

# Mapping the Allosteric Landscape of Full-Length Lyn Kinase:

Integrative Molecular Dynamics and Network Perspectives on  
Cancer-Associated Mutations

Inaugural Dissertation

zur

Erlangung des Doktorgrades

Dr. nat. med.

der Medizinischen Fakultät

und

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

**Mina Rabipour**

aus Babolsar, Iran

Druckerei: Universitäts- und Stadtbibliothek Köln (USB)

Druckort: Köln

2026



**Betreuer/in:**

Dr. Rocio Rebolledo-Rios

**Referent/in:**

Dr. Simon Pöpsel

Prof. Dr. Karsten Niefind

**Datum der mündlichen Prüfung:** 10.07.2026

## *For Those Who Stood By Me*

*To my beloved father, who longed to see this day.  
Though he is no longer here, his love and support remain with me forever.*

*To my beloved Simon, whose patience and devotion  
brought me peace and strength through every step of this journey.*

*This work is dedicated to you both, whose love and care shaped my life.*

# Contents

|                                                                               |             |
|-------------------------------------------------------------------------------|-------------|
| <b>List of Abbreviations</b>                                                  | <b>III</b>  |
| <b>List of Figures</b>                                                        | <b>VI</b>   |
| <b>List of Tables</b>                                                         | <b>VII</b>  |
| <b>Abstract</b>                                                               | <b>VIII</b> |
| <b>Zusammenfassung</b>                                                        | <b>IX</b>   |
| <b>Acknowledgements</b>                                                       | <b>X</b>    |
| <b>1 General Introduction</b>                                                 | <b>1</b>    |
| 1.1 Protein Kinases as Master Regulators of Cellular Signaling . . . . .      | 2           |
| 1.2 Src Family Kinases and the Structure, Function, and Regulation of Lyn . . | 6           |
| 1.3 Allosteric Regulation in Multidomain Kinases . . . . .                    | 12          |
| 1.4 Computational Approaches to Lyn Dynamics . . . . .                        | 13          |
| 1.5 Knowledge Gap and Research Objectives . . . . .                           | 15          |
| <b>2 Molecular Dynamics Methodology for the Study of Biomolecules</b>         | <b>17</b>   |
| 2.1 Molecular Dynamics . . . . .                                              | 20          |
| 2.1.1 Force Field Models for Biomolecular Systems . . . . .                   | 22          |
| 2.1.2 Optimization of $\text{Mg}^{2+}$ Interaction Parameters . . . . .       | 23          |
| 2.1.3 System Preparation and Equilibration . . . . .                          | 24          |
| 2.2 Analysis of Trajectories . . . . .                                        | 26          |
| 2.2.1 Structural Deviation and Fluctuation . . . . .                          | 26          |
| 2.2.2 Interatomic Distance and Angle Calculations . . . . .                   | 27          |
| 2.2.3 Cluster Analysis . . . . .                                              | 28          |
| 2.2.4 Catalytic Site Geometry and Stability . . . . .                         | 29          |
| 2.2.5 Principal Component Analysis . . . . .                                  | 30          |

|          |                                                                                              |            |
|----------|----------------------------------------------------------------------------------------------|------------|
| 2.2.6    | Dynamical Cross-Correlation Analysis . . . . .                                               | 31         |
| 2.2.7    | Interdomain Interface Analysis . . . . .                                                     | 32         |
| <b>3</b> | <b>Machine Learning Concepts and Methods for Structural Biology</b>                          | <b>34</b>  |
| 3.1      | Machine Learning In Structural Biology . . . . .                                             | 35         |
| 3.2      | Supervised Classification Learning . . . . .                                                 | 36         |
| 3.2.1    | Random Forest Classifier . . . . .                                                           | 37         |
| 3.2.2    | Performance evaluation . . . . .                                                             | 39         |
| 3.2.3    | Dealing with Imbalanced Data . . . . .                                                       | 41         |
| <b>4</b> | <b>Conformational Landscape and Dynamic Stability of Full-Length Lyn Kinase</b>              | <b>43</b>  |
| 4.1      | Introduction . . . . .                                                                       | 44         |
| 4.2      | Computational Methods . . . . .                                                              | 45         |
| 4.3      | Results and Discussion . . . . .                                                             | 49         |
| 4.4      | Conclusions . . . . .                                                                        | 74         |
| <b>5</b> | <b>Allosteric Communication Networks and Regulatory Connectors in Full-length Lyn Kinase</b> | <b>76</b>  |
| 5.1      | Introduction . . . . .                                                                       | 77         |
| 5.2      | Computational Methods . . . . .                                                              | 78         |
| 5.3      | Results and Discussion . . . . .                                                             | 82         |
| 5.4      | Conclusions . . . . .                                                                        | 104        |
| <b>6</b> | <b>Structural Determinants of Lyn Kinase Conformational Ensembles</b>                        | <b>107</b> |
| 6.1      | Introduction . . . . .                                                                       | 108        |
| 6.2      | Computational Methods . . . . .                                                              | 109        |
| 6.3      | Results and Discussion . . . . .                                                             | 110        |
| 6.4      | Conclusions . . . . .                                                                        | 117        |
| <b>7</b> | <b>Future Directions in Kinase Regulation</b>                                                | <b>118</b> |
|          | <b>Appendices</b>                                                                            | <b>121</b> |
|          | <b>Publications</b>                                                                          | <b>140</b> |
|          | <b>Bibliography</b>                                                                          | <b>141</b> |

# List of Abbreviations

|                |                                                         |
|----------------|---------------------------------------------------------|
| <b>AML</b>     | Acute Myeloid Leukemia                                  |
| <b>ATP</b>     | Adenosine Triphosphate                                  |
| <b>AUC</b>     | Area Under the Curve                                    |
| <b>BCR</b>     | B Cell Receptor                                         |
| <b>BFKs</b>    | BRK Family Kinases                                      |
| <b>CLL</b>     | Chronic Lymphocytic Leukemia                            |
| <b>CML</b>     | Chronic Myeloid Leukemia                                |
| <b>COSMIC</b>  | Catalogue Of Somatic Mutations In Cancer                |
| <b>CSK</b>     | C-terminal Src Kinase                                   |
| <b>DCCM</b>    | Dynamic Cross-Correlation Matrix                        |
| <b>DFT</b>     | Density Functional Theory                               |
| <b>ESP</b>     | Electrostatic Potential                                 |
| <b>FPR</b>     | False Positive Rate                                     |
| <b>ITAMs</b>   | Immunoreceptor Tyrosine-based Activation Motifs         |
| <b>ITIMs</b>   | Immunoreceptor Tyrosine-based Inhibitory Motifs         |
| <b>kNN</b>     | k-Nearest Neighbors                                     |
| <b>LJ</b>      | Lennard-Jones                                           |
| <b>MD</b>      | Molecular Dynamics                                      |
| <b>ML</b>      | Machine Learning                                        |
| <b>MM</b>      | Molecular Mechanics                                     |
| <b>MM/PBSA</b> | Molecular Mechanics/Poisson–Boltzmann Surface Area      |
| <b>MCC</b>     | Matthews Correlation Coefficient                        |
| <b>MSA</b>     | Multiple Sequence Alignment                             |
| <b>NF1</b>     | Neurofibromin 1                                         |
| <b>N-lobe</b>  | N-terminal lobe of the kinase domain                    |
| <b>C-lobe</b>  | C-terminal lobe of the kinase domain                    |
| <b>NPT</b>     | Constant Number of Particles, Pressure, and Temperature |
| <b>NVT</b>     | Constant Number of Particles, Volume, and Temperature   |
| <b>PAM</b>     | Partitioning Around Medoids                             |
| <b>PBC</b>     | Periodic Boundary Conditions                            |
| <b>PC1</b>     | First Principal Component                               |
| <b>PCA</b>     | Principal Component Analysis                            |
| <b>PDB</b>     | Protein Data Bank                                       |
| <b>PKIDB</b>   | Protein Kinase Inhibitor Database                       |
| <b>PME</b>     | Particle Mesh Ewald                                     |

|               |                                           |
|---------------|-------------------------------------------|
| <b>QM</b>     | Quantum Mechanics                         |
| <b>RF</b>     | Random Forest                             |
| <b>RMSD</b>   | Root Mean Square Deviation                |
| <b>RMSF</b>   | Root Mean Square Fluctuation              |
| <b>ROC</b>    | Receiver Operating Characteristic         |
| <b>RTKs</b>   | Receptor Tyrosine Kinases                 |
| <b>nRTKs</b>  | Non-receptor Tyrosine Kinases             |
| <b>SASA</b>   | Solvent Accessible Surface Area           |
| <b>SFKs</b>   | Src Family Kinases                        |
| <b>SH1</b>    | Src Homology 1 domain                     |
| <b>SH2</b>    | Src Homology 2 domain                     |
| <b>SH3</b>    | Src Homology 3 domain                     |
| <b>SLE</b>    | Systemic Lupus Erythematosus              |
| <b>SMD</b>    | Solvation Model based on Density          |
| <b>SMOTE</b>  | Synthetic Minority Oversampling Technique |
| <b>SVM</b>    | Support Vector Machine                    |
| <b>TCGA</b>   | The Cancer Genome Atlas                   |
| <b>TK</b>     | Tyrosine Kinase                           |
| <b>WT</b>     | Wild-Type                                 |
| <b>WT-ATP</b> | Wild-Type bound to ATP                    |
| <b>WT-DAS</b> | Wild-Type bound to Dasatinib              |

# List of Figures

|      |                                                                                               |     |
|------|-----------------------------------------------------------------------------------------------|-----|
| 1.1  | The human kinome tree . . . . .                                                               | 3   |
| 1.2  | Structural overview of a kinase domain . . . . .                                              | 4   |
| 1.3  | Phylogenetic and structural features of Src family kinases . . . . .                          | 7   |
| 1.4  | Domain organization and conformational states of full-length Lyn kinase. . . . .              | 9   |
| 1.5  | Disease-related consequences of Lyn kinase overexpression in B cells. . . . .                 | 11  |
| 1.6  | Allosteric communication in a multidomain kinase . . . . .                                    | 13  |
| 1.7  | Conceptual overview of analysis strategies applied to molecular dynamics simulations. . . . . | 15  |
| 2.1  | Schematic representation of MD simulations . . . . .                                          | 21  |
| 2.2  | Periodic Boundary Conditions . . . . .                                                        | 24  |
| 4.1  | Computational pipeline for modeling and simulating full-length Lyn kinase. . . . .            | 47  |
| 4.2  | Mutation frequencies and initial simulation model . . . . .                                   | 51  |
| 4.3  | Domain-resolved analysis of structural stability across Lyn kinase systems. . . . .           | 54  |
| 4.4  | Comparative residue-level flexibility changes across systems . . . . .                        | 57  |
| 4.5  | Dominant global motions captured by PCA. . . . .                                              | 59  |
| 4.6  | Heatmap of local binding site contacts in Lyn kinase systems . . . . .                        | 61  |
| 4.7  | ATP and Dasatinib binding in WT and mutant Lyn kinase systems. . . . .                        | 63  |
| 4.8  | Magnesium ions coordination in the Lyn kinase active site . . . . .                           | 65  |
| 4.9  | Distances between magnesium ions and coordinating residues . . . . .                          | 67  |
| 4.10 | Ligand deviation within the catalytic pocket of Lyn kinase. . . . .                           | 69  |
| 4.11 | Comparison of binding pocket surface areas across Lyn systems. . . . .                        | 73  |
| 5.1  | DCCM maps for Lyn kinase systems. . . . .                                                     | 83  |
| 5.2  | Three-dimensional representation of dynamic correlations . . . . .                            | 85  |
| 5.3  | Pairwise Pearson correlation coefficients between DCCMs . . . . .                             | 88  |
| 5.4  | Module composition and structural region purity . . . . .                                     | 90  |
| 5.5  | Spatial and network organization of the 44 allosteric connectors . . . . .                    | 93  |
| 5.6  | Lyn autoinhibition via the proline-rich SH3-linker interface . . . . .                        | 97  |
| 5.7  | Asciminib binding site in ABL1 and predicted allosteric connectors . . . . .                  | 100 |
| 5.8  | RMSF profiles of Lyn WT and inactive-like mutants . . . . .                                   | 103 |
| 6.1  | Receiver Operating Characteristic (ROC) curve for the classifier . . . . .                    | 111 |
| 6.2  | Feature importance based on Gini index and permutation score. . . . .                         | 112 |

|     |                                                                             |     |
|-----|-----------------------------------------------------------------------------|-----|
| 6.3 | Top-ranked geometric features of the Lyn catalytic core . . . . .           | 113 |
| A1  | Molecular structures of the single-point mutations and small molecules . .  | 122 |
| A2  | Comparison of structural quality between AlphaFold and homology models      | 123 |
| A3  | Structural comparison of ligand binding in additional Lyn kinase mutants.   | 123 |
| A4  | Three-dimensional representation of dynamic correlations in apo Lyn systems | 124 |
| A5  | Full-length multiple sequence alignment of Src family kinases . . . . .     | 125 |
| A6  | Cobimetinib binding to an allosteric pocket in MEK1 kinase . . . . .        | 126 |
| A7  | Sequence alignment of SH1 domains from Lyn and MEK1 . . . . .               | 127 |
| A8  | Sequence alignment of SH1 domains from Lyn and ABL1 . . . . .               | 127 |
| A9  | RMSF profiles of Lyn kinase across apo and ATP-bound systems . . . . .      | 128 |



# List of Tables

|     |                                                                          |     |
|-----|--------------------------------------------------------------------------|-----|
| 4.1 | Binding free energy estimates for Lyn systems . . . . .                  | 71  |
| 6.1 | Metrics used to evaluate Random Forest classifier . . . . .              | 111 |
| A1  | Overview of simulation systems . . . . .                                 | 129 |
| A2  | Naturally occurring mutations in Lyn kinase and conservation across SFKs | 129 |
| A3  | Structural and pathogenicity scores for selected mutations . . . . .     | 129 |
| A4  | Wilcoxon Signed-Rank Test Results for SH1 Domain RMSD Comparisons .      | 130 |
| A5  | Wilcoxon Signed-Rank Test Results for SH2 Domain RMSD Comparisons .      | 131 |
| A6  | Wilcoxon Signed-Rank Test Results for SH3 Domain RMSD Comparisons .      | 132 |
| A7  | Clustering summary statistics . . . . .                                  | 133 |
| A8  | Wilcoxon-based significance values for RMSD of ligand . . . . .          | 134 |
| A9  | p-values for Pearson correlation significance between systems . . . . .  | 134 |
| A10 | Properties of Lyn allosteric connectors . . . . .                        | 135 |
| A11 | Src family kinase inhibitors . . . . .                                   | 136 |
| A12 | Similarity of Src family kinases to Lyn (full-length) . . . . .          | 136 |
| A13 | Similarity of Src family kinases to Lyn (kinase domain) . . . . .        | 137 |
| A14 | Structural features selected for supervised classification . . . . .     | 137 |

# Abstract

Protein kinases regulate cellular signal transduction and are key therapeutic targets. Among Src family kinases (SFKs), Lyn plays a critical role in immune signaling and cancer but remains less well characterized. To probe its regulatory mechanisms, a full-length model of Lyn was constructed, comprising the SH3, SH2, and SH1 (kinase) domains. The conformational dynamics of this model were analyzed using molecular dynamics (MD) simulations, network-based analysis, and machine learning.

Apo wild-type, ATP-bound, dasatinib-bound, and six cancer-associated mutants (K275A, E290D/K, I364L/N) were simulated for a total of 78  $\mu$ s. Principal component and cross-correlation analyses revealed that ATP stabilizes domain interfaces and strengthens SH2–SH1 coupling. In contrast, dasatinib and oncogenic mutations weaken long-range correlations and partially decouple SH3 and SH2 modules from the catalytic domain. Local flexibility changes in the  $\alpha$ C helix and hinge region were found to align with Molecular Mechanics Poisson–Boltzmann Surface Area (MM/PBSA) estimates and ligand interactions.

Through network analysis, 44 dynamically central residues were identified, forming a collective allosteric connector spanning the SH3, SH2, SH1, and linker regions. These connectors mediate interdomain communication, preserve activation loop contacts, and anchor critical salt bridges, such as K275–E290 and K275–D385. The fragmentation of these networks under inhibitor binding or mutation reflects a disruption of regulatory communication.

A random forest classifier was trained on MD-derived features, including interdomain distances, regulatory spine integrity, and salt-bridge occupancies, to distinguish active-like from inactive-like ensembles, achieving high predictive accuracy (AUC = 0.98). Interdomain coupling and activation loop contacts were identified as primary contributors by feature importance analysis, consistent with network-derived insights.

Comparative sequence analysis across SFKs showed that many allosteric connectors are conserved, particularly in the activation loop,  $\alpha$ C helix, and SH2–SH1 linker. However, several Lyn-specific residues, such as P73 in the SH3 domain, linker residue I227, and K343, D346, and I475 within the  $\alpha$ E and  $\alpha$ F helices, were found to diverge across the family and serve as high-scoring connectors. Structural mapping revealed overlap between Lyn connectors and a known distant allosteric inhibitor binding pocket in ABL1 kinase (the Asciminib site). The  $\alpha$ E and  $\alpha$ F helices at the SH1C–SH2 interface emerged as Lyn-specific, flexible hotspots responsive to mutations, suggesting potential druggable regulatory sites beyond the traditional ATP cleft.

Together, these findings support the concept of selective allosteric modulation of Lyn kinase. Through the integrated use of long-timescale MD, network analysis, and machine learning, a general framework is provided for uncovering hidden regulatory nodes in kinases and guiding structure-based drug discovery. This approach is broadly applicable to dissecting allosteric control in multidomain signaling proteins across the kinome.

# Zusammenfassung

Proteinkinasen regulieren die zelluläre Signaltransduktion und stellen zentrale therapeutische Angriffspunkte dar. Innerhalb der Src-Familie-Kinasen (SFKs) spielt Lyn eine entscheidende Rolle bei der Immun-Signalisierung und in der Onkogenese, ist jedoch strukturell weniger charakterisiert. Um deren Regulationsmechanismen zu untersuchen, wurde ein Vollstruktur-Modell von Lyn konstruiert, welches die SH3-, SH2- und SH1- (Kinase-) Domänen umfasst. Die Konformationsdynamik dieses Modells wurde mittels Molekulardynamik-Simulationen (MD), netzwerkbasierter Analysen und maschinellem Lernen untersucht.

Apo-Wildtyp, ATP-gebundene und Dasatinib-gebundene Zustände sowie sechs krebsassoziierte Mutanten (K275A, E290D/K, I364L/N) wurden über insgesamt 78  $\mu$ s simuliert. Hauptkomponenten- und Kreuzkorrelationsanalysen zeigten, dass ATP die Domänen-Grenzflächen stabilisiert und die SH2-SH1-Kopplung stärkt. Im Gegensatz dazu schwächen Dasatinib und onkogene Mutationen weitreichende Korrelationen und entkoppeln die SH3- und SH2-Module teilweise von der katalytischen Domäne. Lokale Flexibilitätsänderungen in der  $\alpha$ C-Helix und der Scharnierregion korrelierten mit MM/PBSA-Abschätzungen und Ligand-Interaktionen.

Mittels Netzwerk-Analyse wurden 44 dynamisch zentrale Reste identifiziert, die ein kollektives allosterisches Verbindungsnetzwerk bilden, welches die SH3-, SH2-, SH1- und Linker-Regionen überspannt. Diese Verbindungen vermitteln die Kommunikation zwischen den Domänen, erhalten Kontakte der Aktivierungsschleife aufrecht und verankern kritische Salzbrücken wie K275-E290 und K275-D385. Die Fragmentierung dieser Netzwerke unter Inhibitorbindung oder Mutation spiegelt die Unterbrechung der regulatorischen Kommunikation wider.

Ein Random-Forest-Klassifikator wurde auf MD-basierten Merkmalen trainiert, um aktiv-ähnliche von inaktiv-ähnlichen Ensembles zu unterscheiden (AUC = 0,98). Die Merkmalsrelevanz-Analyse identifizierte die Kopplung zwischen den Domänen und Kontakte der Aktivierungsschleife als primäre Faktoren, was mit den Erkenntnissen der Netzwerkanalyse übereinstimmt.

Sequenzvergleiche innerhalb der SFK-Familie zeigten, dass viele allosterische Verbindungen konserviert sind, insbesondere in der Aktivierungsschleife, der  $\alpha$ C-Helix und dem SH2-SH1-Linker. Dennoch divergieren Lyn-spezifische Reste, wie P73 in der SH3-Domäne, der Linker-Rest I227 sowie K343, D346 und I475 in den  $\alpha$ E- und  $\alpha$ F-Helices, innerhalb der Familie und fungieren als zentrale Knotenpunkte. Strukturelles Mapping ergab Überschneidungen zwischen Lyn-Knotenpunkten und einer bekannten allosterischen Bindungstasche in der ABL1-Kinase (Asciminib-Bindungsstelle). Die  $\alpha$ E- und  $\alpha$ F-Helices an der SH1C-SH2-Schnittstelle erwiesen sich als Lyn-spezifische, flexible Hotspots, was auf potenzielle allosterische Regulationsstellen jenseits der ATP-Bindungstasche hindeutet.

Zusammenfassend unterstützen diese Ergebnisse das Konzept einer selektiven allosterischen Modulation der Lyn-Kinase. Die Integration von Langzeit-MD, Netzwerkanalyse und maschinellem Lernen bietet einen allgemeinen Rahmen zur Identifizierung verborgener regulatorischer Knotenpunkte und zur Unterstützung der strukturbasierten Wirkstoffforschung innerhalb des Kinoms.

## *Acknowledgements*

It has taken a long and changing road to reach this moment. Years ago, I left Iran, my family, my friends, and everything familiar, to build a different life. From Iran to Sweden, and from Sweden to Germany, each step brought its own challenges, but also helped shape who I am today. Two migrations, many difficult days, but a journey that was ultimately meaningful and deeply worth it.

There were painful moments that happened far from home. The day I had no chance to say goodbye to my grandma, and the day I could only whisper to my father through the phone as he quietly slipped away, remain the most difficult moments I carry with me. But life also brought me unexpected gifts. It was here in Germany that I found peace and love. I met my husband, whose support has been one of the most precious parts of this chapter. I am most grateful to you, *Simon*, for finding me, for standing by me, and for bringing light into my life when I needed it the most.

To my loving *Mom*, who gave me the strength to follow my dreams, even when it meant letting me go, thank you for your courage and your silent sacrifices. There were moments you needed me and I was not there. I carry that with me always, with deep love and gratitude. To my dearest brother *Yousef*, who has always protected me, even as a child, always placing me first, your quiet strength and steady presence have been a source of comfort through all these years. From the bottom of my heart, thank you for always standing by my side and for never letting me feel alone.

To *Elena*, my beloved friend who feels like someone I've always known, your calm presence has always brought me peace, and your friendship means more than words can express. To *Sezin* and *Andrea*, thank you for the kindness and warmth you shared during our time together in the lab, and for being there in moments that truly mattered. To my thoughtful colleagues *Michael*, *Mila*, *Heidi*, and *Olaf*, your genuine care and kind words, offered at moments when I needed them most but never expected them, meant more to me than you may realize.

To my academic supervisor, *Rocío*, thank you for your guidance throughout this journey, for the time you dedicated to my project, and for the knowledge I gained under your supervision. To my mentoring supervisors, *Simon* and *Kay*, thank you for your support, your time, and your belief in me. Your encouragement helped me stay focused and strong throughout this project.

To everyone I have mentioned, I wish you a future filled with joy, health, and success.

And to my father and my grandma, may you rest in peace. Your memory gives me strength, and your love is always with me.

# Chapter 1

## General Introduction

*You are the sun, disguised in a veil.  
You are the divine, hiding from yourself.  
If only you could lift that veil,  
you would see,  
even in your loneliest darkness,  
the astonishing light of your own being.*

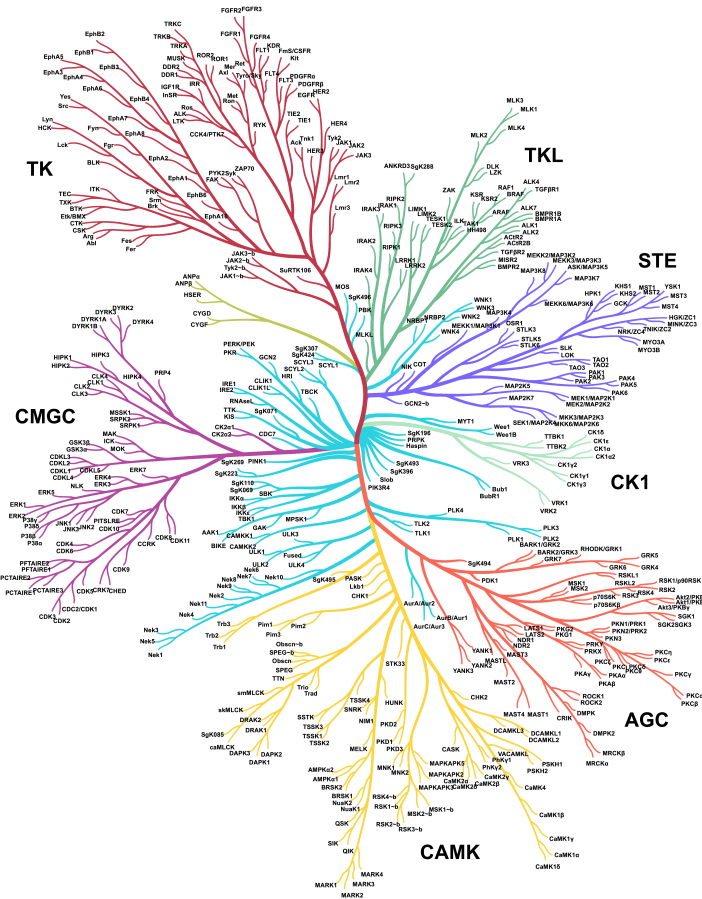
*– Hafiz (14th-century poet and mystic)*

## 1.1 Protein Kinases as Master Regulators of Cellular Signaling

Protein kinases are enzymes that selectively catalyze the transfer of a phosphate group to specific amino acid residues on protein substrates, a reaction known as phosphorylation. The catalytic mechanism relies on adenosine triphosphate (ATP) as the phosphate donor. During phosphorylation, the enzyme cleaves the  $\gamma$ -phosphate from ATP and covalently attaches it to the hydroxyl group of the target amino acids [1–3]. In contrast to other kinase families, such as lipid, carbohydrate, and nucleotide kinases, that typically act on non-protein biomolecules, protein kinases primarily phosphorylate protein substrates. Through this modification, they regulate a wide range of essential cellular processes, including signal transduction, cell-cycle progression, apoptosis, differentiation, metabolism, and survival [1, 4]. Phosphorylation often initiates complex signaling cascades that incorporate multiple regulatory mechanisms, enabling cells to precisely sense and respond to diverse external and internal stimuli. Even subtle alterations in kinase activity can lead to significant downstream effects, underscoring their function as key control points in cellular regulation [5].

The human genome encodes approximately 538 protein kinase genes, collectively referred to as the human kinome, which represents about 2 percent of all human genes [6, 7]. Phosphoproteomic studies estimate that up to one-third of the human proteome undergoes phosphorylation, highlighting the extensive reach of kinase-mediated regulation across diverse signaling pathways [8, 9]. Based on sequence similarity, biochemical properties, and the type of amino acid residues they phosphorylate, these kinases can be classified into groups, families, and subfamilies. A dendrogram known as the human kinome tree illustrates these evolutionary relationships by clustering kinases according to the divergence of their catalytic domain sequences (Figure 1.1).

Seven major groups are typically defined and color-coded in the tree: AGC (protein kinase A, G, and C families), CAMK (calcium/calmodulin-dependent protein kinases), CK1 (casein kinase 1 family), CMGC (CDK, MAPK, GSK3, CLK families), STE (homologs of yeast sterile kinases), TK (tyrosine kinases), and TKL (tyrosine kinase-like kinases) [10]. Each group reflects evolutionary relationships based on conserved catalytic domain sequences and regulatory motifs. Among these, TKs form a distinct and functionally important group that phosphorylates substrate proteins on tyrosine residues, a modification essential for regulating cell proliferation, differentiation, migration, and immune responses [11].



**Figure 1.1: The human kinome tree.** Classification of protein kinases into seven major groups based on sequence similarity of their catalytic domains. Each branch represents a kinase, and branch length reflects the degree of sequence divergence. The tree illustrates the diversity and evolutionary relationships within the kinase superfamily. Illustration reproduced courtesy of Cell Signaling Technology (www.cellsignal.com). The figure was reconstructed using *CORAL* [12], based on the original kinome classification scheme.

possible substrates and biological roles of poorly characterized kinases. In some cases, proximity within the tree has been used to infer functional similarities or predict signaling roles based on known members [17].

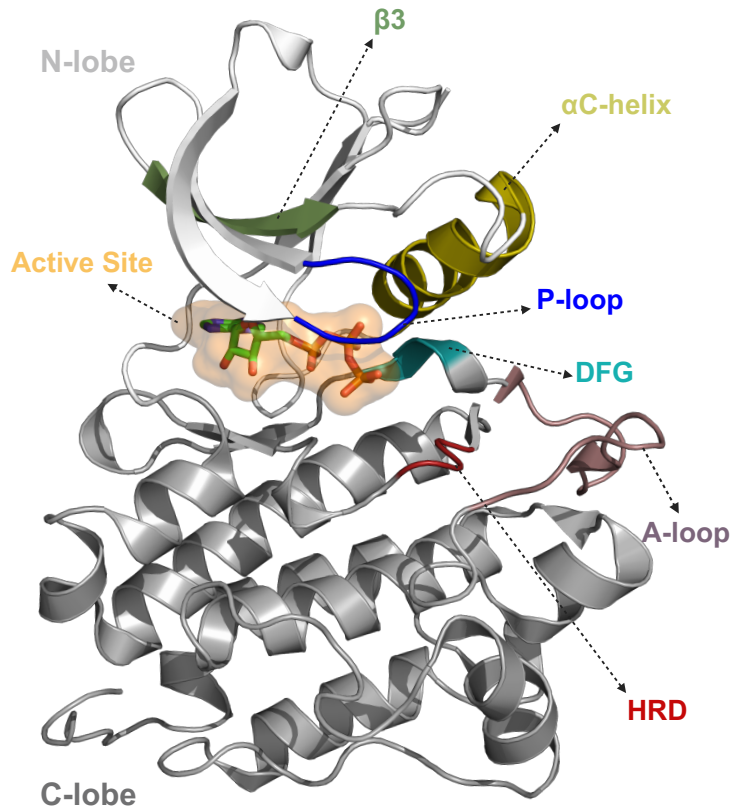
Despite acting on a wide variety of substrates, most protein kinases share a highly conserved catalytic domain architecture composed of two lobes: an N-terminal lobe (N-lobe) and a C-terminal lobe (C-lobe). The N-lobe features a five-stranded  $\beta$ -sheet, a glycine-rich P-loop (GxGxxG) that helps position ATP, and the dynamic  $\alpha$ C-helix, which is essential for aligning catalytic residues. It also contains a conserved lysine in the  $\beta$ 3 strand that participates in ATP binding. The C-lobe is predominantly  $\alpha$ -helical and includes the catalytic loop, the activation loop, and the substrate-binding site [18].

In contrast, kinases in the AGC, CAMK, CK1, CMGC, and STE groups primarily phosphorylate serine or threonine residues. While some dual-specificity kinases, such as MEK within the CMGC group, can phosphorylate both threonine and tyrosine residues, tyrosine phosphorylation is predominantly mediated by TK family members. The TKL group, though structurally related to TKs, largely contains serine/threonine kinases [13, 14]. This functional diversity underscores the evolutionary divergence and specialized roles of each kinase group in cellular signaling. In addition to these typical kinases, the remaining are classified as atypical protein kinases, which lack sequence similarity to conventional kinase families but are predicted or known to possess kinase activity, with some sharing a conserved structural fold [15, 16]. The evolutionary organization of the kinome tree may offer insights into the

A visual representation of the catalytic domain structural features is provided in Figure 1.2. Several highly conserved motifs, including the  $\beta 3$  strand, the HRD motif, and the DFG motif, form the structural and functional core of the active site. Their coordinated arrangement is stabilized by the regulatory spine, a hydrophobic network that spans both lobes and links ATP binding to active-site organization. Activation often involves phosphorylation of the activation loop, promoting a conformational shift that aligns these elements into a catalytically competent state [19–21].

Beyond structural rearrangements, the activity of protein kinases is finely regulated by several additional mechanisms that ensure precise control in response to cellular signals [22,23]. Phosphorylation often serves as a central switch: while modification at activation loops enhances catalytic function, phosphorylation at inhibitory sites can suppress activity. Specific sequence motifs, such as docking sites, facilitate interactions with substrates or regulatory partners. Scaffold and adaptor proteins further organize kinases into signaling complexes, defining when and where they function. Spatial control is also achieved through regulated subcellular localization, with kinases becoming active only upon reaching certain cellular compartments. In some cases, oligomerization into dimers or larger assemblies is necessary for full activation. Furthermore, binding of the modulators at sites distant from the active site (allosteric ligands) can trigger conformational changes that modulate kinase function. Together, these diverse regulatory layers enable kinases to act as context-dependent molecular switches within complex signaling networks [4,24].

The significance of protein kinases arises from their central role in maintaining cellular homeostasis and their frequent dysregulation in human disease. Aberrant kinase activity,



**Figure 1.2: Structural overview of the catalytic domain.** Schematic representation of a canonical protein kinase catalytic domain, highlighting the N-lobe and C-lobe, and the key structural motifs such as the  $\beta 3$  strand, P-loop,  $\alpha C$ -helix, HRD and DFG motifs, and the activation loop (A-loop). A bound ATP analog, shown in stick representation with the surrounding binding surface (highlighted in orange), illustrates the active site nestled between the two lobes. These conserved structural features are essential for coordinating ATP binding, substrate positioning, and catalytic function.



resulting from activating mutations, overexpression, chromosomal translocations, or misregulation by upstream signals, is a hallmark of various pathologies, particularly cancer, inflammatory disorders, and neurodegenerative conditions. These disease-related roles have positioned protein kinases as highly attractive targets for therapeutic intervention. The development of small-molecule kinase inhibitors, particularly ATP-competitive compounds, has significantly reshaped targeted cancer therapy by enabling the modulation of dysregulated signaling pathways. Several clinically approved inhibitors, such as imatinib for chronic myeloid leukemia (CML), capmatinib for non-small cell lung cancer, and tucatinib for HER2-positive breast cancer, have demonstrated therapeutic benefit [25–27]. Beyond oncology, kinase inhibitors have shown efficacy in treating immune-mediated and inflammatory diseases. For instance, tofacitinib and ruxolitinib inhibitors are used to manage rheumatoid arthritis and myelofibrosis, respectively [28, 29].

While these agents have shown substantial efficacy, many exhibit multi-target activity, leading to potential off-target effects that must be carefully considered in clinical application. Although ATP-competitive inhibitors have achieved major clinical successes across various malignancies, the highly conserved architecture of the ATP-binding site presents a fundamental challenge to selectivity [30]. Many kinases possess nearly identical ATP pockets that differ by only a few residues. This high degree of homology makes it difficult to design small molecules that bind selectively to a single kinase. As a result, many inhibitors affect multiple targets, increasing the risk of off-target effects and adaptive resistance through signaling network plasticity [31–33]. Moreover, ATP-competitive inhibitors are designed to fully block kinase activity, but this approach may not be ideal for kinases whose functions vary across different biological contexts. In such cases, therapeutic strategies that fine-tune kinase activity or selectively redirect signaling pathways may offer greater benefit [34, 35].

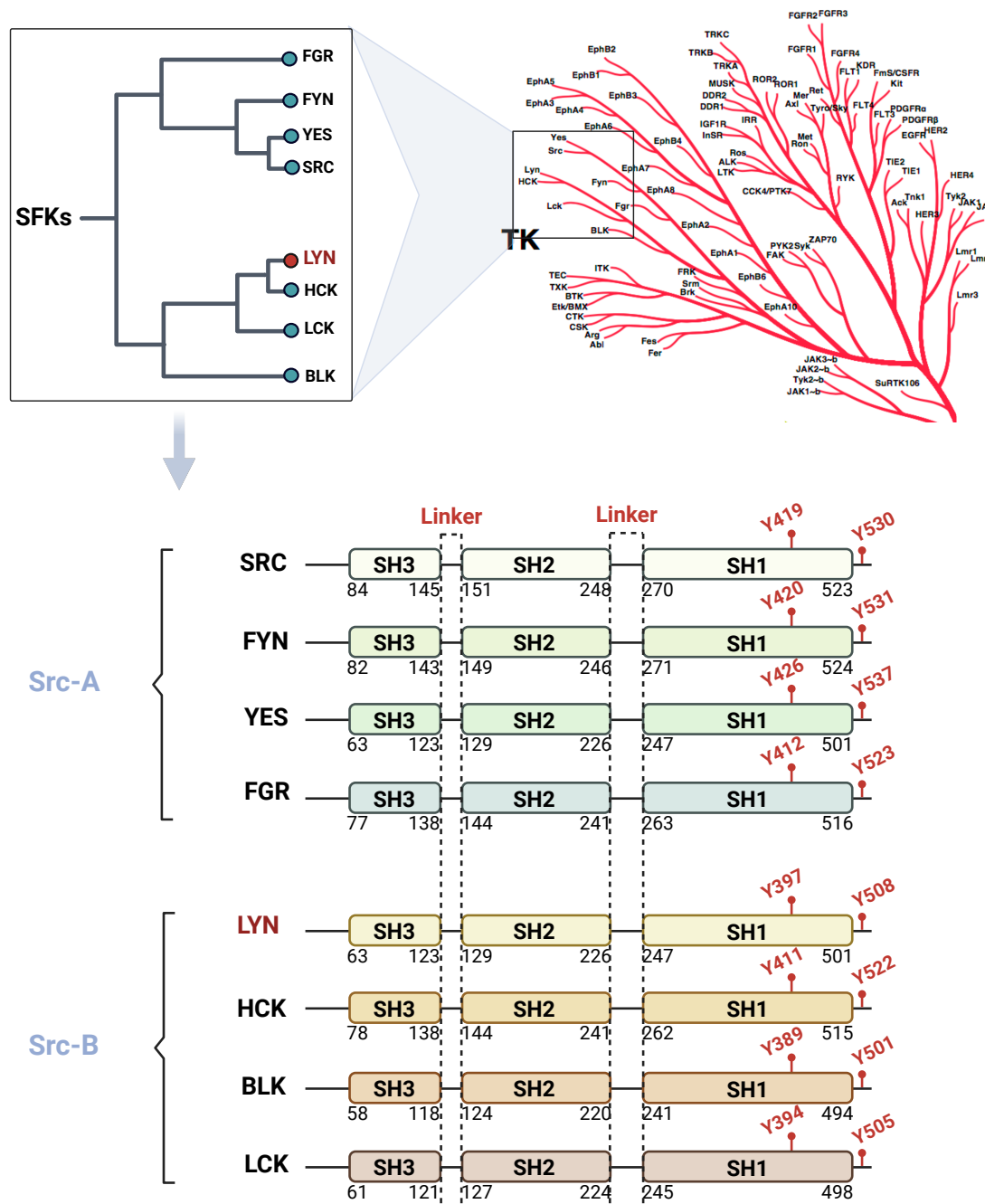
To overcome these challenges, alternative therapeutic strategies are being explored. One prominent example is the allosteric MEK inhibitor trametinib, which binds to a pocket adjacent to the ATP-binding site. By stabilizing MEK in an inactive conformation, trametinib prevents downstream activation without directly competing with ATP, allowing for greater selectivity [36]. Other approaches include the use of HSP90 inhibitors, which indirectly destabilize oncogenic kinases by interfering with their chaperone-mediated folding and stability [37]. Emerging applications extend to rare genetic disorders and neurological diseases. Selumetinib, a MEK inhibitor, was initially approved in 2020 for pediatric patients aged 2 years and older with neurofibromatosis type 1 (NF1) and On November 19, 2025, this indication was expanded by the FDA to include adult patients with the same condition. Beyond NF1, kinase inhibitors such as selumetinib are also under active investigation for non-oncologic diseases, including Alzheimer disease and cardiovascular disorders [38–40]. These emerging modalities highlight the ongoing need to refine kinase-targeted therapies based on biological context and molecular mechanism.

## 1.2 Src Family Kinases and the Structure, Function, and Regulation of Lyn

Among the various subclasses of protein kinases, TKs represent a particularly relevant group due to their roles in controlling key cellular decisions such as growth, survival, and differentiation [41]. These enzymes catalyze the phosphorylation of tyrosine residues on target proteins and are divided into two major categories: receptor tyrosine kinases (RTKs) and non-receptor tyrosine kinases (nRTKs). The human genome contains 90 unique TK genes, of which 58 produce membrane-spanning RTKs that respond to extracellular ligands. The remaining 32 encode nRTKs that mediate intracellular signaling downstream of activated receptors [42, 43].

Within the nRTK category, the Src family kinases (SFKs) represent a canonical group of cytoplasmic nRTKs. First identified through the viral oncogene *v-Src* from Rous sarcoma virus, this family includes nine members in vertebrates: Src, Yes, Fyn, Fgr, Lyn, Hck, Lck, Blk, and Yrk. In humans, eight are expressed (excluding Yrk) [44]. Based on phylogenetic and structural features, SFKs are divided into two major subfamilies: the Src-related subfamily, known as Src-A, that includes Src, Fyn, Yes, and Fgr, the LYN-related subfamily, known as Src-B, which includes Lyn, Hck, Lck, and Blk [45]. Frk (Fyn-related kinase), along with Brk and Srm, is structurally related to SFKs and shares overall domain similarity. However, due to the absence of the N-terminal myristoylation signal required for membrane localization and their distinct regulatory properties, these kinases are classified as BRK family kinases (BFKs) rather than classical members of the Src family [46, 47].

Although SFKs share high sequence conservation and a common domain organization, they display remarkable cell-type specificity and functional divergence. They are widely expressed in hematopoietic cells, neurons, epithelial tissues, and fibroblasts, where they regulate processes as diverse as cell adhesion, motility, proliferation, differentiation, and immune receptor signaling [4, 47]. The classification and evolutionary relationships among SFKs are depicted in Figure 1.3. To further illustrate the internal organization of the SFK members, this figure provides a comparison of their domain architectures and conserved regulatory residues across SFKs. This representation highlights their division into two major subgroups, based on sequence similarity and regulatory features, while emphasizing conserved elements such as the Src homology 3 (SH3) domain, the Src homology 2 (SH2) domain, two interconnecting linkers, and the catalytic Src homology 1 (SH1) or kinase domain, as well as critical tyrosine residues involved in activation and inhibition.



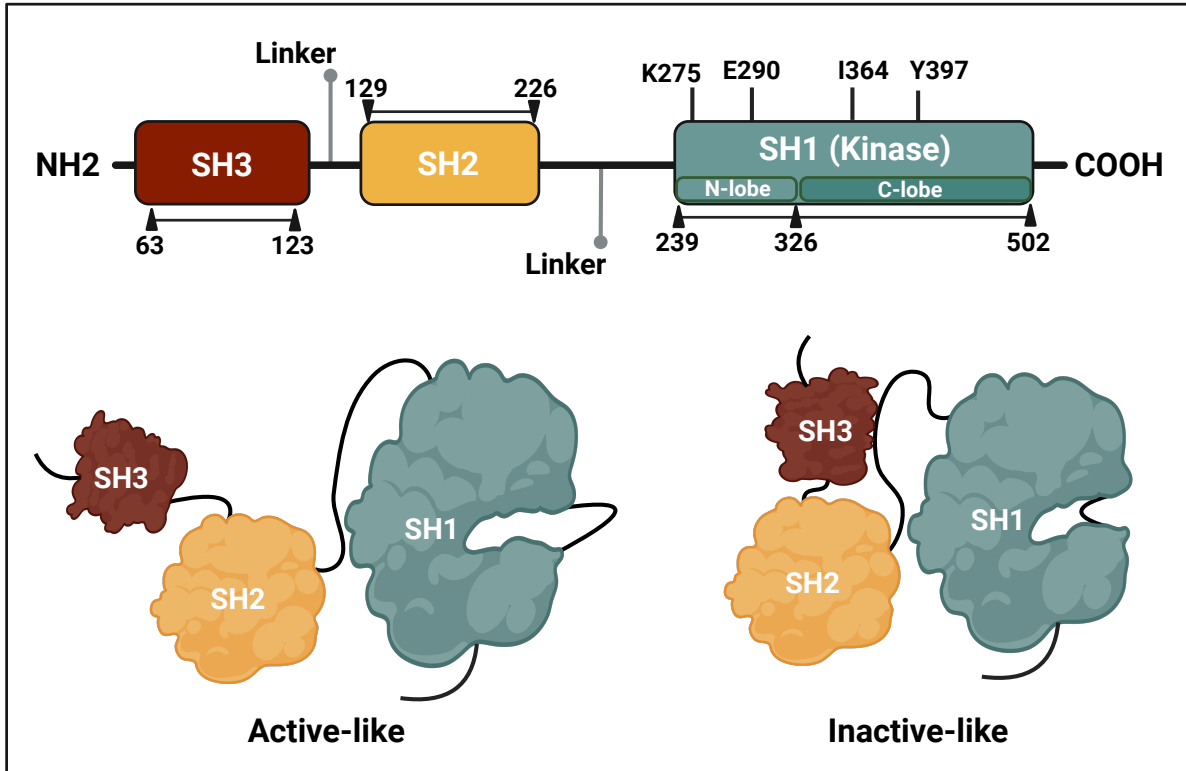
**Figure 1.3: Phylogenetic and structural features of SFKs.** This figure illustrates the classification and domain architecture of SFKs. The upper right panel shows the placement of SFKs within the TK group of the human kinome. The inset on the left highlights their phylogenetic relationships, separating the family into two subgroups: Src-A (SRC, FYN, YES, FGR) and Src-B (LYN, HCK, LCK, BLK), based on sequence similarity [1]. The lower panel depicts the conserved domain organization of each SFK, including the SH3, SH2, and SH1 (kinase) domains, along with interdomain linker regions. Key regulatory tyrosine residues are marked in red (e.g., Src Y419 and Y530; Lyn Y397 and Y508), reflecting conserved phosphorylation sites involved in activation and inhibition. This conserved organization underpins a common regulatory mechanism across SFKs despite their functional divergence.

