nan	9527
9547	Weak	-0.6164	33.33	nan	nan	9527
9537	Strong sender	-0.5055	66.67	nan	nan	9527
9526	Weak	-0.5192	33.33	nan	9536	9527
9528	Weak	-0.4709	33.33	nan	9547	9527
9529	Weak	-0.317	33.33	nan	9547	9527

Table B.41. Classification results for e3osrA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6533	Weak	-0.4042	18.18	20	6878	6679
6875	Receiver with local sender features	-0.1729	27.27	20	6878	6679
6637	Weak	-0.7035	9.09	10	nan	6679
6642	Weak	1.1855	18.18	0	6878	6874
6535	Balanced sender-receiver	-0.1705	27.27	30	6637	6679
6824	Strong receiver	-0.1577	45.45	70	6637	6679
6673	Receiver	-0.4227	18.18	30	6637	6679
6878	Receiver	-0.6723	18.18	40	6637	6679
6643	Receiver	-0.2072	18.18	30	6637	6679
6871	Weak	-0.1572	9.09	20	6637	6679
6530	Receiver	-0.1408	18.18	40	6637	6679
6638	Weak	0.417	9.09	10	6637	6679
6639	Sender with local receiver features	0.0702	9.09	30	6637	6679
6872	Receiver	-0.0074	9.09	30	6637	6679

Table B.41 Classification results for e3osrA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6873	Receiver	-0.1024	9.09	30	6637	6679
6676	Weak	0.1801	9.09	0	6878	6679
6677	Sender	1.0511	45.45	10	6637	nan
6641	Weak	nan	36.36	10	nan	nan
6678	Weak	nan	36.36	10	nan	nan
6679	Sender	1.9803	36.36	10	6637	nan
6783	Weak	nan	36.36	10	nan	nan
6874	Sender	1.7722	36.36	0	6878	nan

Table B.42. Classification results for e3ozyA4 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
1214	Strong sender	0.3111	80	25	9624	nan
1217	Strong sender-receiver	0.0027	100	100	nan	nan
1216	Strong receiver	-0.3487	0	75	nan	1214

Table B.42 Classification results for e3ozyA4 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
9624	Strong receiver	-0.3624	0	75	nan	1214
9614	Strong receiver	-0.2093	20	75	nan	1214
1219	Weak	-0.1045	20	0	1216	1214
1220	Weak	-0.0532	20	0	1216	1214

Table B.43. Classification results for e3pk0A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Receiver with local sender features	-0.0671	28.57	0	2960	2966
2967	Weak	-0.0949	14.29	0	2960	2966
2968	Weak	-0.0633	14.29	0	2939	2966
2969	Weak	-0.0516	14.29	0	2967	2966
2960	Weak	-0.076	0	0	nan	2966
3383	Weak	-0.0583	0	0	nan	2966

Table B.43 Classification results for e3pk0A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2962	Strong receiver	-0.0514	14.29	50	2939	2966
2961	Weak	-0.0716	0	0	nan	2966
3241	Strong receiver	0.0366	28.57	100	2967	nan
3485	Strong receiver	0.0555	28.57	100	2967	nan
2964	Strong receiver	-0.0264	14.29	50	2939	2966
2965	Strong receiver	0.0771	28.57	100	2967	nan
2966	Strong receiver	0.1459	28.57	100	2967	nan

Table B.44. Classification results for e3q25A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6533	Sender	0.0245	41.67	0	6877	6679
6875	Strong sender	0.0108	50	0	6877	6679

Table B.44 Classification results for e3q25A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6637	Sender	0.0438	25	0	6877	6679
6535	Weak	-0.0482	16.67	0	6532	6874
6824	Strong receiver	-0.0356	25	100	6532	6679
6673	Weak	-0.044	8.33	0	6876	6874
6878	Weak	-0.0496	8.33	0	6876	6874
6871	Strong receiver	-0.0258	8.33	50	6878	6874
6530	Strong receiver	-0.028	8.33	100	6878	6679
6638	Strong receiver	-0.0253	8.33	100	6535	6874
6639	Strong receiver	-0.0246	8.33	100	6673	6874
6872	Strong receiver	-0.0264	8.33	100	6878	6874
6873	Strong receiver	-0.0362	0	50	6877	6874
6676	Weak	-0.028	0	0	nan	6533
6677	Sender	0.0814	33.33	0	6877	nan
6678	Sender	0.0645	33.33	0	6877	nan
6679	Sender	0.159	25	0	6877	nan