Among SFKs, Lyn holds a distinctive position due to its pronounced ability to mediate both activating and inhibitory signaling in immune cells [48–50]. While most SFK members, such as Src, Fyn, and Yes, are primarily involved in promoting cell activa-

tion, proliferation, and migration [51], Lyn plays a dual regulatory role that is especially important in immune cells. This includes initiating signaling cascades downstream of immunoreceptors, as well as providing negative feedback to maintain immune balance and prevent overactivation. In contrast, the inhibitory functions of other SFKs are either limited, context-dependent, or not well established. While members like Src, Fyn, and Yes are mainly expressed in epithelial and neuronal tissues and often implicated in solid tumors, Lyn is predominantly expressed in immune cells involved in signaling, such as B lymphocytes, mast cells, and cells of the myeloid lineage. [49, 52].

Among all immune cell types, B lymphocytes exhibit the highest levels of Lyn expression, reflecting its critical involvement in B cell receptor (BCR) signaling [53]. Upon BCR engagement, Lyn acts as one of the initiating kinases by phosphorylating immunoreceptor tyrosine-based activation motifs (ITAMs) on the  $Ig\alpha$  and  $Ig\beta$  chains. This triggers downstream signaling cascades involving the PI3K/AKT and MAPK/ERK pathways. In parallel, Lyn also phosphorylates immunoreceptor tyrosine-based inhibitory motifs (ITIMs) on inhibitory receptors such as CD22 and  $Fc\gamma RIIB$  [54–56]. This regulatory balance distinguishes Lyn as a key modulator of immune cell behavior, integrating activating and inhibitory signals to control B cell activation, maintain immune tolerance, and prevent hyperactivation [57–59]. In contrast, Lyn expression is lower but still functionally relevant in other immune cells such as macrophages, monocytes, neutrophils, mast cells, and platelets, where it regulates inflammatory responses, adhesion, and migration. While Lyn is also detectable in some non-immune tissues, such as epithelial and neural cells, its expression in those contexts is relatively low, and its function is less clearly defined [54].

Consistent with other SFKs, Lyn adopts a modular domain architecture consisting of catalytic domain (SH1 or kinase), two regulatory domains (SH2 and SH3), and two interconnecting linkers (SH3–SH2 linker and SH2–SH1 linker) [60, 61]. This arrangement enables a tightly regulated conformational switch between inactive and active states. In the autoinhibited conformation, the SH2 domain binds to a phosphorylated tyrosine in the C-terminal tail, while the SH3 domain interacts with the SH2–kinase linker, stabilizing an inactive kinase fold. Activation occurs through dephosphorylation of the tail or competitive ligand binding that disrupts these intramolecular contacts, thereby realigning the catalytic motifs for full enzymatic activity [62, 63]. Figure 1.4 illustrates the modular architecture of Lyn and its structural transition between inactive and active conformations.



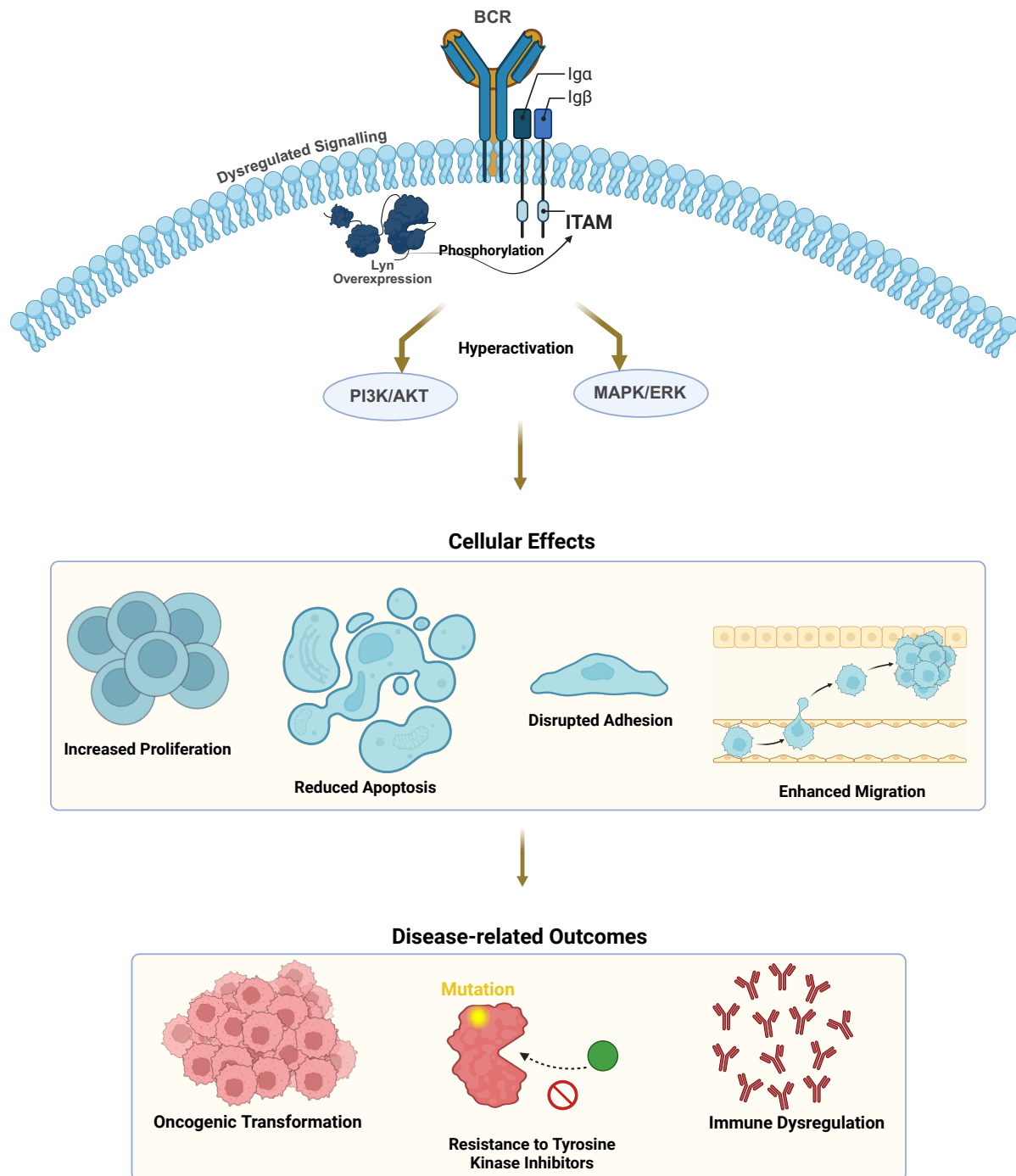
**Figure 1.4: Domain organization and conformational states of full-length Lyn kinase.** Top: Schematic representation of Lyn modular architecture, comprising SH3, SH2, and SH1 (kinase) domains, connected by flexible linkers. Key catalytic residues (e.g., K275 and Y397) are indicated. Bottom: Cartoon depiction of the conformational switch between the active (left) and inactive (right) states. In the inactive-like conformation, SH3 engages the SH2–SH1 linker and SH2 interacts with the phosphorylated C-terminal tail, stabilizing autoinhibition. In the active-like conformation, SH3 is displaced, allowing for an open, catalytically competent conformation.

In human Lyn kinase, autophosphorylation of Lyn Y397 (corresponding to Src Y419) within the activation loop induces a conformational change that enhances nucleotide binding and aligns the catalytic residues for phosphoryl transfer, thereby increasing kinase activity. In contrast, phosphorylation of the C-terminal regulatory site Lyn Y508 (corresponding to Src Y530) by C-terminal Src kinase (Csk) stabilizes a closed, inactive conformation. In this state, the SH2 domain binds intramolecularly to the phosphorylated tail, while the SH3 domain engages the SH2–kinase linker, effectively locking the enzyme in an autoinhibited form. This built-in autoregulatory mechanism highlights how the non-catalytic SH2/SH3 modules exert long-range allosteric control over the SH1 catalytic domain, enforcing precise temporal and spatial regulation of Lyn signaling function [58,64]. This tightly controlled on–off mechanism is essential for maintaining normal immune signaling thresholds. However, when Lyn is overexpressed, the regulatory balance between activation and inhibition can be disrupted. The high expression and functional importance of Lyn in B cells underscore its relevance in B cell–related disorders, including autoimmunity and hematologic malignancies [65,66]. The consequences of disrupted Lyn function are most apparent in disease settings. In autoimmune conditions like systemic lupus erythematosus (SLE), defective negative signaling through Lyn leads to B cell

hyperactivation and loss of self-tolerance. In hematologic malignancies such as chronic lymphocytic leukemia (CLL) and acute myeloid leukemia (AML), Lyn is often overexpressed and constitutively active, promoting continuous survival signaling, resistance to apoptosis, and dysregulated interactions with the tumor microenvironment [65,67]. These aberrant signals not only drive malignant transformation but also reinforce drug resistance and immune dysfunction.

To visualize these interconnected roles, Figure 1.5 illustrates the downstream consequences of Lyn overexpression. The schematic representation integrates molecular events, such as phosphorylation of ITAMs and subsequent hyperactivation of the PI3K/AKT and MAPK/ERK pathways, with key cellular phenotypes, including increased proliferation, apoptosis resistance, disrupted adhesion, and enhanced migration. These aberrant cellular behaviors ultimately contribute to disease-related outcomes such as oncogenic transformation, immune dysregulation, and resistance to tyrosine kinase inhibitors [68]. While the figure highlights BCR signaling as a representative context, these mechanisms are not restricted to B cells alone. Lyn plays a critical regulatory role in multiple immune cell types, where its overexpression can similarly perturb signaling thresholds and immune balance. However, the effects of Lyn are especially significant in B cells, where it shows the highest expression levels. In these cells, the consequences of Lyn dysregulation are well characterized and have been strongly associated with disease [48,69].

Although several ATP-competitive inhibitors targeting SFKs, such as bosutinib and PP2, have demonstrated therapeutic efficacy in both pre-clinical and clinical settings, these compounds often suffer from limited selectivity and carry a high risk of off-target effects due to the conserved nature of the ATP-binding pocket across kinases [70,71]. This presents a major limitation for efforts aimed at specifically modulating Lyn activity without broadly inhibiting other kinases. Dasatinib, a dual Src/Abl inhibitor, binds the ATP-binding site of Lyn with high affinity and is one of the most widely used SFK inhibitors in both research and clinical applications [72]. However, its broad activity profile can interfere with multiple kinases beyond Lyn, complicating its therapeutic precision [73]. One notable alternative strategy involves indirect modulation of Lyn through protein stability pathways. For instance, the HSP90 inhibitor NVP-BEP800 has been reported to reduce Lyn protein levels and inhibit the growth of B cell acute lymphoblastic leukemia by interfering with NF- $\kappa$ B signaling [74]. These findings highlight the therapeutic potential of non-ATP competitive approaches that bypass the challenges of targeting the kinase active site directly.



**Figure 1.5: Disease-related consequences of Lyn kinase overexpression in B cells.** Lyn overexpression enhances ITAM phosphorylation on the BCR complex, resulting in the hyperactivation of downstream PI3K/AKT and MAPK/ERK signaling cascades. This sustained signaling drives altered cellular behaviors, including increased proliferation, reduced apoptosis, disrupted adhesion, and enhanced migration. These changes culminate in disease-associated outcomes such as oncogenic transformation, resistance to tyrosine kinase inhibitors, and immune dysregulation, as observed in hematologic malignancies and autoimmune disorders.

Considering these insights, SFKs represent a conserved yet functionally diverse group of signaling enzymes, and Lyn exemplifies the complex interplay between signal amplifica-

tion and inhibition. Its unique role in immune regulation and leukemogenesis highlights both the therapeutic promise and the challenges of targeting kinases with dual regulatory functions. A deeper understanding of Lyn structural organization, domain communication, and conformational dynamics is therefore essential. Insights into how its regulatory modules cooperate, how residues propagate signals across domains, and how disease-associated perturbations shift these equilibria will be critical for designing interventions that modulate Lyn with precision rather than complete inhibition [75]. Such knowledge has implications not only for basic kinase biology but also for the development of therapeutic strategies that selectively reshape aberrant signaling while preserving essential physiological control. In this sense, Lyn serves as both a model for studying allosteric regulation in multidomain kinases and a compelling target whose complexity mirrors the challenges of modern precision medicine.

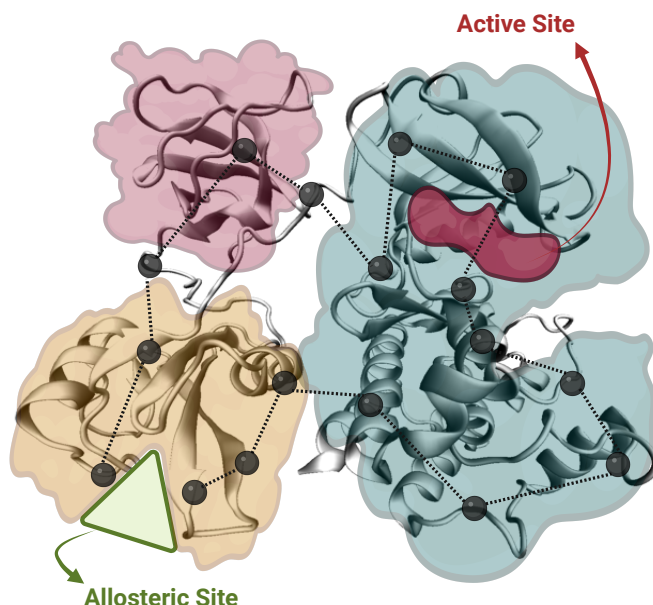
### 1.3 Allosteric Regulation in Multidomain Kinases

Domains are independently folded units within a protein, each capable of performing distinct structural or functional roles. In multidomain kinases, such modular organization enables complex regulatory mechanisms through coordinated domain–domain interactions. Allostery refers to the process by which perturbations at one site of a protein, such as ligand binding, post-translational modifications, or conformational shifts, alter the activity at a distant functional site [76]. In protein kinases, allostery enables spatial and temporal integration of upstream cues to precisely control phosphorylation activity. These principles are illustrated in Figure 1.6, where an allosteric binding site communicates with the active site through an internal network of residues. These networks are composed of dynamically coupled residues that bridge distant domains and enable long-range regulation of catalytic activity. Such allosteric networks represent the structural foundation for fine-tuned control in multidomain kinases. At the molecular level, this regulation relies on a network of non-covalent interactions, including hydrogen bonds, salt bridges, and side-chain contacts. Certain residues act as communication connectors, transmitting allosteric conformational perturbations across the protein and coordinating conformational transitions that influence catalytic output [77].

This is especially relevant for SFKs like Lyn, where signaling output is governed not only by catalytic activity but also by long-range communication between regulatory and catalytic domains. The SH3 and SH2 domains and their interconnecting linkers dynamically communicate with the SH1 domain [78]. Inactive conformations involve intramolecular interactions that constrain kinase activity, while activation disrupts these contacts to permit proper catalytic alignment (see Figure 1.4). This flexible interaction between domains enables the kinase to sample a spectrum of conformations rather than shifting strictly between “on” and “off” states. As a result, structural changes at distant regions



can subtly adjust overall activity, allowing the kinase output to be finely regulated rather than absolutely switched [79].



**Figure 1.6: Allosteric communication in a multidomain kinase.** Schematic representation of a kinase protein illustrating the spatial relationship between the allosteric site (green) and the catalytic active site (red). The protein is organized into modular domains, each shown in a different color. The allosteric site represents a distal pocket where regulatory ligands may bind, away from the catalytic center. Black spheres connected by dashed lines indicate residues forming an allosteric network, a dynamic, non-covalent interaction pathway that transmits conformational changes across domains. This intramolecular network enables communication between distant regions, allowing signals from the allosteric site to modulate activity at the active site and coordinate kinase function across the full protein structure.

In disease contexts, such as CLL, disruption of this allosteric network contributes to aberrant Lyn activation [67]. Therapeutically, allosteric regulation presents a promising opportunity for kinase targeting. Unlike ATP-competitive inhibitors, which frequently bind conserved ATP pockets and affect multiple kinases, allosteric inhibitors interact with less conserved regions. This improves selectivity, reduces toxicity, and allows more precise modulation of signaling pathways [31,80]. Clinically approved examples include trametinib, which targets MEK1/2, and asciminib, an ABL kinase allosteric inhibitor that binds a site distinct from the ATP pocket [81, 82]. Given the reliance of Lyn on allosteric regulation, selectively targeting its regulatory interfaces and intramolecular network connectors offers a more precise therapeutic strategy than conventional ATP-competitive inhibitors such as dasatinib. [49].

Exploring these mechanisms not only enhances understanding of Lyn kinase physiological role in immune homeostasis but also informs the design of next-generation therapeutics capable of selectively modulating kinase function.

## 1.4 Computational Approaches to Lyn Dynamics

Identifying allosteric sites remains a major challenge, particularly when such regions are transient, conformationally masked, or poorly resolved in experimental structures. While crystallography and cryo-electron microscopy have transformed structural biology, they primarily capture static snapshots that often miss the dynamic flexibility essential for

function. This limitation is especially relevant for multidomain kinases like Lyn, where interdomain interactions and long-range allosteric coupling are inherently dynamic processes. The absence of a complete full-length structure of Lyn further complicates conventional structure-based analysis [58].

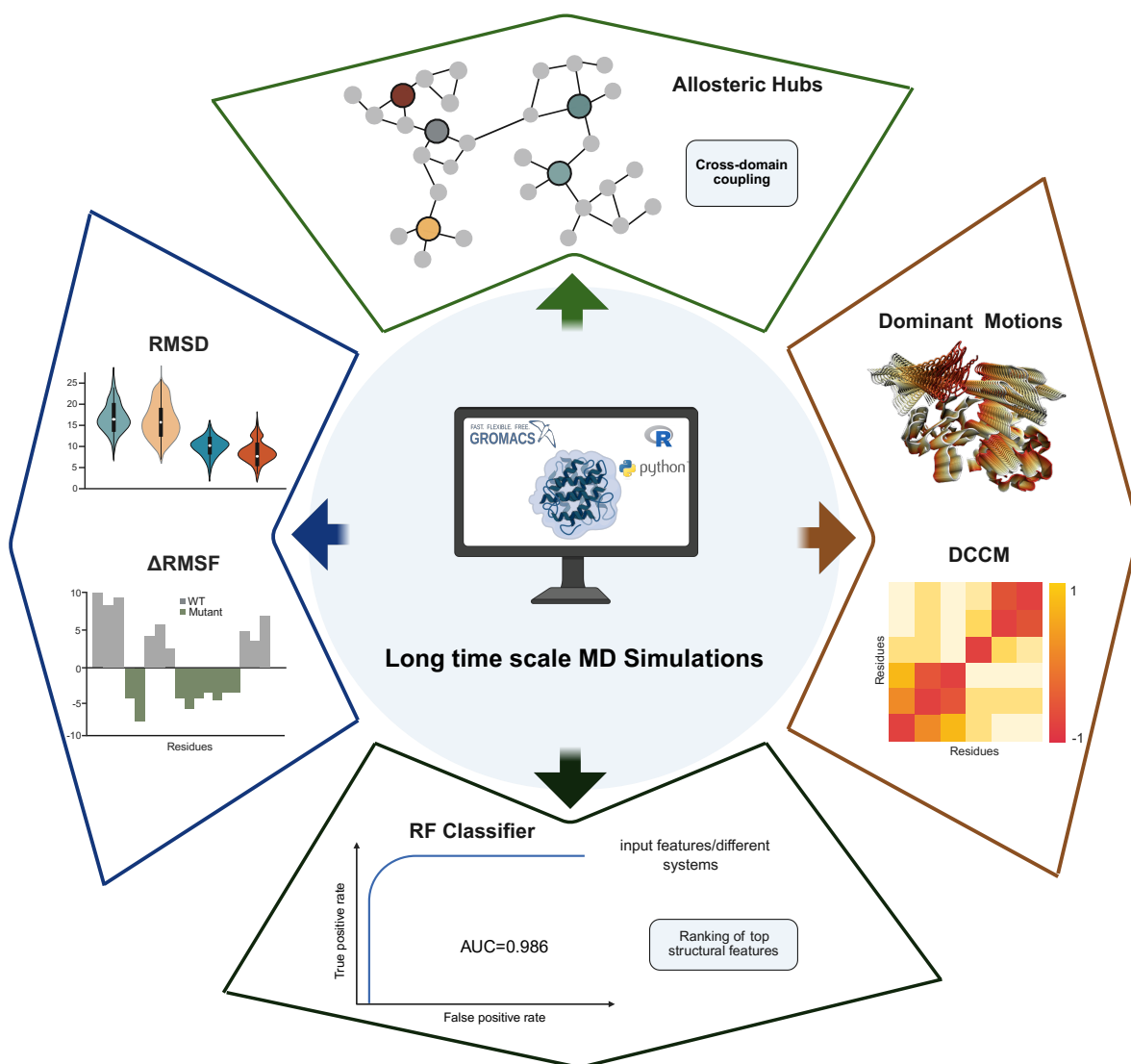
Homology modeling offers a valuable solution by enabling the construction of full-length structural models, which can be further investigated using simulation-based techniques to assess conformational stability, domain crosstalk, and potential allosteric sites. In particular, molecular dynamics (MD) simulations provide a powerful complementary approach by tracking atomic motions over time and revealing dynamic features inaccessible to static methods [83].

The power of MD lies in its ability to reveal hidden dynamic behaviors, transient conformations, and communication pathways that are often invisible in crystallographic data. For protein kinases, MD has been instrumental in uncovering alternative conformations of the activation loop, DFG motif rearrangements, domain movements, and interactions between regulatory elements such as the SH2 and SH3 domains [84]. In particular, MD can capture rare but functionally critical events such as domain opening, allosteric pocket formation, and interdomain uncoupling, which may serve as targets for selective inhibition or modulation.

Analysis of the simulation data helps extract biologically meaningful patterns from high-dimensional data. Statistical tools and structural metrics can quantify motions, identify correlated fluctuations, and compare conformational states. In recent years, network-based and machine learning (ML) approaches have further expanded the ability to interpret complex dynamics and recognize regulatory features within large proteins [85]. An overview of the integrated computational strategy is illustrated in Figure 1.7.

Simulating Lyn in apo, ligand-bound, and mutant forms reveals how perturbations propagate across domains and alter functional states. Given these considerations, computational approaches, anchored by MD simulations and enriched by statistical and network-based analyses, provide a powerful framework for studying protein kinase dynamics [86, 87].

As computational power and algorithmic sophistication continue to grow, these tools will play an increasingly central role in both basic science and drug discovery efforts targeting complex, allosterically regulated enzymes like Lyn. Such insights not only advance the fundamental understanding of kinase biology but also guide the rational design of next-generation therapeutics that achieve selectivity through dynamic, allosteric modulation. A detailed description of the computational methodologies employed in this work is provided in Chapters 2 and 3.



**Figure 1.7: Conceptual overview of analysis strategies applied to molecular dynamics simulations.** The schematic outlines key computational approaches used to extract structural and dynamic information from MD trajectories. These include classical structural metrics, statistical and correlation-based analyses, network modeling of residue communication, and ML methods for state classification.

## 1.5 Knowledge Gap and Research Objectives

Despite decades of work on SFKs, key aspects of Lyn structure and regulation remain poorly characterized. No experimentally determined structure exists for the full-length protein, and available crystallographic data are limited to isolated domains that often lack flexible loops and terminal segments, including the activation loop. These static models fail to capture the coordinated structural changes required for transitions between active and inactive states or to explain how regulatory domains influence the catalytic core. As a result, the effects of ligand binding and disease-associated mutations on Lyn dynamic behavior and interdomain coupling remain unresolved.

Drug development strategies have traditionally targeted the conserved ATP-binding cleft shared among protein kinases. However, achieving selectivity at this site remains a major challenge, particularly within the Src family, due to the high structural similarity of their ATP-binding pockets. As a result, limited attention has been given to the distinct modular regulatory architecture of Lyn kinase or its therapeutic potential. Alternative regulatory regions, including non-canonical binding pockets and interdomain interfaces, have not been thoroughly characterized, yet they may offer a promising route for selective modulation.

To address these unresolved questions, the study is guided by the following objectives:

1. To characterize how ligand binding and single-point mutations impact the conformational dynamics of Lyn.
2. To identify long-range interdomain interactions and key residues that mediate allosteric regulation.
3. To develop a machine learning classifier that evaluates which structural and dynamic features most effectively differentiate functional states.

Taken together, these aims are designed to enhance the understanding of structural flexibility and allosteric communication in Lyn kinase. The integration of molecular dynamics simulations, network-based analysis, and machine learning classification enables the identification of key structural and dynamic features that may support the development of more selective strategies for modulating Lyn function than those targeting the conserved ATP-binding site.

## Chapter 2

# Molecular Dynamics Methodology for the Study of Biomolecules

*Nature! We are surrounded by her, we are embraced by her.  
Powerless to separate ourselves from her, and powerless to  
penetrate beyond her,  
we are drawn into the circle of her dance, She speaks  
unceasingly with us,  
and betrays not her secret,  
All things are she, and she in all things.*

*– Goethe (18th-century poet and philosopher)*

## Parameterization and Molecular Simulation Framework

Understanding biological macromolecules and their interactions requires methods capable of describing both the electronic and atomic levels of organization. Biomolecules such as proteins, nucleic acids, and their complexes display a remarkable range of structural flexibility and chemical reactivity that cannot be captured by a single theoretical approach. To comprehend how local chemical events, such as charge redistribution or hydrogen bonding, influence large-scale conformational changes and biological function, it is essential to combine frameworks that describe different physical scales of matter [88,89].

To overcome this challenge, quantum mechanics (QM) and molecular mechanics (MM) are employed in a complementary manner, each addressing specific aspects of molecular behavior. Their integration provides a conceptual foundation for modeling and simulating complex biological systems with both accuracy and efficiency. Although the simulations in this thesis are based entirely on classical MD, a theoretical overview of QM is included, as its principles underpin the development of MM force fields and were essential for deriving accurate partial charges for the ligand analyzed in later chapters.

**Quantum Mechanics.** QM provides an atomistic framework that explains how electron behavior shapes molecular properties through wave functions and discrete energy levels. In this study, QM calculations were performed for the inhibitor dasatinib to obtain an optimized geometry and electrostatic potential-derived atomic charges, which were later incorporated into the MD framework. In this context, the properties of a molecule, including its geometry, bonding pattern, dipole moment, and reactivity, emerge from solving the Schrödinger equation, which defines the allowed energy levels of the system:

$$\hat{H}\Psi = E\Psi \tag{2.1}$$

Here,  $\hat{H}$  is the Hamiltonian operator representing the total energy of the system,  $\Psi$  is the electronic wave function, and  $E$  is the energy eigenvalue. Because exact solutions are possible only for small systems, practical quantum chemical methods rely on approximations that balance accuracy and computational efficiency [90].

Among these, density functional theory (DFT) is one of the most widely used approaches for molecular systems. DFT replaces the complex many-electron wave function with the electron density  $\rho(r)$ , allowing efficient evaluation of total energy while incorporating electron correlation effects. The total electronic energy can be expressed as:

$$E[\rho] = T[\rho] + V_{ne}[\rho] + J[\rho] + E_{xc}[\rho] \tag{2.2}$$

where  $T[\rho]$  is the kinetic energy,  $V_{ne}[\rho]$  is the electron–nuclear attraction,  $J[\rho]$  is the classical electron–electron repulsion, and  $E_{xc}[\rho]$  is the exchange–correlation functional. This energy functional combines contributions from kinetic, electron–nuclear, Coulomb, and exchange–correlation interactions. Because the exact form of  $E_{xc}[\rho]$  is unknown, approximate functionals must be used to make DFT practically applicable.

Hybrid functionals such as B3LYP combine Hartree–Fock exchange with DFT correlation terms, producing reliable results for organic and biological molecules [91]. DFT is widely used to compute partial charges and optimized geometries for small-molecule ligands, which are required when integrating such molecules into classical force fields. The choice of basis set, such as 6-31G(d), determines the level of mathematical detail in representing atomic orbitals. Because most biological processes occur in aqueous environments, solvation effects are often modeled implicitly using the Solvation Model based on Density (SMD), which treats the solvent as a polarizable continuum surrounding the solute. The choice of basis set influences the accuracy of computed electrostatic potentials, and therefore affects the quality of charge models embedded in MD [92]. Here, the B3LYP hybrid functional, the 6-31G(d) basis set, and the SMD solvation model were chosen to provide a reliable balance between accuracy and computational efficiency in deriving partial charges for the ligand.

These quantum-mechanical principles form the physical basis of classical force fields, which approximate the underlying electronic interactions using simplified empirical potentials. Next, MM formalisms are introduced, which serve as the core of classical biomolecular simulations.

**Molecular Mechanics.** While QM achieves high accuracy, its computational cost increases rapidly with system size. MM provides an efficient alternative for studying large biomolecular systems by describing them with classical physics. Atoms are treated as small spheres connected by springs that represent chemical bonds, and molecular energy is calculated from potential functions based on experimental data and quantum results [93, 94].

In MM, the potential energy of a molecular system is calculated as the sum of bonded and nonbonded interactions. Bonded terms include bond stretching, angle bending, and torsional rotations, while nonbonded terms account for van der Waals and electrostatic interactions between atoms that are not directly bonded. These components together define the approximate potential energy surface that governs atomic interactions and forms the foundation for biomolecular simulations [95]. A more detailed formulation of the energy function used in this work is presented in the section titled Force Field Models for Biomolecular Systems.

MM can be used to minimize the energy of a structure or to compare the stability of

different conformations. Energy minimization yields a static structure on the potential energy surface, which also serves as the foundation for all subsequent molecular dynamics simulations. It allows efficient modeling of large proteins, nucleic acids, and membranes, and helps explore flexibility, folding, and molecular interactions at scales that are not accessible to QM. In this study, force field parameters for the protein and solvent were taken from established parameter sets used in biomolecular simulations (e.g., CHARMM), while ligand parameters were supplemented with atomic charges derived from QM calculations. MM provides the potential energy surface that underlies MD simulations, enabling the exploration of system behavior beyond static structures.

## 2.1 Molecular Dynamics

MD extends MM by adding the dimension of time to the description of molecular systems. It uses the potential energy surface defined by the force field to compute interatomic forces and applies the second law of Newton to determine atomic motion [96]. The motion of each atom is described by the classical equations of motion based on the second law of Newton:

$$F_i = m_i a_i = -\frac{\partial U}{\partial r_i} \quad (2.3)$$

where  $F_i$  is the force acting on atom  $i$ ,  $m_i$  is its mass,  $a_i$  is its acceleration, and  $U$  is the potential energy obtained from MM. These forces are computed from empirical force fields which provide parameters for bonded and nonbonded interactions [97]. The resulting equations of motion are solved iteratively using numerical integration algorithms such as velocity Verlet and leapfrog, which updates the positions and velocities of atoms over very short time steps, typically on the order of one to two femtoseconds. By repeating these calculations for millions of steps, MD generates a time-dependent trajectory that describes how all atoms in the system move and interact under realistic physical conditions [98].

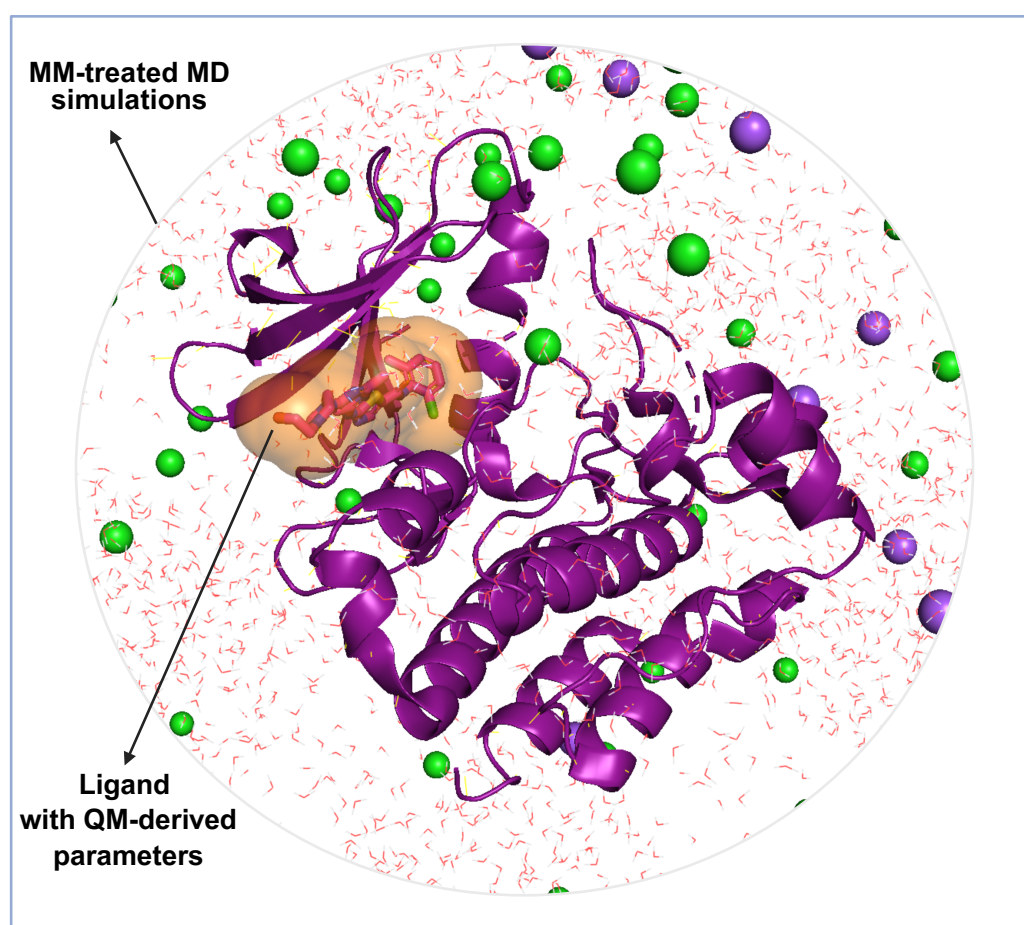
To maintain stability and reproduce the desired thermodynamic ensemble, simulations are usually performed at constant temperature and pressure, controlled by coupling algorithms such as thermostats and barostats. From the resulting trajectories, one can observe molecular motions, domain rearrangements, and conformational transitions that are essential for understanding protein flexibility and function. These simulations also allow the calculation of dynamic properties such as diffusion coefficients, hydrogen bond lifetimes, and conformational free energy changes, linking atomic-level motions to macroscopic observables [89].

The major strength of MD lies in its ability to bridge structure and function. It reveals how



interactions between atoms, driven by forces derived from the potential energy surface, lead to collective structural changes over time. For proteins, this includes domain motion, allosteric signaling, and the dynamic stabilization of ligand binding events. Through these simulations, the temporal evolution of biomolecules can be studied at an atomistic level with realistic environmental effects, including explicit solvent and ions, which are essential for capturing biologically relevant behavior [94,98].

The schematic illustration in Figure 2.1 summarizes the overall concept of the multiscale modeling strategy. In this schematic representation, the ligand is treated with QM to obtain accurate atomic parameters, which are then incorporated into the MD simulation using MM. This integration of QM precision with MM and MD efficiency provides the conceptual foundation for realistic atomistic simulations of biomolecular systems [99].



**Figure 2.1: Schematic representation of a multiscale molecular simulation system.** The figure illustrates a protein system simulated in a fully solvated and ionized environment, reflecting biologically relevant conditions. The ligand was parameterized using QM calculations for geometry optimization and charge assignment. These QM-derived parameters were then incorporated into the MM framework to enable efficient MD simulations.

In this study, MD simulations were performed using the GROMACS 2022.4 software package [100], which provides an efficient and widely validated framework for simulating biomolecular systems. The key parameters and methodological choices applied in

GROMACS for the present work are described in the following subsections.

### 2.1.1 Force Field Models for Biomolecular Systems

The reliability of MD simulations depends critically on the accuracy of the force field used to describe atomic interactions. A force field defines how the potential energy of a molecular system varies with the positions of all atoms and provides the mathematical framework for calculating the forces that drive motion during simulation [93, 94]. Each force field consists of mathematical functions and parameter sets derived from experimental data and high-level QM calculations.

Although the potential energy function is based on MM principles, its use in MD requires carefully developed and validated parameters to reproduce realistic physical behavior. Several widely used biomolecular force fields exist, each built on similar physical principles but tuned differently to achieve the best balance between accuracy, transferability, and computational efficiency. For example, AMBER [101] and CHARMM [94] are optimized mainly for proteins and nucleic acids, while OPLS [102] is designed for small organic molecules and condensed-phase systems. GROMOS [103] emphasizes compatibility with experimental thermodynamic data and is often used for simulations in solution. Together, these force fields provide the practical foundation for most atomistic simulations in structural biology [93, 95].

In this work, the CHARMM36 all-atom additive force field was employed to describe proteins, ligands, and solvent molecules. CHARMM36 has been extensively optimized and validated for a wide range of biological systems, offering accurate treatment of bonded and nonbonded interactions [94]. Its parameters are derived from high-level QM calculations and experimental data to reproduce both bonded and non-bonded interaction energies with high accuracy. This force field is fully compatible with the TIP3P water model and includes parameter sets for ions, enabling faithful reproduction of solvated protein–ligand systems under physiological conditions. Partial atomic charges for the ligand were generated from QM electrostatic potential fitting and were incorporated into the CHARMM36 framework following the standard ligand parameterization protocol.

The CHARMM36 force field follows the general MM functional form but includes additional terms, such as improper dihedrals and refined Lennard–Jones parameters, to better reproduce experimental geometries and interaction energies. Its total potential energy is

given by:

$$\begin{aligned}
U_{\text{CHARMM}} = & \sum_{\text{bonds}} k_b(l - l_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} k_\phi[1 + \cos(n\phi - \delta)] \\
& + \sum_{\text{impropers}} k_\omega(\omega - \omega_0)^2 \\
& + \sum_{i < j} \left[ \epsilon_{ij} \left( \left( \frac{R_{\min,ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{R_{\min,ij}}{r_{ij}} \right)^6 \right) + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right]
\end{aligned} \tag{2.4}$$

Compared with the general MM expression, CHARMM introduces two additional terms. The *improper dihedral* term,  $k_\omega(\omega - \omega_0)^2$ , is used to preserve planarity in peptide bonds and aromatic rings. The nonbonded term uses the parameters  $R_{\min,ij}$  and  $\epsilon_{ij}$  to define van der Waals interactions more precisely, ensuring an accurate balance between attraction and repulsion. These refinements improve the structural stability and energetic consistency of biomolecular simulations [104].

The quality of a force field determines how faithfully simulations reproduce molecular structure, energetics, and motion. When parameters are well calibrated, the resulting trajectories capture realistic conformational dynamics and allow direct comparison with experimental observations, providing a reliable foundation for understanding biomolecular mechanisms.

### 2.1.2 Optimization of $\text{Mg}^{2+}$ Interaction Parameters

Classical force fields are parameterized using empirical and theoretical data to reproduce macroscopic thermodynamic properties, such as solvation free energies or hydration enthalpies, rather than microscopic coordination behavior. As a result, the default nonbonded parameters for divalent cations can sometimes lead to unrealistic ion–ligand distances or coordination numbers in MD simulations. This limitation arises because the Lennard–Jones (LJ) potential used to describe van der Waals interactions is typically derived from empirical combination rules that do not accurately capture specific ion–oxygen interactions in biological systems [93, 98].

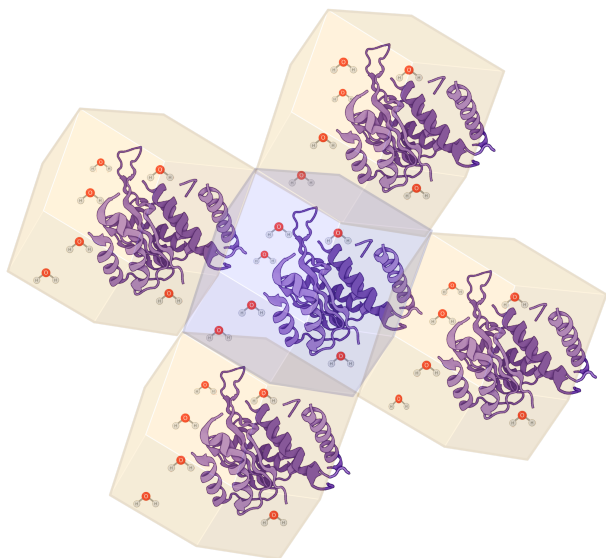
In the case of magnesium ions, standard parameters often overestimate the interaction strength with negatively charged residues or underestimate hydration shell stability. To achieve more realistic coordination behavior, optimized LJ parameters have been developed based on high-level QM calculations and experimental solvation data [105, 106]. These revised parameters adjust the depth of the LJ potential well ( $\epsilon$ ) and the atomic size parameter ( $\sigma$ ) for  $\text{Mg}^{2+}$ , allowing improved representation of ion–water and ion–protein interactions.

In this study, optimized parameters for  $\text{Mg}^{2+}$  were employed to ensure accurate modeling of its first solvation shell and coordination geometry within the Lyn kinase systems. The revised LJ terms reproduce the experimentally observed octahedral coordination with six water or oxygen atoms and yield realistic Mg–O distances of approximately 2.0 Å. This refinement enhances the physical realism of the simulated systems by maintaining proper electrostatic balance and preserving the stability of the metal coordination environment throughout the trajectory [107].

### 2.1.3 System Preparation and Equilibration

**Periodic Boundary Conditions.** In MD simulations, periodic boundary conditions (PBC) are applied to make a small system behave as if it were part of an infinite environment. The simulation box is virtually repeated in all directions, so when an atom leaves one side of the box, it reappears from the opposite side. This prevents artificial edge effects and allows the system to mimic bulk behavior rather than behaving like an isolated cluster (Figure 2.2).

Nonbonded interactions are computed within a defined cutoff distance using the minimum image convention, which ensures that each atom interacts only with the closest periodic image of every other atom. Long-range electrostatic interactions are handled using the Ewald summation method, which separates the Coulomb potential into short- and long-range components for efficient computation under PBC.



**Figure 2.2: Illustration of periodic boundary conditions using a rhombic dodecahedron simulation box.** The central unit (blue) is replicated in all directions, allowing atoms to move continuously across boundaries and eliminating artificial edge effects.

In practice, the Particle Mesh Ewald (PME) algorithm is used to achieve this efficiently while preserving accuracy [108,109]. Van der Waals interactions are smoothly truncated according to the force field parameters to prevent energy discontinuities at the cutoff distance. The simulation box is constructed to be sufficiently large to avoid nonphysical interactions between a solute and its own periodic images. Under PBC, atoms that cross a boundary continue their trajectory seamlessly into the adjacent periodic image, maintaining a continuous and physically realistic representation of molecular motion within a finite, periodically replicated system [110].

**Defining the Solvent Environment.** Accurate representation of water is essential in MD simulations, as it mediates hydrogen bonding, solvation, and electrostatic screening in biomolecular systems. Solvent can be represented either explicitly, by modeling individual water molecules, or implicitly, by treating the solvent as a continuous dielectric medium. In this study, explicit solvent models were used, where water molecules are described atomistically through parameterized potential functions. Because fully quantum mechanical treatment of bulk water is computationally infeasible, empirical water models are employed to approximate its structural and dynamical properties through simple functional forms.

Water models differ in the number of interaction sites, treatment of polarization, and parameterization targets such as density, dielectric constant, or diffusion coefficient. The most commonly used fixed-charge models include SPC [111], SPC/E [112], TIP3P [113], TIP4P [114], and TIP5P [115], which represent water molecules with three, four, or five interaction sites. These models define water geometry through bond lengths and angles, while nonbonded interactions are represented using Lennard–Jones and Coulombic terms centered on oxygen and hydrogen atoms or additional massless sites.

In this study, the TIP3P water model was used to solvate the simulation systems. TIP3P is a three-site rigid model originally developed to reproduce the experimental liquid density and heat of vaporization of water [116]. Despite its simplicity, it provides a good balance between computational efficiency and realistic hydrogen-bonding behavior, which makes it compatible with most biomolecular force fields, including CHARMM36.