Table B.44 Classification results for e3q25A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6783	Sender	0.085	33.33	0	6877	nan
6874	Sender	0.1693	33.33	0	6877	nan
6876	Weak	-0.0386	0	0	nan	6679
6532	Weak	-0.0592	0	0	nan	6874
12238	Receiver	-0.0779	0	0	nan	6874
6877	Receiver	-0.0975	0	0	nan	6874

Table B.45. Classification results for e3rg1A6 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15670	Weak	0.2714	20	0	15510	15514
15509	Sender	0.5189	30	0	15510	15514
15595	Balanced sender-receiver	0.414	30	25	15510	15514
15518	Sender with local receiver features	0.1867	10	25	15481	15514
15593	Receiver	-0.6213	0	37.5	15507	15514

Table B.45 Classification results for e3rg1A6 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15519	Receiver	-0.675	0	37.5	15507	15514
15481	Receiver	-0.81	0	25	nan	15514
15596	Receiver	-0.4548	0	25	nan	15514
15592	Weak	-0.2272	0	12.5	15670	15514
15507	Receiver	-0.3812	0	25	15481	15509
15506	Weak	-0.0983	0	12.5	15481	15670
15508	Receiver	-0.3389	10	37.5	15481	15509
15520	Receiver	-0.5018	10	37.5	15481	15509
15510	Receiver	-0.523	10	25	15481	15509
15511	Balanced sender-receiver	0.1886	30	25	15481	15509
15521	Strong sender	0.6687	50	12.5	15481	nan
15514	Strong sender	0.8426	50	0	15481	nan
15515	Strong sender	0.6114	60	0	15481	nan
15516	Sender	0.4676	40	0	15481	15509
15594	Sender	0.4891	30	0	15481	15521
15591	Sender	0.3349	30	0	15481	15521
15517	Weak	0.1666	20	0	15481	15521

Table B.46. Classification results for e3rg1E1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15512	Weak	-0.0776	10	0	15510	15514
15509	Weak	0.1736	20	0	15510	15514
15595	Sender	0.25	40	0	15510	15514
15518	Sender	0.4337	30	0	15510	15514
15593	Receiver	-0.1051	20	33.33	15510	15514
15519	Strong receiver	-0.39	20	50	15507	15514
15481	Receiver	-0.6577	0	33.33	nan	15514
15596	Receiver	-0.4373	0	33.33	nan	15514
15592	Weak	-0.3486	0	16.67	15512	15514
15506	Weak	-0.0832	0	0	15481	15514
15507	Receiver	-0.2449	0	33.33	15592	15518
15508	Strong receiver	-0.2652	0	50	15481	15514
15520	Strong receiver	-0.3708	0	50	15481	15518
15510	Receiver	-0.4175	0	33.33	15481	15518
15511	Balanced sender-receiver	0.1278	40	33.33	15481	15518

Table B.46 Classification results for e3rg1E1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15513	Balanced sender-receiver	0.1264	40	33.33	15481	15518
15521	Strong sender	0.5929	60	16.67	15481	nan
15514	Sender	0.8184	30	0	15481	nan
15515	Sender	0.5084	40	0	15481	nan
15516	Weak	0.3321	20	0	15481	15518
15591	Weak	0.1703	10	0	15481	15514
15517	Weak	-0.1081	0	0	15481	15514

Table B.47. Classification results for e3ro3A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
13394	Strong receiver	-0.1944	0	66.67	nan	13643
13643	Strong sender	0.0542	50	33.33	13394	nan

Table B.47 Classification results for e3ro3A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
13689	Balanced sender-receiver	-0.0718	33.33	33.33	13394	13644
13386	Strong receiver	-0.0999	33.33	66.67	nan	nan
13390	Strong sender-receiver	-0.0505	50	100	13689	13643
13396	Strong receiver	-0.1126	0	66.67	nan	nan
13644	Strong receiver	-0.0441	16.67	66.67	13396	13643