Each water molecule in TIP3P consists of one LJ site on the oxygen atom and three partial charges distributed over the oxygen and two hydrogens. The fixed O–H bond length and H–O–H angle are maintained using constraint algorithms such as LINCS [117] or SHAKE [118], ensuring numerical stability during MD integration. TIP3P reproduces the key structural features of liquid water and allows accurate representation of solvation dynamics, ion hydration, and protein–water interfaces within the CHARMM–GROMACS framework.

Overall, the use of TIP3P water in combination with the CHARMM36 force field ensures a consistent and well-validated physical model for simulating the solvated protein and protein–ligand systems in this study.

**System Equilibration.** After system setup and solvation, MD simulations begin with equilibration steps to relax the system and remove unfavorable contacts introduced during preparation. In GROMACS, equilibration is performed in two main stages: constant volume (NVT) and constant pressure (NPT). These steps ensure that temperature, density, and pressure reach stable values before the production simulation starts [119, 120].

In the first stage, the NVT ensemble (constant Number of particles, Volume, and Temperature) is used to gradually bring the system to the target temperature. The simulation box volume remains fixed while the thermostat controls the temperature by adjusting atomic velocities. This step allows the solvent and solute to adapt to each other under controlled thermal conditions, without fluctuations in system size. Common thermostats used in GROMACS include the V-rescale (modified Berendsen) and Nosé–Hoover algorithms, both providing smooth and stable temperature coupling across all atoms.

Once the system has reached the desired temperature, the equilibration continues under the NPT ensemble (constant Number of particles, Pressure, and Temperature). In this phase, both temperature and pressure are controlled simultaneously. The barostat regulates the box dimensions to achieve the target pressure, ensuring that the system density matches experimental conditions. The Parrinello–Rahman barostat is often used in GROMACS for accurate pressure coupling, allowing anisotropic box fluctuations when needed.

Throughout equilibration, positional restraints are typically applied to the heavy atoms of the protein or ligand, allowing the solvent and ions to relax around a stable solute structure. This avoids large deviations or structural distortions before the system is thermodynamically equilibrated. Temperature and pressure values are monitored to confirm convergence before proceeding to the unrestrained production run [121].

Together, the NVT and NPT equilibration steps prepare the molecular system for reliable production simulations by ensuring stable thermodynamic conditions consistent with the chosen ensemble.

## 2.2 Analysis of Trajectories

In this study, MD trajectories were analyzed using a combination of GROMACS tools and custom workflows in R Statistical Software (v4.4.1; R Core Team, 2021) [122]. These tools were used to examine the structural and dynamic features of the simulated systems over time. The most important analysis methods applied in this work are described below, each focusing on a specific aspect of molecular behavior such as stability, flexibility, or atomic-level changes during the simulation.

### 2.2.1 Structural Deviation and Fluctuation

**Root Mean Square Deviation.** Root mean square deviation (RMSD) is a widely used metric for assessing the structural stability of biomolecules over time during MD simulations. It measures the average positional deviation of the structure at a given time

point relative to a reference structure, typically the initial conformation. RMSD allows tracking of structural changes throughout the simulation and helps determine whether the system achieves stability over time.

For a system with  $N$  atoms, the RMSD at time  $t$  is calculated as:

$$\text{RMSD}(t) = \sqrt{\frac{1}{N} \sum_{i=1}^N \|\vec{r}_i(t) - \vec{r}_i^{\text{ref}}\|^2} \quad (2.5)$$

where  $\vec{r}_i(t)$  is the position vector of atom  $i$  at time  $t$ , and  $\vec{r}_i^{\text{ref}}$  is the position of the same atom in the reference structure.

By quantifying overall atomic displacements, RMSD helps compare the conformational behavior of different systems and identify major structural shifts during the simulation.

**Root Mean Square Fluctuation.** Root mean square fluctuation (RMSF) is used to evaluate the flexibility of individual atoms or residues throughout a MD simulation. While RMSD measures global structural changes over time, RMSF focuses on local fluctuations, providing insight into which regions of a biomolecule are more dynamic or stable.

For a given atom  $i$ , the RMSF is defined as:

$$\text{RMSF}(i) = \sqrt{\frac{1}{T} \sum_{t=1}^T \|\vec{r}_i(t) - \langle \vec{r}_i \rangle\|^2} \quad (2.6)$$

where  $T$  is the total number of time frames,  $\vec{r}_i(t)$  is the position of atom  $i$  at time  $t$ , and  $\langle \vec{r}_i \rangle$  is the average position of that atom over the simulation.

RMSF analysis highlights regions of high or low atomic mobility and is especially useful for comparing the dynamic behavior of different residues or domains under varying simulation conditions.

## 2.2.2 Interatomic Distance and Angle Calculations

Geometric measurements such as interatomic distances and bond angles are fundamental descriptors of structural dynamics in molecular simulations. They provide precise information on how specific atomic relationships evolve over time and can reveal local structural rearrangements that contribute to larger conformational transitions.

In this study, distances and angles were computed from MD trajectories using GRO-

MACS analysis tools. Distance analysis enables monitoring of key interactions such as hydrogen bonds, salt bridges, or metal coordination, while angle measurements capture geometric variations in backbone or side-chain orientations. Together, these parameters offer detailed insight into local conformational changes, ligand-induced rearrangements, and coordination stability within binding or catalytic regions.

### 2.2.3 Cluster Analysis

Cluster analysis is used to identify groups of similar conformations sampled during the simulation based on structural similarity. By comparing the backbone RMSD between frames, the trajectory can be partitioned into clusters, each representing a distinct conformational state. This approach simplifies the complex conformational landscape and allows for the identification of dominant structural patterns.

In this study, clustering was performed using the partitioning around medoids (PAM) algorithm [123]. PAM is a robust alternative to centroid-based methods such as k-means. Instead of using abstract centroids, PAM selects actual trajectory frames, called medoids, as representatives of each cluster. The algorithm works by first selecting an initial set of medoids and then iteratively refining them to minimize the total dissimilarity between all points in a cluster and their corresponding medoid. Dissimilarity was measured using pairwise backbone RMSD, ensuring that clustering reflected meaningful structural differences.

This method is particularly interpretable, as each cluster is represented by an actual trajectory frame rather than an abstract structural average, allowing direct examination of representative conformations. It also avoids sensitivity to outliers and non-Euclidean distance metrics, making it particularly suitable for analyzing protein structures.

To evaluate the quality of the clustering, silhouette coefficients were calculated [124]. For each trajectory frame  $i$ , the silhouette score is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2.7)$$

where  $a(i)$  is the average RMSD between  $i$  and all other points within the same cluster, and  $b(i)$  is the minimum average RMSD between  $i$  and all points in the closest neighboring cluster. The silhouette value ranges from  $-1$  to  $1$ . A value close to  $1$  indicates that the point is well matched to its own cluster and distinctly separated from neighboring clusters. Values near  $0$  suggest that the point lies at the boundary between two clusters and may not be strongly associated with either. Negative values indicate that the point may have been assigned to the wrong cluster, showing stronger similarity to a different group.



Silhouette analysis was used to guide the choice of the optimal number of clusters and to confirm the separation between distinct conformational states. This combination of RMSD-based clustering and silhouette evaluation provides a compact and interpretable summary of the structural diversity sampled during the simulation.

## 2.2.4 Catalytic Site Geometry and Stability

**Binding Pocket Surface Area.** The size of the ligand binding pocket was estimated by calculating the solvent-accessible surface area (SASA) of the ligand-surrounding region. This metric provides a quantitative measure of how exposed the binding site is to the solvent, offering insight into pocket compaction or expansion under different simulation conditions.

SASA quantifies the surface area of a molecule that is accessible to a solvent probe, typically represented by a rolling sphere of a given radius. It is computed by tracing the surface formed by the probe as it rolls over the van der Waals surface of the atoms. For a system of  $N$  atoms, the total SASA is given by:

$$\text{SASA} = \sum_{i=1}^N A_i \quad (2.8)$$

where  $A_i$  is the solvent-accessible area of atom  $i$ , taking into account its radius and the occlusion caused by neighboring atoms.

In this study, the SASA of the ligand-defined region was used as an approximation of pocket surface area. Changes in SASA values across systems reflect variations in pocket openness and help characterize structural effects induced by mutations or ligand binding.

**Binding free energy estimation.** Binding free energy estimation provides a comparative measure of the interaction strength between a protein and its ligand. In this study, the molecular mechanics Poisson–Boltzmann surface area (MM/PBSA) method was used to calculate the energy of binding from MD trajectories.

MM/PBSA is a post-processing approach that combines MM energy terms with solvation contributions to approximate the free energy difference between bound and unbound states. It is particularly useful for comparing relative binding affinities across different systems without requiring extensive sampling of the unbound protein or ligand. However, it generally neglects entropic contributions and is therefore limited to estimating relative rather than absolute binding free energies. More rigorous approaches, such as alchemical free energy methods, are required for accurate absolute binding calculations.

The binding free energy  $\Delta G_{\text{bind}}$  is estimated using the equation:

$$\Delta G_{\text{bind}} = \langle E_{\text{complex}} \rangle - (\langle E_{\text{protein}} \rangle + \langle E_{\text{ligand}} \rangle) \quad (2.9)$$

where  $\langle E \rangle$  denotes the average free energy of each component. Each term is computed as the sum of MM energies (bonded, electrostatic, and van der Waals interactions) and solvation energies (polar and nonpolar contributions).

Although MM/PBSA does not fully account for entropic effects, it offers a practical and widely used framework for estimating relative binding strengths. In this work, MM/PBSA helped assess how mutations or ligand differences influence binding energetics across systems.

**Mg<sup>2+</sup> Coordination Analysis.** Coordination analysis of metal ions such as Mg<sup>2+</sup> provides a detailed view of how these ions interact with their surrounding environment throughout an MD simulation. This technique involves identifying and monitoring the atoms, typically electronegative atoms such as oxygen, that remain within a defined cutoff distance from the ion over time.

By examining the frequency and stability of coordinating atoms or molecules, this approach helps characterize the dynamic coordination number and the geometry of the coordination shell. It can reveal how the presence of ligands or mutations influences ion binding and whether key coordination patterns are preserved or altered during the simulation. In particular, Mg<sup>2+</sup> coordination analysis is useful for evaluating structural or catalytic roles of metal ions and for detecting changes in ion–protein or ion–solvent interactions that may affect the functional behavior of the system.

### 2.2.5 Principal Component Analysis

Principal component analysis (PCA) is a dimensionality reduction technique used to identify dominant collective motions within molecular dynamics trajectories. It simplifies the complex atomic displacements of a biomolecular system by transforming them into a new set of orthogonal coordinates that describe the principal directions of motion. Each direction, or principal component, corresponds to an eigenvector of the covariance matrix, and its amplitude is given by the corresponding eigenvalue, which represents the variance along that mode.

The method is based on constructing the covariance matrix of atomic positional fluctuations. For a system with  $N$  atoms, the covariance matrix  $C$  is defined as:

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \quad (2.10)$$

where  $x_i$  and  $x_j$  are the Cartesian coordinates of atoms  $i$  and  $j$ , and the angle brackets denote time averages over the trajectory. Diagonalization of  $C$  yields a set of eigenvectors and eigenvalues according to:

$$C\vec{v}_k = \lambda_k \vec{v}_k \quad (2.11)$$

Here,  $\vec{v}_k$  represents the  $k$ -th eigenvector (the principal mode of motion), and  $\lambda_k$  is the corresponding eigenvalue, which quantifies the magnitude of variance captured by that mode. The first few eigenvectors typically describe the largest and most collective conformational changes, while higher-order modes capture localized or less significant fluctuations.

The atomic motions along each principal component can be visualized by projecting the trajectory onto the corresponding eigenvector, expressed as:

$$p_k(t) = \sum_i \vec{v}_{k,i} \cdot (x_i(t) - \langle x_i \rangle) \quad (2.12)$$

where  $p_k(t)$  is the projection value of frame  $t$  along the  $k$ -th principal component. This projection enables visualization of conformational transitions and quantification of the contribution of each mode to the overall motion.

In this study, PCA was applied to the positional fluctuations of the  $C_\alpha$  atoms of protein to characterize large-scale domain rearrangements and to compare the dominant motions across systems. The first principal component were used to represent the essential subspace of motion and to visualize conformational distributions in reduced dimensional space.

### 2.2.6 Dynamical Cross-Correlation Analysis

Dynamical cross-correlation analysis is a widely used technique for quantifying how atomic motions are correlated during a molecular dynamics simulation. It provides a residue-residue level view of how different parts of a protein move in relation to one another over time, helping to identify regions that move cooperatively or in opposite directions.

For a pair of atoms  $i$  and  $j$ , the correlation coefficient  $C_{ij}$  is defined as:

$$C_{ij} = \frac{\langle \Delta \vec{r}_i \cdot \Delta \vec{r}_j \rangle}{\sqrt{\langle \|\Delta \vec{r}_i\|^2 \rangle \langle \|\Delta \vec{r}_j\|^2 \rangle}} \quad (2.13)$$

where  $\Delta\vec{r}_i = \vec{r}_i(t) - \langle\vec{r}_i\rangle$  is the fluctuation of atom  $i$  from its average position over time, and similarly for atom  $j$ . The value of  $C_{ij}$  ranges from  $-1$  to  $+1$ , where:  $+1$  indicates perfectly correlated motion (moving in the same direction),  $-1$  indicates perfectly anti-correlated motion (moving in opposite directions), and values near  $0$  indicate uncorrelated movement.

The matrix of these pairwise correlation values, known as the dynamical cross-correlation matrix (DCCM), can be visualized as a heatmap, revealing coordinated dynamic patterns across the protein. Positive correlations often reflect domains or secondary structure elements that move together, while negative correlations may suggest opposing motions.

DCCM analysis helps uncover functionally relevant coupling between distant regions of the protein and provides insight into how ligand binding or mutations can reshape long-range communication pathways.

### 2.2.7 Interdomain Interface Analysis

To identify residues involved in domain–domain interactions, interface analysis was performed by evaluating changes in SASA. This approach estimates the extent to which the surface of each residue becomes buried when two protein domains associate to form an interface.

The method compares the total SASA of the isolated domains with that of the full complex. When two domains interact, portions of their surfaces become shielded from solvent, and this buried surface area is quantified as:

$$\Delta\text{SASA} = \sum_{i \in D_1} \text{SASA}_i + \sum_{j \in D_2} \text{SASA}_j - \sum_{k \in D_{1+2}} \text{SASA}_k \quad (2.14)$$

where  $\text{SASA}_i$  and  $\text{SASA}_j$  denote the solvent-accessible surface areas of atoms or residues in domain 1 and domain 2, respectively, calculated in isolation.  $\text{SASA}_k$  represents the corresponding values when the two domains are present together in the full complex (i.e.,  $D_{1+2}$ ). The summations are typically performed at the residue level, although atom-level resolution can also be used.

The difference  $\Delta\text{SASA}$  reflects the total surface area buried upon domain association. To localize interface regions, the per-residue change in SASA is computed by subtracting the SASA of each residue in the complex from its value in the isolated domain. Residues exhibiting a reduction in solvent exposure greater than a defined threshold (commonly  $1.0 \text{ \AA}^2$ ) are classified as interface residues. This analysis enables the identification of contact surfaces and reveals structural coupling between distinct regions of the protein.

## Chapter Summary

This chapter presented the computational framework used to investigate the conformational dynamics of full-length Lyn kinase. Beginning with the theoretical foundations of quantum and molecular mechanics, the chapter outlined a multiscale strategy for ligand parameterization and classical MD simulations. Key methodological components, including force field selection, system setup, system equilibration, and trajectory analysis, were described in detail. Together, these approaches establish a robust and reproducible platform for exploring conformational changes, interdomain communication, and allosteric regulation. In Chapters 4 and 5, this simulation framework is employed to examine how ligand binding and cancer-associated mutations influence the dynamic behavior of full-length Lyn kinase and to distinguish the underlying allosteric communication networks and regulatory connectors.

## Chapter 3

# Machine Learning Concepts and Methods for Structural Biology

*You are as water, flowing quietly through the veins of  
the earth, unseen, yet sculpting the world into form.*

*Each drop carries a secret from the hidden realm,  
each ripple a whisper of silent wisdom.*

*Though your journey leaves no trace on the surface,  
its echo lingers forever in the soul of existence.*

– Sohrab Sepehri (19th-century poet and painter)

### 3.1 Machine Learning In Structural Biology

Machine learning (ML) has become an important tool in structural biology, particularly for understanding how biomolecules change shape, interact, and transition between conformational ensembles. In recent years, ML has advanced many areas of structural biology, including protein structure prediction, identification of functional sites, and characterization of conformational ensembles [125–127].

ML is especially valuable for analyzing structural and dynamic data from MD simulations, which produce thousands of frames capturing the time-resolved motion of atoms in a protein [128]. Each frame can be described by structural features such as interatomic distances, backbone angles, local environments, or domain-level movements. ML models can detect patterns in these features that correlate with conformational ensembles, even when such patterns are subtle or dispersed across the structure [129]. This is particularly important in allosteric systems, where local perturbations can influence distant regions through long-range communication. By learning how structural features differ between conformational ensembles, ML can help uncover the molecular determinants of conformational switching and signal propagation [130, 131].

In this study, structural features refer to quantitative measurements extracted from each MD frame, including interatomic and interdomain distances and backbone angles. These features encode key aspects of protein conformation in a form suitable for ML-based classification. As MD datasets continue to grow through longer simulations and improved sampling techniques, ML offers a scalable and efficient strategy to identify patterns linked to function.

ML approaches are generally divided into three main categories: *supervised learning* (learning from labeled data) [132], *unsupervised learning* (discovering hidden patterns in unlabeled data) [133], and *reinforcement learning* (optimizing decisions through trial and error guided by rewards) [134]. Each of these paradigms has found applications in structural biology. Unsupervised learning is often used to cluster conformational states from simulations or identify intrinsic structural variability in cryo-EM maps. Reinforcement learning has been explored for tasks such as protein folding or optimizing molecular docking strategies. However, supervised learning remains the most commonly applied approach when a reference annotation is available. In this work, supervised classification was used to analyze MD-derived features of Lyn kinase and to uncover patterns associated with functionally relevant conformational changes. The next section introduces the classification framework applied in this study.

## 3.2 Supervised Classification Learning

Supervised classification involves training a model to map input data to predefined categories using labeled examples. This requires constructing a dataset in which each sample is represented by a structured set of features and paired with a corresponding class label. The model is trained to recognize statistical patterns that distinguish one class from another and is then used to classify new, unlabeled data.

In a typical classification task, each input frame is described by a set of features  $\mathbf{x}_j = (x_{j1}, x_{j2}, \dots, x_{jd})$ , where  $d$  is the number of features and  $j$  indexes the frames. The associated label  $y_j$  is drawn from a discrete set of categories  $\mathcal{C} = \text{active-like, inactive-like}$ . The training dataset can be expressed as:

$$X = (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n), \quad (3.1)$$

where  $n$  is the total number of labeled frames. The goal of the classification algorithm is to learn a function  $f(\mathbf{x})$  that can accurately assign labels to new, unseen inputs.

Supervised classifiers typically operate in two stages: a training phase, in which the model learns patterns by minimizing classification error over the training data, and a prediction phase, in which the trained model is applied to new samples. Model performance is evaluated using cross-validation or independent test data, with metrics such as accuracy, precision, recall, and area under the ROC curve (AUC), which are described in a later section. Importantly, in the context of structural biology, model interpretability plays a crucial role. Many classifiers, including decision tree-based methods, offer feature importance scores that indicate how strongly each input variable contributes to the classification decision. These scores are especially valuable when the features represent biologically meaningful quantities.

A variety of classification algorithms are available for this purpose, including Support Vector Machines (SVM) [135], k-Nearest Neighbors (kNN) [136], and ensemble-based approaches such as Random Forest (RF) [137]. In this work, Random Forest was selected due to its robustness, ability to handle correlated features, and strong generalization performance on moderate-sized datasets. Moreover, its capacity to generate interpretable feature importance profiles makes it well-suited for biological applications. The next section details the Random Forest implementation and its application to MD-derived structural data.



### 3.2.1 Random Forest Classifier

RF is a supervised learning algorithm that constructs an ensemble of decision trees during training and combines their predictions to make a final decision [137]. Each decision tree is trained on a randomly selected subset of the data and considers a random subset of features when evaluating split decisions. This internal variability improves generalization and reduces overfitting by introducing diversity across trees.

RF is particularly well-suited for analyzing MD data due to several key advantages. First, it handles heterogeneous features without assumptions about their distributions or scales, which is important when combining structural features such as distances, angles, and domain-level motions. Second, RF is robust to feature collinearity, a common occurrence in MD datasets where multiple descriptors may capture overlapping aspects of protein motion. Third, the algorithm supports non-linear decision boundaries, which allows it to capture complex structure–function relationships that may not be linearly separable. Finally, RF provides interpretable feature importance scores based on impurity reduction, enabling direct insight into which structural features best differentiate between conformational ensembles.

In this study, RF was applied not to create a predictive model for deployment, but to evaluate which structural features derived from MD trajectories contribute most strongly to distinguishing protein conformations. This supports the broader goal of identifying mechanistic indicators of long-range allosteric coupling and dynamic switching.

An additional practical advantage of RF is that it does not require feature scaling or normalization. Tree-based methods are invariant to monotonic transformations of input features, making them particularly convenient for mixed-unit or non-standardized structural descriptors commonly encountered in MD analysis.

RF operates by aggregating the predictions of multiple decision trees. In the case of classification, each tree produces a class label, and the final output is determined by majority voting:

$$\hat{y} = \text{majority\_vote}(h_1(x), h_2(x), \dots, h_T(x)) \quad (3.2)$$

Here,  $\hat{y}$  is the predicted class label for an input feature set  $x$ , each  $h_t(x)$  denotes the prediction from the  $t^{\text{th}}$  tree, and  $T$  is the total number of trees in the forest [138]. This ensemble strategy reduces overfitting and improves the model’s generalization ability.

**Impurity-Based Splitting.** Each decision tree in the forest is grown by recursively partitioning the data based on the feature that yields the greatest reduction in class

impurity at each node. A common impurity measure is the *Gini impurity*, defined as:

$$G = 1 - \sum_{i=1}^C p_i^2 \quad (3.3)$$

where  $C$  is the number of classes and  $p_i$  is the proportion of frames assigned to class  $i$  at the node. The best split is selected to maximize the impurity decrease:

$$\Delta i = i(\text{parent}) - \left( \frac{n_L}{n} i(\text{left}) + \frac{n_R}{n} i(\text{right}) \right) \quad (3.4)$$

where  $n_L$  and  $n_R$  are the number of frames in the left and right child nodes, and  $n$  is the total number of frames in the parent node [139].

**Feature Importance.** RF enables quantification of feature importance by measuring the cumulative impurity reduction associated with each feature across all decision trees:

$$\text{Importance}(f_j) = \sum_{t=1}^T \sum_{s \in \mathcal{S}_j^t} \Delta i_s \quad (3.5)$$

Here,  $\mathcal{S}_j^t$  denotes the set of all nodes in tree  $t$  where feature  $f_j$  is used for splitting, and  $\Delta i_s$  is the impurity reduction achieved at split  $s$  [140]. This ability to trace feature contributions enables mapping of functional differences to specific structural changes, in line with the broader goal of understanding molecular determinants of conformational ensembles.

**Applying Random Forest to MD Derived Features.** Structural features were extracted from MD trajectories of full-length Lyn kinase based on mechanistic relevance. These features include interatomic distances within the catalytic site, distances between regulatory domains, interdomain angles, and geometric descriptors of the ATP-binding cleft conformation. Each feature was computed for every simulation frame, resulting in a time-resolved dataset that captures the conformational behavior of the protein under different conditions.

These features were chosen to reflect known or hypothesized aspects of allosteric regulation. For example, changes in the angle between domains may indicate shifts in compactness or accessibility, while specific interresidue distances can report on the formation or disruption of salt bridges and hydrogen bonds critical for catalytic function. By quantifying these structural relationships across the trajectory, a functional fingerprint was obtained for each frame that reflects its underlying structural conformation.

In RF, each tree is trained on a random subset of the training data and selects feature-based thresholds that split the data to maximize class separation. At each decision node, the algorithm chooses the feature and threshold that most reduces class impurity (often measured by the Gini index). Because each tree sees slightly different data and selects different splitting criteria, the final model represents an average over many diverse hypotheses. This ensemble strategy enhances generalization and reduces overfitting.

Importantly, RF offers a way to quantify the relative importance of each feature in driving the classification. This is achieved by measuring how much each feature contributes to impurity reduction across all trees and nodes where it is used. Features that consistently help split the data into more homogenous classes are assigned higher importance scores. In this context, the goal is to identify which structural features most effectively distinguish active-like from inactive-like conformations.

Such insights are critical in the context of allostery. Through the identification of features influencing conformational classification, insight is gained into the structural determinants that may mediate long-range communication between regulatory and catalytic domains. This helps bridge the gap between dynamic motion and functional output. Moreover, identifying dominant features provides potential targets for mutational studies or allosteric drug design, allowing specific structural elements to be perturbed in a controlled manner. In this way, supervised ML not only classifies protein states but also enhances mechanistic understanding of allosteric regulation through data-driven structural analysis.

### 3.2.2 Performance evaluation

The discriminative performance of the RF classifier was evaluated using several standard classification metrics. These metrics quantify the quality of classification by comparing the predicted and actual labels across all MD frames in the test set [141].

The accuracy is the proportion of total correct predictions and is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.6)$$

where  $TP$  is true positives,  $TN$  is true negatives,  $FP$  is false positives, and  $FN$  is false negatives.

The precision measures how many of the predicted positive (e.g., active-like) frames were actually correct:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.7)$$

The recall (also known as sensitivity or true positive rate) reflects how many of the actual

positive frames were correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.8)$$

The F1-score is the harmonic mean of precision and recall, balancing both metrics:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.9)$$

To evaluate the model across all possible classification thresholds, the receiver operating characteristic (ROC) curve was also generated [142], which depicts the relationship between the true positive rate (recall) and the false positive rate:

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN} \quad (3.10)$$

The AUC summarizes the ability of the classifier to distinguish between the two classes across all thresholds. AUC values close to 1.0 indicate excellent classification performance, while a value of 0.5 corresponds to random guessing.

In addition to these metrics, the Matthews Correlation Coefficient (MCC) was also computed, as it is a robust and informative metric particularly well-suited for binary classification with imbalanced class distributions. Unlike accuracy or F1-score, the MCC takes into account all four categories of the confusion matrix (TP, TN, FP, FN) and provides a single score that reflects the overall quality of the classification:

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3.11)$$

The MCC ranges from  $-1$  to  $+1$ , where  $+1$  indicates perfect classification,  $0$  corresponds to no better than random, and  $-1$  indicates total disagreement between prediction and observation. Due to its balanced nature, even when class sizes differ, the MCC is considered one of the most reliable indicators of a model's predictive power in real-world biological data.

Together, these evaluation metrics provide a comprehensive quantitative picture of the classification ability of the RF model, highlighting its capacity to correctly and consistently distinguish between the two conformational ensembles based solely on MD-derived structural descriptors.

### 3.2.3 Dealing with Imbalanced Data

A common issue in supervised learning is class imbalance, where one category appears far more frequently than the other in the dataset. This can cause a model to become biased toward the majority class, resulting in high overall accuracy but poor performance on the minority class [143].

In MD-derived datasets, class imbalance often arises when the simulation samples one conformational state more extensively than others. This occurs either due to the intrinsic equilibrium behavior of the system or the design of the simulations. In this case, a larger number of trajectories represent inactive-like conformations, leading to an imbalance between conformational ensembles. If left unaddressed, this imbalance can reduce the ability of the model to recognize structural features specific to the less frequent state.

To mitigate this, the Synthetic Minority Oversampling Technique (SMOTE) [144] was applied only to the training data, ensuring that no synthetic examples were introduced into the test set. SMOTE generates new data points for the minority class by interpolating between existing ones. Given a frame  $\mathbf{x}_i$  from the minority class and one of its  $k$  nearest neighbors  $\mathbf{x}_{zi}$ , SMOTE creates a synthetic point  $\mathbf{x}_{\text{new}}$  as follows:

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda \cdot (\mathbf{x}_{zi} - \mathbf{x}_i) \quad (3.12)$$

Here,  $\lambda$  is a random number drawn uniformly from the interval  $[0, 1]$ . By interpolating between existing minority frames, SMOTE enriches the underrepresented region of feature space without duplicating data. This helps the model learn a more balanced decision boundary.

To ensure stability and minimize sampling bias, all model training was performed with cross-validation. This allowed each fold to benefit from balanced training while preserving an unbiased evaluation set.

## Chapter Summary

In this chapter, an ML framework was presented for the classification of MD-derived conformations of Lyn kinase and the identification of structural features associated with its conformational ensembles. The supervised classification approach, implemented using an RF algorithm, enabled both accurate frame-level labeling and the extraction of biologically meaningful feature importance scores. These insights support the broader aim of understanding allosteric regulation through dynamic structural relationships. This framework is further applied in Chapter 6 to evaluate how specific geometric descriptors contribute to distinguishing active-like and inactive-like Lyn conformations. The top-

ranked features identified through this analysis provide a structural basis for interpreting conformational shifts in full-length Lyn kinase.

## Chapter 4

# Conformational Landscape and Dynamic Stability of Full-Length Lyn Kinase

*A body in motion  
reveals the hidden harmony of the cosmos,  
a silent music that binds the stars and atoms alike.  
Even the smallest shift carries meaning,  
echoing through the vastness of being.*

– Tusi (13th-century polymath and astronomer)

## 4.1 Introduction

Protein kinases function as dynamic molecular switches, with activity regulated by conformational changes and domain-level interactions [145]. In multidomain kinases such as Lyn, structural flexibility and coordinated motion between regulatory and catalytic domains are essential for effective signal transduction [146]. These conformational dynamics underlie transitions between inactive and active conformations and determine how ligands or mutations influence the enzyme regulatory architecture [147].

Static structures provide only snapshots of specific conformations and often fail to capture the full range of motions that govern kinase function [148]. In contrast, the dynamic ensemble of conformations sampled by a protein reveals how flexibility, domain coupling, and long-range communication contribute to regulation [149]. This dynamic perspective is particularly important for full-length Lyn kinase, where regulatory control depends on subtle changes in domain orientation, activation loop positioning, and the spatial arrangement of allosteric centers [150].

Building on the structural and regulatory framework presented in Chapter 1, this chapter examines the dynamic behavior of Lyn kinase through a computational strategy that links conformational dynamics to regulatory mechanisms. Alongside the apo form, ATP-bound and dasatinib-bound systems were analyzed to assess how ligand binding alters flexibility and interdomain coordination. Dasatinib, a broad-spectrum ATP-competitive inhibitor originally developed for BCR-ABL, also binds several SFKs, including Lyn, with high affinity and has been shown to modulate their conformational ensemble [151, 152].

To further dissect the role of specific residues in modulating Lyn kinase behavior, clinically observed cancer-associated mutations were investigated in both apo and ATP-bound contexts. These mutations can perturb conformational dynamics, resulting in altered flexibility, domain stability, and shifts in the conformational ensemble. In kinases, such mutations may stabilize inactive or active forms, interfere with ligand binding, or disrupt spatial coordination among regulatory elements [153–155].

This chapter begins with the structural modeling and domain-resolved stability assessment of Lyn kinase systems, followed by an investigation of residue-level flexibility and dynamic features. To capture large-scale conformational changes, principal component analysis is applied to extract dominant motions. Finally, the impact of ligand binding on catalytic site geometry is examined, and binding energetics are evaluated. Together, these analyses provide a detailed view of Lyn conformational dynamics, directly addressing Aim 1 of this study, to determine how ligand interactions and single-point mutations influence its regulatory architecture.



## 4.2 Computational Methods

**Homology Modeling.** A full-length model of human Lyn kinase (residues 60–502) was constructed using comparative homology modeling with MODELLER [156]. Because a complete experimental structure for Lyn is not available, high-resolution crystal structures of closely related SFKs were selected as templates, and the canonical LYN A (p56lyn) sequence was used for modeling. Specifically, the SH3 domain was modeled using human Lyn (PDB ID: 6NMW) [157], the SH2 domain from mouse Lyn (PDB ID: 4TZI) [158], and the SH1 domain was derived from the full-length human Hck structure (PDB ID: 5H0B) [159], a close Src family homolog with high sequence identity and conserved regulatory elements. These domain-specific templates were aligned to the human Lyn sequence, and multiple models were generated. The model with the best statistical score was selected for further assessment. The structure was validated using the ERRAT server [160], which evaluates the overall quality of atomic interactions. For simulation, the selected model was subjected to energy minimization using the steepest descent algorithm in GROMACS 2022.04 [100]. This step removed unfavorable contacts or artifacts and optimized the geometry for downstream MD analysis.

**Selection of Mutants.** To identify biologically relevant mutations in Lyn kinase for simulation, cBioPortal [161], a platform aggregating somatic mutation data from large-scale cancer genomics studies, was queried. This search revealed four naturally occurring missense mutations, E290D, E290K, I364L, and I364N, observed in multiple tumor types. All four mutations affect conserved residues within the kinase domain.

To further evaluate their relevance, each mutation was analyzed using ProtVar [162]. This platform integrates genomic and structural information from The Cancer Genome Atlas (TCGA) [163], the Catalogue Of Somatic Mutations In Cancer (COSMIC) [164], UniProt [165], and AlphaFold [166]. Structure-aware pathogenicity annotations were extracted directly from ProtVar, which incorporates predictions from tools including PolyPhen [167], CADD [168], AlphaMissense [169], and ESM 1b [170]. These mutations were consistently identified as somatic and predicted to be functionally significant. Based on this evidence, the four mutations were selected for MD simulations.

In addition, mutation of the conserved catalytic lysine in related tyrosine kinases abolishes catalytic activity, reflecting the essential role of this residue in ATP binding and phosphotransfer. This principle justifies the inclusion of the K275A mutant as a catalytic-inactive reference in Lyn simulations [171, 172]. Together, this mutation set supports systematic evaluation of both clinically observed and experimentally validated perturbations in Lyn kinase.

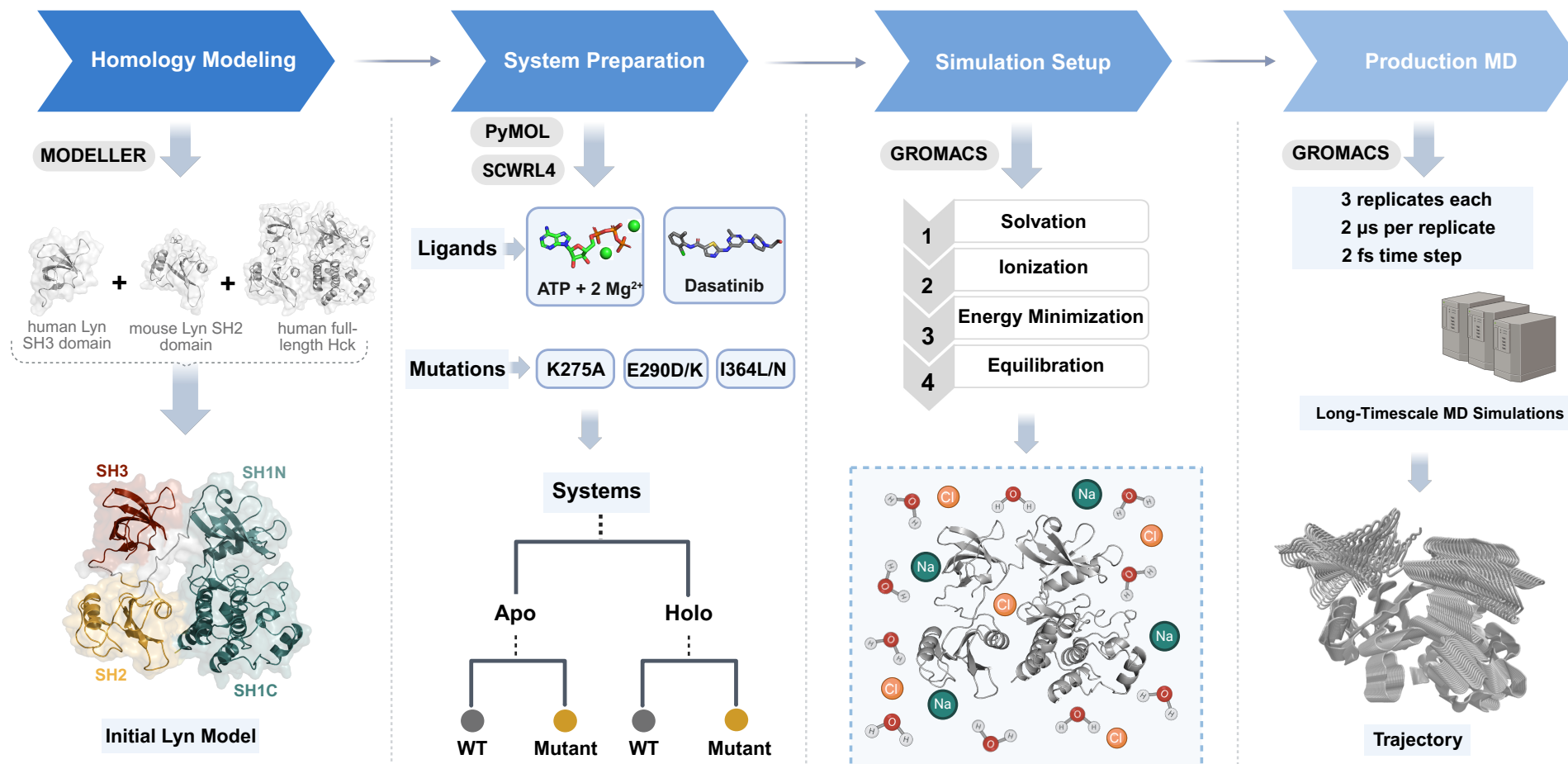
**Mutant Generation and Small Molecule Incorporation.** Single-point mutations

at residues I364 (I364L and I364N), E290 (E290D and E290K), and K275 (K275A) were introduced into the full-length Lyn model using SCWRL4 program with default parameters [173]. For ATP-bound systems, one ATP molecule and two  $\text{Mg}^{2+}$  ions were positioned in the catalytic cleft by structural alignment to the active kinase complex (PDB ID: 1ATP) [174]. For dasatinib-bound simulations, ligand coordinates were extracted from the kinase-inhibitor crystal structure (PDB ID: 2ZVA) [72] and superimposed onto the modeled apo Lyn structure. Energy minimization was then performed to relieve steric clashes and refine local binding interactions. The apo, ATP-bound, and dasatinib-bound forms of WT and mutant systems were used as initial structures for MD simulations. Figure A1 summarizes the amino acid substitutions and ligand structures used throughout this study.

**Quantum Mechanical Calculations.** To obtain accurate electrostatic parameters for MD simulations, QM calculations on dasatinib were performed using Gaussian 16 [175]. Geometry optimization, vibrational frequency analysis, and electrostatic potential (ESP) charge derivation were performed at the B3LYP/6-31G(d) level using the SMD solvation model with water as the solvent [176]. All structures were confirmed as local minima through frequency analysis. The resulting ESP-derived atomic charges were incorporated into the subsequent simulation setup.

**MD Simulations.** A total of 78  $\mu\text{s}$  of MD simulations was performed using the GRO-MACS 2022.04 software package [100] and the all-atom additive CHARMM36 force field [94], which includes optimized parameters for magnesium ion coordination [107]. To ensure reproducibility, each system was simulated in triplicate (Table A1).

An overview of the computational workflow, including modeling, system preparation, and MD setup, is presented in Figure 4.1. The protein complexes were placed at the center of a rhombic dodecahedron box with a minimum distance of 10 Å from the solute to the box boundaries. Solvation was performed using the TIP3P explicit water model [116], and sodium and chloride ions were added to achieve charge neutrality and a physiological ionic strength of 0.15 mol  $\text{L}^{-1}$ . Energy minimization was followed by two equilibration phases: 5 ns in the NVT ensemble with positional restraints on heavy atoms, and 5 ns in the NPT ensemble without restraints. The temperature was maintained at 310 K using the V-rescale thermostat [177] and the pressure at 1 atm using the Parrinello–Rahman barostat [178]. Electrostatic interactions were computed using the PME method [108], and all covalent bonds involving hydrogen atoms were constrained using the LINCS algorithm [117]. Production runs were carried out for 2  $\mu\text{s}$  per replicate with a 2 fs time step on an Nvidia GeForce RTX 4090 GPU (CUDA 12.2, Koeln, Germany). Coordinates, velocities, and energies were recorded every 1 ns for downstream analysis.



**Figure 4.1: Computational pipeline for modeling and simulating full-length Lyn kinase.** The workflow begins with homology modeling using MODELLER, followed by ligand placement and mutation construction. Systems were prepared in apo and holo (ligand-bound) forms for WT and mutant models. MD simulations were conducted in GROMACS, including solvation, energy minimization, equilibration, and production runs of 2 μs per replicate.

**Trajectory Preparation and Structural Analysis.** All trajectories were preprocessed using the `gmx trjconv` tool to remove artifacts caused by periodic boundary conditions. This step ensured proper continuity of molecular motion and enabled accurate downstream structural analyses.

Trajectory analyses were performed using GROMACS tools to assess structural stability and residue-level fluctuations. Backbone RMSD was calculated with `gmx rms` following domain-wise alignment to minimize the influence of global rigid-body motions, allowing the conformational stability of individual domains to be evaluated over the simulation time. RMSF was computed on  $C_\alpha$  atoms using `gmx rmsf` to quantify residue-level flexibility. For comparative analysis, differential RMSF values ( $\Delta\text{RMSF}$ ) were calculated as  $\Delta\text{RMSF} = \text{RMSF}_{\text{reference}} - \text{RMSF}_{\text{comparison}}$ , where the reference corresponds to WT or WT-ATP simulations, and the comparison refers to mutant or ligand-bound systems. To explore local structural variations, key interatomic distances and internal angles were measured using `gmx distance` and `gmx angle`, respectively. All resulting outputs were processed and analyzed in R Statistical Software (v4.4.1; R Core Team 2021) [122].

To identify representative conformations, trajectory snapshots were clustered based on mutual backbone RMSD values. Clustering was carried out using the PAM algorithm implemented in the `cluster` R package (v2.1.8.1) [123]. Clustering quality was evaluated using the silhouette coefficient, calculated with the `fpc` R package (v2.2-13) [124], which measures how well frames fit within their assigned clusters and how distinct each cluster is from the others. For each system, the most populated cluster was identified, and its medoid, defined as the frame with the lowest average RMSD to all other frames in the cluster, was selected as the representative structure.

**Ligand Positional Stability.** Positional stability of bound ligands was assessed by calculating RMSD values using `gmx rms` in GROMACS 2022.04. Heavy atom index groups were defined to exclude hydrogens, ensuring accurate measurement of ligand displacement relative to the protein backbone. RMSD profiles were compared across WT and mutant systems to examine the impact of structural perturbations on ligand accommodation.

**Magnesium Ion Coordination Analysis.** Coordination of  $\text{Mg}^{2+}$  ions in ATP-bound systems was evaluated using MDAnalysis library [179]. For each trajectory frame, distances between the magnesium ions and surrounding protein oxygen atoms were computed using a 4.0 Å cutoff. Coordination events were determined per frame, and atom-wise coordination frequencies were calculated across the trajectory. This analysis provided insight into the stability and preferred binding environment of divalent cations within the catalytic cleft.

**Binding Pocket Surface Area.** Pocket geometry was characterized by SASA analysis. Residues within 5 Å of each ligand in the most populated cluster representative structure were selected to define the binding site. Their SASA values were then computed per frame

across the trajectory using `gmx sasa` with a 1.4 Å probe radius. Time-resolved surface area trends were analyzed in R to compare the dynamic expansion or contraction of the pocket surface across systems.

**Binding Free Energy Estimation.** Binding free energies for ATP and dasatinib were estimated using the MM/PBSA approach implemented via `gmx MMPBSA`. Calculations included molecular mechanics energy components (electrostatic and van der Waals), polar solvation energy based on the Poisson–Boltzmann equation, and nonpolar solvation energy derived from SASA-based surface tension. For each system, MM/PBSA analysis was performed across the full 6 μs trajectory, corresponding to 6000 frames per system.

**Principal Component Analysis.** To investigate large-scale conformational dynamics, PCA was performed on the atomic displacements of C<sub>α</sub> atoms for each system. The positional covariance matrix was computed using `gmx covar`, and the resulting eigenvectors were analyzed with `gmx anaeig`.

To visualize dominant motions, trajectory frames were projected along the first principal component (PC1), which captures the largest variance in atomic positions. Extreme conformations along PC1 were identified to represent the maximal deviations along this mode. A series of interpolated structures were constructed between these extremes to illustrate the conformational transition pathway. These structures were visualized using VMD [180], allowing for qualitative comparison of system-specific dynamic behaviors.

**Statistics.** Statistical significance of pairwise comparisons was assessed using the two-sided Wilcoxon signed-rank test (threshold  $p < 0.05$ ). All analyses were performed in R Statistical Software (v4.4.1; R Core Team 2021). Detailed numerical data supporting the statistical comparisons are provided in the Appendices.

## 4.3 Results and Discussion

### System Preparation

The full-length Lyn kinase structure was generated by homology modeling and rigorously evaluated for structural quality. The final model achieved a global quality factor of 99.1, as assessed by ERRAT2, with scores above 95 indicating well-validated atomic interactions and reliable model quality, supporting suitability for downstream MD simulations. In comparison, the AlphaFold-predicted model showed a lower ERRAT2 score of 92.1, with localized deviations particularly in the linker and SH2 domains. Corresponding ERRAT2 validation plots are shown in Figure A2.