Table B.48. Classification results for e3v30A2 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
13355	Sender	0.0236	44.44	0	13336	nan
13330	Strong sender	0.0219	77.78	33.33	13336	nan
13344	Strong sender	0.0141	77.78	33.33	13336	nan

Table B.48 Classification results for e3v30A2 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
13356	Strong sender	0.0101	77.78	33.33	13336	nan
13368	Balanced sender-receiver	0.0049	33.33	33.33	13336	nan
13340	Balanced sender-receiver	0.0046	33.33	33.33	13336	nan
13326	Weak	-0.0111	11.11	0	nan	13355
13333	Strong receiver	-0.0224	11.11	66.67	nan	13355
13334	Receiver	-0.0092	0	33.33	nan	13355
13328	Strong receiver	-0.0148	0	66.67	nan	13355
13327	Strong receiver	-0.0169	22.22	66.67	nan	13355
13336	Strong receiver	-0.0341	11.11	66.67	nan	13355
13323	Strong receiver	-0.0209	22.22	66.67	13326	13355
13357	Receiver	-0.023	22.22	33.33	13326	13355
13324	Weak	-0.0097	22.22	0	13326	13355

Table B.49. Classification results for e4c1dA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2921	Strong receiver	-0.0161	20	57.14	nan	2823
2825	Sender	0.0198	20	28.57	2826	2823
2824	Receiver	-0.0176	10	28.57	2826	2823
2893	Receiver	-0.0237	10	28.57	nan	2823
2894	Receiver	-0.0359	10	42.86	nan	2823
2839	Receiver	-0.0308	10	42.86	nan	2823
2826	Receiver	-0.0514	0	14.29	nan	2823
2862	Strong sender	0.0219	80	28.57	2894	nan
2820	Weak	0.0117	20	0	2824	2825
2821	Strong sender	0.0099	60	28.57	2894	nan
2822	Strong sender	0.0129	70	28.57	2893	nan
2823	Strong sender	0.0488	60	28.57	2824	nan

Table B.50. Classification results for e4c1hA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2921	Strong receiver	-0.011	22.22	62.5	nan	2822
2825	Receiver	-0.0001	11.11	37.5	2826	2895
2840	Sender with local receiver features	0.0054	11.11	25	2826	2862
2824	Sender with local receiver features	0.0161	22.22	25	2838	2895
2893	Receiver	-0.0003	22.22	37.5	2838	2895
2894	Strong receiver	-0.0122	22.22	50	nan	2820
2839	Strong receiver	-0.0101	22.22	50	nan	2895
2826	Receiver	-0.0217	11.11	25	nan	2820
2828	Balanced sender-receiver	0.0048	33.33	37.5	nan	nan
2895	Sender	0.0217	33.33	25	2823	nan
2820	Sender	0.024	33.33	25	2894	nan

Table B.50 Classification results for e4c1hA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2836	Strong sender	0.0104	55.56	37.5	2921	nan
2862	Strong sender	0.019	77.78	25	2894	nan
2821	Strong sender	0.0164	55.56	25	2894	nan
2822	Strong sender	0.0228	66.67	12.5	2921	nan
2823	Sender	0.0036	44.44	12.5	2826	2895
2838	Weak	-0.0067	22.22	12.5	2826	2824

Table B.51. Classification results for e4fn4A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Sender	0.0225	33.33	0	2967	2962
2967	Receiver	-0.0451	11.11	0	3314	2939
2968	Sender	0.043	33.33	0	3314	nan
2969	Sender	0.0156	33.33	0	2966	nan
2960	Strong receiver	-0.038	33.33	50	2966	2968

Table B.51 Classification results for e4fn4A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2962	Strong receiver	0.0265	44.44	100	2966	2968
2961	Strong receiver	0.0282	33.33	100	2967	2968
3314	Receiver	-0.0773	0	0	nan	2968
2964	Strong receiver	0.0076	0	50	2939	2968
2966	Receiver	-0.0552	0	0	nan	2939