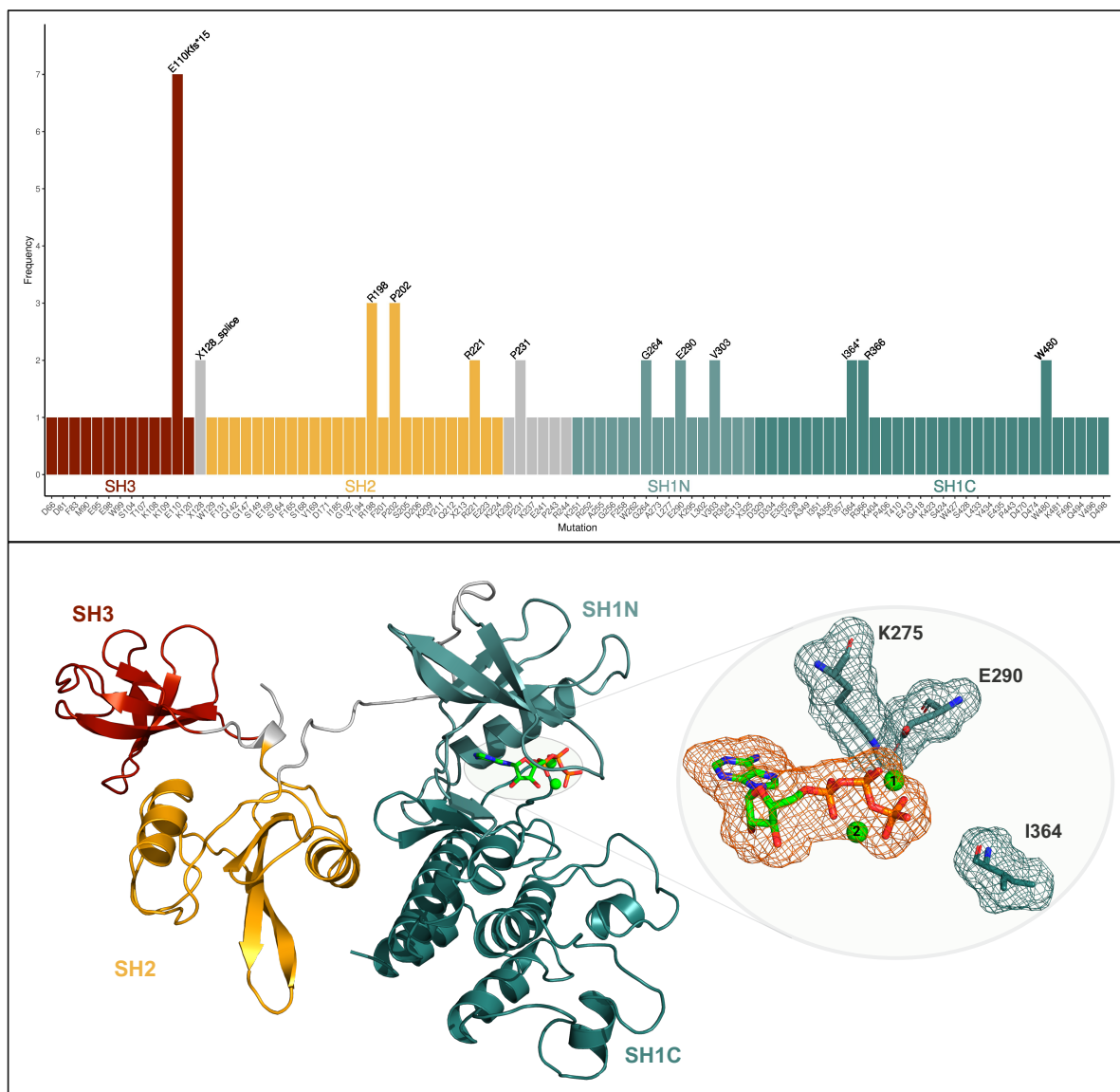
With a reliable structural model established, clinically observed Lyn mutations were

selected for simulation-based investigation. A comprehensive search in cBioPortal for naturally occurring alterations in Lyn and related SFKs identified I364L and I364N as statistically significant hotspots, particularly in bladder urothelial carcinoma. While experimental validation is currently lacking, their recurrence and predicted pathogenicity support a role in modulating kinase function in cancer contexts. E290D and E290K, observed in stomach and colon adenocarcinomas, target a catalytic residue essential for ATP stabilization within the active site. The strong conservation of E290 and I364 across the Src family, summarized in Table A2, underscores their likely importance for kinase activity and regulation.

Figure 4.2 presents a domain-resolved distribution of naturally occurring Lyn mutations, highlighting the positional frequency across SH3, SH2, SH1N, and SH1C domains. The four selected mutations (E290D, E290K, I364L, and I364N) are annotated as somatic in both TCGA and COSMIC databases. According to PolyPhen-2 predictions [167], all are classified as “probably damaging”, with scores exceeding 0.95, suggesting a high likelihood of structural or functional disruption. Pathogenicity predictions and structural impact scores obtained from the ProtVar database, which integrates results from multiple *in silico* tools, are summarized in Table A3. Moreover, substitution of the conserved catalytic lysine (equivalent to K275 in Lyn) to alanine has been shown to abolish kinase activity in protein TKs by disrupting ATP coordination, consistent with the conserved catalytic mechanism of this enzyme family [171, 172].

To visualize the spatial positioning of the selected mutations, Figure 4.2 also depicts the initial energy-minimized structure of Lyn with ATP and two  $\text{Mg}^{2+}$  ions incorporated into the catalytic cleft. The K275 and E290 residues are positioned within the active site and are involved in stabilizing ATP through electrostatic and polar interactions with the phosphate moieties [58]. In contrast, I364 is located in the C-lobe of the SH1 domain (SH1C), spatially distant from the nucleotide. This positioning suggests a potential role in long-range conformational regulation rather than direct coordination with ATP.

Building on the prepared systems, Lyn kinase dynamics were analyzed under apo conditions and in the presence of ATP or dasatinib to assess the effects of ligand binding and cancer-associated mutations. In this study, the terms ‘active-like’ and ‘inactive-like’ conformations are used based on characteristic structural features defined for SFKs. Active-like conformations are typically characterized by a well-aligned catalytic spine, an inward-oriented  $\alpha\text{C}$  helix, coordinated ATP and two  $\text{Mg}^{2+}$  ions, along with an extended activation loop positioned away from the active site [79]. In contrast, inactive-like conformations display a displaced  $\alpha\text{C}$  helix, altered or disrupted coordination geometry, and a misaligned or collapsed activation loop [181, 182]. These structural signatures serve as a reference for interpreting the dynamic responses of Lyn under different conditions. The following sections present a detailed analysis of conformational stability, residue-level flexibility, domain motions, and the active site architecture across all simulated systems.



**Figure 4.2: Mutation frequencies and initial simulation model of Lyn kinase.** **Top:** Domain-resolved distribution of naturally occurring mutations in Lyn kinase. Mutation frequencies were calculated based on data retrieved from cBioPortal. Mutations with a frequency greater than one are labeled above their corresponding bars. This domain-resolved view highlights mutational hotspots that may underlie functional or regulatory perturbations. **Bottom:** Initial structural model used in MD simulations, showing ATP and two Mg<sup>2+</sup> ions placed in the catalytic cleft. A magnified view highlights key residues K275 and E290 in proximity to ATP, and I364 located in the SH1C domain and spatially distant from the nucleotide. The two Mg<sup>2+</sup> ions are shown as green spheres (Mg1 and Mg2). Mg1 lies adjacent to K275 and E290, contributing to ATP coordination and catalytic geometry, while Mg2 occupies the secondary coordination site within the active site.

## Global and Local Dynamics of Lyn Protein Across Domains

With all systems established and relevant conformational landmarks defined, global conformational changes were initially evaluated using backbone RMSD analysis. Backbone RMSD values were calculated over the complete 6  $\mu$ s trajectory of each system, and domain-specific profiles were extracted for the SH1, SH2, and SH3 domains (Figure 4.3). The RMSD distributions revealed distinct fluctuation patterns and consistent differences

in structural stability among the three domains. These results underscore the conformational heterogeneity of Lyn across various ligand-binding and mutational conditions, highlighting that each domain exhibits a distinct response to structural or chemical perturbations.

Among the three domains, the SH3 domain consistently exhibited the highest backbone RMSD values across all systems, reflecting its intrinsically flexible nature and its well-established role in regulatory switching. In the WT system, the SH3 domain displayed a bimodal RMSD distribution pattern, suggestive of multiple conformational subpopulations. This observation supports the notion that SH3 undergoes dynamic switching events, potentially linked to allosteric regulation or intramolecular ligand engagement. In contrast, the SH3 domain of the I364N mutant exhibited a narrower RMSD distribution with a single dominant peak, indicating reduced conformational variability and stabilization into a more restricted structural ensemble. This constrained behavior may reflect enhanced local stabilization, potentially due to changes in hydrogen bonding or steric effects introduced by the polar side chain. While reduced backbone fluctuations are often associated with inactive-like conformations, further validation through binding pocket analysis is required to substantiate this interpretation.

ATP binding conferred a strong stabilizing effect on the kinase architecture. Across all three domains, ATP-bound systems consistently exhibited lower RMSD values relative to their ATP-free counterparts, highlighting the role of ligand engagement in reducing conformational variability. This stabilizing influence was most prominent in the SH1 domain, where ATP interacts directly with conserved catalytic residues, thereby reinforcing compactness and structural order. These observations align with the established role of ATP in promoting the active-like conformation of SFKs [183], and underscore its contribution to maintaining structural integrity within the catalytic core.

In contrast, the dasatinib-bound WT system (WT-DAS) exhibited a divergent domain-specific pattern. Specifically, the SH1 and SH2 domains showed significantly higher median RMSD values (3.98 Å and 3.39 Å, respectively) compared to WT-ATP, indicating increased conformational variability. This elevated flexibility may reflect the broad binding profile and relatively lower specificity of dasatinib, which does not constrain the kinase architecture as tightly as ATP. As a result, Lyn may access a wider range of conformational ensembles in the presence of dasatinib. Interestingly, the SH3 domain in WT-DAS displayed a reduced median RMSD (7.5 Å) compared to WT-ATP, suggesting a selective stabilizing effect on this regulatory domain. This observed decrease in SH3 flexibility may reflect an allosteric restriction mechanism, whereby dasatinib limits conformational heterogeneity and favors a more rigid, potentially inactive structural ensemble.

Mutation-related changes also affected the structural stability of Lyn domains in distinct ways. Among the simulated mutants, E290D and I364L exhibited higher RMSD values,

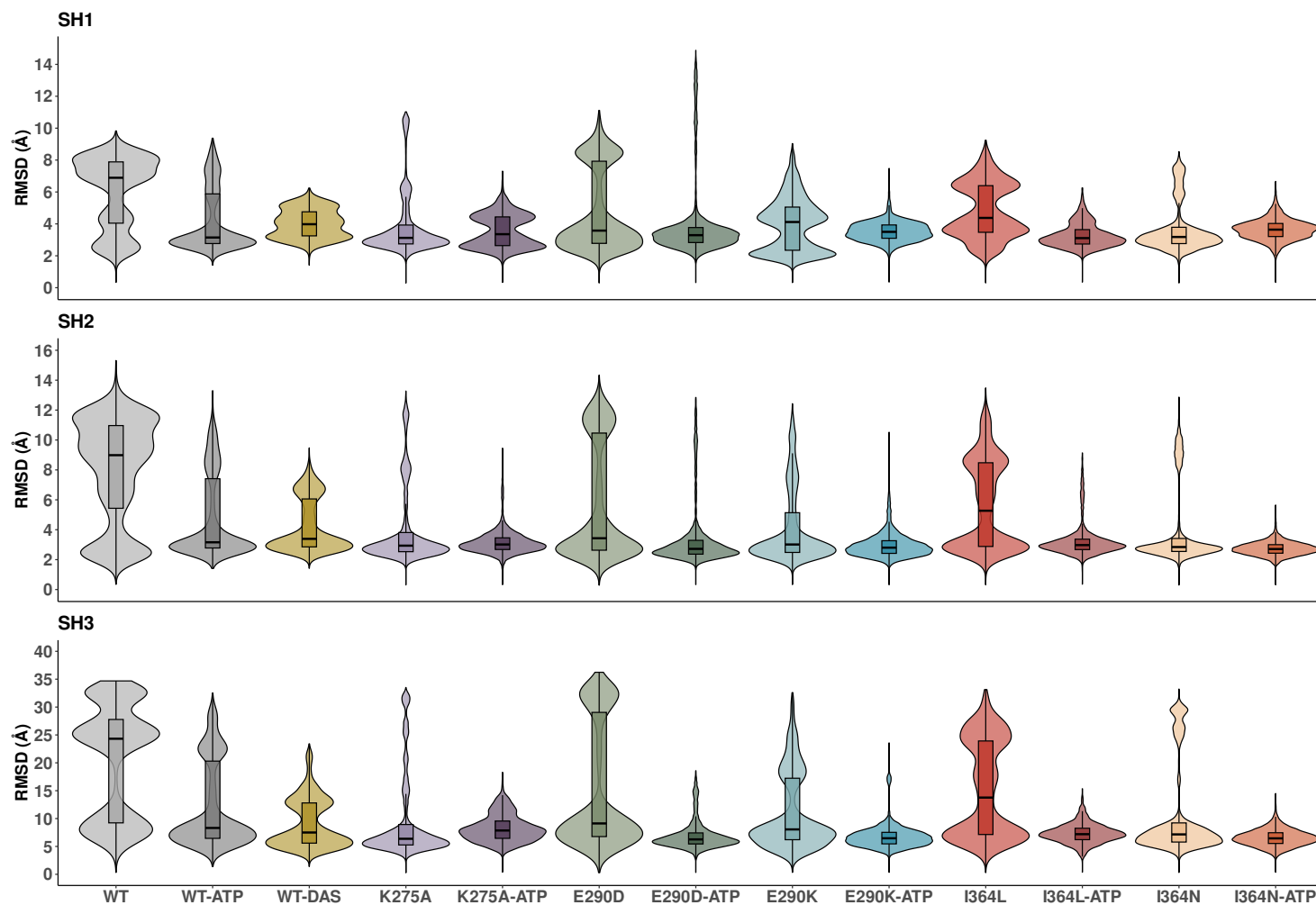


particularly in the SH3 domain, when compared to E290K and I364N. These observations indicate that substitutions involving residues with similar physicochemical properties, such as replacing glutamate with aspartate while preserving negative charge (E290D) or replacing isoleucine with leucine while preserving side-chain size and hydrophobic character (I364L), resulted in greater conformational deviation. In contrast, substitutions introducing charge reversal (E290K) or increased polarity (I364N) were associated with lower RMSD values across multiple domains, reflecting greater structural stability. These findings emphasize that the specific physicochemical nature of the substituted residue plays a decisive role in shaping domain-level structural responses to mutational perturbations.

Mutations that modify the local electrostatic environment, such as E290K and I364N, appeared to promote more compact and restricted dynamic behavior, while conservative replacements such as E290D and I364L retained slightly higher structural variability. Overall, the mutant systems exhibited lower RMSD values relative to the WT, although the degree of structural stabilization varied depending on how each mutation altered local chemical properties. Although increased backbone flexibility within catalytic or regulatory regions has been linked to the emergence of activation-associated structural features, experimental validation would be required to determine whether the observed differences correspond to altered enzymatic function.

The K275A-ATP system exhibited moderately elevated RMSD values across all three domains relative to the WT-ATP system, consistent with structural destabilization following removal of the conserved lysine residue. This lysine has been established to participate in essential catalytic interactions within the ATP-binding pocket [184], and its substitution with alanine was expected to disrupt the local electrostatic environment and hydrogen bonding network. The resulting loss of stabilizing contacts may have contributed to increased flexibility within the catalytic cleft and neighboring regions, leading to the observed structural deviation.

Together, these analyses demonstrate that both ligand binding and single-point mutations exert domain-specific influences on the overall structural stability of Lyn. In particular, the SH3 domain was identified as a significant contributor to structural variability across systems, reflecting its pronounced dynamic behavior in full-length SFKs [185–187]. To evaluate the statistical significance of differences in structural deviation among the systems, two-sided Wilcoxon rank-sum tests were performed for each domain. The results are provided in Tables A4, A5, and A6, corresponding to the SH1, SH2, and SH3 domains, respectively.



**Figure 4.3: Domain-resolved analysis of structural stability across Lyn kinase systems.** Violin plots with embedded boxplots depict backbone RMSD distributions for the SH1, SH2, and SH3 domains across all 13 simulated systems. Each violin illustrates the distribution of RMSD values, while the internal boxplot indicates the interquartile range and median. RMSD values represent deviations from the initial conformation, providing a measure of domain-specific structural stability and conformational variability under different ligand-binding and mutational conditions.

While RMSD analysis provided insights into domain-resolved structural stability, it did not capture subtle, residue-specific fluctuations that may reflect altered structural behavior. To investigate local flexibility changes induced by ligand binding and cancer-related mutations, the RMSF of C $\alpha$  atoms was computed for all simulated systems. RMSF quantifies the average positional deviation of each residue along the simulation trajectory, and variations in these values can reveal localized flexibility changes arising from structural perturbations.

Figure 4.4 displays the  $\Delta$ RMSF profiles across the Lyn kinase sequence. Positive  $\Delta$ RMSF values indicate that the reference system exhibits greater flexibility at a given residue, whereas negative values reflect increased flexibility in the compared system. This differential analysis was used to map mutation- and ligand-induced changes in local dynamics across the 443-residue Lyn sequence. The resulting data were processed in R, and  $\Delta$ RMSF values were plotted along the SH3, SH2, and SH1 domains. This analysis provided residue-level resolution of dynamic alterations and complemented the domain-resolved stability trends identified through RMSD.

The conservative substitutions E290D and I364L induced substantial increases in structural flexibility relative to WT-ATP, particularly within the SH1 and SH3 domains. These negative  $\Delta$ RMSF values indicate enhanced residue-level fluctuations in the mutant systems, which may reflect a shift toward greater dynamic variability that permits access to geometries favorable for catalytic activity. In contrast, the more chemically disruptive mutations E290K and I364N produced widespread reductions in flexibility across the entire protein, with markedly positive  $\Delta$ RMSF values observed in all three domains. These mutations appeared to stabilize Lyn in a more compact and rigid conformation, as indicated by reduced atomic fluctuations. This increased rigidity was interpreted to limit the conformational freedom typically required for catalytic function. In particular, I364N led to a pronounced decrease in fluctuations within the SH1 domain, suggesting a rigidified catalytic core that may hinder substrate accommodation.

The K275A mutation, in which the conserved lysine required for ATP binding is removed, exhibited a distinct  $\Delta$ RMSF profile. Most prominently, it was associated with strongly positive  $\Delta$ RMSF values in the SH1 domain and C-terminal region, indicating reduced stability and consistent with local destabilization in and around the catalytic site. Interestingly, residues between positions 390 and 430 displayed mildly negative  $\Delta$ RMSF values, which were interpreted to suggest a localized increase in flexibility that may reflect compensatory motion or disruption of intramolecular packing.

These results reveal distinct flexibility patterns associated with each mutation. Conservative substitutions such as E290D and I364L promoted broader backbone fluctuations, consistent with increased structural adaptability that may support geometries compatible with catalytic activity. In contrast, charge-reversing or polarity-altering mutations such as

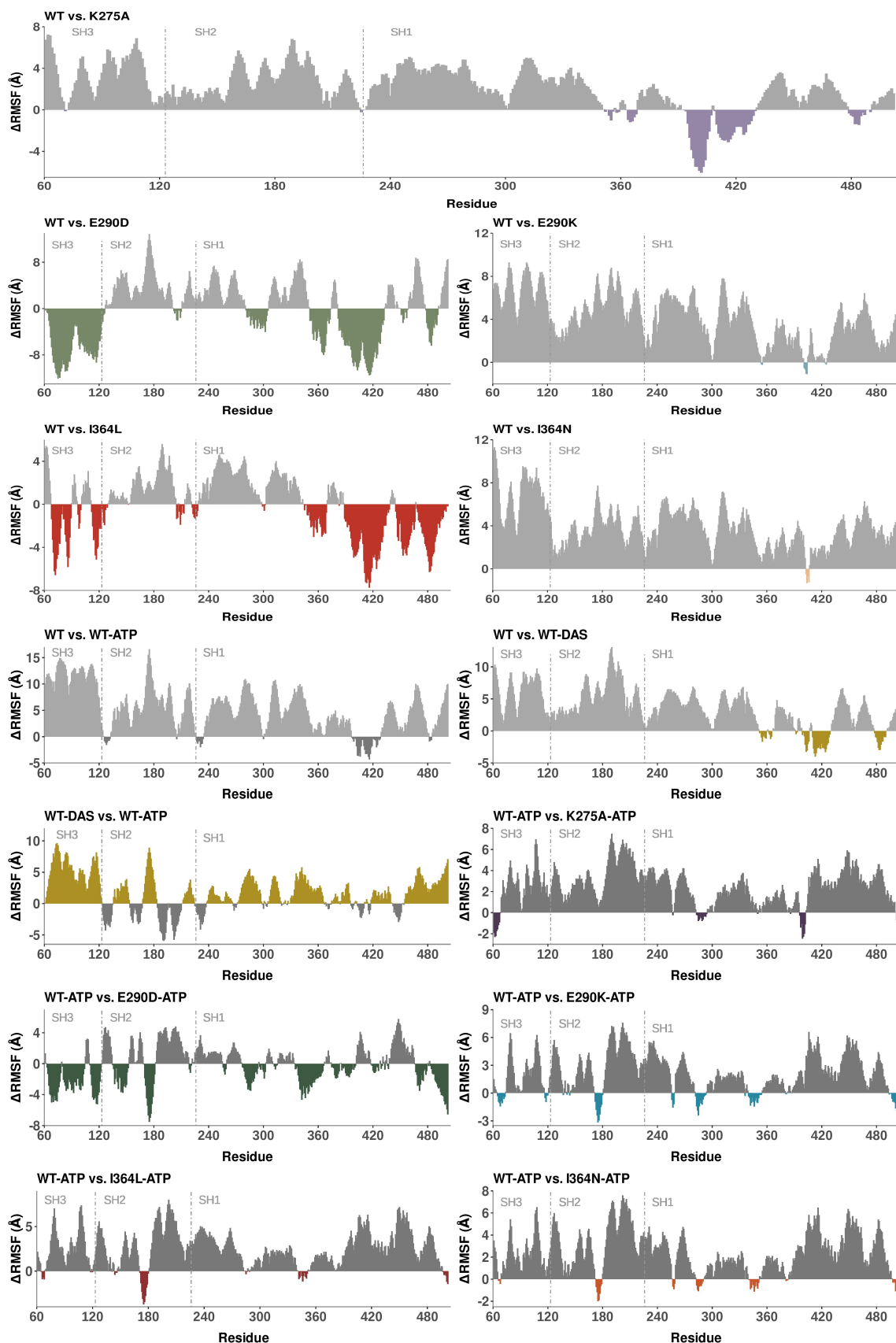
E290K and I364N reduced residue-level fluctuations across multiple domains, indicating a shift toward more structurally restrained behavior commonly associated with reduced catalytic efficiency. The K275A mutation altered flexibility in a domain-specific manner, combining rigidification near the catalytic site with increased motion in neighboring segments.

In addition to mutation-induced perturbations, the effects of dasatinib binding on the dynamic behavior of Lyn were examined. Compared to WT-ATP, the WT-DAS system was associated with increased flexibility in the SH1 and SH3 domains, particularly within regulatory loops and linker regions. This contrasted with the overall rigidity observed upon ATP binding and suggested that dasatinib may induce a redistribution of dynamic fluctuations that modulate communication between domains.

These results are consistent with prior findings showing that dasatinib disrupts features of the inactive conformation, enhancing sampling of more open, flexible conformations. The comparison between WT-DAS and WT-ATP further revealed increased fluctuations in the activation loop and regions adjacent to the  $\alpha$ C helix, which are known allosteric nodes involved in kinase activation. Together, these observations suggest that dasatinib binding influences not only the ATP-binding pocket but also distal regulatory elements, modifying the broader conformational ensemble. In contrast, E290K-ATP and I364N-ATP continued to exhibit globally reduced flexibility, mirroring their previously observed rigidity in the absence of ligand and reinforcing their association with structurally constrained, inactive-like conformations. The K275A-ATP mutant again displayed hybrid behavior, with pronounced rigidity at the catalytic core and increased flexibility in peripheral regions. Notably, this mutant retained fluctuations within interdomain linkers such as SH2-SH1, indicating partial preservation of allosteric communication despite disruption of catalytic site architecture.

Prominent  $\Delta$ RMSF changes were observed in loop and linker segments across all domains, including residues 60–126 in SH3, 202–210 in SH2, 281–305 and 345–369 in SH1, along with regions of the activation loop (384–407, particularly 401–404) and the distal C-terminal segment (479–487). These regions represent structurally and functionally important positions that appear particularly responsive to perturbations introduced by oncogenic mutations [188].

Overall, the  $\Delta$ RMSF profiles illustrate how specific substitutions and ligand interactions alter the conformational dynamics of Lyn kinase. Conservative mutations promote broader backbone fluctuations and may support conformations compatible with catalytic function, whereas charge-reversing mutations appear to stabilize more conformationally restricted ensembles. Dasatinib binding enhances flexibility in regulatory regions, unlike the conformational restriction observed with ATP binding.



**Figure 4.4: Comparative residue-level flexibility changes across WT, mutant, and ligand-bound Lyn kinase systems.** Each panel shows  $\Delta\text{RMSF}$  values representing differences in  $\text{C}\alpha$  atomic fluctuations between two systems, with the reference listed first. Positive values indicate greater flexibility in the reference system, whereas negative values indicate increased fluctuations in the compared system. Domain boundaries for the SH3, SH2, and SH1 regions are indicated by vertical dashed lines.

These findings emphasize the importance of local dynamics in shaping the broader conformational behavior of protein kinases and provide mechanistic insight into how structural alterations modulate the conformational landscape of Lyn. Taken together, the RMSD and RMSF analyses reveal how different mutations and ligands affect the conformational stability and local flexibility of Lyn across domains. While these analyses provide valuable insight into structural fluctuations, they do not capture how such motions are coordinated across the protein.

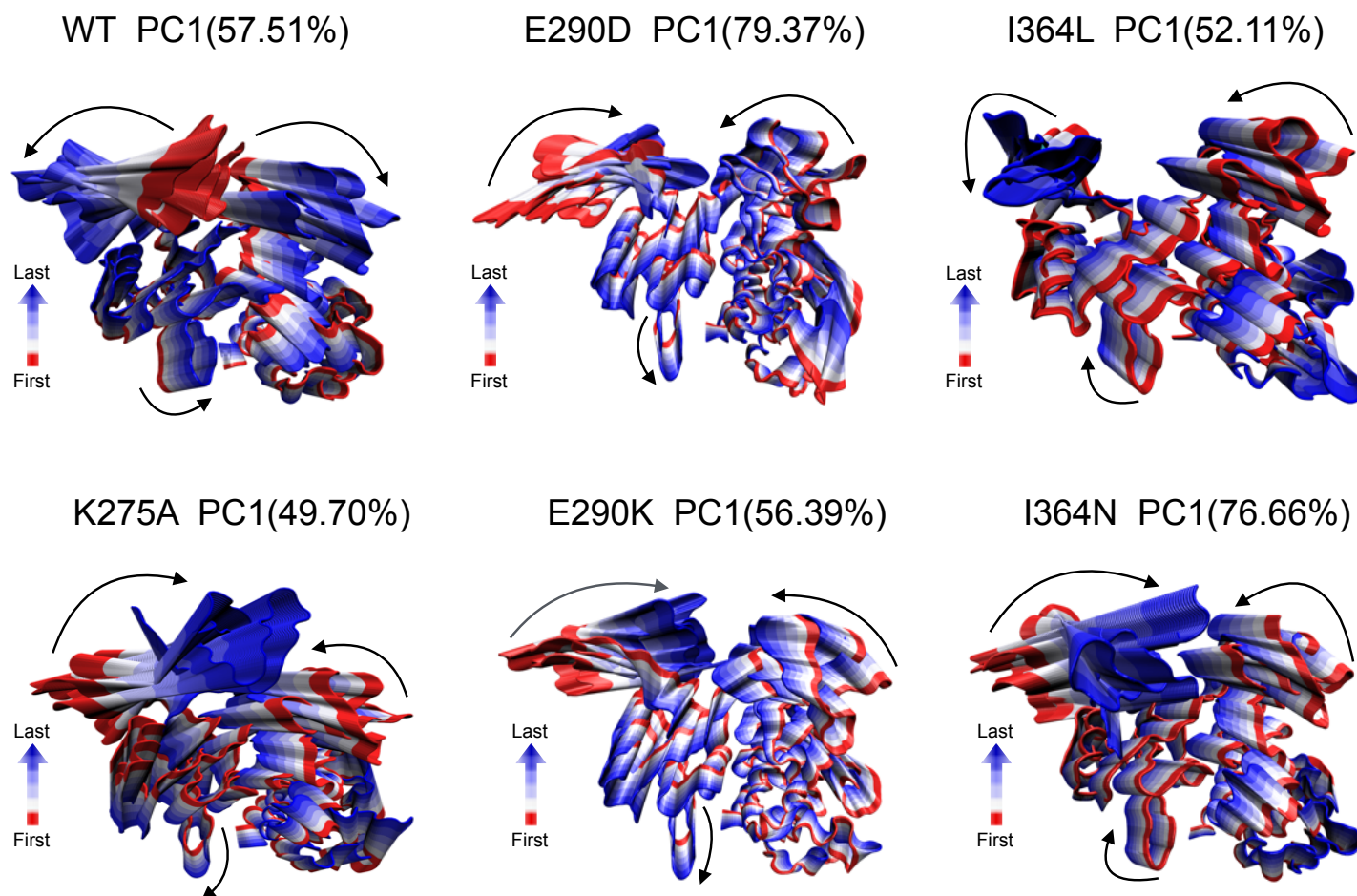
## Coordination of Motions

Since coordinated atomic displacements often underlie biologically relevant conformational transitions [189,190], PCA was applied to the MD trajectories of the WT and mutant apo systems, including K275A, E290D, E290K, I364L, and I364N. This approach captures the dominant modes of collective motion within the multidomain structure, providing deeper insight into how single-point mutations influence the global conformational landscape of Lyn kinase.

The analysis focused on the covariance matrix of C $\alpha$  atomic positional fluctuations, enabling the extraction of dominant modes of motion that reflect collective and functionally relevant structural rearrangements. The PC1, which captured the highest variance in atomic displacements, was used to represent the dominant global motion characteristic of each system. Across the systems analyzed, PC1 accounted for approximately 49% to 79% of the total variance, highlighting its central role in describing large-scale conformational changes.

Figure 4.5 presents dominant motions projected along PC1, illustrating the most prominent conformational transitions observed across the simulations. In the WT system, PC1 captured a coordinated yet oppositely directed displacement between the SH1 and SH3 domains, indicating a flexible interdomain dynamic mode reminiscent of motions typically linked to active-like conformations. This interdomain displacement is proposed to facilitate allosteric communication and maintain catalytic accessibility at the ATP binding site, reflecting the inherent adaptability of the native Lyn architecture.

In contrast, the mutant systems displayed markedly altered motion signatures along PC1. The K275A system, lacking the conserved catalytic lysine, displayed constrained and non-canonical displacement, consistent with impaired large-scale domain coupling. The E290K mutation, involving charge reversal at a conserved catalytic position, and the I364N mutation, introducing polarity into a hydrophobic region, both produced highly localized atomic displacements with minimal propagation across domains. These restricted motion profiles suggest a loss of long-range dynamic coordination and diminished interdomain coupling.



**Figure 4.5: Dominant global motions captured by PCA.** Snapshots representing extreme projections along PC1 are shown for WT and mutant systems (K275A, E290D, E290K, I364L, and I364N). PC1 accounts for the highest variance in atomic fluctuations and captures large-scale domain movements. Structures are colored from red (start of motion) to blue (end of motion), illustrating the directionality and amplitude of motion. Arrows indicate the dominant direction of displacement along PC1 for SH3, SH2, and SH1 domains. Percentage values in parentheses denote the variance explained by PC1 in each system.

The conservative substitutions E290D and I364L, by contrast, retained broader domain-scale motion and exhibited greater structural displacement along PC1 than their disruptive counterparts. These systems preserved more of the structural flexibility observed in the WT, suggesting that local physicochemical similarity may support a more adaptable dynamic profile.

Altogether, these findings suggest that single-point mutations can substantially reshape the dominant motion landscape of Lyn kinase, modifying the distribution of flexibility across its multidomain architecture. The directionality and extent of displacement along PC1 emphasize how local residue changes spread through the multidomain architecture, shifting the balance between flexibility and rigidity. These system-specific motion patterns reinforce the idea that the conformational ensemble of Lyn is closely tied to its collective dynamics. Similar to observations in Lyn, MD-based studies of c-Src kinase revealed that large-scale domain rearrangements and intermediate conformational ensembles can be sampled *in silico*, highlighting collective motions as key determinants of functional transitions [191–193].

Analyses of global and local dynamics, along with the principal modes of motion, revealed distinct alterations in overall stability and domain-specific flexibility of the Lyn protein. However, the mechanistic consequences of these perturbations within the catalytic core required further investigation. To clarify how structural changes influence the capacity of the catalytic region to accommodate ATP, and to contrast these effects with the binding of the ATP-competitive inhibitor dasatinib, multiple structural and energetic parameters derived from long MD trajectories were systematically evaluated. These included local residue proximity, coordination of catalytic magnesium ions, ligand positional stability, binding pocket size, and binding free energies. Through the integration of these analyses, changes in the plasticity of the active site, ATP interaction strength, and distribution of conformational effects across the catalytic cleft were dissected. Moreover, substrate- and inhibitor-bound ensembles were compared to identify structural determinants associated with active-like conformations, as opposed to those stabilized upon ATP-competitive inhibition. Structural organization and flexibility of the catalytic region under distinct binding conditions were systematically evaluated through this comparison.

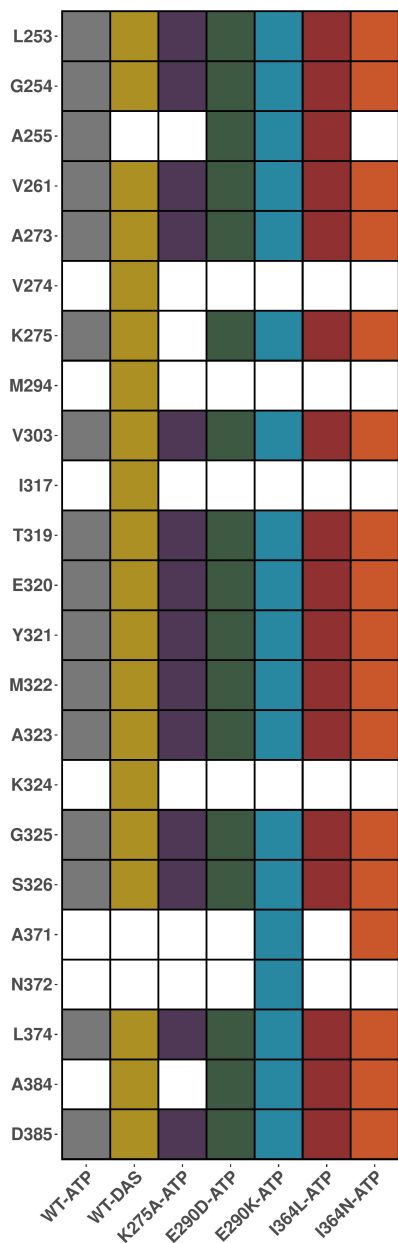
## Local Residue Proximity and Binding Environment

To examine differences in ligand-interacting residues, proximity analysis was performed across all ATP- and dasatinib-bound systems. Residues located within 5 Å of the ligand were identified to define the immediate binding environment. Figure 4.6 presents a heatmap summarizing these contacts, capturing residue-level variations in ligand proximity across the seven holo systems.

A common set of ligand-proximal residues was identified across all ATP-bound systems,



including L253, G254, V261, A273, V303, the T319–A323 loop, G325, S326, L374, and D385. These contacts define the canonical SH1 pocket and contribute to stable ATP accommodation. Additional residues such as A255, A371, N372, and A384 appeared variably in specific ATP-bound mutants, reflecting context-dependent rearrangements of the binding site. Notably, the WT–ATP, E290D–ATP, and I364L–ATP systems exhibited highly similar contact patterns.



**Figure 4.6: Composition of the local ligand-binding environment.** Heatmap of residues within 5 Å of the ligand across ATP- and dasatinib-bound systems.

In contrast, a rearrangement of ligand-proximal contacts was observed in the E290K–ATP system. Residues such as A371 and N372, not present in the WT–ATP system, were identified in proximity to the ligand. These differences may reflect altered ligand positioning, potentially influenced by electrostatic effects arising from the lysine substitution at position 290. A similar pattern was observed in the I364N–ATP system, where A371 was present and A255 was absent. Although the catalytic core, including the T319–A323 loop and D385, was preserved in both systems, the redistribution of residue contacts suggests changes in the geometry of the ligand-binding region, with possible implications for  $\text{Mg}^{2+}$  coordination and catalytic alignment.

In the K275A–ATP system, the loss of proximity to K275 was not accompanied by the formation of new contacts. No compensatory residues were identified, and a reduction in overall contact frequency was observed compared to the WT–ATP system, including the absence of A255. The core architecture of the SH1 pocket appeared preserved, suggesting that no major rearrangement occurred in the local structure. These observations may indicate that ATP remains accommodated by the scaffold, albeit with reduced electrostatic stabilization due to the missing lysine interaction.

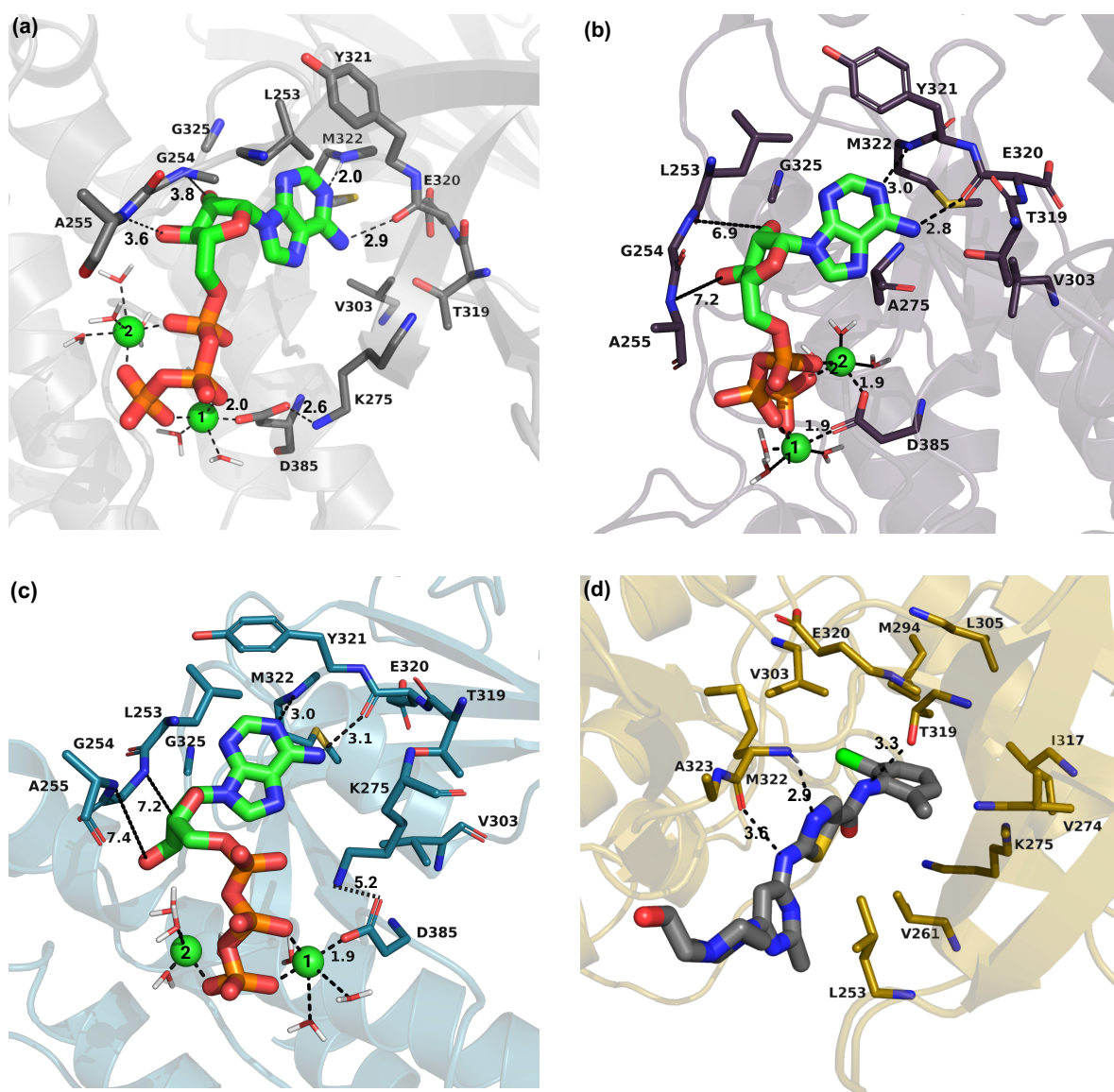
In the WT–DAS system, a broader distribution of ligand-proximal residues was observed, including V274, M294, I317, K324, and A384, several of which extended beyond the canonical ATP-binding cleft. This expanded contact pattern was consistent with the larger and more elongated structure of dasa-

tinib, enabling interactions with loop and beta sheet regions not engaged during ATP binding. Residues uniquely contacted by dasatinib were absent in all ATP-bound systems, indicating a non-canonical binding profile and highlighting structural differences between substrate and inhibitor interactions.

To complement the residue proximity analysis across systems, representative structures from the most populated conformational clusters were analyzed. The clustering results for all systems are summarized in Table A7. Figure 4.7 visualizes ligand positioning within the ATP-binding site for four selected systems that exhibited greater variation in ligand-proximal contacts, highlighting differences in coordination geometry and local residue engagement. In the WT-ATP structure, panel (a), ATP is positioned in a compact, curved conformation within the SH1 pocket. Key hydrogen bonds are maintained with catalytic and metal-coordinating residues, including D385-Mg1 (2.0 Å), D385-K275 (2.6 Å), and E320-adenine (2.0 Å). Both  $\text{Mg}^{2+}$  ions are stably coordinated by D385, the ATP triphosphate moiety, and surrounding water molecules. Coordination of Mg1 involves the carboxyl oxygen atoms of D385, while both magnesium ions engage with the phosphate groups of ATP, supporting stable ligand accommodation.

Panels (b) and (c) correspond to the K275A-ATP and E290K-ATP systems which illustrate the structural effects of mutations at the ATP-binding site. In both systems, the interaction between K275 and D385 that was observed in the WT-ATP structure was no longer maintained. In the K275A model, substitution of the catalytic lysine with alanine resulted in the complete loss of the positively charged side chain, eliminating direct interactions with the ATP phosphate groups. In the E290K-ATP system, K275 remained present but was displaced, increasing its distance from D385 and reducing the electrostatic interaction between these catalytic residues. In both cases, this disruption corresponded with reduced coordination of the ATP phosphate tail.

Moreover, the conformation of the ATP molecule is visibly altered in the mutants. Compared to the relatively compact horseshoe-like shape in the WT system, ATP appears more bent and distorted in both K275A-ATP and E290K-ATP. In the K275A-ATP system, this effect is especially pronounced, with ATP undergoing a clear rearrangement and rotation within the binding pocket. These structural changes reflect a loss of anchoring interactions and are consistent with an impaired, inactive-like conformation of the catalytic site. Such loss of electrostatic stabilization and phosphate coordination has been previously associated with inactive or catalytically impaired kinase conformations in structural and mutational studies [18, 194, 195]. Figure A3 provides an overview of the binding geometries observed across other analyzed mutants, illustrating how different mutations alter ATP placement and interactions with surrounding residues.



**Figure 4.7: Structural representations of ligand positioning within the ATP-binding site of Lyn kinase.** Shown are representative structures from the most populated conformational clusters of each system.. (a) WT-ATP: ATP adopts a compact, curved conformation within the SH1 pocket, stabilized by key interactions with D385, K275, E320, and two Mg<sup>2+</sup> ions. (b) K275A-ATP: Substitution of the catalytic lysine eliminates the K275-D385 interaction and weakens ATP anchoring, leading to a bent ATP geometry. (c) E290K-ATP: K275 remains but is displaced from D385, weakening electrostatic coupling and reducing phosphate coordination. (d) WT-DAS: Dasatinib engages a broader and more spatially dispersed residue network within the cleft, forming van der Waals and  $\pi$ -stacking interactions distinct from those stabilizing ATP. Atoms are colored as follows: carbon (ligand) in green/gray, nitrogen in blue, oxygen in red, and Mg<sup>2+</sup> ions in green spheres. Dashed lines indicate coordination distances with catalytic residues and metal ions.

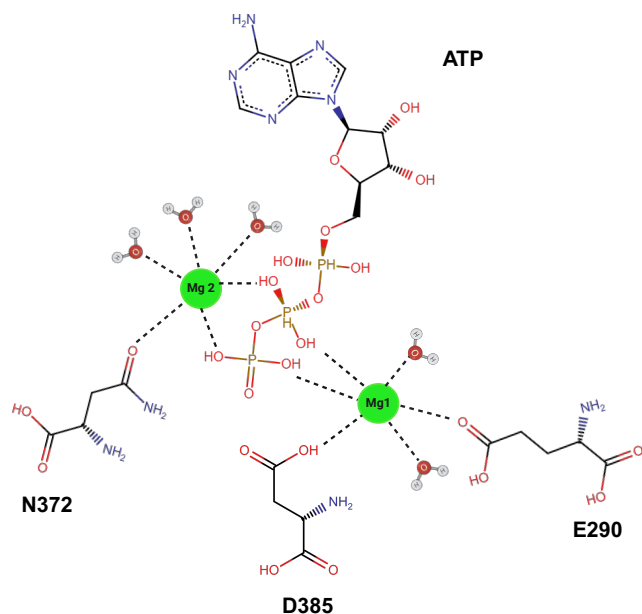
In contrast, dasatinib was observed to engage the ATP-binding cleft in a distinct manner, as shown in panel (d). Although it occupies the same pocket, its interactions extend to a broader and more spatially distributed set of residues, particularly in the  $\beta$ -sheet and loop regions of the SH1 domain. This expanded contact pattern is consistent with the larger and more rigid molecular structure of dasatinib. Both the crystal structure (PDB: 2ZVA) [72] and the representative MD structures indicated placement of dasatinib

within the hydrophobic pocket formed by residues of the  $\beta$ -sheet and  $\alpha$ C-helix, including K275, M294, V303, and I317. Unlike ATP, which is stabilized through hydrogen bonding and magnesium coordination, dasatinib engages primarily through  $\pi$ -stacking and van der Waals interactions. These included persistent contacts with M322 (2.9 Å), A323 (3.6 Å), and T319 (3.3 Å), in agreement with previous structural studies [72,196]. Although largely nonpolar, dasatinib was occasionally observed to form transient hydrogen bonds with backbone atoms in the hinge region. While the overall binding pose remained consistent with crystallographic data, flexible adjustments in the adjacent loop regions and  $\alpha$ C-helix were detected during the simulations. These local motions were more apparent in the full-length context and under solvent conditions, potentially reflecting contributions not captured in static crystal structures.

Dasatinib interactions with the binding pocket were observed to persist in approximately 80% of simulation frames, indicating stable residue engagement over time. Based on distance measurements, approximately 44% of the molecule extended beyond the canonical ATP pocket, with an overall length of 19.9 Å and an estimated 11.2 Å segment buried within the cleft. The remaining portion remained solvent-exposed and formed transient contacts with surrounding loop regions. This extended binding footprint, together with the rigid aromatic core of dasatinib, supported a distinct interaction pattern compared to ATP. The observed geometry and spatial distribution of contacts were consistent with an inactive-like conformation of the kinase domain, similar to those described in crystallographic studies of other inhibitor-bound TKs, including Abl and c-Src [197–199].

Taken together, dasatinib was observed to engage the ATP-binding site through a distinct and spatially extended interaction network, differing substantially from the compact binding mode of ATP. This non-canonical orientation was associated with broader residue engagement and a reduced reliance on classical catalytic contacts. Differences in ligand structure, in combination with specific kinase mutations, collectively altered the local binding environment, disrupted phosphate group stabilization, and reshaped the geometry of the catalytic cleft. Notably, disruption of the K275–D385 interaction, a hallmark of the active kinase architecture [18,19], correlated with structural distortion of the catalytic cleft, highlighting increased plasticity and a shift toward inactive-like geometries.

Given that ATP positioning is tightly governed by the coordination of catalytic  $\text{Mg}^{2+}$  ions, the stability and geometry of these interactions were next assessed across systems. Proper  $\text{Mg}^{2+}$  coordination is a defining structural feature of active kinases and plays a central role in anchoring ATP within the cleft. In canonical kinase mechanisms, each  $\text{Mg}^{2+}$  ion typically adopts an octahedral geometry coordinated by oxygen atoms from the ATP phosphates, water molecules, and side chains from conserved residues [200]. Figure 4.8 illustrates the representative  $\text{Mg}^{2+}$  coordination environment in Lyn kinase.



**Figure 4.8: Representative schematic of coordination of magnesium ions bound to ATP in the Lyn kinase active site.** Coordinating atoms from ATP phosphate groups, water molecules, and selected side chains of protein residues were included. D385 was depicted as the most frequent  $\text{Mg}^{2+}$  ligand, while E290 and N372 contributed to coordination under conditions associated with more saturated ligand environments.

in persistent  $\text{Mg}^{2+}$  interactions, indicating their structural importance within the local coordination network.

To gain deeper insight into the coordination environment of the  $\text{Mg}^{2+}$  ions, Figure 4.9 further illustrates the coordination geometry of Mg1 and Mg2 across WT and mutant systems. Each panel shows the distribution of distances between the ions and coordinating residues throughout the MD trajectories, revealing how mutations affect the integrity of the coordination network essential for kinase function. In the WT–ATP system (panel a), both Mg1 and Mg2 are consistently coordinated by oxygen atoms of D385, with low median distances and frequent detection across frames (e.g., Mg1 is coordinated by D385-OD2 in over 5800 frames). The WT system displays stable and symmetric magnesium coordination, representing the active form of the kinase. This serves as a reference for comparing the mutant systems.

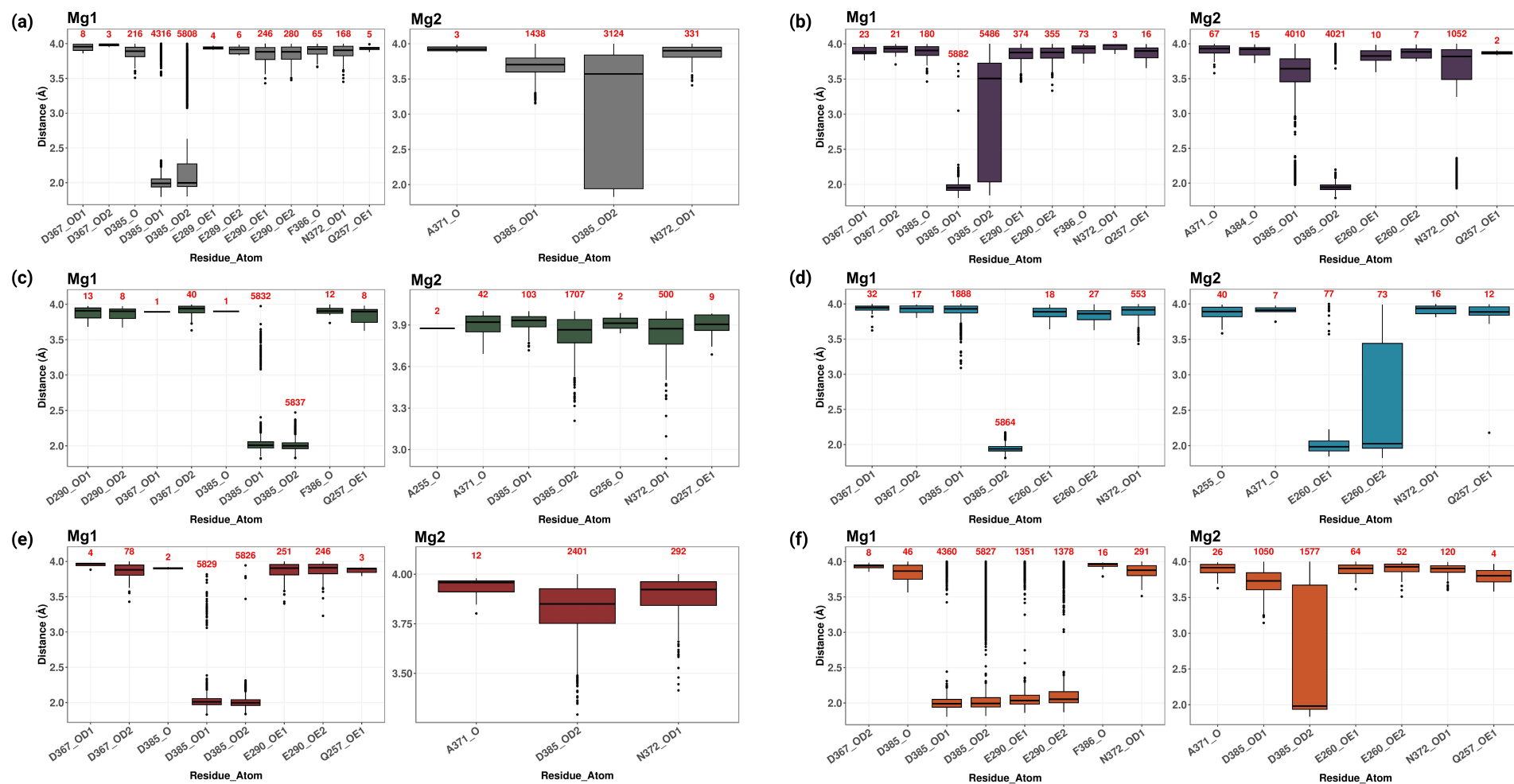
Panel (b) shows the K275A–ATP, which preserves the overall coordination pattern of Mg1 and Mg2 but displays a slightly reduced contact frequency with D385, suggesting minor alterations in the geometry of the catalytic region. A more pronounced disruption is observed in the E290K–ATP (panel d), where coordination of Mg2 is markedly impaired. In particular, D385 displays a substantial increase in median distance from the ion ( $>$

In this illustration, the  $\text{Mg}^{2+}$  ions are coordinated by oxygen atoms from ATP phosphate groups, surrounding water molecules, and side chains of residues D385, E290, and N372. Analysis of MD trajectories indicated that D385 primarily interacted with Mg1 and, in certain instances, also contributed to the coordination of Mg2. In specific frames, both  $\text{Mg}^{2+}$  ions were simultaneously coordinated by oxygen atoms of this residue, suggesting a central role within the binding environment. Residues E290 and N372 also participated in  $\text{Mg}^{2+}$  coordination, particularly in frames characterized by more saturated coordination geometries. E290 predominantly interacted with Mg1, while N372 contributed to coordination of Mg2. These residues were observed more frequently than others to engage

4 Å), indicating a loss of stable interaction. This reduction in D385 coordination is associated with local structural distortion near the catalytic residues. In the I364N-ATP (panel f), coordination of Mg<sup>2+</sup> is further destabilized. Interactions with D385 and E290 are significantly weakened, accompanied by broader distance distributions and reduced contact persistence (e.g., the D385-OD2 interaction occurs in 5808 frames in WT-ATP, compared to only 1577 frames in I364N-ATP). The side chain of asparagine at position 364 may compromise local structural stability, further weakening magnesium binding. In panels (c) and (e), the E290D and I364L mutants maintain relatively stable Mg<sup>1+</sup> coordination, while Mg<sup>2+</sup> coordination appears partially reduced. In both systems, D385 displays a decreased number of interactions with Mg<sup>2+</sup>.

Overall, these results indicate that specific mutations at catalytic or structurally coupled residues produce varying degrees of disruption in Mg ion coordination. The loss or asymmetry of these interactions is associated with reduced catalytic capacity, providing a mechanistic explanation for the structural alterations observed in the E290K-ATP and I364N-ATP systems. ATP adopts a tightly coordinated binding mode stabilized by conserved contacts and interactions with catalytic Mg<sup>2+</sup>, whereas dasatinib engages in a more spatially dispersed orientation with limited coordination. These distinct interaction profiles reflect the structural differences underlying enzymatic catalysis versus inhibitor binding, emphasizing the molecular basis of regulatory specificity in Lyn kinase.

Multiple structural and biochemical studies have demonstrated that complete coordination of two Mg<sup>2+</sup> ions is essential for maintaining catalytic function in protein kinases. In particular, the presence of two properly coordinated Mg<sup>2+</sup> ions has been shown to be critical for efficient phosphoryl transfer [200,201]. This requirement is conserved among tyrosine kinases, including members of the Src family. For example, the Csk from this family requires dual Mg<sup>2+</sup> binding to support enzymatic function [202]. Loss or asymmetry in Mg<sup>2+</sup> coordination reduces catalytic efficiency and destabilizes the positioning of ATP within the active site. Recent computational and structural analyses further indicate that Mg<sup>2+</sup> binding induces a favorable reorganization of the catalytic region and stabilizes ATP geometry required for phosphoryl transfer [203]. In kinases lacking canonical Mg<sup>2+</sup> binding motifs, such as CASK, enzymatic activity is severely reduced, underscoring the structural role of metal coordination [204]. In the systems examined here, mutations such as E290K and I364N disrupted coordination geometry and reduced or unbalanced Mg<sup>2+</sup> interactions. This disruption was associated with increased structural distortion of the catalytic cleft and supported a mechanistic link between impaired metal binding and the adoption of inactive-like conformations.



**Figure 4.9: Distances between magnesium ions and coordinating residues in ATP-bound systems.** Boxplots show the distribution of distances (in Å) between magnesium ions Mg1 (left) and Mg2 (right) and their coordinating atoms throughout MD trajectories for (a) WT-ATP, (b) K275A-ATP, (c) E290D-ATP, (d) E290K-ATP, (e) I364L-ATP, and (f) I364N-ATP systems. Red numbers above each box indicate the number of trajectory frames in which the corresponding interaction was detected.

Ultimately, elucidating how alterations in the catalytic cleft environment, particularly those affecting the coordination of ATP and  $\text{Mg}^{2+}$  ions, influence the spatial orientation and stability of ATP provides mechanistic insight into how these disruptions interfere with the alignment of reactive groups required for efficient phosphoryl transfer. To further assess the structural consequences of impaired coordination, ligand flexibility and spatial accommodation within the binding site were subsequently investigated.

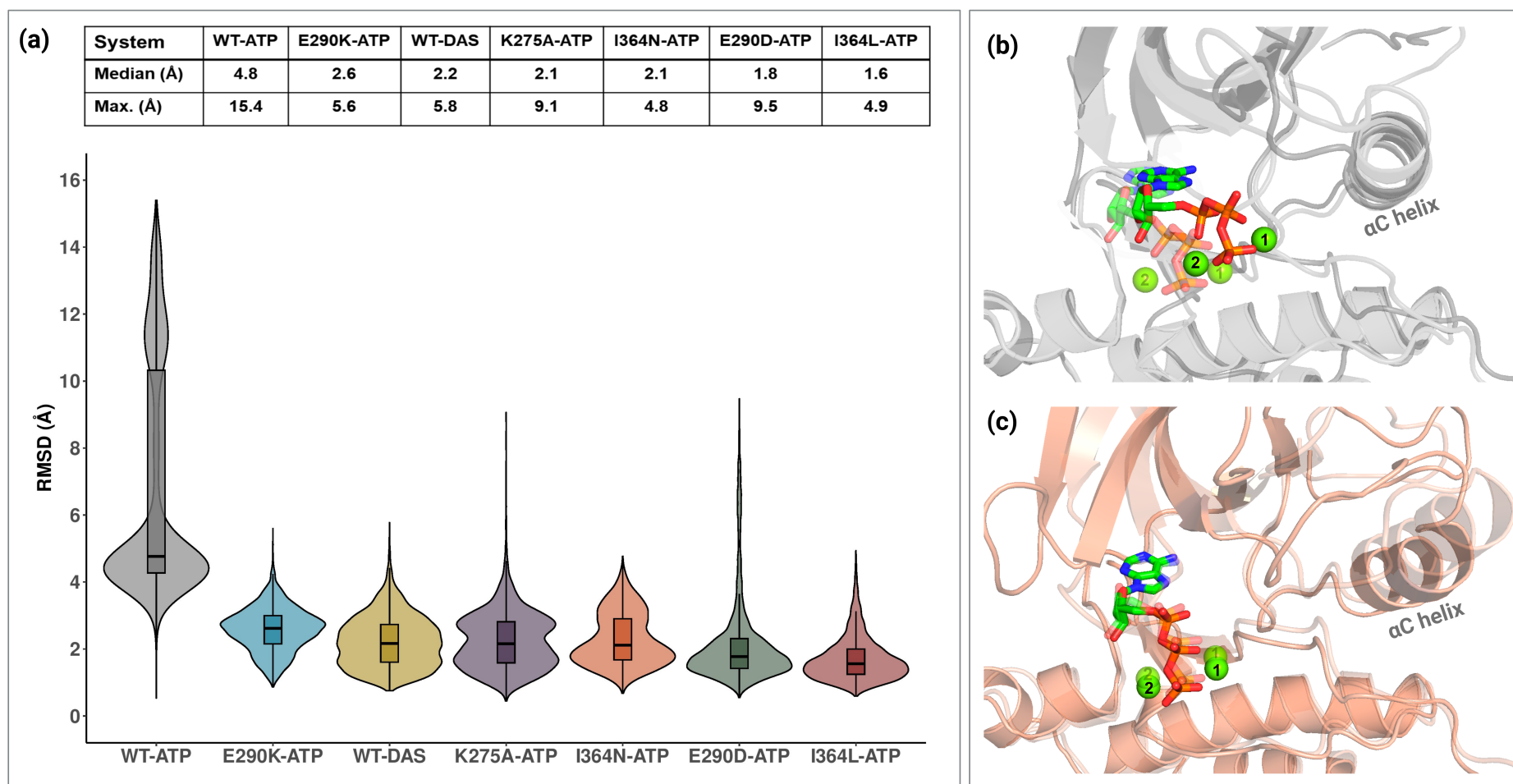
## Dynamic Remodeling of the Catalytic Binding Site

In light of the observed coupling between coordination of ATP and  $\text{Mg}^{2+}$  ions and the geometry of the catalytic cleft, the effect of ligand remodeling across systems was subsequently evaluated. Such remodeling directly influences catalytic efficiency, as precise ligand positioning is essential for proper alignment of the phosphate groups and effective phosphoryl transfer [205,206]. As shown in Figure 4.10a, the ligand RMSD distributions demonstrated that ATP sampled a more restricted range of positions in all mutant systems compared to the WT-ATP complex. Median RMSD values were consistently lower across every mutant, reflecting reduced internal flexibility and a more tightly anchored placement of ATP within the catalytic pocket. This stabilization suggested that the introduced mutations imposed spatial constraints or altered the shape and accessibility of the cleft, thereby limiting ATP displacement during the trajectory. These effects implied that local rearrangements around the binding site restricted ligand breathing motions and confined its conformational ensemble to a narrower spatial range within the pocket.

Closer inspection of representative structures (Figure 4.10, panels b and c) was performed to obtain a more nuanced interpretation of ligand flexibility. The structures shown correspond to the medoids of the most populated trajectory clusters and were used as the structural basis for subsequent analyses of ligand geometry and pocket organization. Details regarding the number of clusters and their relative populations across systems are provided in Table A7. In the WT-ATP system (panel b), ATP was observed to undergo gradual reorientation toward the  $\alpha\text{C}$  helix during the simulation, adopting a position closer to the catalytic core and aligning along an axis consistent with an active-like conformation [20,207]. This shift reflected the intrinsic adaptability of the WT binding pocket, enabling ATP to sample catalytically favorable conformations. Similar ligand transitions have been reported in active Src and Abl kinases, where interactions between ATP and the  $\alpha\text{C}$  helix help stabilize the active conformation [208,209].

In contrast, ATP in all mutant systems remained more rigidly positioned and did not exhibit the reorientation observed in the WT system. Instead, it persisted in a location farther from the  $\alpha\text{C}$  helix, as illustrated by the I364N-ATP structure (panel c). The ligand in this mutant was confined to a deeper and more static pose that limited conformational breathing and prevented the repositioning observed in the WT.





**Figure 4.10: Ligand deviation within the catalytic pocket of Lyn kinase.** (a) Violin plots of ligand RMSD distributions for ATP and dasatinib across all systems, depicting the variability of ligand positions during the simulations. Boxplots inside each violin indicate the median and interquartile range. The WT-ATP complex exhibits the greatest ligand flexibility, while all mutant systems and WT-DAS show more restricted ligand motion. The summary table above reports median and maximum RMSD values for each system. (b) Representative structures from the most populated clusters of the WT-ATP system, with ATP gradually shifting toward the  $\alpha$ C helix, reflecting a catalytically permissive orientation and enhanced positional adaptability. (c) Representative structures from the most populated clusters of the I364N-ATP system, where ATP remains rigidly positioned away from the  $\alpha$ C helix, consistent with an inactive-like conformation and reduced flexibility required for catalytic alignment.

This restricted geometry reduced the approach toward the  $\alpha$ C helix and thereby hindered the establishment of the spatial arrangement characteristic of active-like conformations. The observed reduction in ligand flexibility was consistent with an inactive-like geometry or catalytically incompetent conformation, in which the active site was misaligned for efficient phosphoryl transfer.

The RMSD of dasatinib was observed to be significantly lower than that of ATP in the WT system, reflecting more rigid binding and reduced structural displacement during the simulation. As a competitive inhibitor, dasatinib occupied the same catalytic site as ATP but exhibited minimal positional fluctuation throughout the trajectory. This reduced motion reflected a more fixed orientation within the pocket, with limited structural adjustment over time. Such rigid binding behavior contrasted with the flexible accommodation observed for ATP and suggested that dasatinib restricted the range of motion typically sampled within the catalytic cleft. Through this stabilization, a reduction in local adaptability was induced, contributing to the formation of a more rigid structural ensemble. Statistical comparisons of ligand RMSD values across all systems are summarized in Table A8.

These findings suggest that while conservative substitutions may have restricted ATP movement without fully eliminating catalytic potential, mutations characterized by alterations in charge or polarity were associated with a more rigid local environment that limited ligand repositioning toward the  $\alpha$ C helix, thereby impeding the transition required for catalysis. In contrast, the native WT environment supported dynamic realignment of ATP toward key catalytic regions, a feature likely essential for kinase activation and effective substrate positioning. Although the structural analyses provided clear evidence that mutations and inhibitor binding altered ATP positioning and restricted ligand flexibility within the catalytic cleft, such geometric constraints alone were insufficient to explain the observed reduction in catalytic efficiency and inhibitor impact. To further investigate the energetic consequences of these interactions, the following section examines the binding free energies of ATP and dasatinib across all systems.

## Evaluation of Ligand–Protein Binding Free Energies

Distinct RMSD profiles of ATP and dasatinib reflected differing dynamic behaviors and degrees of conformational stabilization within the catalytic pocket. These positional fluctuations were shown to influence not only local structural rearrangements but also the thermodynamic favorability of ligand binding [210, 211]. To assess these contributions across systems, ligand–protein binding free energies were estimated using the MM/PBSA method. The resulting values aligned with the structural trends observed in ligand flexibility and residue-level contacts, thereby providing thermodynamic context to the observed conformational distinctions. It should be noted that MM/PBSA estimates provide qualitative comparisons rather than absolute affinities, as they are influenced by systematic

uncertainties related to entropic treatment and dielectric parameter selection.

Table 4.1 summarizes the ligand–protein binding free energies calculated via the MM/PBSA method. The first column reports the average total binding free energy ( $\Delta G_{\text{total}}$ ) across all analyzed trajectory frames, while the second column shows the median value of the energy distribution, which captures central tendencies while mitigating the influence of fluctuations and outliers. In this analysis, the WT–ATP complex exhibited the most favorable binding energy ( $-28.2$  kcal/mol), consistent with strong ligand engagement in a catalytically aligned geometry. Conservative mutations such as I364L and E290D moderately reduced the interaction strength ( $-22.9$  and  $-24.6$  kcal/mol, respectively), in agreement with their more rigid yet still competent binding poses.

In contrast, the charge-altering or polarity-reversing mutations I364N and E290K led to significantly weakened ATP binding, with total binding free energies of  $-18.1$  and  $-17.9$  kcal/mol. These reductions suggest that such mutations disrupted key electrostatic interactions and compromised the capacity of the catalytic site to accommodate ATP. Similarly, the K275A mutation resulted in a less favorable binding energy ( $\Delta G_{\text{total}} = -22.1$  kcal/mol), likely due to the loss of stabilizing contacts between ATP and the conserved catalytic lysine. This disruption further weakened the electrostatic interaction network required for optimal ligand engagement.

The WT–DAS complex produced a binding free energy of  $-24.2$  kcal/mol, approximately 4 kcal/mol weaker than WT–ATP. While this value still indicated a relatively stable interaction, it reflected the broader and more nonspecific binding profile of dasatinib. These findings supported the interpretation that although dasatinib engaged Lyn in a conformationally stable manner, its interactions were more diffuse and less targeted than those of ATP, in line with its broad-spectrum inhibitory mechanism.

Table 4.1: Binding free energy estimates for ligand-bound Lyn kinase systems from MM/PBSA calculations. Mean total and median values of the energy distributions are reported in kcal/mol.

| System    | $\Delta G_{\text{total}}$ (kcal/mol) | Median (kcal/mol) |
|-----------|--------------------------------------|-------------------|
| WT–ATP    | $-28.2$                              | $-22.6$           |
| WT–DAS    | $-24.2$                              | $-24.2$           |
| K275A–ATP | $-22.1$                              | $-22.8$           |
| E290D–ATP | $-24.6$                              | $-24.4$           |
| E290K–ATP | $-17.9$                              | $-18.0$           |
| I364L–ATP | $-22.9$                              | $-23.3$           |
| I364N–ATP | $-18.1$                              | $-19.3$           |

Overall, the MM/PBSA results underscored how differences in ligand stabilization and the surrounding electrostatic environment translated into measurable variations in binding strength across systems. However, binding affinity is also influenced by the structural accessibility and spatial organization of the catalytic cleft, which govern how effectively a ligand can be accommodated and repositioned over time. To assess how these structural features varied across WT and mutant systems, the geometric openness of the catalytic pocket was subsequently evaluated.

## Active Site Accessibility

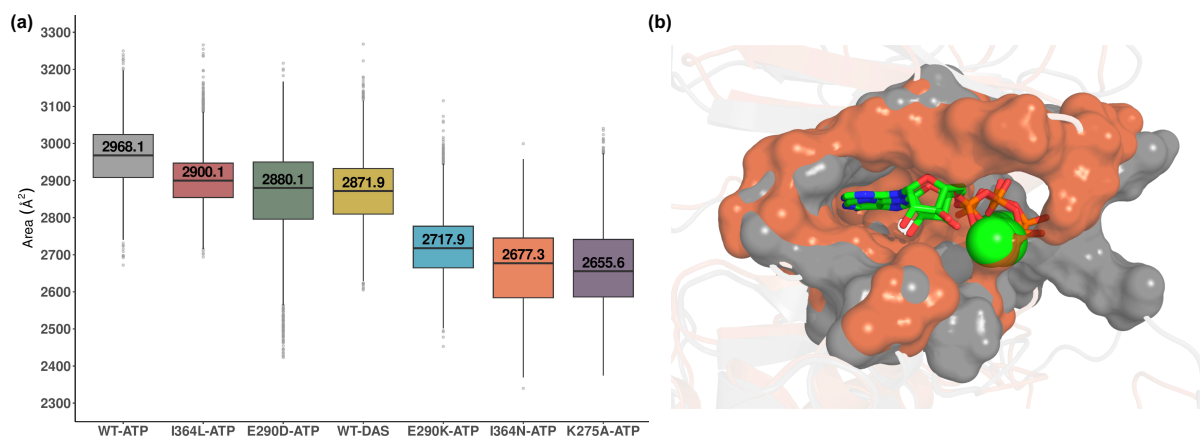
To assess how ligand type and cancer-associated substitutions influenced the geometric accessibility of the catalytic site, the SASA of the ATP-binding pocket was computed for all systems. As shown in Figure 4.11, the WT-ATP complex exhibited the most open catalytic pocket, with a surface area of 2968.1 Å<sup>2</sup>. This expanded geometry reflected a conformation favorable for substrate engagement and efficient phosphoryl transfer.

In contrast, all ATP-bound mutant systems displayed reduced SASA values, indicating narrowing of the catalytic cleft and potential restriction of ligand access and conformational adaptability. Among the mutants, K275A-ATP, I364N-ATP, and E290K-ATP exhibited the most pronounced reductions, with surface areas of 2655.6, 2677.3, and 2717.9 Å<sup>2</sup>, corresponding to decreases of approximately 10.5%, 9.8%, and 8.4% relative to WT. These reductions aligned with prior flexibility analyses, suggesting that these mutations induced a more rigid and spatially constrained active site. E290D-ATP displayed a modest reduction (~3.0%), while I364L-ATP retained a relatively open pocket (~2.3% decrease), consistent with its more flexible and catalytically permissive behavior.

The WT-DAS system showed an intermediate pocket size (2871.9 Å<sup>2</sup>), smaller than that of WT-ATP but more open than the most rigid ATP-bound mutants. This observation suggests that dasatinib, although occupying the catalytic cleft, stabilized a distinct conformation that is neither fully active nor fully inactive, in agreement with its non-catalytic binding mode. Visual comparison of pocket surfaces in WT-ATP and I364N-ATP systems (Figure 4.11, panel b) further highlighted the extent of spatial narrowing induced by the I364N mutation.

Overall, the findings emphasized the critical role of spatial accessibility and conformational adaptability at the ATP-binding site in maintaining catalytically favorable conformations of Lyn kinase. Disruption of this environment, whether through mutations that rigidify the binding cleft or through inhibitors that restrict the catalytic geometry, was associated with reduced ligand flexibility, impaired metal coordination, and a shift toward inactive-like conformations. In contrast, the WT system preserved a more open and flexible pocket architecture, which not only enabled efficient ligand accommodation but also supported the structural rearrangements necessary for maintaining global conformational

adaptability.



**Figure 4.11: Comparison of binding pocket surface areas across Lyn systems.** (a) Boxplots show the SASA of the catalytic pocket across all systems. The WT-ATP system exhibits the largest average surface area, while all ATP-bound mutants display varying degrees of reduction, suggesting a narrowing or partial collapse of the binding environment. (b) Structural overlay of the pocket surface in WT-ATP (gray) and I364N-ATP (burnt orange). The mutant displays a visibly more constricted pocket around the ATP molecule, indicating a reduction in accessible space that may limit ligand accommodation and positional flexibility.

Taken together, these complementary analyses revealed that catalytic function in Lyn kinase depended on the coordinated tuning of multiple structural and energetic features at the active site. High local residue proximity, stable  $\text{Mg}^{2+}$  coordination, elevated ligand flexibility, favorable binding free energy, and an open pocket geometry collectively characterized an active-like conformation. The WT-ATP system consistently exhibited all of these characteristics, supporting efficient substrate accommodation and dynamic adaptability within the catalytic cleft. In contrast, inactive-like conformations emerged in systems where these features were compromised. For instance, mutations such as E290K and I364N simultaneously weakened magnesium coordination, reduced ligand RMSD, lowered binding strength, and constricted the pocket surface, thereby locking the kinase into a more rigid and catalytically incompetent conformation. Even conservative mutations like I364L and E290D, while partially preserving native features, induced modest impairments in ATP positioning and electrostatic stabilization. Notably, the dasatinib-bound system illustrated a distinct inhibitory mechanism, characterized by low ligand RMSD and intermediate SASA, reflecting a stable but non-productive interaction. These findings underscore that no single structural descriptor fully defines catalytic capacity. Rather, it is the concerted interplay among residue proximity, metal ion coordination, ligand dynamics, pocket openness, and thermodynamic favorability that determines the kinase ability to adopt and sustain an active-like conformation [5,212]. Disruption of any one element may propagate through the structural network and shift the energetic equilibrium toward an inactive-like conformation, highlighting the complexity and interdependence of features governing functional regulation in Lyn.

## 4.4 Conclusions

This chapter presented a comprehensive analysis of the conformational dynamics of full-length Lyn kinase, integrating domain-resolved structural features, ligand interactions, and system-specific motion profiles across WT, mutant, and ligand-bound systems. Distinct flexibility patterns were observed across domains, with the SH3 domain emerging as a prominent modulator of overall structural adaptability. RMSD and RMSF analyses demonstrated that both cancer-associated mutations and ligand binding reshaped the dynamic behavior of individual domains and interdomain linkers, thereby influencing the broader conformational landscape.

Principal component analysis reinforced these findings by highlighting mutation-induced alterations in the vector and magnitude of dominant motions. While the WT system retained a coordinated dynamic mode that supported interdomain transitions, mutant systems exhibited reduced or spatially decoupled displacements, consistent with impaired long-range communication. These results emphasized how localized residue perturbations can propagate through the protein and reprogram both local interactions and global conformational behavior.

ATP binding was found to stabilize the kinase structure, especially in the SH1 domain, maintaining features consistent with a catalytically competent arrangement. In contrast, dasatinib increased flexibility in regulatory domains while restricting conformational movement at the catalytic cleft, consistent with a stabilizing influence that may suppress catalytic alignment. Mutational effects were context-dependent: conservative substitutions (E290D, I364L) preserved dynamic features and surface accessibility, whereas charge-reversing and polarity-altering mutations (E290K, I364N) produced globally rigidified conformations associated with impaired ATP binding and disrupted magnesium coordination.

Binding site analyses further demonstrated how specific substitutions altered the catalytic environment, impacting ligand positioning, residue contacts, and  $Mg^{2+}$  coordination geometry. These local perturbations corresponded to measurable differences in surface area and predicted binding free energy, reinforcing the connection between flexibility and ligand accommodation. Among all systems, WT retained the most adaptable and catalytically favorable geometry, while the I364N and E290K systems consistently adopted conformations consistent with impaired catalytic potential.

These findings demonstrated that enzymatic function in Lyn kinase depends on the concerted alignment of structural and energetic features within the catalytic region. However, certain limitations must be acknowledged. The simulation time, although extensive, remains finite and may not fully capture slower or rare events. The use of classical force fields introduces approximations in describing electronic effects, particularly for metal

coordination and ligand polarization. Moreover, direct experimental validation of the observed conformational behaviors was not performed. Despite these constraints, the consistency of structural trends across replicate trajectories and systems reinforces the robustness of the interpretations.

Importantly, this chapter has established a rigorous structural and dynamic framework for understanding how local perturbations, whether through mutation or ligand engagement, reshape the broader conformational landscape of Lyn kinase. The insights derived here serve as a foundation for the next chapter, which shifts focus from local catalytic rearrangements to long-range communication pathways within the protein. Chapter 5 explores how spatially distant regions interact through coordinated motions, leading to the emergence of regulatory connectors that mediate allosteric control in the multidomain context of full-length Lyn.

## Chapter 5

# Allosteric Communication Networks and Regulatory Connectors in Full-length Lyn Kinase

*The one who knows not,  
and knows not that he knows not, is lost.  
Knowledge is born of questions,  
and questions arise from doubt.  
Only the one who dares to doubt  
will ever come to truly know.*

– Ibn Sina (11th-century philosopher and polymath)



## 5.1 Introduction

Allosteric regulation enables proteins to integrate signals across distant regions, allowing localized perturbations such as small-molecule association or amino acid substitutions to influence global conformational behavior. In multidomain TKs such as Lyn, this long-range communication cannot be fully understood from static structural snapshots. Instead, it arises from distributed patterns of correlated motion that interconnect the SH3, SH2, and kinase domains and govern the conformational ensembles adopted by the protein [213].

MD simulations reveal the temporal coordination of motions between structurally distant regions of a protein and dynamic cross-correlation analysis offers a powerful means to identify patterns of coupled motion between residues. These correlations provide a system-level view of how domains interact and how structural or chemical perturbations might alter this balance [214]. Within these networks of correlated motion, interface residues play a particularly important role in transmitting long-range influence across domains. These residues can act as bridges, helping to transmit dynamic information flow from one region to another. Identifying which of these residues are central in the network of communication provides a deeper understanding of how the protein maintains its coordinated behavior, and how this coupling may break down in disease-related mutations or inhibitor-bound conformations [215].

The communication networks in Lyn are likely to be relevant beyond this single kinase, given that many SFKs share similar structural organization and regulation mechanisms [216]. By comparing these networks across different Src family members, common allosteric elements may be identified. Such comparative observations provide a foundation for exploring alternative control mechanisms in SFKs and may support the development of strategies aimed at achieving selective modulation through distant, non-catalytic regions.

To build on these insights and move from pairwise correlations to higher-order communication analysis, a network-based approach was applied. Correlation matrices derived from MD trajectories were converted into residue-level networks that reflect the architecture of dynamic coupling across the protein. These networks were analyzed using Louvain community detection to identify groups of residues with internally coordinated motion. Betweenness centrality was computed to prioritize residues positioned at the interface between communities, representing potential communication bridges. These residues are defined as cross-domain allosteric connectors and are hereafter referred to as “connectors” for brevity. The evolutionary relevance of these connectors was evaluated by examining sequence conservation across the Src family using multiple sequence alignment. This analysis enabled the distinction between conserved connectors and those specific to Lyn.

This chapter addresses Aim 2 of this study by identifying allosteric connectors in full-

length Lyn kinase. These connectors were evaluated for their role in organizing long-distance coupling and influencing conformational ensembles. The results shed light on how internal communication in Lyn was redistributed under perturbation.

## 5.2 Computational Methods

**Dynamical Cross-Correlation Analysis.** To investigate coordinated motions within the protein, dynamical cross-correlation analysis was performed for all systems. This approach quantifies how fluctuations in the position of one residue are related to those of another over the course of the simulation. The analysis was carried out using the `dccm` function from the Bio3D package [217].

Each trajectory was aligned to the first frame of its respective simulation based on the C $\alpha$  atoms to remove overall translational and rotational motions. Based on C $\alpha$  atom displacements, residue pair correlations were computed to generate an  $N \times N$  matrix, where  $N$  is the number of residues. Each element in the matrix reflects the degree of dynamic correlation between two residues, with values ranging from  $-1$  to  $+1$ . Positive values indicate motion in the same direction, while negative values reflect motion in opposite directions. For comparative analysis, the upper triangle of each matrix (excluding the diagonal) was vectorized to obtain a one-dimensional representation of residue-residue correlation patterns. Pearson correlation coefficients and associated p values were then computed between these vectorized profiles to compare systems.

The resulting matrices were visualized to identify characteristic patterns of domain-level coupling and to compare the effect of mutations or ligand binding on inter-residue communication. For structural projections and PyMOL visualizations, only residue pairs with absolute correlation values  $|C_{ij}| \geq 0.9$  were included, highlighting the strongest dynamic couplings.

**Community Detection and Integration Module Identification.** To investigate how dynamic correlations are organized within the protein, community detection was performed using the Louvain algorithm [218]. This method groups residues into modules based on their connectivity within the residue-residue correlation network derived from the DCCM. For each system, an undirected weighted graph was constructed from the upper triangle of the DCCM. Residues were represented as nodes, and edges were defined between residue pairs with absolute correlation values of  $|C_{ij}| \geq 0.8$ . Edge weights reflected the magnitude of the dynamic correlation between residues. This threshold was chosen to focus on the strongest correlated motions and to avoid over-connected graphs that compromise the resolution of modular structure. Communities were then identified by applying the Louvain algorithm, which partitions the graph into densely connected

clusters.

To explore how these dynamically derived communities relate to protein structure, the module membership of each residue was mapped to its structural region (SH1N, SH1C, SH2, SH3, or linker). For each community, a purity score defined as the fraction of residues coming from the most dominant structural region was computed. Modules with high purity reflected localized domain-level dynamics, while those with low purity included residues from multiple domains, suggesting coupling between distinct structural regions. The module with the lowest purity in each system was labeled as the integration module and selected for further analysis. This module typically included residues spanning three to five structural regions, depending on the system, and was interpreted as a potential connector for allosteric communication.

**Interface Residue Identification.** To determine which residues participate in interdomain interactions, interface residues were identified on representative structures for each system using the `InterfaceResidues` plugin in PyMOL v3.1.0 [219,220]. This method estimates changes in SASA to detect contacts formed between protein regions. For each calculation, the total SASA of the full protein was compared to the sum of the SASAs of individual regions treated separately. The difference, referred to as  $\Delta$ SASA, reflects the surface area buried upon domain association. Residues with a  $\Delta$ SASA greater than  $1.0 \text{ \AA}^2$  were classified as interface residues, a commonly used threshold that indicates loss of solvent exposure when two regions interact and are therefore likely to mediate physical contact between structural domains. To systematically detect all key interdomain interfaces, three separate calculations were performed. In each case, one domain was defined as the region of interest, while the remaining structural regions were grouped to form the complementary segment. Specifically, SH1 was defined as residues 247 to 502, and its interface was calculated against the complementary region comprising residues 60 to 246. Similarly, SH2 was defined as residues 129 to 226, with the complementary segment including residues 60 to 128 and 227 to 502. For SH3, residues 60 to 123 were selected, and the complementary region included residues 124 to 502.

To assess the persistence of interface residues across the full trajectory, a distance-based vicinity analysis was performed using `gmx select` from GROMACS 2022.04 [100]. For each system, residues from the domain of interest were selected if their center of mass remained within  $5 \text{ \AA}$  of the complementary region. Residue-wise contact occurrence was recorded per frame and analyzed in R. Only residues present within the defined vicinity in at least 80% of frames were retained, ensuring selection of consistently maintained interdomain contacts. The resulting interface residues were later used to annotate allosteric connector positions and interpret their spatial relevance in interdomain communication.

### **Identification of Allosteric Connectors from Network and Interface Features.**

To identify residues with potential allosteric relevance, betweenness centrality values across all MD systems were computed for each residue within the residue–residue correlation networks. These networks were constructed using edges with absolute correlation values above a defined threshold, and centrality was calculated only within the integration module of the system. Residues were then ranked by their centrality values to assess their roles as connectors.

Three complementary features were integrated to derive an allosteric relevance score: (i) the extent to which residue centrality differed between active-like and inactive-like systems, (ii) the frequency with which a residue participated in mixed-domain modules, and (iii) the frequency with which it appeared at interdomain interfaces across systems. These three features were combined into a single interface-weighted score, and residues exceeding the predefined threshold of 0.5 on the final score were classified as putative allosteric connectors. The full mathematical definitions and scoring formulas are provided in Equations (A1–A7).

**Conservation Analysis of Allosteric Connectors Within Src Family.** To assess the evolutionary conservation of the identified allosteric connectors in Lyn kinase, a set of comparative sequence analyses was performed across eight SFK members. Pairwise sequence identity between Lyn and each homolog was first evaluated using BLAST [221], separately for full-length sequences and for isolated kinase domains. These similarity measures provided a quantitative overview of evolutionary relatedness and informed the interpretation of connector conservation across regions.

Full-length protein sequences of eight SFKs were retrieved from UniProt and aligned using MAFFT v7 [222]. The resulting multiple sequence alignment (MSA) was used for conservation scoring. Two complementary gap-aware conservation metrics were computed for each connector position in Lyn: identity-based conservation and normalized Shannon entropy. The identity score was defined as the fraction of aligned sequences that contained the same amino acid as Lyn at the given position and was expressed as a percentage, while the entropy-based metric accounted for chemically similar substitutions and alignment diversity and was a normalized value between 0 and 1 [223]. To visualize conservation patterns, the full-length MSA was annotated and displayed using Jalview (v2.11.5) [224], providing a graphical representation of aligned sequences across the Src family. Together, these analyses enabled a systematic evaluation of Lyn allosteric connectors in an evolutionary context, distinguishing residues that were conserved across SFKs from those likely to represent Lyn-specific regulatory elements. The mathematical definitions and formulas are provided in Equations (A8–A9).

**Mapping of Allosteric Drug-Binding Sites in Lyn Kinase.** To assess whether the dynamically central allosteric connectors identified in Lyn kinase overlap with known

allosteric drug-binding sites in other kinases, a comparative structural analysis was performed. A curated dataset of kinase inhibitors was retrieved from the Protein Kinase Inhibitor Database (PKIDB) [225], which includes approved and clinical-stage inhibitors annotated by binding mode and target kinases. Particular focus was placed on type 3 (distant allosteric) inhibitors, as well as those with undefined binding types that could represent uncharacterized allosteric interactions, to identify potential overlap with predicted Lyn connectors. Structural data for selected inhibitors were obtained from the Protein Data Bank (PDB) [226].

Structural alignments were performed in PyMOL, with superposition restricted to the SH1 (kinase) domains of Lyn and reference kinase-inhibitor complexes. Alignments were carried out using backbone C $\alpha$  atoms, and iterative outlier rejection was applied to optimize structural overlap and minimize alignment bias. Two types of structural comparisons were performed: domain-level alignment to assess global kinase fold similarity, and local alignment of residues within a 5 Å radius of the bound allosteric inhibitor to evaluate correspondence at the inhibitor-binding region. RMSD values were calculated following each alignment to quantify structural similarity at both global and local levels. To complement the structural analysis, sequence similarity between Lyn and reference kinases was evaluated at the kinase domain level. SH1 domain sequences were aligned using MAFFT, and the resulting MSA was visualized and inspected in Jalview. This combined structural and sequence-based approach enabled evaluation of both global kinase fold conservation and local correspondence of allosteric binding regions.