Table B.52. Classification results for e4h4rA3 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Strong sender	0.1284	50	28.57	3624	nan
3519	Strong sender-receiver	0.1648	100	57.14	3624	2939
3624	Receiver	-0.2376	0	14.29	nan	3519
3499	Receiver	-0.135	0	28.57	nan	3519

Table B.53. Classification results for e4i3vA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
7107	Weak	-0.5913	0	12.5	nan	7088
7106	Weak	-0.5316	0	0	7069	7019
7104	Weak	-0.4756	0	0	7069	7088
7069	Receiver	-0.7625	0	12.5	nan	7088
7105	Receiver	-0.6642	0	0	nan	7088
7088	Sender	1.0356	29.41	0	7028	nan
7019	Weak	0.4827	17.65	0	7021	7026
7020	Weak	0.2623	23.53	0	7069	7026
7021	Weak	-0.3274	5.88	0	7069	7019
7100	Weak	-0.3892	5.88	0	7069	7019
7099	Receiver	-0.3971	5.88	25	7069	7019
7101	Receiver	-0.1354	11.76	25	7069	7026
7022	Sender with local receiver features	0.071	11.76	25	7028	7026
7023	Receiver	-0.1479	5.88	25	7069	7026
7024	Sender	0.9606	23.53	37.5	7027	7088
7025	Sender	0.7585	29.41	37.5	7027	7088
7102	Sender	1.1284	23.53	12.5	7028	7088

Table B.53 Classification results for e4i3vA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
7103	Sender with local receiver features	0.1127	23.53	25	7069	7088
7027	Receiver	-0.8842	5.88	12.5	7069	7024
7026	Sender	1.6845	23.53	12.5	7028	7088
7028	Receiver	-1.0844	0	0	7069	7026
7029	Sender	0.3186	29.41	12.5	7028	7026
7049	Receiver with local sender features	-0.0525	29.41	12.5	7028	7026
7030	Balanced sender-receiver	0.0225	35.29	25	7028	7026
7031	Weak	-0.2218	5.88	12.5	7028	7026
7051	Weak	0.063	0	12.5	7028	7088

Table B.54. Classification results for e4k5uB1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15904	Sender	0.0239	30	0	15914	15891
15301	Strong receiver	-0.0625	10	75	nan	15891
15350	Receiver with local sender features	-0.0066	30	0	15914	15891
15907	Receiver with local sender features	-0.0376	30	0	15586	15891
15435	Receiver with local sender features	-0.0123	30	0	15586	15891
15333	Receiver with local sender features	-0.0718	30	0	15301	15891
15421	Weak	-0.061	20	0	15586	15891
15590	Weak	-0.0279	20	0	15586	15891
15914	Receiver	-0.1218	0	25	nan	15891

Table B.54 Classification results for e4k5uB1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15680	Strong receiver	-0.0765	0	50	nan	15891
15723	Strong receiver	-0.0729	10	75	nan	15891
15586	Receiver	-0.1468	0	25	nan	15891
15351	Receiver	-0.0602	0	25	nan	15891
15434	Weak	-0.0189	0	0	nan	15891
15587	Weak	-0.0044	0	0	15586	15590
15330	Strong receiver	-0.0297	10	75	15586	15904
15881	Strong receiver	0.056	40	75	15586	nan
15809	Strong receiver	-0.0261	10	50	15586	15904
15347	Strong receiver	-0.0123	10	75	15586	15904
15875	Strong receiver	-0.0258	10	50	15586	15904
15471	Strong sender-receiver	0.0526	50	75	15586	nan

Table B.54 Classification results for e4k5uB1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15890	Strong sender-receiver	0.0798	70	75	15586	nan
15420	Strong receiver	0.0733	40	75	15586	15587
15452	Strong sender	0.1216	60	25	15586	nan
15891	Strong sender	0.1525	60	25	15586	nan
15555	Sender	0.1289	30	0	15586	nan
15343	Strong sender	0.1417	50	0	15586	15881