**Flexibility Analysis of Lyn-Specific Allosteric Connectors.** The local flexibility of Lyn-specific allosteric connectors was quantified using C $\alpha$  RMSF values derived from MD trajectories. RMSF was calculated for the SH1 domain across all 13 simulated systems to characterize residue-level fluctuations associated with top-ranked connectors in both active-like and inactive-like ensembles. A detailed comparative analysis was subsequently conducted for the active-like WT Lyn construct and two representative inactive-like mutants, E290K and I364N, both of which consistently demonstrated lower activation metrics and restricted sampling of conformational space in prior evaluations. C $\alpha$  RMSF values were extracted using `gmx rmsf` from GROMACS. The resulting profiles were examined to assess flexibility changes at selected Lyn-specific, non-conserved connectors. This analysis aimed to determine whether these regions exhibit dynamic features suggestive of active involvement in allosteric control.

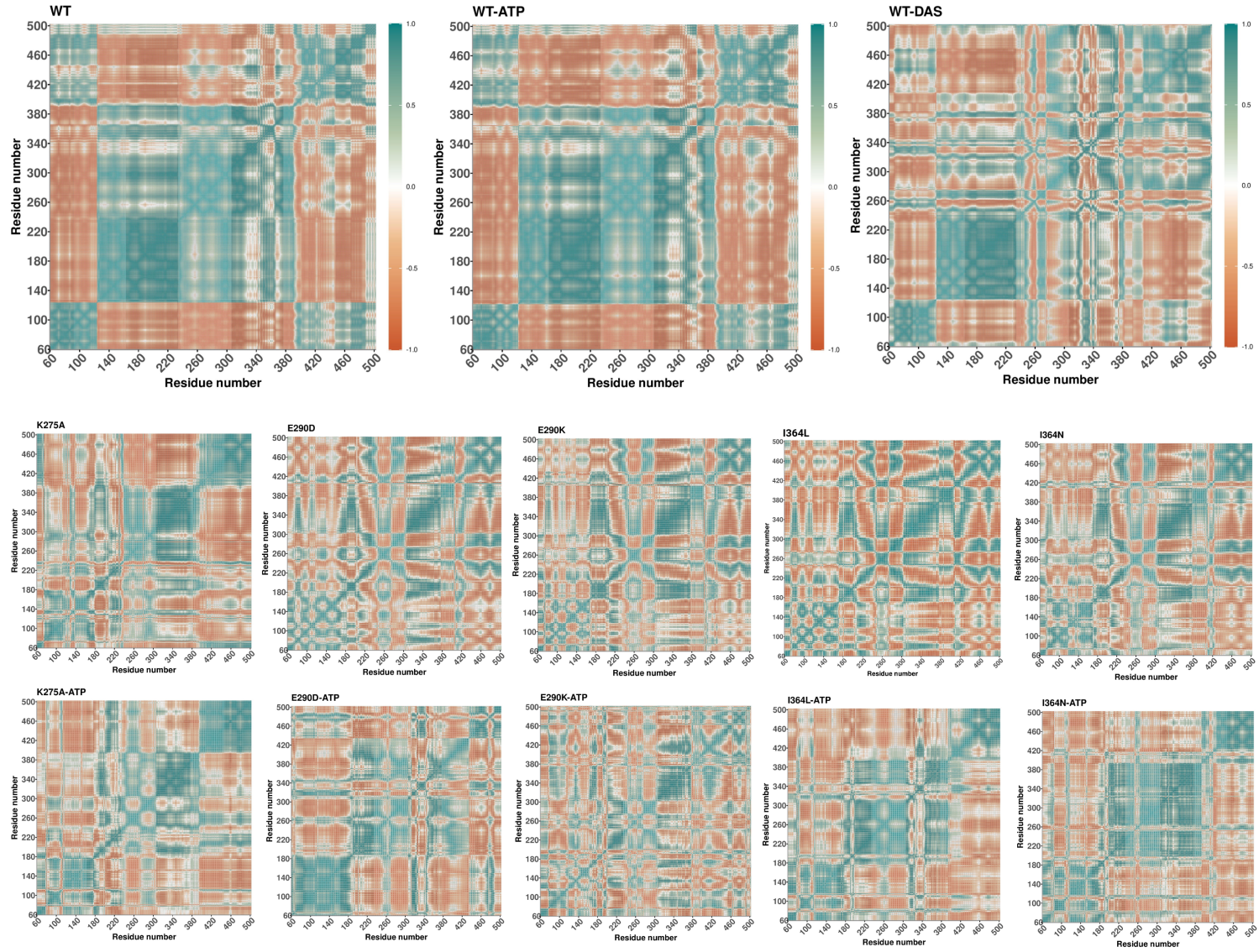
## 5.3 Results and Discussion

### Residue Correlation Across Lyn Kinase Systems

While PCA identified dominant directions of global motion, it did not describe how specific residue pairs fluctuate in relation to one another. To address this, DCCMs were computed for all systems to capture patterns of residue-residue coordination. In these matrices, each element reflected the extent to which two residues moved in a coordinated manner throughout the simulation. Positive values indicate motion in the same direction, while negative values reflect motion in opposite directions.

Figure 5.1 displays the DCCMs for WT, ligand-bound, and mutant Lyn systems. In the WT system, the correlation matrix exhibited a well-structured pattern, characterized by strong positive correlations between residues from SH2 and SH1N, and between SH3 and SH1C, as well as strong negative correlations between SH3 and SH1N, between SH2 and SH1C, and between SH2 and SH3. A highly organized pattern of internal coordination was also observed within the SH1 domain, particularly around flexible loops. These interdomain and intradomain correlations were consistent with an active-like structural organization. The WT-ATP system similarly preserved this correlation architecture, indicating that ATP binding did not disrupt the intrinsic dynamic coupling. In contrast, WT-DAS, although it maintained correlations between SH2 and SH3, showed a distinct shift, with less consistent correlations within SH1 and between SH1 and the regulatory domains. This reduction in long-range coordination suggested a ligand-induced decoupling of interdomain dynamics.

All mutants displayed a similar trend of reduced SH1-SH2 and SH1-SH3 correlation, accompanied by more fragmented regions of short-range coordination within individual domains. E290K and I364N showed an even more disrupted pattern of correlated motion with diminished long-range coupling and loss of the clear SH2-SH3 axis, suggesting impaired dynamic communication. E290D and I364L retained moderate domain-level organization, but coupling was generally weaker compared to WT. Notably, I364N displayed the most disrupted pattern among all mutations. Compared to the WT system, this mutant exhibited a pronounced reduction in interdomain correlations, particularly between SH2 and SH3 as well as SH1 and SH2. The SH3 domain was observed to be dynamically isolated, lacking coordinated motion with the rest of the protein. Consistent with this observation, the correlation map was characterized by overall lower correlation values relative to WT, indicating weakened dynamic coupling across the structure. This pattern indicated a breakdown of long-range communication networks that were preserved in WT and partially retained in other mutants.



**Figure 5.1: DCCM maps for 13 full-length Lyn kinase systems.** Each panel represents the residue-residue dynamic cross-correlation matrix computed from MD simulations. Correlation values range from +1 (strong positive correlation, shown in teal) to -1 (strong negative correlation, shown in burnt orange), with white indicating weak or no correlation.



These disrupted interdomain correlations were consistent with the structural constraints identified in Chapter 4, where RMSF and PCA analyses showed that E290K and I364N reduced the flexibility of regulatory domains and restricted the breathing motions of the catalytic cleft. These localized rigidities extended across domains, limiting the global rearrangements that normally sustained SH3–SH2–SH1 communication. The fragmented DCCM signatures observed here, therefore, reflected the same mutation-driven loss of long-range coupling identified previously.

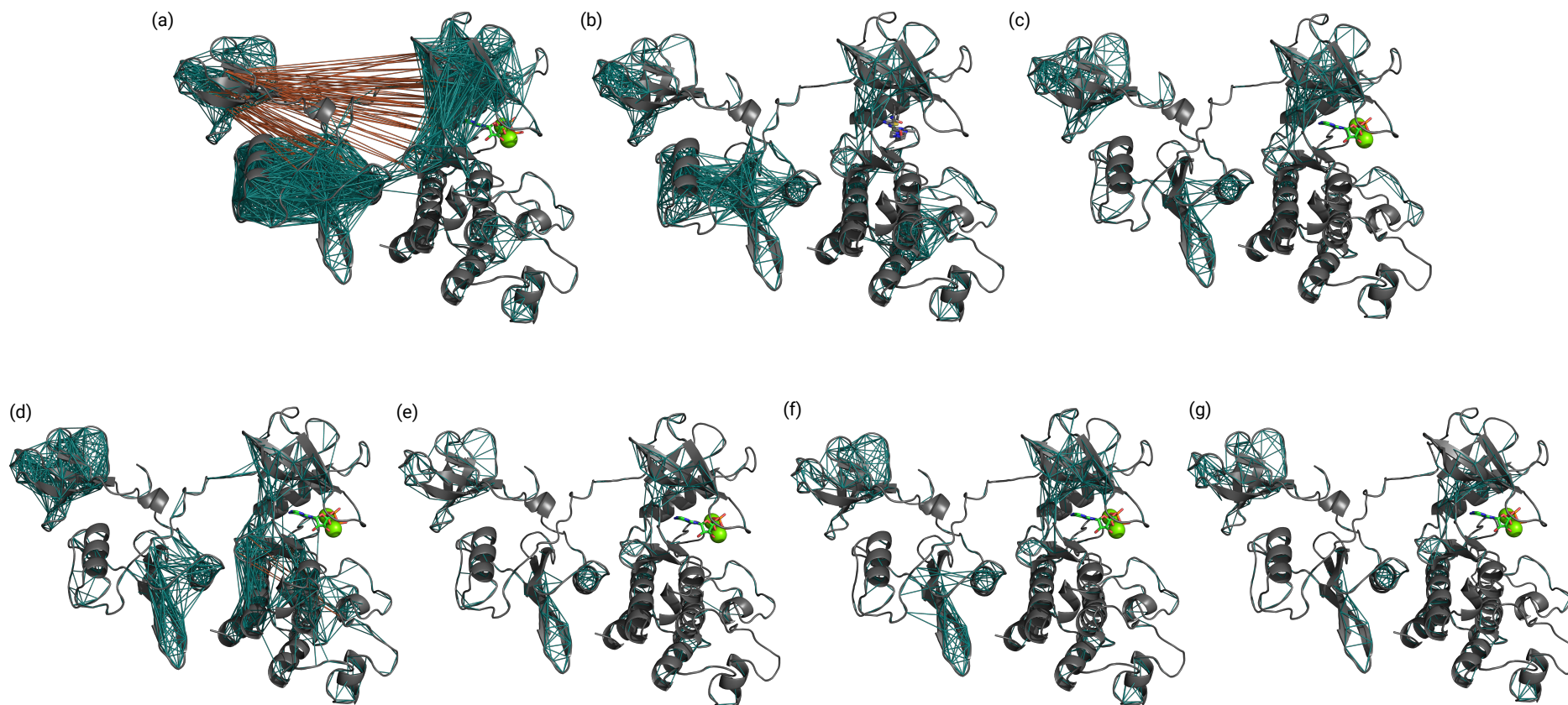
The ATP-bound mutants further revealed system-specific responses. E290D–ATP exhibited partial recovery of interdomain coordination, particularly between SH2 and SH1, while SH3 remained weakly coupled. E290K–ATP retained the disrupted correlation patterns of its apo form, suggesting a dominant effect of the mutation over ligand binding. I364L–ATP regained substantial coordination across domains, especially between SH1 and SH2, resembling a partially restored dynamic network. In contrast, I364N–ATP remained fragmented, with very limited coupling observed between SH3 and the linker region.

In the K275A system, intradomain correlations within SH1 were largely preserved relative to the WT system, whereas interdomain coupling was markedly reduced. Its ATP-bound counterpart, K275A–ATP, displayed further disruption, with interdomain coordination largely diminished. Together, these matrices demonstrated that while some mutations allowed partial recovery of communication upon ATP binding, others, particularly E290K and I364N, imposed strong structural constraints that were easily rescued. This analysis provided a comprehensive system-level view of how single-point mutations and ligand binding reshaped the internal dynamics of Lyn. The structured positive correlations characteristic of the WT system were variably disrupted in mutants, often leading to domain decoupling and altered communication networks.

To further examine how residue–residue correlations are distributed in three-dimensional space, DCCM values were mapped onto the full-length Lyn structure for all simulated systems. This spatial visualization complements the matrix-based analysis by illustrating dynamic coupling across domains. Structural projections for the ligand-bound systems are shown in Figure 5.2, while those for the apo systems are included in Figure A4. In these representations, teal lines indicate strong positive correlations ( $\geq +0.9$ ), while burnt orange lines denote strong negative correlations ( $\leq -0.9$ ) between residue pairs.

In the WT–ATP system, the structure was densely connected, with extensive teal and burnt orange edges spanning from the SH3 domain to SH2, Linker, SH1N, and SH1C. These long-range correlations illustrated robust interdomain coupling, consistent with an active-like ensemble and aligned with the ATP-stabilized conformations described in Chapter 4.





**Figure 5.2: Three-dimensional representation of residue-residue dynamic correlations in ligand-bound Lyn systems.** Spatial projections of DCCM-derived correlations are shown for ligand-bound systems: (a) WT-ATP, (b) WT-DAS, (c) K275A-ATP, (d) E290D-ATP, (e) E290K-ATP, (f) I364L-ATP, and (g) I364N-ATP. Colored lines denote strong correlations between residue pairs, with teal lines indicating strong positive correlations ( $\geq +0.9$ ) and burnt orange lines indicating strong negative correlations ( $\leq -0.9$ ). These visualizations depict differences in long-range dynamic coupling across systems and illustrate how conformational perturbations affect global communication.

In contrast, the WT–DAS system exhibited a much sparser correlation network, with most connections restricted to individual domains and limited evidence of long-range coordination. This pattern supports previous observations that dasatinib binding reduces dynamic coordination and favors an inhibitor-bound, inactive-like conformation.

Among the mutant systems, the long-range interdomain correlations present in WT–ATP were absent, indicating a consistent disruption of coordinated communication between SH3, SH2, and SH1. K275A–ATP, E290D–ATP, and I364L–ATP retained partial long-range correlations within individual domains, particularly in SH2 and SH1N. Notably, E290D–ATP also exhibited both positive and negative long-range correlations within SH1C, distinguishing it from the other mutants.

In contrast, E290K–ATP and I364N–ATP displayed correlation patterns dominated by short-range, intradomain interactions, with no extended correlations observed in SH2 or SH3. In both systems, the domains appeared dynamically decoupled, with the SH3 domain fully uncoupled from the rest of the structure, reflecting the lowest level of long-range coordination observed. These observations indicate that ATP binding was insufficient to restore global coordination in the presence of certain mutations. This dynamic decoupling reflects the broader effect of these substitutions on domain flexibility and interdomain communication, as previously observed at both structural and catalytic levels.

These spatial correlation maps reinforced the patterns observed in the DCCM matrices: the WT–ATP system preserved an extended, organized pattern of interdomain communication, whereas mutant systems such as I364N–ATP and E290K–ATP exhibited dynamic isolation of domains and a breakdown of global coupling.

While the DCCM matrices and their spatial projections helped localize regions of altered communication, a system-level comparison was required to evaluate the overall similarity or divergence of correlation networks across systems. To this end, pairwise Pearson correlation coefficients were computed between the entire DCCMs of all systems (Figure 5.3), enabling a quantitative assessment of how mutations and ligand binding influenced the overall patterns of coordinated residue motions in Lyn kinase.

As anticipated, the highest correlation coefficient ( $r = 0.95$ ) was observed between WT and WT–ATP, indicating that ATP binding preserved the core dynamic communication network of the WT protein. This high similarity confirms that ATP binding maintains the overall pattern of residue–residue correlations characteristic of the native conformational ensemble. In contrast, WT–DAS displayed a markedly lower similarity to both WT and WT–ATP ( $r = 0.45$  and  $r = 0.42$ , respectively). This outcome aligns with the fragmented and disordered DCCM pattern of WT–DAS seen in Figure 5.1, and suggests that dasatinib binding induces a distinct reorganization of the dynamic correlation network, consistent with an altered pattern of interdomain communication.

Among mutant systems, strong correlations were observed between E290K and I364L ( $r = 0.91$ ), E290D and I364L ( $r = 0.89$ ), E290K and I364N ( $r = 0.87$ ), and E290D and I364N ( $r = 0.86$ ). These observations suggest that, despite arising at distinct spatial sites, E290 within the catalytic region and I364 in a distal segment, these mutations may perturb similar global communication networks. The convergence of correlation profiles across these systems reinforces the concept that localized structural changes, whether proximal or distal to the active site, can exert a comparable influence on the allosteric communication network of Lyn.

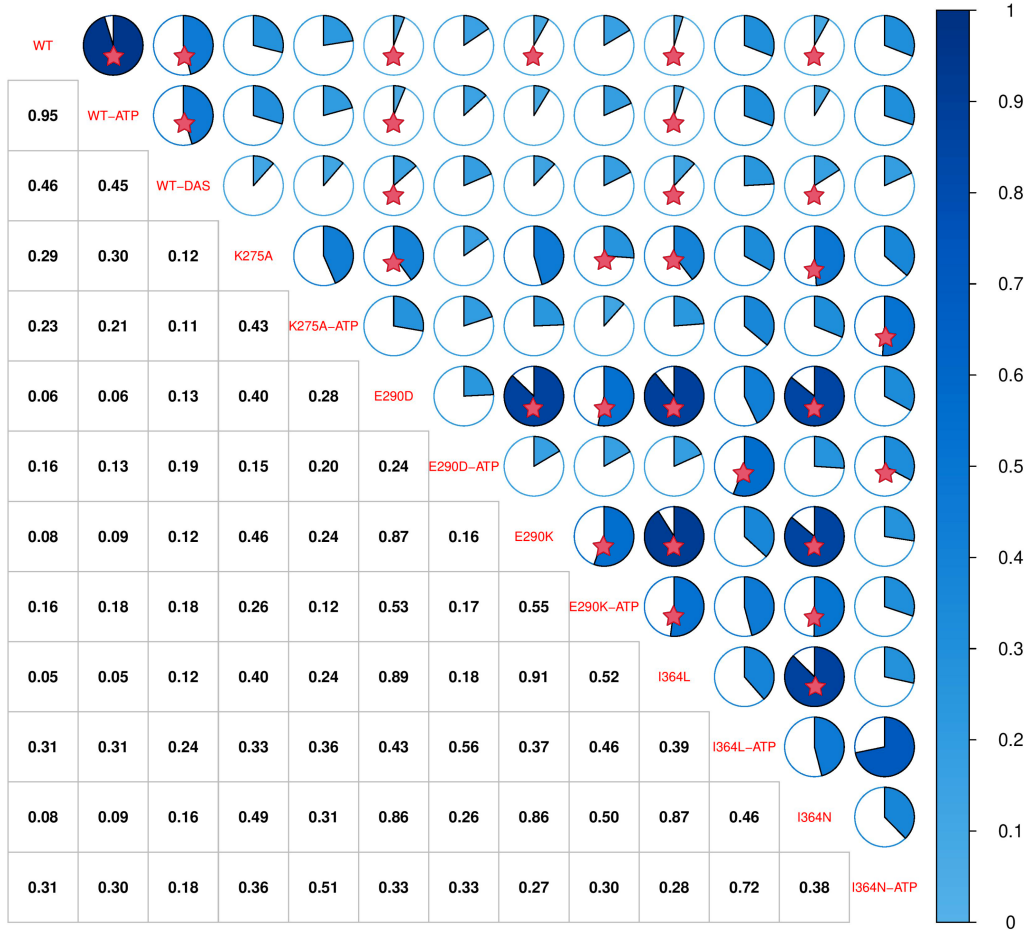
The K275A system showed moderate correlations with E290D ( $r = 0.49$ ), E290K ( $r = 0.46$ ), and I364L ( $r = 0.40$ ), suggesting partial disruption of dynamic connectivity without complete interdomain decoupling. However, the ATP-bound counterpart, K275A-ATP, exhibited weak similarity with all other systems ( $r \leq 0.51$ ), indicating a more isolated pattern of correlated motions upon ligand binding in this background.

Interestingly, the ATP-bound forms of mutant systems such as I364N-ATP, E290K-ATP, and E290D-ATP showed generally low similarity to both the WT and other mutant systems. In many comparisons, the Pearson correlation between vectorized DCCMs was below  $r = 0.55$ , and several did not meet the threshold for statistical significance ( $p < 0.05$ ). These results indicate that ATP binding does not consistently restore WT-like communication patterns in the presence of mutations. Instead, each mutant exhibits distinct changes in residue-residue correlation behavior, exposing mutation-specific disruptions in interdomain communication.

The correlation heatmap revealed that while some mutants shared similar patterns of altered dynamic coupling, others differed substantially, especially under ligand-bound conditions. This diversity highlighted the complex interplay between point mutations, domain motions, and the stabilizing or disruptive effects of ligand binding on allosteric communication networks. The corresponding  $p$  values for all pairwise system comparisons are provided in Table A9.

Overall, the DCCM-based analyses provided a detailed view of how dynamic correlations were redistributed across the Lyn structure in response to ligand binding and mutation. Matrix-level comparisons revealed that certain systems maintained coherent long-range coordination, while others exhibited more fragmented patterns, particularly in the presence of disruptive substitutions. Spatial projections further emphasized how these alterations affected interdomain communication, with several mutant systems displaying restricted connectivity across regulatory and catalytic regions. Although Pearson correlation analysis offered a system-level comparison of overall similarity in dynamic coupling, it did not directly identify which specific residues mediated the observed changes in long-range communication. To enable a mechanistic interpretation of how information might be transmitted across the protein structure, a network-based approach was applied. This

method captured the topology of residue interactions derived from DCCM data. Dynamic correlation networks were constructed and analyzed for their community structures, allowing the identification of critical residues, referred to as allosteric connectors, that bridged structural regions and contributed to the modulation of Lyn allosteric behavior.



**Figure 5.3: Pairwise Pearson correlation coefficients between DCCMs of all 13 systems.** Each circle represents the correlation between two systems, with the shaded area proportional to the correlation strength (ranging from 0 to 1, as shown in the scale bar). Red stars mark statistically significant correlations, indicating meaningful similarity in dynamic coupling patterns between the respective systems. This visualization demonstrates how specific systems exhibit shared or divergent communication patterns relative to one another.

## Network-Based Identification of Allosteric Connectors

To extend the residue-level correlation analysis toward functional interpretation, dynamic residue interaction networks were constructed from the DCCM data, and community detection was subsequently performed. This network-based approach allowed modular patterns of communication within Lyn kinase to emerge, and residues likely to serve as allosteric connectors to be identified.

Based on structural and dynamic features observed in previous analyses, WT and WT-ATP were classified as active-like systems, whereas all mutants and WT-DAS were categorized

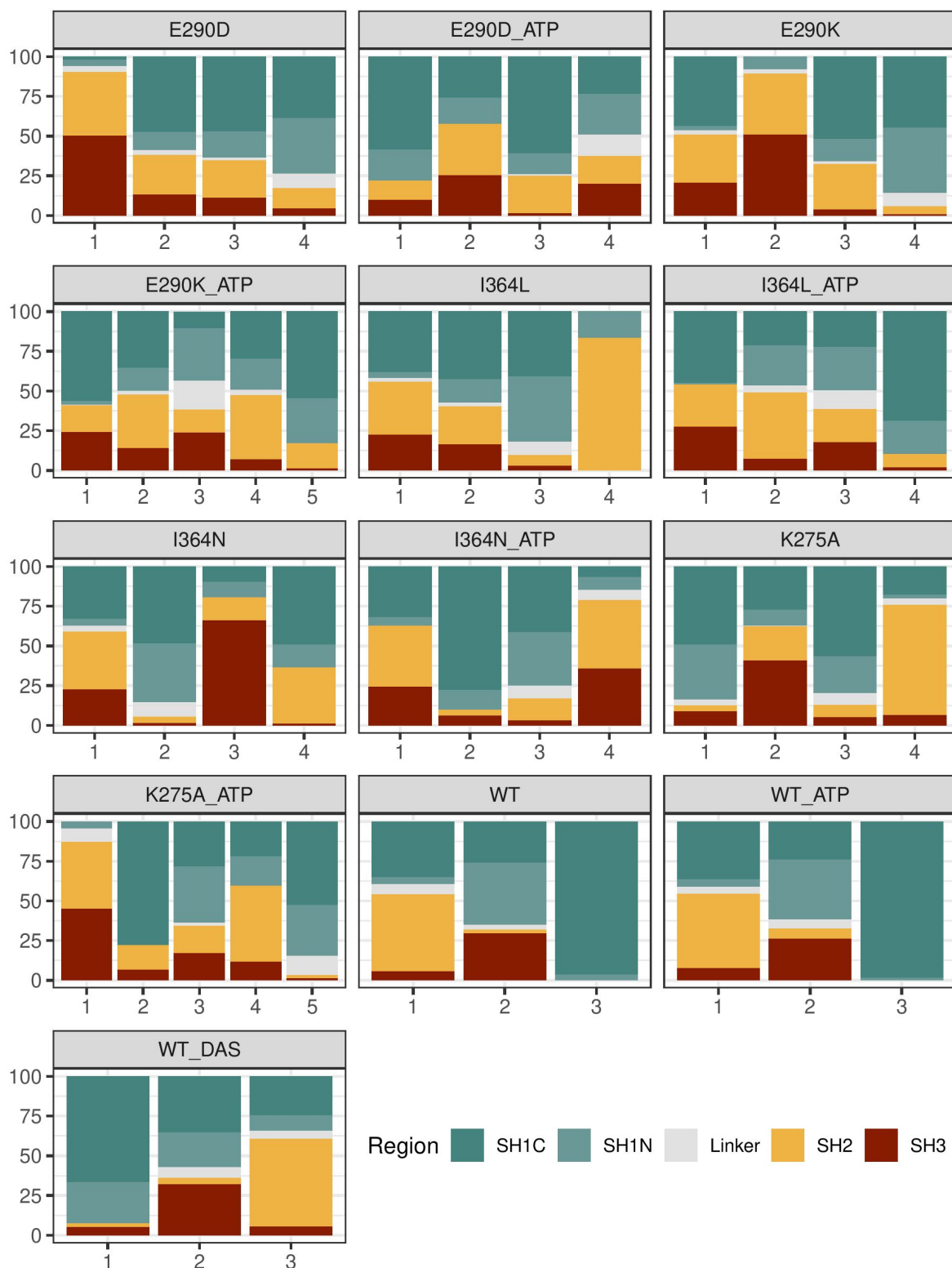
as inactive-like. This classification was guided by metrics such as domain-level RMSD and RMSF, binding site compaction, and global correlation patterns. Each DCCM was transformed into a network in which residues were represented as nodes, and connections reflected the strength of their dynamic correlations. Community detection applied to these networks enabled the identification of modules, groups of residues with coordinated motions, likely involved in shared communication pathways. This method is particularly informative in multidomain proteins such as Lyn, where cross-domain coordination serves an essential regulatory function [227].

To evaluate whether a given module remained confined to a single structural domain or spanned multiple domains, a parameter called *module purity* was introduced. This parameter was defined as the fraction of residues within a module that originated from the same structural region. A high purity score (close to 1.0) was interpreted as a structurally localized module, whereas a low purity score was associated with cross-domain communication. Modules with the lowest purity in each system were defined as *integration modules*, representing the most likely sites of long-range allosteric coupling.

For this analysis, Lyn was partitioned into five regions: SH3, SH2, the interdomain linker, SH1N, and SH1C. The SH3 and SH2 domains were considered to comprise the regulatory N-terminal region of Lyn, while the interdomain linker was described as a flexible segment connecting SH2 to the SH1 domain. The SH1N lobe (N-terminal lobe of the SH1 domain) included the ATP-binding pocket and regulatory  $\alpha$ C-helix, whereas the SH1C lobe (C-terminal lobe of the SH1 domain) was defined by the catalytic loop and activation segment. These regions were introduced to enable a direct interpretation of how each module aligned with the domain architecture.

To visualize how residues from different regions contribute to each module, the composition of every community was examined using stacked bar plots (Figure 5.4). Each bar represented one module, with colors corresponding to the proportion of residues originating from the five defined regions. This visualization allowed a determination of whether a module is confined to a single domain or distributed across multiple domains.

A key insight from Figure 5.4 was that in most systems, at least one module contained residues from all five regions, reflecting extensive cross-domain communication. In contrast, several mutant systems exhibited reduced integration. Specifically, in E290K, I364N, and K275A, the integration modules included only four regions rather than five. The most pronounced case was observed in the I364N system, where the modules were overall less mixed and lacked contribution from the interdomain linker. This suggested that the linker may have become dynamically decoupled in these mutants, possibly weakening interdomain coordination.



**Figure 5.4: Module composition and structural region purity across all systems.** Stacked bar plots show the distribution of structural regions (SH3, SH2, linker, SH1N, SH1C) within each detected community module per system, as identified by Louvain clustering on dynamic correlation networks. Each bar corresponds to a distinct module, and region-specific colors indicate the degree of compositional heterogeneity. Modules containing residues from multiple structural regions suggest strong cross-domain integration.

Notably, the WT and WT-ATP systems exhibited highly similar patterns, with their integration modules drawing from all five regions in comparable proportions. This supported the conclusion that ATP binding preserved the dynamic architecture of the active-like conformation. Interestingly, the WT and WT-ATP systems displayed only three community modules in their dynamic correlation networks, whereas many mutant systems exhibited four or five, indicating increased fragmentation of communication pathways.

This difference suggested that the WT and ATP-bound forms of Lyn maintained a more integrated and cohesive dynamic architecture, with broader regions of the protein moving in concert. Fewer modules implied stronger long-range coordination, which aligned with their classification as active-like systems. In contrast, the higher number of modules observed in most mutants reflected a more fragmented dynamic network, where structural regions moved more independently. This modular fragmentation may have underlain the loss of efficient interdomain communication seen in inactive-like conformations. Although WT-DAS also showed only three modules, previous correlation analyses (Figures 5.1 to 5.3) indicated that these modules were less structured and displayed disrupted domain-level communication, consistent with an allosterically inhibited conformation. Additionally, 17 residues were not assigned to any integration module across all systems, likely reflecting regions with limited or flexible dynamics.

To identify which residues might act as allosteric connectors, *betweenness centrality*, a graph-theoretic measure that captures nodes lying on the shortest communication paths between other nodes, was computed. The top 25 residues with the highest centrality in each system were selected as candidate connectors. Although variation was observed in the specific residues identified across systems, several positions were found to appear consistently in multiple networks, suggesting a conserved role in mediating residue-residue communication. To further explore their potential importance within the dynamic network architecture, centrality scores were examined in relation to system-level behavior (active-like vs. inactive-like). Residues with strong positive or negative correlations were regarded as important for distinguishing between conformational ensembles. To refine the list further, a *consistency score* was calculated, measuring how often a residue ranked as central across different network thresholds, along with a *mixing score* indicating how many regions were represented in each integration module. These values were then combined into a unified score.

Finally, to incorporate structural context, the locations of candidate residues at interdomain interfaces were analyzed based on residue-residue contact frequencies across the MD trajectory. Each residue was annotated according to its presence at SH1, SH2, SH3, or linker interfaces. By integrating dynamic centrality and interface location, an *interface-weighted score* was defined, which enabled the prioritization of residues most likely to serve as functional connectors. Residues with a final score above 0.5 were defined as *allosteric connectors* ( $n = 44$ ). These connectors were distributed across all five

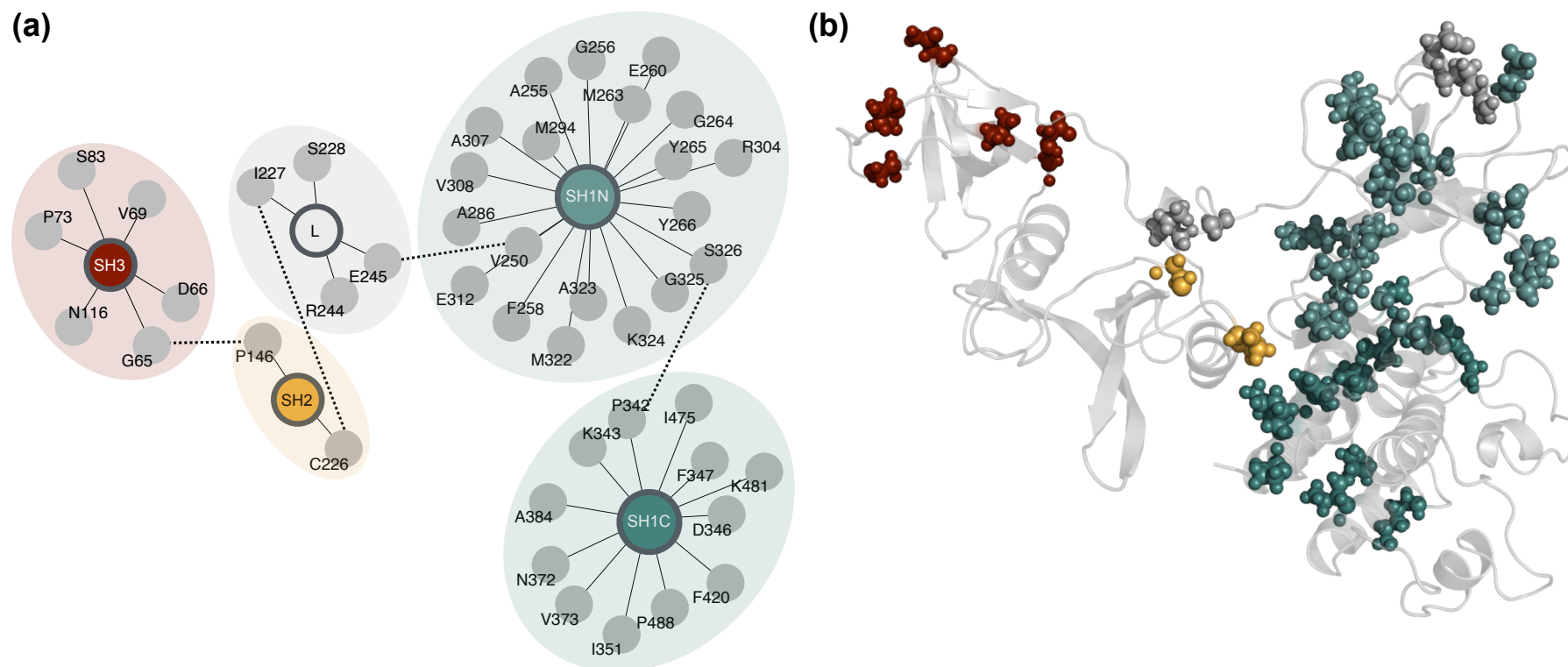
structural regions, though fewer were identified in SH2. Notably, many of the connectors were located at domain interfaces, including the SH3-SH1C boundary, SH2-SH1 linker, and SH1N lobe. This distribution is consistent with a model in which long-range allosteric control in Lyn is mediated through structurally embedded networks that span multiple domains. The complete dataset of allosteric connectors, including all calculated scores, is publicly available via Zenodo [228] (RebollidoRiosLab, 2025; <https://doi.org/10.5281/zenodo.16986533>). A summary is provided in Table A10.

The domain-level organization of connector residues is depicted in Figure 5.5a, illustrating their distribution across SH3, SH2, SH1N, SH1C, and the interdomain linker. Rather than remaining confined within individual regions, these residues form cross-domain clusters. Dashed connections between modules indicate putative interdomain communication paths derived from the correlation network. When projected onto the three-dimensional structure (Figure 5.5b), these residues form a spatially distributed network spanning the full protein architecture. This organization supports the interpretation that these positions constitute an interdomain allosteric wiring system capable of mediating long-range coupling and responding to both mutation- and ligand-induced perturbations.

The allosteric connectors identified here were more consistently retained in the active-like systems (WT and WT-ATP) than in the mutant systems and WT-DAS. This suggests that dasatinib binding, like disruptive mutations, destabilizes the native allosteric network rather than shifting it to an alternative conformation. The loss of key connector residues across some inactive-like conformations underscores a breakdown in long-range coordination and emphasizes the distinction between catalytically competent and inhibited conformations.

The identification of allosteric connectors in full-length Lyn revealed a structurally integrated communication network capable of mediating long-range dynamic coupling across multiple domains. These connectors, distributed across interface regions and enriched in active-like systems, offer valuable perspectives into the internal communication architecture of Lyn. To place these findings in a broader biological and pharmacological context, two complementary comparative analyses were performed. First, the evolutionary conservation of the identified allosteric connectors across other Src family members was assessed to determine whether these dynamic connectors represent generalizable features of SFK regulatory architecture or are specific to Lyn. Second, the spatial distribution of these connectors was compared to known allosteric binding sites in other kinases, including those targeted by FDA-approved allosteric inhibitors. This cross-comparison enabled an evaluation of whether the dynamically important regions revealed in the simulations coincide with surfaces that have already been explored pharmacologically. Together, these analyses extended the implications of the network-based insights, reinforced the potential for conserved allosteric mechanisms to be exploited in drug discovery, and established a foundation for the future exploration of non-catalytic regulatory sites in kinases.





**Figure 5.5: Spatial and network organization of the 44 allosteric connectors identified across all simulated systems.** (a) Network-based schematic showing the distribution of allosteric connectors across Lyn kinase domains. Residues are grouped and colored by structural region (SH3, SH2, interdomain linker [L], SH1N, SH1C). Dashed edges represent putative interdomain communication paths inferred from correlation network topology. (b) Three-dimensional mapping of the same 44 residues onto the WT full-length Lyn structure. Connector residues are shown as spheres and colored according to domain as in (a), illustrating their widespread distribution and role in bridging regulatory and catalytic regions. This architecture supports a model of structurally embedded allosteric communication responsive to mutation and ligand binding.

## Conservation and Divergence of Allosteric Connectors Across the Src Family

To determine whether the allosteric connectors identified in Lyn kinase represent conserved structural features within the Src family or reflect characteristics unique to Lyn, a comparative conservation analysis was performed across eight Src family members. Two related questions were addressed: which of the dynamic connectors are broadly conserved across the family, and which show divergence specific to Lyn that may contribute to its distinct regulatory properties. By combining residue-level conservation with the previously calculated allosteric scores, shared structural elements were distinguished from subtype-specific adaptations. These findings provide insight into how evolutionary pressures have shaped the dynamic architecture of SFKs and reveal opportunities for the selective targeting of Lyn through its distinct allosteric connectors.

As discussed in Chapter 1, SFKs are broadly classified into two subgroups based on sequence and functional similarity: the Lyn-related subfamily, comprising Lyn, Hck, Lck, and Blk; and the Src-related subfamily, including Src, Fyn, Yes, and Fgr. These groups reflect distinct evolutionary branches within the family and may exhibit differences in regulatory mechanisms [229]. To provide context for comparing allosteric connectors, sequence similarity between Lyn and other SFKs was assessed, both for full-length proteins and isolated kinase domains. Hck exhibited the highest sequence identity to Lyn at 70.5%, while Src displayed the lowest, approximately 60.3% (see Tables A12 and A13). These values are consistent with the subgroup classification and support a more detailed analysis of conservation patterns.

To evaluate the conservation of the 44 allosteric connectors across the Src family, a multiple sequence alignment was conducted. As summarized in Table A10, the aligned residues from other SFKs are listed alongside those of Lyn, providing a position-specific overview of residue-level variation. Two complementary conservation metrics were included to capture distinct aspects of evolutionary constraint. The Lyn identity score quantifies how many SFKs retain the exact same residue as Lyn, indicating strictly preserved positions. In contrast, the normalized Shannon entropy score accounts for chemically similar substitutions that may preserve function despite sequence differences. Using both metrics helps distinguish residues that are strictly conserved from those that are more flexible but still influential in maintaining structural or regulatory roles. Combined with the previously described dynamic and interface scores, these values provide an integrated view of the evolutionary and structural relevance of the identified connectors.

A majority of the connectors displayed moderate to high conservation, with 30 out of 44 positions showing identical residues in more than half of the family members. Several residues, including G256, M294, M322, N372, A384, and F420 were completely conserved across all aligned sequences. Despite this high degree of evolutionary conservation, these

positions received some of the lowest interface-weighted scores in the dynamic network analysis. All are located near the ATP-binding pocket or within the activation loop, suggesting that they belong to a structurally rigid and functionally essential core. Their low dynamic centrality indicates that, although they are critical for maintaining catalytic structure and ligand coordination, they contribute minimally to long-range allosteric communication. This contrast underscores the distinction between residues conserved to stabilize the enzymatic core and those involved in dynamic coordination across domains. Among these fully conserved sites, both N372 and M322 were shown in Chapter 4 to participate directly in ligand coordination: N372 coordinates  $Mg^{2+}$  in the ATP-binding site, while M322 forms direct contacts with both ATP and dasatinib. The strict conservation and functional relevance of these residues corroborate the view that this catalytic core represents a shared regulatory scaffold maintained throughout the Src family.

In contrast to these fully conserved residues, several other positions showed subgroup-specific conservation. A subset of connectors, including Y266, V250, Y265, A255, and A323, were conserved only within the Lyn-related subgroup but differed in all members of the Src-related subgroup. This selective conservation suggests that distinct sets of residues may contribute to regulatory functions unique to each subgroup. Such divergence supports the possibility that the two evolutionary branches within the Src family have adopted different allosteric coupling strategies, potentially reflecting subgroup-specific functional roles in cellular regulation.

Notably, 9 of the 44 connectors, K343, D66, S228, R304, P73, I227, G65, A255, and I475, showed both low Lyn identity scores and low Shannon scores. This combination indicates that these positions are not only divergent from Lyn in other SFKs but are also substituted with biochemically distinct residues rather than chemically similar alternatives. Strikingly, these sites received the highest interface-weighted scores in the allosteric network analysis, indicating that they occupy central positions in interdomain communication and contribute significantly to Lyn dynamic coupling behavior. Such pronounced divergence implies that these residues may represent Lyn-specific allosteric adaptations. Therefore, it can be concluded that although the main principles of dynamic allostery are conserved across the Src family, Lyn possesses unique coupling features that may underlie its specific roles in cellular signaling and cancer.

Interestingly, the four highest-scoring connectors, K343 (in SH1C,  $\alpha$ E helix), D66 (in SH3), S228 (in the linker), and I475 (in  $\alpha$ F helix), were among the least conserved. These residues are located at critical domain interfaces, including SH1C–SH2, SH3–SH2, and linker–SH3 boundaries, and their sequence divergence may contribute to differences in dynamic regulation among SFKs.

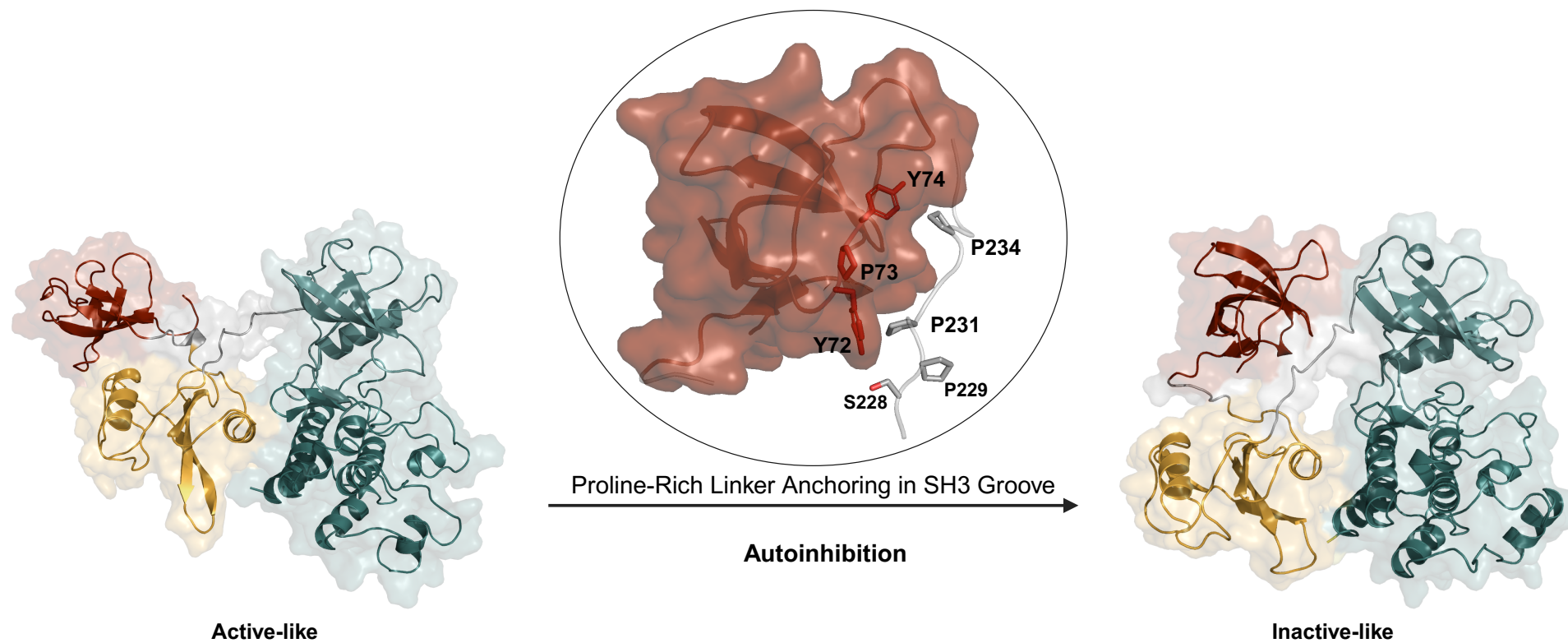
Of particular interest are Lyn-specific connectors located in regions known to mediate autoinhibition. For instance, residue P73 in the SH3 domain is embedded within a con-

served YPY motif (Y72–P73–Y74), which forms the canonical proline-rich binding groove that engages PxxP motifs from the SH2–kinase linker (Figure 5.6). This groove enables intramolecular SH3–linker interactions that are critical for maintaining the inactive-like conformation of SFKs. The two tyrosines form a hydrophobic aromatic shelf, while the central proline introduces rigidity, optimally positioning the motif for recognition of linker prolines [230, 231]. While this motif is highly conserved in Lyn, the central proline is notably absent in other SFKs, where an aspartate is present instead. This suggests that Lyn may possess a more structurally rigid SH3–linker interaction, potentially contributing to its distinct autoinhibitory behavior.