Table B.55. Classification results for e4kliA3 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
9358	Receiver	-0.2619	0	0	nan	9357
9371	Receiver	-0.2101	0	0	9360	9357
9361	Strong receiver	-0.1539	25	66.67	9358	9357

Table B.55 Classification results for e4kliA3 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
9360	Strong receiver	-0.0951	25	66.67	9358	9357
9365	Strong sender-receiver	0.0391	50	100	9358	nan
9362	Strong sender-receiver	0.1586	87.5	66.67	9358	nan
9363	Sender	0.1756	37.5	33.33	9358	nan
9370	Weak	-0.0016	12.5	0	9358	nan
9357	Strong sender	0.2229	75	33.33	9358	nan

Table B.56. Classification results for e4lscA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15325	Sender	0.0334	44.44	0	15794	nan
15609	Receiver	-0.043	11.11	33.33	nan	15817
15631	Receiver	-0.022	11.11	33.33	nan	15817
15663	Receiver	-0.0444	0	33.33	nan	15817

Table B.56 Classification results for e4lscA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15818	Strong receiver	-0.0312	0	100	nan	15325
15619	Receiver	-0.0167	0	33.33	nan	15325
15649	Receiver	-0.0051	0	33.33	nan	15816
15664	Strong receiver	-0.019	0	66.67	nan	15325
15620	Weak	-0.0068	0	0	nan	15816
15813	Receiver	-0.0205	0	33.33	15663	15325
15621	Strong receiver	-0.0253	0	66.67	15663	15325
15301	Strong receiver	-0.0258	0	66.67	15663	15325
15847	Sender	0.0044	0	0	15609	15325
15856	Strong receiver	-0.0009	33.33	66.67	15609	15325
15794	Strong receiver	-0.0355	0	66.67	15663	15325
15816	Strong receiver	-0.0137	22.22	66.67	15609	15325
15852	Receiver	-0.008	11.11	33.33	15609	15325
15817	Sender	0.0027	22.22	33.33	15609	15325

Table B.57. Classification results for e4mn8B1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15325	Strong sender	0.0628	62.5	0	15620	nan
15617	Strong receiver	-0.0294	12.5	66.67	nan	15857
15631	Strong receiver	-0.0185	12.5	66.67	nan	15857
15853	Strong receiver	-0.0287	12.5	100	nan	15325
15859	Strong receiver	-0.0346	12.5	100	nan	15325
15818	Strong receiver	-0.0223	12.5	100	nan	15325
15663	Receiver	-0.0269	0	33.33	nan	15857
15847	Receiver	-0.0141	0	33.33	nan	15857
15854	Strong receiver	-0.0247	0	66.67	nan	15325
15649	Weak	-0.0115	0	0	nan	15325
15664	Receiver	-0.0287	0	33.33	nan	15325
15844	Receiver	-0.0195	0	33.33	15617	15325
15301	Receiver	-0.0292	0	33.33	nan	15325
15855	Receiver	-0.0182	0	33.33	nan	15325
15646	Receiver	-0.0289	0	33.33	15663	15325

Table B.57 Classification results for e4mn8B1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
15856	Balanced sender-receiver	-0.0131	25	33.33	15617	15325
15620	Receiver	-0.0457	0	33.33	nan	15325
15621	Balanced sender-receiver	-0.0182	25	33.33	15617	15325
15860	Receiver	-0.0391	0	33.33	15617	15325
15816	Receiver with local sender features	-0.0124	25	0	15617	15325
15791	Weak	-0.0167	0	0	15663	15325
15793	Weak	-0.0111	0	0	15617	15325
15787	Weak	-0.0086	0	0	15617	15325
15857	Sender	0.0077	25	0	15617	15325
15817	Sender	0.0109	25	0	15617	15325

Table B.58. Classification results for e4pf4A1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
1829	Sender	0.2539	18.18	0	1835	1832
1837	Strong sender-receiver	0.2401	54.55	80	1835	1832
1838	Strong receiver	0.1036	27.27	80	1835	1832
1962	Strong receiver	0.0229	27.27	100	1835	1829
1963	Sender with local receiver features	0.2039	18.18	40	1835	nan
1977	Sender	0.2186	18.18	40	1835	nan
1826	Weak	-0.149	9.09	20	1835	1829
1916	Receiver	-0.1786	9.09	20	nan	1829
1973	Weak	-0.1309	18.18	20	nan	1829
1833	Sender	0.0127	27.27	0	1835	1977
1832	Sender	0.2889	27.27	0	1835	1963
1834	Receiver with local sender features	-0.0466	36.36	0	1916	1977

Table B.58 Classification results for e4pf4A1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
1974	Receiver with local sender features	-0.0193	27.27	0	1916	1977
1835	Receiver	-0.2495	9.09	0	1916	1832
1836	Receiver	-0.2131	9.09	0	1916	1832

Table B.59. Classification results for e4psxB1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
14940	Weak	0.0196	7.14	0	14937	14946
14945	Weak	0.0192	7.14	0	14816	15222
14946	Weak	0.0124	0	0	14816	15222
14942	Receiver	-0.0379	14.29	28.57	14946	15222
14944	Balanced sender-receiver	-0.0013	42.86	28.57	14941	15222
14812	Sender	0.0367	28.57	0	14942	14816
14813	Sender	0.0069	35.71	14.29	14937	15222
14937	Weak	-0.0332	7.14	0	15219	14946

Table B.59 Classification results for e4psxB1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
14939	Weak	-0.0512	7.14	0	nan	14946
15218	Weak	-0.0298	7.14	0	15219	14946
14938	Weak	-0.039	7.14	0	15219	14945
15220	Weak	-0.0394	7.14	0	nan	14946
15221	Weak	-0.048	14.29	14.29	nan	14946
14949	Receiver	-0.0554	0	0	nan	14946
14815	Receiver	-0.0618	0	0	nan	14946
14816	Receiver	-0.063	0	0	14812	14946
15219	Receiver	-0.072	0	0	14812	14946
14950	Strong receiver	0.0454	35.71	57.14	14942	nan
14861	Sender	0.0554	21.43	28.57	14942	nan
15222	Sender	0.0799	28.57	42.86	14942	nan
14951	Sender	0.0568	21.43	28.57	14942	nan
14941	Receiver	-0.0184	14.29	28.57	14939	14812

Table B.60. Classification results for e4qedA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Receiver with local sender features	-0.0495	27.27	20	2967	2966
2968	Receiver	-0.04	18.18	40	2960	2966
2967	Receiver	-0.0874	18.18	40	2960	2966
3380	Receiver	-0.0572	18.18	40	2939	2966
2960	Receiver	-0.0812	0	0	nan	2966
2962	Weak	-0.0447	9.09	0	2939	2966
2961	Weak	-0.0571	9.09	0	2939	2966
3485	Strong sender	0.0698	54.55	40	2967	nan
3379	Strong sender	0.0898	54.55	40	2967	nan
2964	Weak	0.0376	9.09	20	2967	2966
2965	Sender	0.0981	45.45	40	2967	nan
2966	Sender	0.1488	45.45	20	2967	nan

Table B.61. Classification results for e4qgeA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
12902	Receiver with local sender features	-0.0173	40	0	12916	12952
12885	Receiver with local sender features	-0.0144	40	0	12916	12952
12916	Receiver	-0.0353	0	0	nan	12929
12906	Strong receiver	-0.0281	20	50	12899	12929
12932	Strong receiver	-0.0228	20	50	12899	12910
12917	Receiver	-0.0191	20	25	12899	12910
12899	Receiver	-0.0375	0	25	nan	12929
12900	Receiver	-0.032	0	25	nan	12910
12901	Receiver	-0.0293	0	25	nan	12910
12903	Receiver	-0.0294	0	25	nan	12929
12952	Weak	-0.0217	0	0	12902	12891
12914	Receiver	-0.0105	0	25	nan	12929
12954	Receiver	-0.0012	0	25	12932	12894
12908	Sender	0.0094	20	25	12902	nan

Table B.61 Classification results for e4qgeA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
12891	Receiver	-0.0042	0	25	12952	12910
12915	Receiver	-0.0059	0	25	12952	12910
12909	Sender with local receiver features	0.0061	0	25	12916	12894
12929	Sender	0.0208	20	0	12916	nan
12910	Sender	0.0256	20	0	12916	nan
12894	Sender	0.0208	20	0	12916	nan