Residues I227 and S228 (the third highest-scoring connector) in the linker region represent additional Lyn-specific connectors that are not conserved in other SFKs. Structural models show that S228 is positioned near Y72 in the SH3 domain, where it may stabilize SH3–linker docking through hydrogen bonding or polar interactions. This suggests the presence of a tightly coupled linker–SH3 module unique to Lyn, with implications for its dynamic switching between inactive and active conformations.

Compared to other SFKs, Lyn exhibits markedly different domain behavior during activation. In particular, the SH3 domain plays a dominant regulatory role in Lyn, whereas regulation in other SFKs, such as Src, is more strongly influenced by SH2 or SH1C domains. Studies using HDX-MS and FRET-based conformational reporters have shown that Lyn SH3 domain undergoes distinct conformational changes upon ligand binding and phosphorylation, indicating a more active role in modulating kinase activity [232]. This domain-specific regulatory behavior may explain why certain connectors in the SH3 domain and linker are unique to Lyn and may contribute to its specialized roles in immune signaling and cancer.

Collectively, this conservation analysis reinforces the notion that while a common allosteric framework is preserved across the Src family, particularly in structurally critical regions such as the active site and activation loop, Lyn has evolved distinct allosteric couplings that differentiate it from other family members. These Lyn-specific residues, many of which occupy central positions in the dynamic network, may reflect specialized regulatory functions adapted to the unique biological roles of Lyn, especially in immune signaling and oncogenesis. Their divergence from the broader SFK consensus suggests potential vulnerabilities that could be exploited for selective therapeutic targeting. In this context, evaluating whether these dynamically important regions coincide with known allosteric binding sites in other kinases may help assess their druggability. The following section addresses this question by comparing the predicted Lyn allosteric connectors to established non-catalytic binding pockets targeted by FDA-approved allosteric kinase inhibitors.



**Figure 5.6: Structural basis of Lyn autoinhibition via the proline-rich SH3-linker interface.** Left: Active-like conformation of full-length Lyn kinase, with SH3 (red), SH2 (orange), SH1 (teal), and linker (gray). Right: Inactive-like conformation featuring a more compact spatial arrangement of SH3 and SH2 in proximity to the SH1 domain. Center: Close-up view of the proline-rich groove in the SH3 domain (Y72-P73-Y74) interacting with linker prolines (P231, P234) and S228. This intramolecular interaction stabilizes the inactive-like conformation and is further supported in Lyn by the presence of S228 and P73, which are not conserved in other Src family kinases.

## Mapping Allosteric Connectors to Known Kinase Inhibitor Pockets

To assess the broader relevance of the allosteric connectors identified in Lyn kinase, the spatial relationship between these dynamically central residues and known kinase inhibitor binding sites was examined. Specifically, the potential overlap between the predicted connectors and the binding regions of approved or clinical-stage kinase inhibitors was evaluated. The PKIDB classifies inhibitors based on their binding location and the conformational ensemble of the kinase they stabilize [233,234]. From the subset of inhibitors annotated for SFKs, nine compounds currently approved or under clinical investigation were identified as ATP-competitive molecules (types 1, 1.5, or 2), as summarized in Table A11. Notably, no type 3 inhibitors, defined as binding to distant ATP-independent allosteric pockets, have been reported or approved for SFKs to date, highlighting a potential opportunity to explore alternative modes of inhibition. In the absence of SFK-specific distant allosteric inhibitors, comparative structural analyses across other kinase families provide a rational framework for identifying and evaluating analogous regulatory pockets.

Structural comparisons of allosteric binding pockets across kinases have been widely employed to investigate the conservation, variability, and druggability of non-catalytic regulatory sites. Large-scale surveys have mapped allosteric inhibitor binding regions across diverse kinase structures, identifying shared spatial features and revealing family-specific differences in pocket geometry [235]. These analyses commonly rely on structural alignment and superposition of kinase domains to evaluate whether predicted allosteric sites spatially coincide with known ligand binding positions. Reviews of kinase inhibitor binding sites further illustrate how similarities and differences in pocket architecture can inform the exploration of alternative inhibitory strategies across related kinases [236]. Such observations underscore the value of comparative structural frameworks for examining distal regulatory pockets, especially when investigating whether dynamically important residues coincide with known drug-binding pockets in related kinases.

As an initial step toward assessing the druggability of the identified connectors, the binding patterns of FDA-approved type 3 kinase inhibitors were examined. Among these, cobimetinib, selumetinib, and trametinib, selective MEK1 inhibitors, were found to bind to a distinct allosteric pocket near but not intersecting with the ATP-binding site. MEK1 is a dual-specificity kinase (capable of phosphorylating both serine/threonine and tyrosine residues) that belongs to the STE kinase group. Structural alignment of the cobimetinib-bound MEK1 with the Lyn SH1 domain is shown in Figure A6 as a representative example, with details reported for both global and local comparisons.

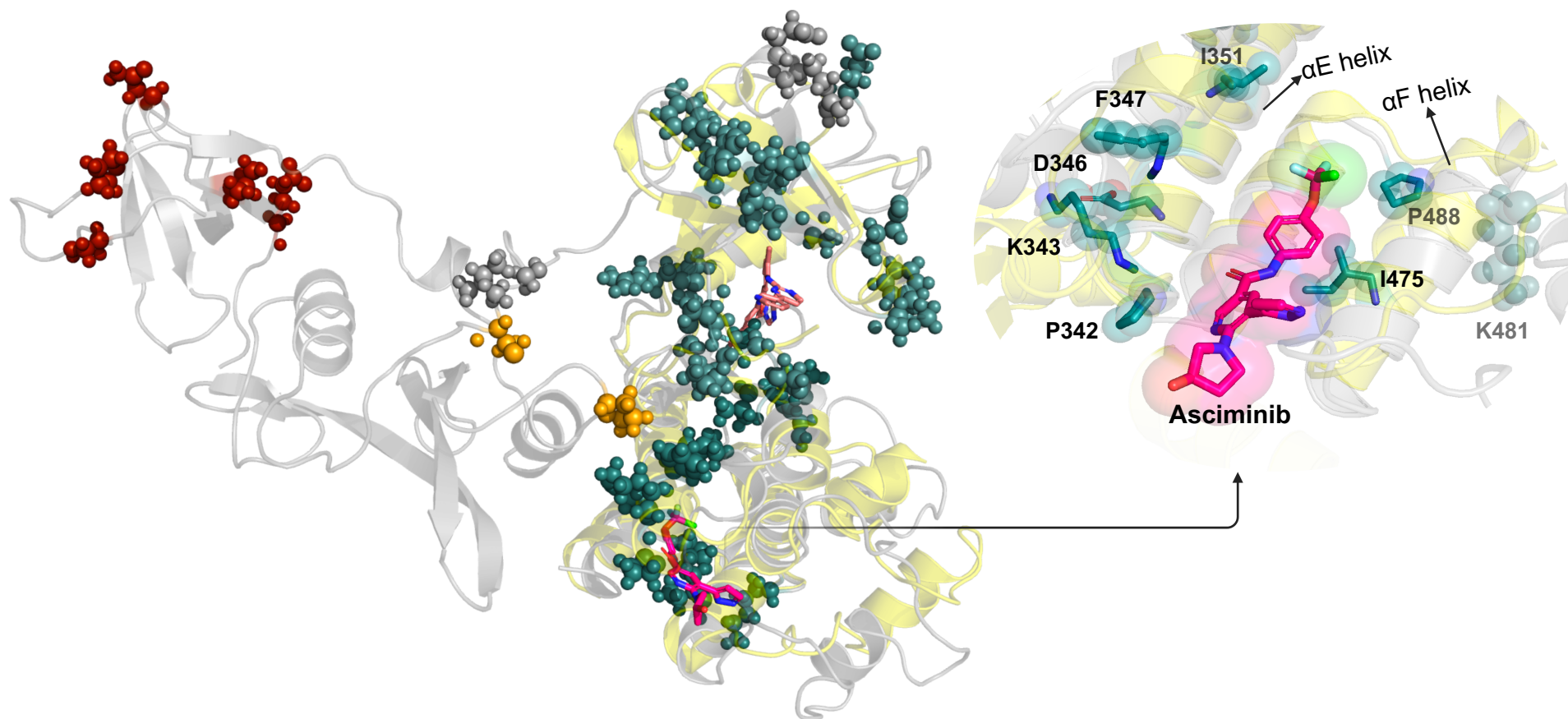
Within the vicinity of the MEK1 allosteric pocket, four predicted allosteric connectors in Lyn, M294, R304, N372, and A384, were found to be spatially co-localized with the bound inhibitor. This convergence reinforces the functional relevance of the region as a potential druggable pocket. In parallel, several residues directly interacting with cobimetinib in

MEK1 exhibited clear structural or functional equivalence in Lyn. These include F290 (equivalent to Lyn F386) within the DFG motif, V127 (Lyn V303), which lies in the ATP-binding cleft and is also a naturally occurring mutation site in Lyn (see Table A2), and K97 (Lyn K275), the conserved catalytic lysine essential for ATP coordination. Notably, binding was also observed at N195 (corresponding to Lyn N372), a conserved allosteric connector in SFKs and previously implicated in  $Mg^{2+}$  coordination in Chapter 4.

A similar spatial overlap with predicted Lyn connectors was observed for the two additional allosteric inhibitors. The ATP binding site and its surrounding residues are known to be structurally and functionally conserved across a wide range of kinases, including those belonging to distant branches of the kinome [237]. As a consequence, inhibitors targeting this region are often associated with limited selectivity and may inhibit multiple kinase families. In contrast, higher specificity can be achieved by targeting distal regulatory pockets, where allosteric inhibitors engage less conserved regions that play key roles in modulating long-range communication dynamics [238].

Beyond structurally annotated inhibitors, asciminib was identified as an FDA-approved distant allosteric inhibitor among PKIDB entries with undefined binding types. Asciminib targets the ABL1 kinase, a member of the TK family, by binding to the myristoyl pocket within the SH1C domain, which is spatially distant from the ATP-binding site. The co-crystal structure of ABL1 with asciminib (PDB: 5MO4 [82]) reveals a dual-ligand configuration, with asciminib occupying the allosteric site and a second ligand bound in the ATP pocket. This architecture exemplifies the strategy of double-drugging, in which orthosteric and allosteric inhibitors are co-administered to achieve enhanced selectivity and therapeutic efficacy [239]. Clinically, asciminib has been approved for the treatment of Philadelphia chromosome-positive chronic myeloid leukemia (Ph+ CML), including cases resistant to ATP-competitive inhibitors such as those harboring the T315I mutation [240, 241]. As a validated distant allosteric modulator, asciminib underscores the therapeutic potential of non-catalytic pockets in kinase drug development.

The SH1 domains of ABL1 kinase (PDB: 5MO4) and Lyn kinase were structurally aligned and a total of 183 residues were retained after five iterative outlier rejection cycles, yielding a global RMSD of 1.708 Å. Consistent with this structural correspondence, sequence alignment of the SH1 domains revealed strong conservation across the kinase core, including the DFG, HRD, APE motifs, as well as the catalytic lysine, reflecting the shared architecture of TK family kinases (see Figure A8). In addition, a focused alignment of the asciminib binding site resulted in an RMSD of 1.861 Å over 12 C $\alpha$  atoms. These values indicate strong structural similarity at both the global fold level and the local allosteric binding region, supporting the comparative mapping of allosteric pockets between the two proteins. As shown in Figure 5.7, a striking spatial overlap was observed between the asciminib-bound allosteric pocket in ABL1 and several of the highest scoring allosteric connectors identified in Lyn.



**Figure 5.7: Structural overlap between asciminib binding site in ABL1 and predicted allosteric connectors in Lyn kinase.** The full-length Lyn kinase model is shown as a gray cartoon, with predicted allosteric connectors color-coded by structural region: SH3 (dark red), SH2 (orange), SH1 (teal), and linker (gray). The SH1 domain of ABL1 kinase (yellow cartoon, PDB: 5MO4 [82]), co-crystallized with the allosteric inhibitor asciminib (magenta sticks), was structurally aligned to the SH1 domain of Lyn. Asciminib binds a distal site between the  $\alpha$ E and  $\alpha$ F helices, where several high-scoring Lyn connectors, including K343, D346, F347, P342, I351, and I475, are spatially co-localized. This structural convergence supports the potential for targeting conserved, noncatalytic regulatory pockets in kinases.



In ABL1, asciminib is bound at a distal site situated between the  $\alpha$ F and  $\alpha$ E helices within the SH1C region. Notably, the  $\alpha$ E helix also contributes to the SH1–SH2 interface in Lyn, which was previously identified as a key region of dynamic interdomain coupling. Through structural alignment, key residues comprising this pocket, N355, L359, L360, and A363 (from  $\alpha$ E), and V487, L490, and M491 (from  $\alpha$ F), were found to correspond to several major dynamic connectors in Lyn. These include K343, the top-scoring connector in the analysis and located at the SH1–SH2 interface, D364, another prominent interface residue, and nearby contributors such as P342, F347, I351, I475, and P488. As illustrated in Figure 5.7, this notable structural correspondence reinforces the functional relevance of the identified dynamically central connectors in Lyn kinase. The alignment reveals that the SH1–SH2 interface in Lyn, corresponding to the asciminib-binding pocket in ABL1, is not only structurally conserved but also critical for dynamic regulation. These findings underscore the broader potential of targeting conserved interdomain allosteric sites as druggable interfaces for distal inhibition, and emphasize the strength of network-based approaches in uncovering functionally critical regions beyond the catalytic site.

Overall, these insights demonstrate that Lyn kinase contains a distinct set of dynamically central allosteric connectors that show both structural divergence and spatial convergence with known distant inhibitor-binding sites. Several residues at the SH1–SH2 interface, including K343, D364, and I475, were identified as top-ranking connectors overlapping the asciminib-binding pocket in ABL1. This structural correspondence underscores the druggability of non-catalytic interdomain regions in Lyn, particularly those that are evolutionarily distinct and dynamically influential. To further assess their potential regulatory role, the following section evaluates whether these Lyn-specific connectors exhibit local flexibility and responsiveness to perturbation, biophysical properties commonly associated with allosteric control.

## Dynamic Flexibility of Lyn-Specific Allosteric Connectors

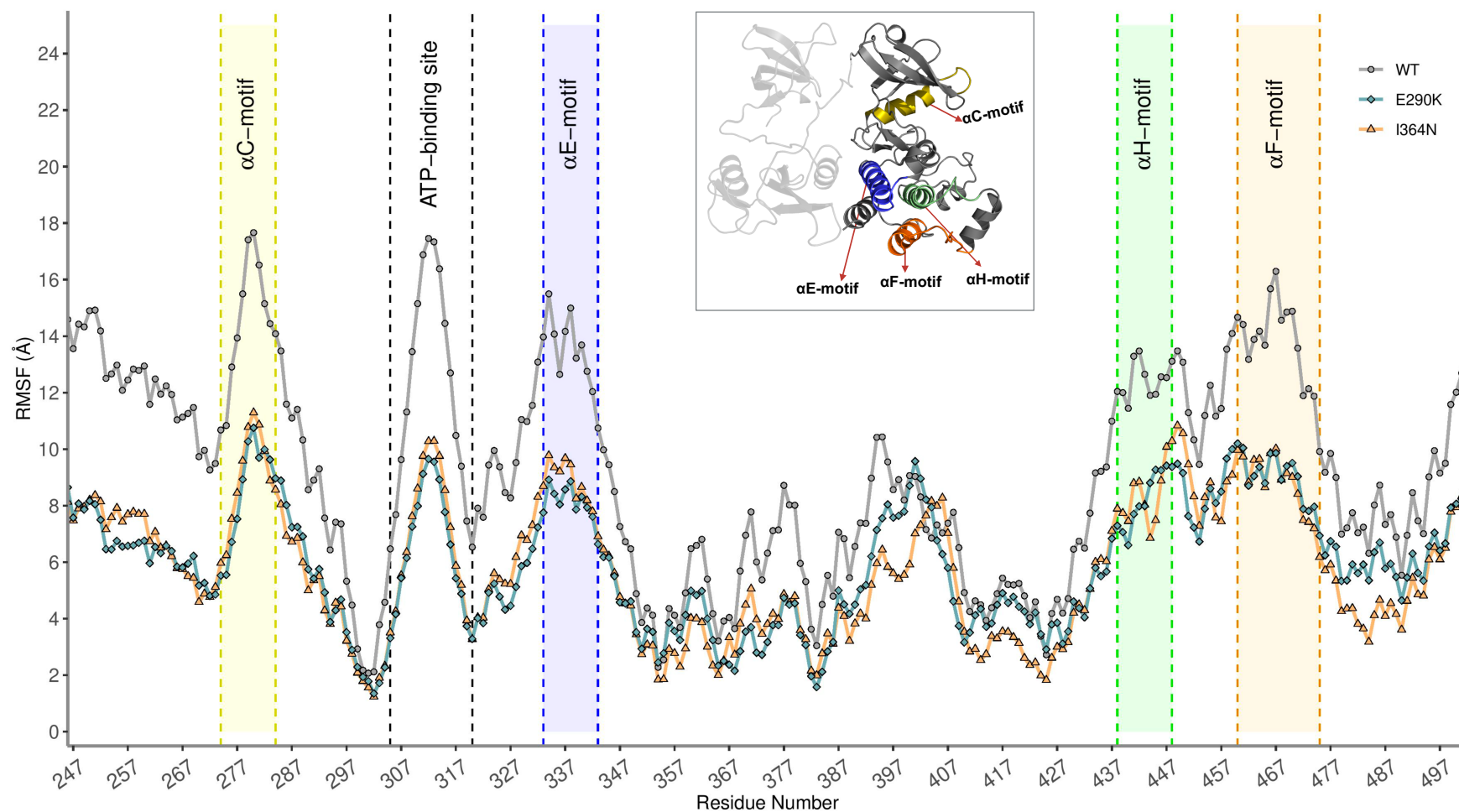
To integrate insights from network analysis, conservation profiling, and inhibitor-binding comparisons, Lyn-specific allosteric connectors were examined for local dynamic properties indicative of regulatory elements. Despite their lack of conservation across SFKs, these residues consistently ranked among the top interface-weighted connectors and overlapped structurally with established allosteric inhibitor-binding pockets. This convergence across independent analyses suggested their relevance as regulatory elements unique to Lyn. The final analysis aimed to determine whether these connectors also exhibit the dynamic plasticity characteristic of functional allosteric regions, given that flexibility is widely recognized as a hallmark of sites mediating long-range communication and ligand responsiveness [242]. In active-like conformations, such regions often display enhanced adaptability, enabling accommodation of small molecules or responsiveness to perturba-

tions [77, 243].

To evaluate this possibility, residue-wise C $\alpha$  RMSF values were calculated from MD simulations, as such analyses have been widely used to identify potential allosteric sites in protein kinases [244, 245]. RMSF profiles for ten apo and holo systems are shown in Figure A9, providing a broad comparison across Lyn mutants and ligand-bound conditions. A focused analysis of three representative systems, WT Lyn, and the E290K and I364N mutants, is presented in Figure 5.8. WT Lyn was previously characterized by an active-like conformation, while both mutants consistently exhibited reduced activation features and altered dynamics.

In the mutant systems, a pronounced reduction in flexibility was observed across the  $\alpha$ C,  $\alpha$ E, and  $\alpha$ F motifs relative to WT. Notably, this stabilization is most prominent at the RMSF maxima corresponding to the flexible loops adjacent to these helices, indicating a significant attenuation of conformational sampling in these regions. This behavior parallels the reduced flexibility seen in the ATP-binding site under inactive-like conditions, reinforcing the role of these coupled helix-loop segments in long-range regulatory communication. These findings are consistent with the established model of kinase allostery, where the structural orientation of a core helix is intrinsically coupled to the dynamics of its tethered loops, creating functional modules that respond as a single unit to mutational stress [215]. Comparable patterns have been reported in other kinases, where motifs involving the  $\alpha$ F and  $\alpha$ C helices function as dynamic centers of allosteric communication and frequently serve as binding sites for non-catalytic ligands [246–249].

The  $\alpha$ E-motif, located at the SH1C–SH2 interface, includes several top-ranked connectors identified in the analysis, such as P342 and the highest-scoring connector K343. Similarly, the  $\alpha$ F-motif contains I475, another key residue located within the region targeted by allosteric inhibitors like asciminib in ABL1. The observed reduction in flexibility across both motifs in the mutant systems suggests that their dynamic behavior, is highly sensitive to conformational perturbations, a defining feature of functionally relevant regulatory elements. Structurally, the  $\alpha$ H-motif lies between the  $\alpha$ E and  $\alpha$ F regions and appears to be influenced by dynamic coupling with these adjacent segments. Such inter-motif communication has been well documented in other kinases, where shifts in local flexibility propagate through spatially proximal elements via backbone or side-chain interactions, particularly within the densely packed catalytic core [237, 250].



**Figure 5.8: RMSF profiles of Lyn WT and inactive-like mutants.** Residue-wise  $C\alpha$  RMSF values for the SH1 domain are shown for WT, E290K, and I364N systems. Shaded regions highlight key regulatory motifs, defined as the coupled  $\alpha$ -helices and their adjacent flexible loops, which exhibit significant suppression of fluctuation peaks in mutant systems. These regions encompass the  $\alpha$ C,  $\alpha$ E,  $\alpha$ H, and  $\alpha$ F motifs, as well as the ATP-binding site. The observed reduction in flexibility across these motifs reflects a stabilization toward inactive-like conformations. Notably, the  $\alpha$ E and  $\alpha$ F motifs contain Lyn-specific allosteric connectors K343 and I475, reinforcing their potential as candidate sites for selective allosteric modulation. The inset depicts the spatial localization of these dynamically coupled motifs within the kinase domain architecture.

On the other hand, low conservation at a residue site indicates greater sequence variability across SFKs, which can be advantageous for developing selective allosteric inhibitors by reducing the likelihood of off-target interactions [77, 236]. Regions that diverge substantially across the family are more likely to present unique structural environments suitable for specific molecular recognition [238]. Among the regions evaluated, the  $\alpha$ C-motif, although dynamically active and showing significant peak suppression in the RMSF analysis, exhibits high evolutionary conservation throughout the SFK family [250]. This conservation likely stems from its essential role in maintaining the structural integrity of the ATP-binding pocket and supporting catalytic function [251]. Notably, allosteric connectors such as A286, situated within the  $\alpha$ C-motif, display some of the highest conservation scores across the family. Consequently, despite its dynamic nature and sensitivity to the mutations, the  $\alpha$ C-motif is less likely to represent a viable target for selective allosteric inhibition due to this lack of sequence divergence.

Taken together, these analyses indicate that the  $\alpha$ E and  $\alpha$ F motifs constitute Lyn-specific allosteric regions that are both dynamically adaptable and evolutionarily divergent from the highly conserved catalytic core. Their localization at interdomain interfaces, combined with the coupled flexibility of their helix-loop architecture and their lack of sequence conservation, underscores their potential as candidate sites for selective allosteric modulation. Importantly, several residues within these motifs overlap with known allosteric pockets in other kinases and display high sensitivity to activation-linked structural rearrangements in Lyn, as evidenced by the suppressed fluctuation peaks in the mutant systems. These properties collectively reinforce their relevance not only for understanding Lyn regulation but also for guiding future structure-based drug discovery efforts. Targeting such non-catalytic, conformationally responsive motifs may enable the development of selective allosteric inhibitors with reduced off-target effects across the Src family.

## 5.4 Conclusions

In this chapter, long-range communication and allosteric regulation in full-length Lyn kinase were systematically characterized through the combined use of dynamic cross-correlation analysis, network-based residue interaction mapping, evolutionary conservation assessment, and structural flexibility evaluation. This integrated framework enabled the identification of dynamically relevant regions that govern interdomain coordination and regulatory communication.

The WT active-like conformation of Lyn was found to maintain a highly organized dynamic architecture, marked by strong interdomain coupling between the SH3, SH2, and SH1 regions. Upon dasatinib binding and in several point mutants, particularly E290K and I364N, this coordination was disrupted, resulting in fragmented correlation networks

and reduced global communication. Although partial recovery of domain-level communication was observed for some mutant systems upon ATP binding, others, most notably I364N-ATP, remained dynamically decoupled, indicating a limited capacity for structural adaptation.

Through network-based analysis of dynamic correlation patterns, a total of 44 allosteric connector residues were identified. These connectors were distributed across SH3, SH2, SH1N, SH1C, and the interdomain linker, and were characterized by high betweenness centrality, frequent localization at domain interfaces, and cross-domain connectivity. The presence of these connectors in active-like systems and their loss or attenuation in inactive-like systems support a model in which long-range allosteric regulation is mediated by an embedded communication network that is reorganized across structurally distinct states. Network fragmentation in mutant and inhibitor-bound systems was consistently associated with the disruption of these connectors, reinforcing their functional relevance.

Evolutionary conservation analysis revealed a dual pattern among the identified connectors. While a subset of residues was broadly conserved across Src family members, several high-scoring connectors, including D66, P73, S228, and K343, were found to be Lyn-specific and localized at key interdomain interfaces. These divergent residues participate in regulatory interactions such as SH3-linker docking and SH1-SH2 coupling, which differ structurally from corresponding mechanisms in other SFKs. Such divergence likely contributes to Lyn-specific regulatory behavior in immune signaling and oncogenic contexts.

The structural and pharmacological significance of the identified connectors was further evaluated through comparison with known allosteric inhibitor-binding sites in other kinases. Notably, a strong spatial and sequential correspondence was observed between high-scoring Lyn connectors and the distant allosteric pocket targeted by the FDA-approved inhibitor asciminib in ABL1. This pocket, located between the  $\alpha$ E and  $\alpha$ F helices, overlaps with several Lyn connectors, including K343, F347, and I475, supporting the interpretation that this region represents a conserved and druggable regulatory site within kinase SH1C domains.

Assessment of local structural flexibility demonstrated that Lyn-specific connectors within the  $\alpha$ E and  $\alpha$ F motifs exhibit pronounced reductions in flexibility in inactive-like mutants. Importantly, this stabilization was most evident at the RMSF maxima corresponding to the loops adjacent to the helices, confirming that these regions function as dynamically coupled motifs where helix orientation modulates loop entropic freedom. This behavior mirrors that of the ATP-binding site and is consistent with a role in allosteric regulation mediated by conformationally responsive segments. In contrast, although the  $\alpha$ C-motif displayed comparable dynamic features and peak suppression, its high evolutionary conservation across kinases limits its suitability as a selective allosteric target.

Taken together, these conclusions demonstrate that long-range allosteric regulation in Lyn kinase is governed by a distributed network of 44 dynamically central residues that form a distributed regulatory architecture across domains. This network exhibits clear structural correspondence with known distant allosteric inhibitor-binding sites, reveals Lyn-specific divergence at critical interdomain interfaces, and highlights the suppression of motif flexibility as a defining feature that distinguishes regulatory allosteric regions from conserved catalytic elements.

These results provide a mechanistic basis for long-range allosteric regulation in Lyn kinase and define structurally and dynamically informed targets for selective allosteric modulation. To further evaluate which structural features most effectively discriminate between active-like and inactive-like ensembles, the next chapter applies a supervised machine learning approach to the MD trajectories. By quantifying the relative contribution of individual geometric descriptors, this analysis reveals which catalytic, interdomain, and regulatory features serve as key determinants of Lyn kinase activity and complements the network-based findings presented here.

## Chapter 6

# Structural Determinants of Lyn Kinase Conformational Ensembles

*Only through the morning gate of Beauty  
Did you pass into the land of knowledge.  
To the temple of the spirit, there leads no way  
But through the path of Beauty.  
What we here worship as Beauty,  
will one day be understood as truth.*

– Schiller (18th-century poet and philosopher)

## 6.1 Introduction

Understanding how proteins transition between distinct conformational ensembles is fundamental to elucidating their biological roles. This is especially true for multidomain regulatory proteins like kinases, where activity is modulated not by rigid structural changes alone, but through subtle rearrangements in interdomain contacts and intramolecular interactions [32, 252]. In such systems, small variations in structural features, such as the distance between two residues, the angle formed between three atoms, or the relative positioning of functional domains, can have significant consequences for activity. For instance, the formation or disruption of a salt bridge between catalytic residues can switch the kinase between an active and inactive conformation. Similarly, angular shifts in key helices or loops may allow or prevent substrate binding or allosteric communication. These geometrical parameters act as molecular readouts of the underlying conformational ensemble and provide a mechanistic basis for understanding regulation [18]. Such features can capture both local interactions, such as catalytic residue coupling, and global changes, including domain rearrangements. However, not all features are equally informative. Redundant, noisy, or non-discriminative fluctuations must be filtered out to isolate the core interactions that define activity [253].

MD simulations are particularly well suited to capture these subtleties by providing a dynamic view of protein flexibility and motion. The central challenge lies in converting the high-dimensional output into biologically meaningful insight, identifying which structural features best distinguish between conformational ensembles, and how they are affected by perturbations such as ligand binding or mutation.

Earlier chapters characterized Lyn using MD-based trajectory analysis, residue communication networks, and evolutionary comparisons. These analyses revealed how regulation is altered by perturbations and highlighted interdomain coupling that defines active-like and inactive-like ensembles. The present chapter builds on that foundation with a data-driven test, addressing Aim 3 of this project through a supervised machine learning (ML) approach to evaluate which MD-derived features best differentiate these conformational ensembles. A set of structural features was extracted from MD trajectories, guided by prior mechanistic hypotheses and known regulatory principles in SFKs [254–256]. An interpretable Random Forest (RF) classifier was then applied to classify simulation frames and rank structural features based on their contribution to the classification [257]. This approach enables a quantitative assessment of how specific geometric parameters shape the conformational landscape and identifies which features most effectively distinguish active-like from inactive-like ensembles in Lyn kinase.



## 6.2 Computational Methods

**Structural Feature Extraction.** Sixteen structural features were extracted from the MD trajectories of full-length Lyn kinase using GROMACS tools [258]. These features were grouped into four geometric categories, as summarized in Table A14: catalytic region, activation segment, intradomain architecture, and interdomain coupling. Features assigned to the catalytic region included the distances between K275–E290, D385–K275, and E290–D385, reflecting geometric relationships between the  $\alpha$ C helix and the catalytic loop. Activation segment features captured contacts involving the activation loop, specifically Y397–R366 and Y397–E289, which indicate either a stabilized or displaced conformation of this segment. Intradomain architecture was described by features internal to individual domains, comprising SH1 lobe compaction measured by the SH1N–SH1C distance. In addition, the regulatory spine (R-spine) angle was calculated based on residues F386, M294, and L305. Interdomain coupling was quantified using distances linking the regulatory domains to the kinase core, including D206–K89, D498–K137, I77–N449, and additional features measured coupling between SH3, SH2, and SH1 lobes based on large-scale domain separation.

Inter-residue and interdomain distances were calculated using `gmx distance`, while backbone and side-chain angles relevant to catalytic geometry were computed using `gmx angle`. Domain-level distances were calculated using center-of-mass positions of the SH3, SH2, and SH1 lobes. All these biologically interpretable features were extracted from approximately 78,000 simulation frames and subsequently processed and analyzed using R Statistical Software (v4.4.1; R Core Team, 2021) [122]. Complete definitions and values of all sixteen features are provided in the Supplementary Materials of the published study arising from this project [259], and the full dataset is publicly available via Zenodo [228] (RebollidoRiosLab, 2025; <https://doi.org/10.5281/zenodo.16986533>). These features were used as input to the classification model presented in the following section.

**Random Forest Classification.** A supervised machine learning workflow based on Random Forest classification was implemented using the `scikit-learn` library in Python [260]. Each simulation frame was assigned a class label according to its system of origin, with WT and WT–ATP assigned to the active-like class and WT–DAS together with all mutant systems assigned to the inactive-like class. All features were treated as continuous variables and were scaled to a common range using Min–Max normalization. After pre-processing, the dataset was randomly shuffled and divided into a training set containing 80% of the data and a testing set containing the remaining 20%. Stratified sampling was applied to preserve the class proportions in both subsets.

Class imbalance was addressed using SMOTE [144], which was applied within the train-

ing procedure to generate synthetic samples of the minority class. SMOTE was integrated with the Random Forest classifier as part of a single modeling pipeline. Model parameters were optimized through a grid search combined with five-fold cross-validation, which ensured that the selected settings generalized well across the data. The final optimized parameters used for training were `n_estimators = 1000`, `max_depth = None`, `class_weight = None`, and `smote_sampling_strategy = 0.5`. Model performance was evaluated on the held-out test set using several standard classification metrics, including accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC). Feature importance values were computed using two complementary approaches, namely the Gini index and permutation importance. The resulting importance scores were reported using the original feature definitions.

## 6.3 Results and Discussion

### Key Structural Features Distinguishing Lyn Conformational Ensembles

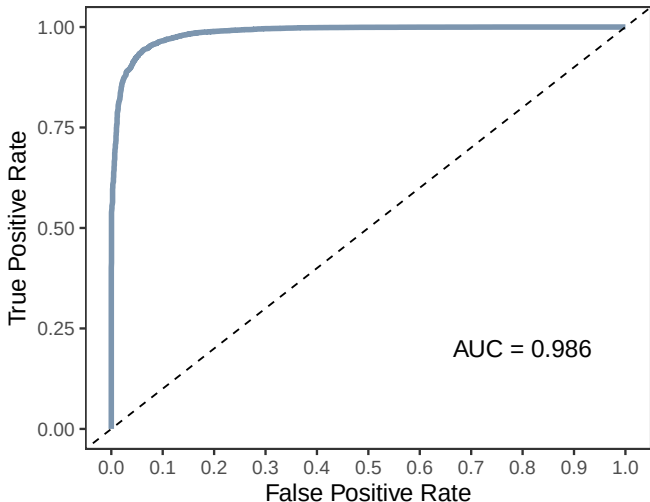
An RF classifier was trained using sixteen interpretable features extracted from each simulation frame. Although the objective of this analysis was not to develop a predictive model for future classification tasks, the use of an RF classifier enabled a systematic evaluation of which features were most informative in distinguishing active-like from inactive-like conformations. The model was trained on a subset of the data and evaluated on an independent test set. An overall classification accuracy of 96% was achieved, indicating that the selected features effectively captured the conformational differences between the two ensembles across all systems and simulation time frames.

In addition to accuracy, the model also exhibited high precision, recall, and F1-score for both classes, as summarized in Table 6.1. These metrics provide complementary insights into the quality of classification. Precision quantifies the proportion of frames assigned to a given class that were truly members of that class, reflecting the reliability of positive predictions produced by the model. Recall, on the other hand, measures the proportion of true frames in that class that the model successfully retrieved, measuring its sensitivity. The F1-score represents the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives. Importantly, the classifier achieved an MCC of 0.85. Given that MCC is robust to class imbalance and penalizes discordant predictions, this value indicates a reliable and well-generalized separation of conformational ensembles, further supporting that the model is not overfitting. The consistently high values for all four metrics across both classes suggest that the classifier performs reliably, without favoring one conformational ensemble over the other, and is capable of distinguishing subtle structural differences with high confidence.

Table 6.1: Performance metrics for the Random Forest classifier applied to MD-derived features of full-length Lyn kinase. Per-class scores for active-like and inactive-like frames are shown, along with macro and weighted averages across both classes.

| Label            | Precision | Recall | F1-Score |
|------------------|-----------|--------|----------|
| Active-like      | 0.91      | 0.84   | 0.87     |
| Inactive-like    | 0.97      | 0.98   | 0.98     |
| Accuracy         |           | 0.96   |          |
| MCC              |           | 0.85   |          |
| Macro average    | 0.94      | 0.91   | 0.92     |
| Weighted average | 0.96      | 0.96   | 0.96     |

To further evaluate the ability of the model to distinguish between conformational ensembles, the ROC curve was examined, as shown in Figure 6.1. The ROC curve provides a comprehensive picture of classifier performance by plotting the trade-off between the true positive rate (sensitivity) and the false positive rate across a continuum of classification thresholds. A model that performs no better than random chance is expected to yield a diagonal curve, with an AUC of 0.5. In contrast, a well-performing classifier is characterized by a curve that arcs toward the upper left corner, indicating both high sensitivity and specificity.



**Figure 6.1: ROC curve.** Receiver operating characteristic curve for the Random Forest classifier trained on MD-derived features of full-length Lyn kinase. The model distinguishes the ensembles with an area under the curve of 0.986.

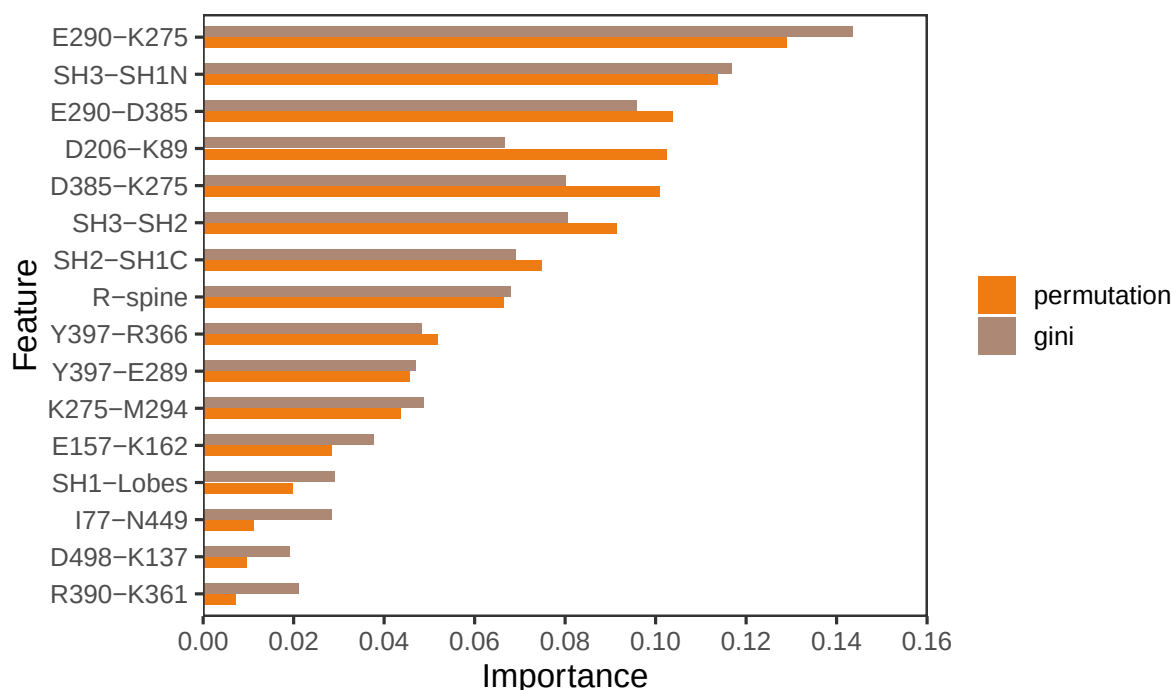
In this analysis, an AUC of 0.986 was achieved, reflecting excellent discriminative power. This near-perfect score indicates that active-like and inactive-like conformations were reliably and consistently separated by the RF classifier across the full distribution of simulation frames. Importantly, the relevance and quality of the selected structural features were underscored by this result, demonstrating that biologically meaningful distinctions between the two ensembles were effectively encoded. The high AUC not only supports the robustness of the model, but also reinforces the hypothesis that conformational coupling and

specific residue-level interactions serve as strong determinants of kinase activity.

Beyond overall classification accuracy, an important advantage of the RF framework is its ability to quantify the relative contribution of each input feature to classification. In this study, feature importance was evaluated using two complementary approaches: the Gini importance (also known as mean decrease in impurity) and the permutation importance.

The Gini index measures how much each feature decreases the impurity (or uncertainty) within decision trees. Features that consistently result in more homogeneous node splits across the forest are assigned higher importance. In other words, a feature that helps partition the data into more distinct active-like and inactive-like subsets will receive a higher Gini score. The permutation importance approach, on the other hand, provides a model-agnostic measure by evaluating how much the accuracy of the model drops when a single feature is randomly shuffled. If shuffling a feature significantly degrades performance, it indicates that the model relied strongly on that feature for classification. This method offers a more direct estimate of how much each feature contributes to predictive performance.

As shown in Figure 6.2, both feature importance approaches converged on a consistent ranking, identifying a small set of structural descriptors as key discriminators between active-like and inactive-like conformations.

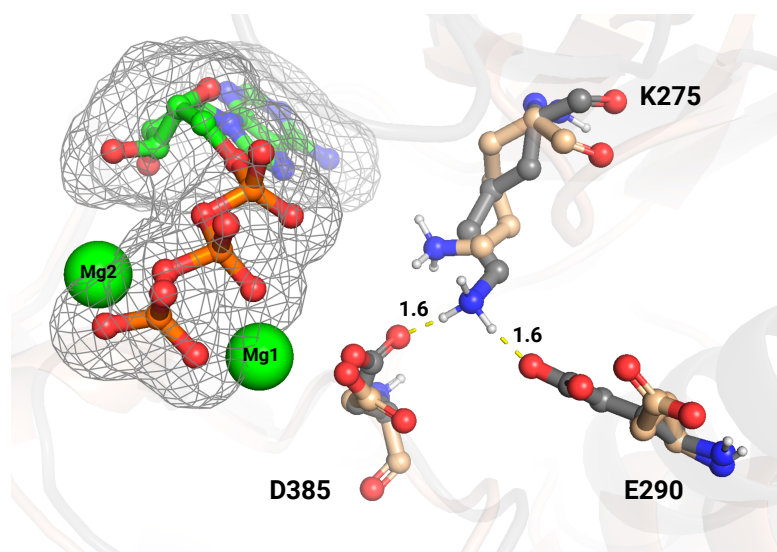


**Figure 6.2: Feature importance based on Gini index and permutation score.** Key predictive features include K275–E290, E290–D385, SH3–SH1N separation, and D385–K275, all of which participate in catalytic regulation or interdomain interactions.

Among the highest-ranked features were three distances from the catalytic region, K275–E290,

E290–D385, and D385–K275, which together form a local geometric motif that emerged as a key signature of activation in Lyn. These distances help stabilize the catalytic core and regulate the accessibility and positioning of the  $\alpha$ C helix relative to the active site. Contacts involving D385 (from DFG motif) and both K275 (catalytic lysine) and E290 (from  $\alpha$ C helix) preserve the structural arrangement of the catalytic loop and facilitate productive phosphoryl transfer. The catalytic (DFG-in/ $\alpha$ C-in) conformation has been observed in structural studies of Lyn, where the inward orientation of the  $\alpha$ C helix and the presence of the K275–E290 salt bridge characterize the active conformation [58].

The structural behavior of these catalytic features is shown in Figure 6.3, comparing the WT–ATP and I364N–ATP systems. In the WT–ATP system, all three distances remain compact, reflecting an active-like conformation in which the inward positioning of the  $\alpha$ C helix is stabilized. The K275–E290 salt bridge, a hallmark of kinase activation, anchors the  $\alpha$ C helix and aligns the catalytic lysine for ATP coordination. These observations are consistent with active site features previously discussed in Chapter 4, specifically the accessible and adaptable positioning of the  $\alpha$ C helix, where the catalytic core remains open and permissive to productive interactions. In this conformation, electrostatic coupling between K275, E290, and D385 supports a coherent active site architecture.



**Figure 6.3: Top-ranked geometric features of the Lyn catalytic core.** The three catalytic features, K275–E290, D385–K275, and E290–D385, were identified by Random Forest classification as the most discriminative features in the catalytic region for distinguishing active-like from inactive-like ensembles in Lyn. The triangle formed by these residues is shown for comparison in WT–ATP (gray) and I364N–ATP (coral). The K275–E290 and D385–K275 distances are highlighted in WT–ATP and increase to 7.8 Å and 5.4 Å, respectively, in I364N–ATP, reflecting disruption of the catalytic core.

In contrast, the I364N–ATP mutant exhibits a clear expansion of all three distances, indicating a breakdown of this coupling network. The opening of the catalytic triangle corresponds to reduced accessibility of the  $\alpha$ C helix and a loss of structural integration between the helix and the catalytic core. This behavior aligns with previous observations in Chapter 4, where mutants displayed a more constrained active site environment and diminished conformational adaptability. The prominence of these three distances in the RF ranking, therefore, reflects their role as sensitive reporters of  $\alpha$ C he-

lix engagement and catalytic core accessibility, key features that define the separation between active-like and inactive-like Lyn conformations.