Table B.62. Classification results for e4qvhB1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6875	Balanced sender-receiver	-0.0767	27.27	25	6878	6874
6533	Balanced sender-receiver	-0.0617	27.27	25	6878	6874
6637	Receiver	-0.1755	18.18	25	6876	6874

Table B.62 Classification results for e4qvhB1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
6535	Receiver	-0.4377	9.09	25	6876	6874
6673	Receiver	-0.4851	9.09	25	6876	6874
6878	Receiver	-0.568	9.09	25	6876	6874
6871	Weak	0.4353	18.18	12.5	6535	6679
6873	Receiver	-0.2235	18.18	37.5	6878	6874
6676	Weak	-0.1058	9.09	0	6878	6874
6677	Sender	0.4595	36.36	12.5	6878	nan
6678	Sender	0.2881	27.27	12.5	6878	6871
6679	Sender	1.2784	27.27	0	6878	nan
6783	Sender	0.741	27.27	0	6535	6877
6874	Sender	1.5085	27.27	0	6535	nan
6876	Weak	-0.4366	0	0	nan	6874
6532	Weak	-0.1532	0	0	6673	6679
6877	Weak	0.0498	0	0	6673	6679

Table B.63. Classification results for e4urgA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
10467	Receiver	-0.0523	16.67	28.57	10435	10445
10446	Strong sender	0.1156	50	28.57	10438	nan
10456	Strong sender-receiver	0.0507	50	57.14	10438	nan
10440	Strong sender	0.1827	50	14.29	10438	nan
10439	Strong sender	0.1176	50	42.86	10438	nan
10445	Sender	0.1711	33.33	28.57	10438	nan
10460	Balanced sender-receiver	0.1298	33.33	28.57	10438	nan
10435	Receiver	-0.0842	0	42.86	nan	10440
10436	Receiver	-0.0822	0	42.86	nan	10440
10447	Receiver	-0.0432	16.67	28.57	10467	10440
10438	Receiver	-0.0814	33.33	28.57	10435	10440

Table B.64. Classification results for e4wecA1 domain.

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
2939	Receiver	-0.0979	11.11	0	nan	3493
2967	Strong receiver	-0.0817	11.11	50	2960	3493
2968	Strong receiver	-0.0508	11.11	50	2961	3493
3034	Strong receiver	-0.0484	11.11	50	2960	3493
2969	Strong receiver	-0.0517	11.11	50	2939	3493
2960	Weak	-0.0688	0	0	3472	3493
2961	Weak	-0.0555	0	0	nan	3493
2962	Strong receiver	-0.0576	11.11	50	2939	3493
3471	Strong sender-receiver	0.0502	66.67	50	2939	nan
2963	Strong sender-receiver	0.0674	77.78	50	2939	nan
3395	Strong receiver	-0.0442	11.11	50	2939	3493
3472	Weak	-0.0373	0	0	2939	3493
3313	Weak	-0.0467	0	0	2939	3493

Table B.64 Classification results for e4wecA1 domain. (cont.)

Theme ID	**Class**	**netTE of Theme**	**% Sources**	**% Sinks**	**Top Receiver Theme**	**Top Sender Theme**
3396	Strong receiver	-0.0341	11.11	50	2939	3493
2964	Strong receiver	-0.029	11.11	50	2939	3493
2965	Strong sender-receiver	0.1026	77.78	50	2939	nan
2972	Strong receiver	0.06	33.33	50	2939	2966
3493	Strong sender-receiver	0.185	66.67	50	2967	nan
2966	Sender	0.1643	44.44	0	2967	nan

APPENDIX C: FIGURE PERMISSION

The figures presented in this thesis have been reused in accordance with the publisher's policy on "reuse of author-generated content and graphics", as stated on the publisher's official website. These figures originate from a published article for which the copyright has been transferred to the publisher. The reuse of these figures complies with the applicable licensing terms outlined below.

Creative Commons License: This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. No additional permission is required to reuse this article.