In addition to the catalytic triangle, interdomain interactions were also among the most influential features. The SH3–SH1N and SH3–SH2 distances capture the relative positioning of the SH3 domain with respect to the N-lobe of the kinase domain and the SH2 domain, reflecting long-range structural organization. Larger distances are characteristic of active-like conformations, whereas inactive-like ensembles exhibit a more compact arrangement. This behavior is fully consistent with findings from earlier chapters, where SH3 motion was shown to play a decisive role in reorganizing interdomain interactions and triggering global conformational shifts. Given the established regulatory role of SH3 displacement in Lyn, its strong contribution to classification confirms its mechanistic importance and highlights the coherence between previous structural analysis and ML-based feature selection.

Another highly ranked interdomain feature in the classification analysis was the SH2–SH1C distance. This measurement captures the spatial relationship between the SH2 domain and the C-lobe of the kinase domain, a region previously identified as a hotspot for dynamic allosteric communication in Lyn. In Chapter 5, network-based analysis revealed that several of the highest-scoring allosteric connectors were located at the SH2–SH1C interface, including K343 and D346. These residues form a dense interdomain cluster that bridges the regulatory SH2 domain and the kinase C-lobe, mediating long-range coupling. The strong contribution of the SH2–SH1C distance is consistent with this observation, indicating that dynamic variations at this interface reflect shifts in the allosteric communication network and help define distinct conformational ensembles. The convergence between ML-based feature ranking and network-based connector identification provides robust support for the functional relevance of the SH2–SH1C interface in Lyn. This coupling appears to play a central role in modulating kinase activation by coordinating structural integration and long-range communication across domains.

Among intradomain descriptors, the RF model ranked the feature corresponding to the R-spine within the SH1 domain among the most informative. The R-spine represents a set of hydrophobic residues that align across the N- and C-lobes of the kinase domain and is essential for maintaining a catalytically competent architecture. In Lyn, stabilization of this intradomain spine reflects tight coupling between the N-lobe, the activation loop, and the catalytic core. This finding is consistent with the SH1-focused analyses presented in Chapter 4. RMSD results showed that ATP binding markedly stabilizes the SH1 domain, while disruptive mutations such as E290K and I364N lead to reduced flexibility and a more compact, rigid catalytic core. Complementary RMSF profiles further revealed that these mutations dampen local fluctuations in regions adjacent to the R-spine, indicating restricted intradomain adaptability. Moreover, PCA analysis demonstrated that active-like Lyn conformations are characterized by coordinated SH1 motions that are diminished

or lost in inactive-like mutants. Together, these findings indicate that RF correctly identified the R-spine feature as a sensitive reporter of intradomain organization and catalytic readiness in Lyn. Rather than acting independently, R-spine integrity in Lyn appears tightly linked to both local catalytic geometry and long-range interdomain communication, reinforcing its central role in defining the active-like conformational ensemble.

Together, the top seven features shown in Figure 6.2 contributed approximately 66.7% of the total permutation importance, highlighting that only a fraction of the input space was required to effectively differentiate active-like from inactive-like ensembles. Overall, this supervised learning analysis complements and reinforces insights from structural, dynamical, and network-based investigations by providing quantitative validation of the critical factors underlying Lyn activation. In particular, the model confirms that catalytic site coherence and interdomain alignment across SH3, SH2, and SH1 represent the primary structural determinants underlying conformational regulation in Lyn kinase.

## Lyn Regulatory Features in a Broader Kinase Context

Understanding whether the structural features that define the conformational ensembles of Lyn kinase are broadly conserved or represent Lyn-specific adaptations is essential for evaluating the generalizability of activation mechanisms across TKs. Comparative analysis with other SFKs and structurally related proteins provides a valuable framework for identifying common regulatory strategies and unique subtype-specific features. In particular, the top-ranked classification features identified in this study, centered on catalytic geometry, interdomain coupling, and regulatory spine integrity, reflect conserved elements of kinase architecture, while also revealing regulatory divergence across kinase families.

Within SFKs, many of the key structural motifs identified in Lyn are also conserved. The catalytic triangle formed by the conserved lysine, the  $\alpha$ C glutamate, and the DFG motif aspartate is a well-established hallmark of the active conformation in c-Src, Hck, and Fyn. Disruption of this geometry, whether by mutation or inhibitor binding, consistently results in impaired kinase activity and conformational inactivation [198]. Similarly, R-spine stabilization is required for full activation in both c-Src and Hck, where proper alignment of hydrophobic residues spanning the N- and C-lobes supports catalytic core integrity and  $\alpha$ C helix positioning [191, 261]. However, most computational studies on SFKs to date have been limited to the isolated SH1 domain [262, 263], which restricts insight into long-range regulation. Only a limited number of studies have incorporated full-length constructs, including MD simulations of full-length Src and Lck that reveal domain-specific allosteric communication networks and regulatory couplings otherwise absent in truncated models [193, 264, 265]. In comparison, the Lyn model captures dynamic interdomain couplings in more detail, particularly through SH3–SH1 and SH2–SH1 interfaces.

Interestingly, Lyn differs from other SFKs in the relative importance of specific interdomain contacts. While SH2–SH3 coupling is the dominant regulatory axis in c-Src and Hck, Lyn exhibits a stronger reliance on SH3–SH1N and SH2–SH1C contacts to stabilize its active conformation [146, 266, 267]. This divergence supports a subtype-specific model of SFK activation, where different family members deploy alternative interdomain strategies to achieve functional switching [268].

Beyond the Src family, similar structural determinants have been observed across other tyrosine kinases. In Abl kinase, the catalytic K–E–D triangle is tightly associated with ATP binding, and its disruption, either by mutation or through imatinib binding, leads to a collapse of the catalytic geometry and misalignment of the R-spine [269]. Consistently, supervised learning models such as SVM and RF trained on MD simulations have identified catalytic loop distances,  $\alpha$ C helix positioning, and N–C lobe alignment as the most informative structural features for distinguishing activation ensembles in Abl [270, 271]. EGFR exhibits similar mechanistic dependence on R-spine integrity and DFG motif orientation. Proper alignment enables ligand-induced dimerization, while displacement of these elements leads to inactive or drug-resistant conformations [272]. RF-based studies in EGFR have highlighted the K–E salt bridge, DFG motif flexibility, and activation loop dynamics as key discriminants between active and inactive ensembles [273]. Related features have also been identified in the insulin receptor kinase, where  $\alpha$ C positioning and DFG motif orientation are central to activation and substrate recognition [274].

Comparable analyses in serine/threonine kinases such as BRAF and CDKs have revealed similar structural determinants of activation. In these systems, R-spine integrity and  $\alpha$ C– $\beta$ 4 loop alignment have emerged as central regulators of conformational transitions. These elements not only influence catalytic geometry but also serve as reliable markers in data-driven models. Supervised learning approaches trained on dynamic structural features consistently identify these motifs as top-ranked discriminants between active and inactive conformations [275, 276].

Taken together, these studies support the broader relevance of the findings in Lyn. While the precise residue identities and domain interactions vary across kinase families, the recurring importance of catalytic geometry and interdomain organization points to a conserved set of structural principles underlying kinase regulation. The RF-based analysis presented here not only highlights these features in Lyn, but also aligns them with global regulatory strategies observed across the kinase landscape, reinforcing the utility of structural classification frameworks in understanding kinase activity and regulation.



## 6.4 Conclusions

This chapter demonstrated that structural features derived from MD simulations can robustly distinguish between active-like and inactive-like conformational ensembles of full-length Lyn kinase. Top-ranked features, such as the K275–E290–D385 catalytic triangle, SH3–SH1N and SH2–SH1C interdomain distances, and regulatory spine geometry, closely mirrored regions previously implicated in allosteric communication, network centrality, and ligand-responsive flexibility. Their consistent emergence across orthogonal analyses reinforces their mechanistic relevance and highlights a coherent regulatory architecture.

Notably, the alignment between ML-based feature rankings and prior DCCM and network analyses supports a model in which Lyn activation is governed by both conserved catalytic elements and subtype-specific interdomain couplings. While the catalytic features align with established activation signatures in tyrosine kinases, the prominence of Lyn-specific interdomain distances suggests a distinct mode of long-range structural coordination mediated by the SH3 and SH2 domains.

From a therapeutic perspective, the results provide a structural rationale for exploring non-catalytic regulatory sites as allosteric targets. Key classification features overlap with previously identified allosteric connectors, particularly at SH2–SH1C interfaces, which are structurally adaptable and dynamically responsive. These regions may offer promising entry points for selective modulation of Lyn function, especially in disease contexts where ATP-site inhibitors show limited efficacy or resistance. Although this study does not directly assess binding pocket properties, it highlights structural determinants that are both mechanistically meaningful and potentially druggable.

It should be noted that the classification relied on a defined set of geometric features and binary ensemble labeling, which may not capture the full complexity of intermediate ensembles or context-specific modulation. Nonetheless, the high agreement between biologically grounded features and classification performance suggests that these descriptors capture the dominant axes of functional variation in Lyn.

Taken together, this chapter integrates data-driven classification with prior structural and network-based analyses to reinforce the key determinants of Lyn kinase regulation. By identifying conserved catalytic features and subtype-specific interdomain couplings, the results establish a structural basis for understanding how dynamic architecture shapes functional behavior in multidomain kinases. These findings lay the groundwork for future directions, which are explored in the next chapter through both computational and experimental perspectives.

## Chapter 7

# Future Directions in Kinase Regulation

*A man sees in the world what he carries in his heart.  
To understand the universe, one must first listen  
to the quiet truth that stirs within.  
Only the inward eye, made bright by wonder,  
can perceive what lies beyond the veil of appearances.*

– Goethe (18th-century poet and philosopher)

The integrative analysis presented in this thesis provides a mechanistic framework for understanding long-range regulation in full-length Lyn kinase. By combining long-timescale MD simulations, network theory analysis, and ML-based feature selection, this work connects dynamic behavior with regulatory architecture in a way that is both predictive and structurally grounded. These results open several clear directions for future investigation, both in deepening mechanistic understanding and advancing therapeutic applications.

Enhancing biological fidelity in MD simulations remains essential for capturing the full spectrum of Lyn regulatory dynamics. Current analyses rely on classical force fields, limited sampling timescales, and binary labeling of conformational ensembles, all of which constrain the representation of intermediate or partially active conformations. Enhanced sampling techniques, continuous activation metrics, and inclusion of biologically relevant contexts, such as tyrosine phosphorylation, membrane anchoring, or scaffold protein interactions, would provide a more comprehensive view of conformational transitions in Lyn.

A critical distinction moving forward lies in separating the identification of dynamic communication networks from the validation of structurally defined allosteric binding pockets. The current work maps communication networks through connector residues, many of which localize to protein interfaces. However, their druggability and functional accessibility remain unresolved. Several allosteric connectors identified here therefore provide concrete candidates for experimental validation, which could be pursued using targeted mutagenesis and conformation-sensitive approaches such as hydrogen–deuterium exchange, FRET-based biosensors, or covalent fragment screening to assess their potential for selective allosteric modulation.

Structurally, the convergence of Lyn dynamic connectors with known allosteric pockets in other kinases highlights an opportunity for rational design of distant inhibitors in SFKs. These regions, particularly the regulatory motifs characterized by coupled dynamics, low sequence conservation, and consistent influence on Lyn conformational ensembles, represent promising yet underexplored targets. A focused search for scaffolds that engage these dynamically active, non-catalytic sites could yield kinase-selective chemical probes with therapeutic relevance in immune signaling and hematologic malignancies.

Beyond Lyn, the methodological framework developed here can be applied to other kinases with unresolved regulatory features. Future studies could integrate evolutionary models, machine-learned embeddings, or ligand-bound experimental structures to expand predictive accuracy and generalizability. This is particularly relevant for SFKs, where subtle subtype differences in interdomain architecture likely shape isoform-specific regulation and drug response.

In summary, this work lays a foundation for mechanistic modeling of multidomain kinase regulation, with implications for both basic science and therapeutic design. By bridg-

ing conformational dynamics, network communication, and data-driven classification, it enables a shift from static structural models to dynamic regulatory landscapes that are accessible to both computation and experiment.

# Appendices

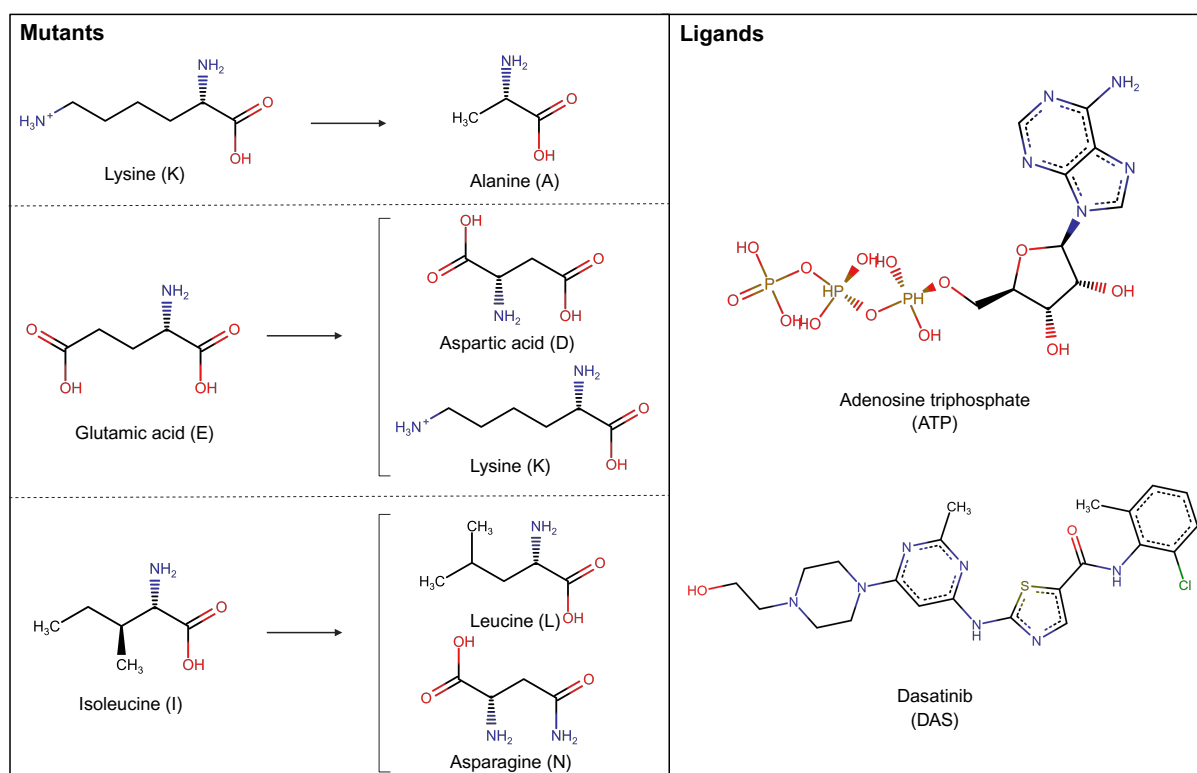
*When the path disappears, the seeker does not stop  
but becomes the path.*

*The destination was never a place  
but a transformation of the self.*

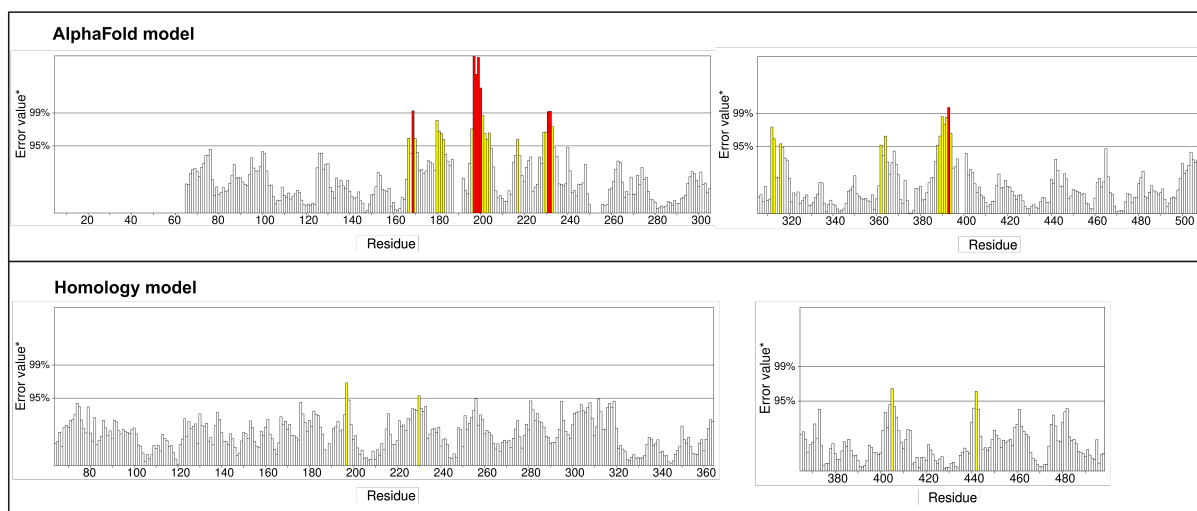
*What ends outside continues within, quietly and endlessly.*

– Attar of Nishapur (12th-century mystic and poet)

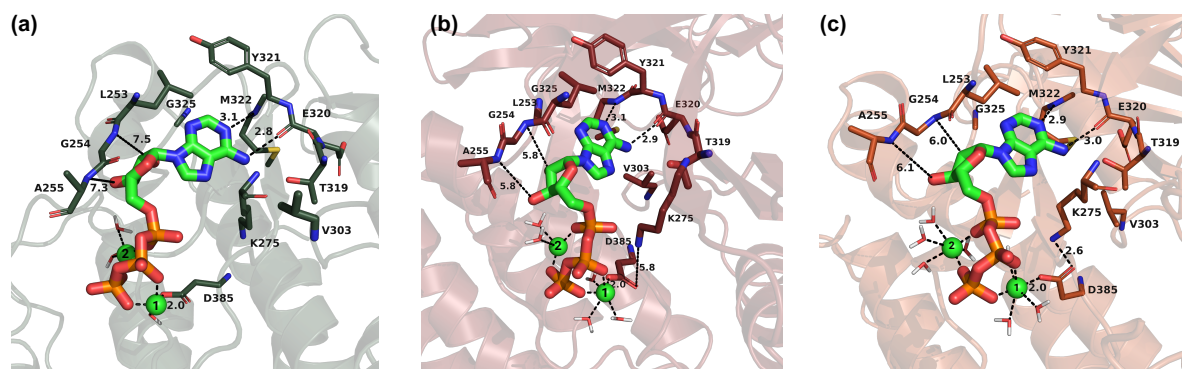
## Appendix A: Figures



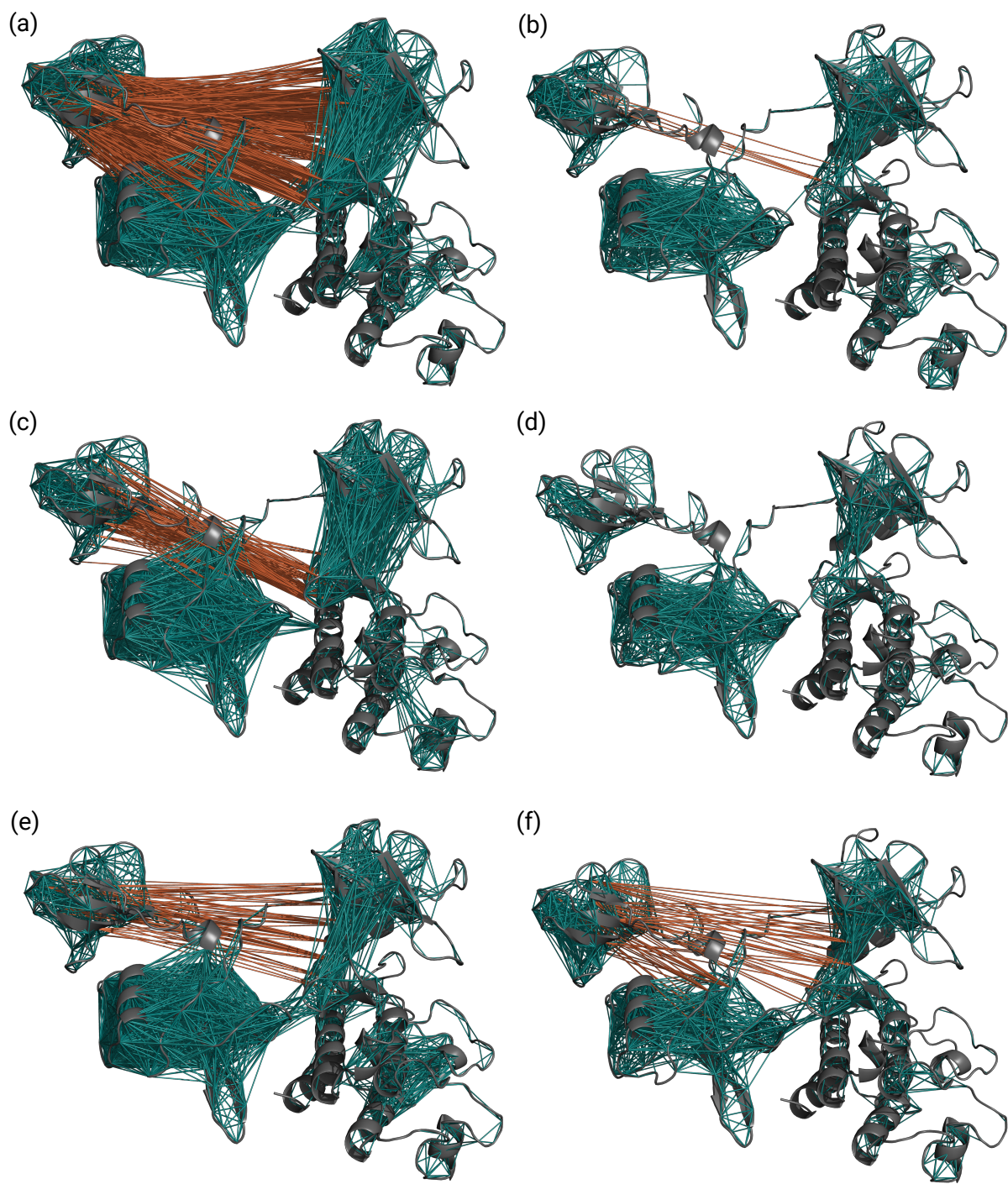
**Figure A1: Molecular structures of the single-point mutations and small molecules used in this study.** **Left panel:** Single-point substitutions were introduced at three key residues in Lyn kinase. At position K275, lysine was replaced by alanine (K275A), removing the positively charged side chain. At position E290, glutamic acid was substituted by either aspartic acid (E290D), which shortens the side chain while preserving negative charge, or by lysine (E290K), resulting in a charge reversal. At position I364, isoleucine was substituted by either leucine (I364L), a conservative hydrophobic replacement, or by asparagine (I364N), introducing a polar side chain. These substitutions were selected based on observed mutation frequencies and relevance to kinase regulation. **Right panel:** Ligand molecules used in the MD simulations include adenosine triphosphate (ATP), the physiological nucleotide substrate of kinases, and dasatinib (DAS), a clinically approved tyrosine kinase inhibitor. Both molecules were modeled in the catalytic cleft to evaluate their impact on structural dynamics and inter-residue communication.



**Figure A2: ERRAT2 structural quality assessment of AlphaFold- and homology-modeled full-length Lyn kinase structures.** The homology model achieved a higher overall quality factor (99.1) with only minor deviations across the sequence, indicating high structural integrity suitable for MD simulations. In contrast, the AlphaFold model yielded a lower quality factor (92.1) and exhibited several high-error regions, particularly within the SH2 domain and linker regions. ERRAT2 error values are shown as vertical bars along the sequence: gray bars indicate residues with acceptable quality, yellow bars represent moderate deviations, and red bars correspond to regions with high error values suggestive of potential structural inaccuracies. Both models are displayed in two panels due to the default plotting behavior of ERRAT2, which splits sequences longer than approximately 300 residues. The underlying structural models remain continuous and contain no missing regions.

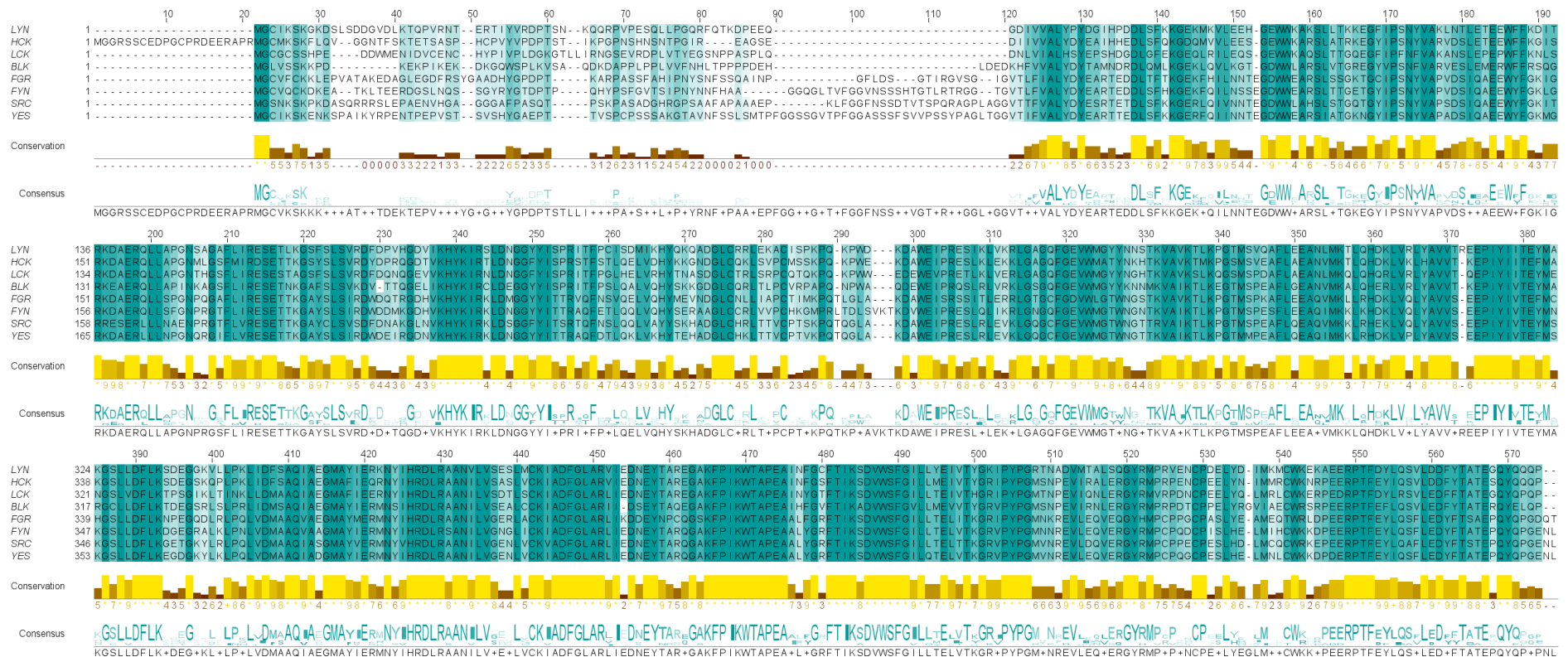


**Figure A3: Structural representations of ligand binding in the ATP-binding site of Lyn kinase mutants.** (a) E290D-ATP. The conservative substitution of glutamate by aspartate at position 290 causes a slight repositioning of catalytic residues. The interaction between K275 and D385 is retained, although the coordination of ATP appears marginally destabilized. (b) I364L-ATP. Replacement of isoleucine by leucine at position 364 leads to minimal changes in ATP binding. The overall geometry and catalytic coordination remain similar to the WT system. (c) I364N-ATP. Substitution by asparagine alters the local polarity and disrupts hydrophobic contacts. ATP adopts a more extended conformation, and its coordination with surrounding residues is weakened. Atom colors are as follows: carbon atoms in the ligand are shown in green, nitrogen in blue, oxygen in red, and  $\text{Mg}^{2+}$  ions as green spheres. Dashed lines represent coordination distances with catalytic residues and metal ions.

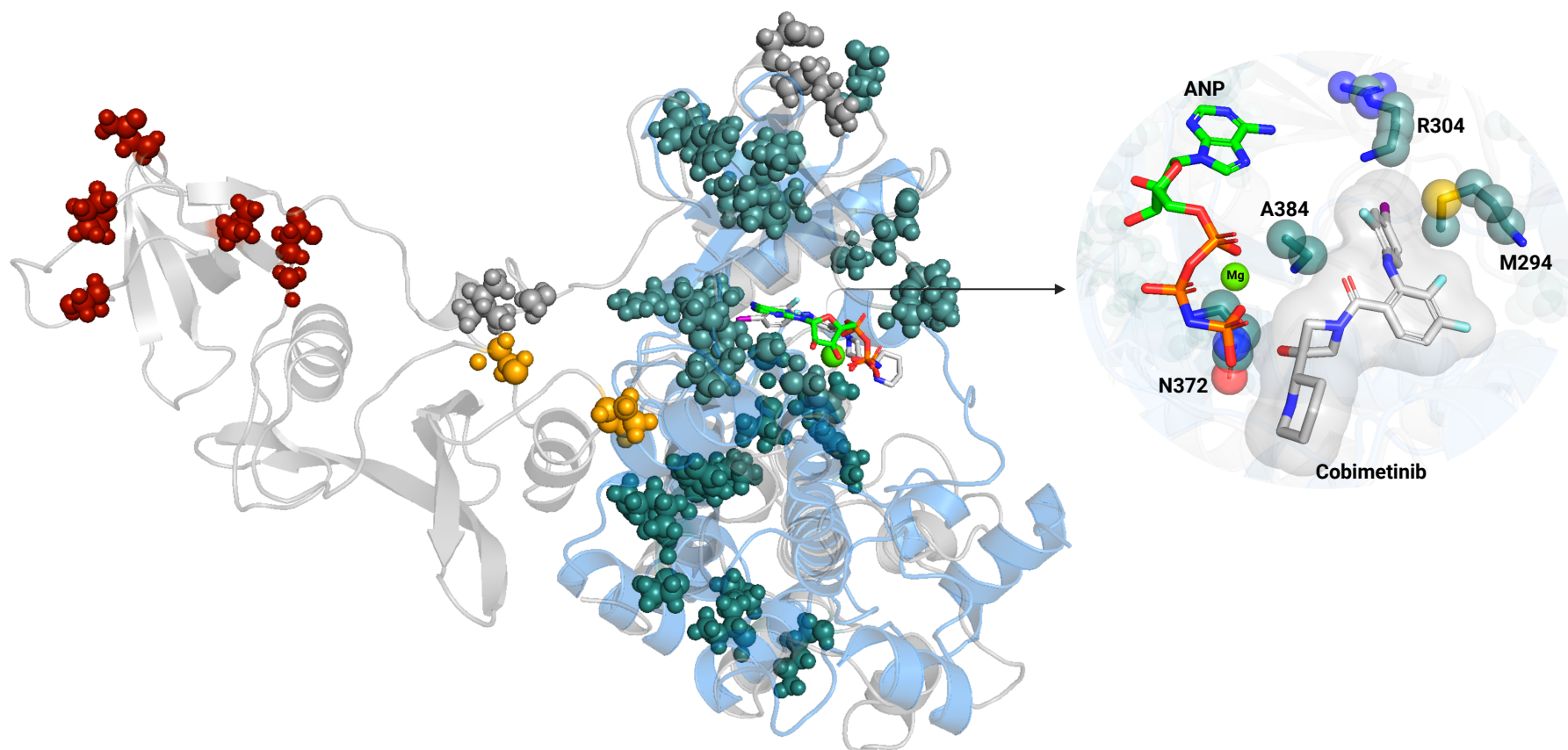


**Figure A4: Three-dimensional representation of residue–residue dynamic correlations in apo Lyn systems.** Spatial projections of DCCM-derived correlations are shown for the apo forms of Lyn: (a) WT, (b) K275A, (c) E290D, (d) E290K, (e) I364L, and (f) I364N. Colored lines denote strong correlations between residue pairs, with teal lines indicating strong positive correlations ( $\geq +0.9$ ) and burnt orange lines indicating strong negative correlations ( $\leq -0.9$ ). These structural correlation networks illustrate mutation-induced alterations in long-range coupling in the absence of ligand binding.

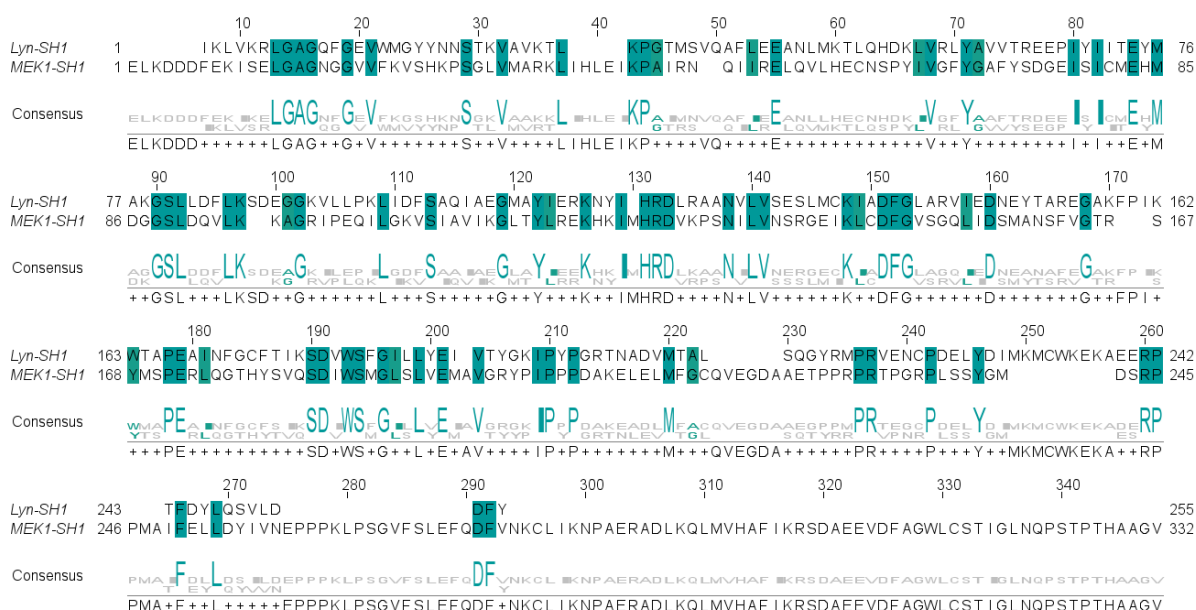




**Figure A5: Full-length multiple sequence alignment of eight Src family kinase members visualized in Jalview.** Residue conservation across the aligned sequences is represented. The Conservation bar (scaled 0–10) reflects amino acid similarity calculated using the BLOSUM62 substitution matrix, which incorporates both sequence identity and biochemically tolerated substitutions. The Consensus line depicts the most frequent residues at each position using sequence logos, indicating conserved motifs. This visualization complements the computational conservation analysis by providing a graphical overview of sequence-level conservation patterns in the Src family.



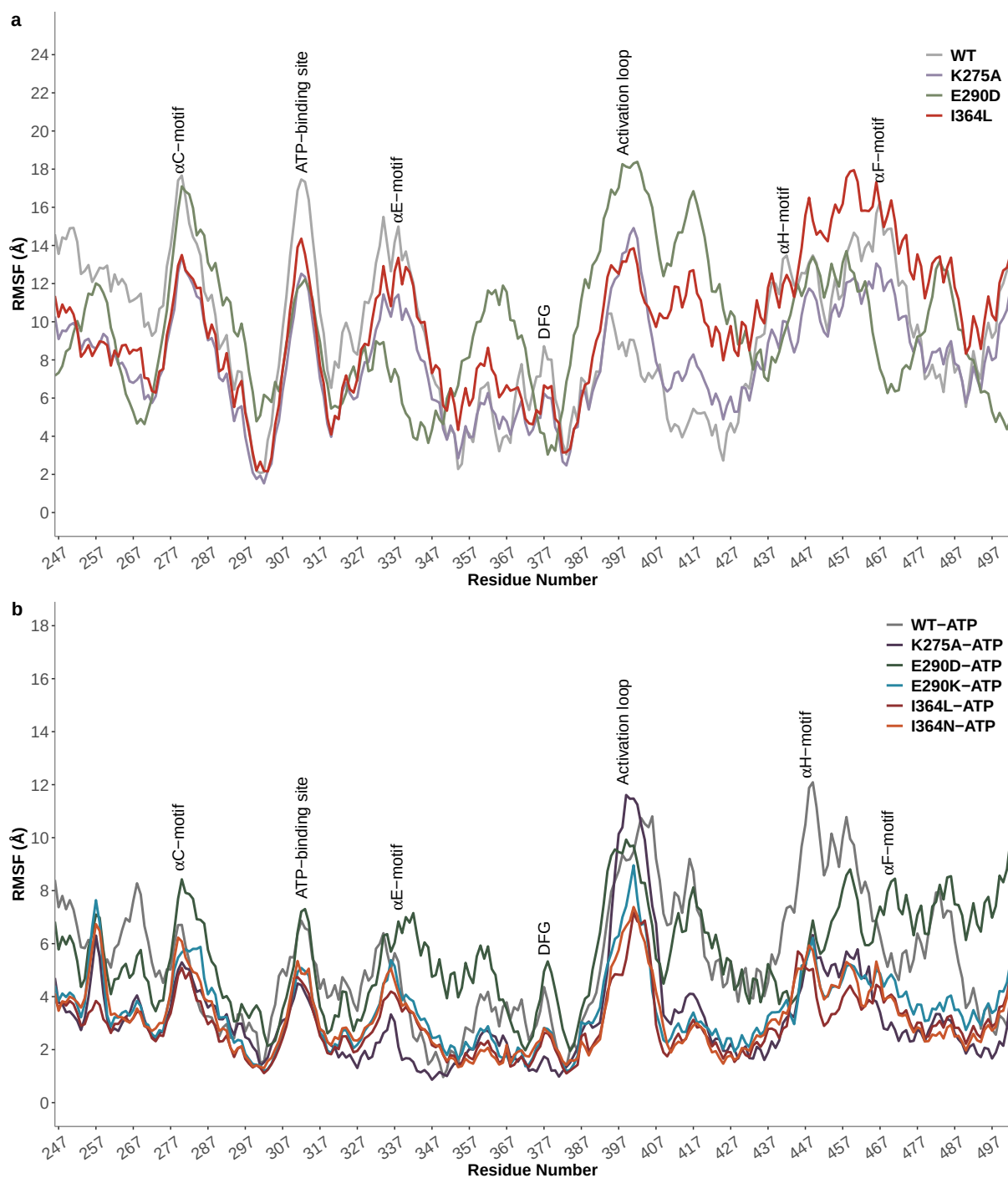
**Figure A6: Structural alignment of MEK1 and Lyn highlighting the cobimetinib-binding pocket.** The full-length Lyn kinase model is shown as a gray cartoon, with predicted allosteric hubs colored by domain: SH3 (dark red), SH2 (orange), SH1 (teal), and linker (gray). The MEK1 kinase domain (PDB: 4LMN), co-crystallized with the type 3 inhibitor cobimetinib (shown as white sticks) and the ATP analog ANP (green), is overlaid in light blue. A domain-level alignment of the SH1 regions yielded an RMSD of 2.135 Å, indicating global similarity in the kinase fold. A focused structural alignment within 5 Å of cobimetinib resulted in an RMSD of 1.232 Å over spatially matched residues, supporting local correspondence of the allosteric pocket between MEK1 and Lyn. The inset highlights the region surrounding cobimetinib, including four aligned Lyn connectors (M294, R304, N372, A384).



**Figure A7: Multiple sequence alignment of the SH1 domains from Lyn and MEK1 kinases visualized in Jalview.** Conserved and similar residues are highlighted according to physicochemical properties. Although Lyn and MEK1 belong to different kinase groups, TK and STE kinases, respectively, several residues within the ATP-binding site are conserved. Additionally, key structural motifs such as the DFG and HRD motifs are retained, supporting the observed structural alignment between the two SH1 domains.



**Figure A8: Multiple sequence alignment of the SH1 domains from Lyn and ABL1 kinases visualized in Jalview.** Conserved and similar residues are highlighted according to physicochemical properties. Both kinases belong to the TK group and display strong sequence similarity. Notably, key motifs such as the DFG loop, catalytic lysine, HRD, and APE motifs are preserved across the aligned region, reflecting the highly conserved architecture of the SH1 domain within this kinase family.



**Figure A9: Residue-wise C $\alpha$  RMSF profiles of Lyn kinase across multiple systems.** (a) Comparison of RMSF profiles for apo Lyn WT SH1 domain and selected mutant systems (K275A, E290D, I364L). (b) RMSF profiles for ATP-bound systems, including WT and five mutants (K275A, E290D, E290K, I364L, I364N). Key structural elements are annotated for reference, including the ATP-binding site, the  $\alpha$ C,  $\alpha$ E,  $\alpha$ F, and  $\alpha$ H motifs, the DFG motif, and the activation loop. The plots highlight region-specific differences in C $\alpha$  flexibility across mutations and ligand-binding conditions.

## Appendix B: Tables

Table A1: Overview of MD simulation systems, replicates, and trajectory lengths.

| System    | Type | Number of Replicates | Time ( $\mu$ s) | Number of Frames |
|-----------|------|----------------------|-----------------|------------------|
| WT        | Apo  | 3                    | 6               | 6000             |
| K275A     |      | 3                    | 6               | 6000             |
| E290D     |      | 3                    | 6               | 6000             |
| E290K     |      | 3                    | 6               | 6000             |
| I364L     |      | 3                    | 6               | 6000             |
| I364N     |      | 3                    | 6               | 6000             |
| WT-ATP    | Holo | 3                    | 6               | 6000             |
| WT-DAS    |      | 3                    | 6               | 6000             |
| K275A-ATP |      | 3                    | 6               | 6000             |
| E290D-ATP |      | 3                    | 6               | 6000             |
| E290K-ATP |      | 3                    | 6               | 6000             |
| I364L-ATP |      | 3                    | 6               | 6000             |
| I364N-ATP |      | 3                    | 6               | 6000             |

Table A2: Naturally occurring mutations in Lyn kinase retrieved from cancer genomics datasets in cBioPortal and compared to aligned residues in other SFKs to assess conservation. Mutation sites are highlighted, with orange-colored substitutions indicating reported amino acid changes.

| Protein | Subfamily | Sequence | Residue 1    | Residue 2 | Residue 3 | Residue 4 | Residue 5 | Residue 6 |
|---------|-----------|----------|--------------|-----------|-----------|-----------|-----------|-----------|
| LYN     | Src-B     | 247–501  | G264S/splice | E290D/K   | V303L/M   | I364L/N   | R366Q     | W480L     |
| HCK     | Src-B     | 262–515  | A279         | E305      | V318      | I378P     | R380      | W494      |
| LCK     | Src-B     | 245–498  | G262W        | E288      | V301      | I361      | R363      | W477      |
| BLK     | Src-B     | 241–494  | G258         | E284Q     | V297      | I357V     | R359H     | W473      |
| SRC     | Src-A     | 270–523  | G287splice   | E313      | V326      | V386I     | R388      | W502      |
| FYN     | Src-A     | 271–524  | G288         | E314      | V327      | I387      | R389Q     | W503shift |
| YES     | Src-A     | 277–530  | G294splice   | E320      | V333      | I393      | R395Q     | W509      |
| FGR     | Src-A     | 263–516  | G280         | E306K     | V319M     | I379      | R381S     | W495      |

Table A3: Structural and pathogenicity prediction scores for selected Lyn kinase mutations. The table summarizes computational predictions for four cancer-associated missense mutations in Lyn kinase.  $\Delta\Delta G$  values (in kcal/mol) estimate changes in protein stability upon mutation, where positive values suggest destabilization. CADD scores  $> 20$  indicate potentially deleterious effects. AlphaMissense scores (ranging from 0 to 1) reflect predicted pathogenicity, with values near 1 denoting high probability of disease relevance. ESM-1b scores (more negative = more deleterious) are derived from deep learning models that capture functional disruption.

| Mutation | $\Delta\Delta G$ (kcal/mol) | CADD | AlphaMissense | ESM-1b |
|----------|-----------------------------|------|---------------|--------|
| I364L    | 0.58                        | 27.3 | 0.91          | -9.6   |
| I364N    | 1.92                        | 31.0 | 1.00          | -16.7  |
| E290D    | 1.82                        | 20.3 | 1.00          | -14.7  |
| E290K    | 1.77                        | 33.0 | 1.00          | -15.8  |

Table A4: Wilcoxon signed-rank test results for pairwise RMSD comparisons within the SH1 domain of Lyn kinase. The p-values indicate the statistical significance of differences in structural fluctuations between WT, mutant, and ligand-bound systems, with p-values less than 0.05 considered significant.

| Comparison              | p-Value   | Significance    | Comparison              | p-Value   | Significance    | Comparison              | p-Value   | Significance    |
|-------------------------|-----------|-----------------|-------------------------|-----------|-----------------|-------------------------|-----------|-----------------|
| WT vs. K275A            | <2.2e-16  | Significant     | WT vs. E290D            | <2.2e-16  | Significant     | WT vs. E290K            | <2.2e-16  | Significant     |
| WT vs. I364L            | <2.2e-16  | Significant     | WT vs. I364N            | <2.2e-16  | Significant     | WT vs. WT-DAS           | <2.2e-16  | Significant     |
| WT vs. WT-ATP           | <2.2e-16  | Significant     | WT vs. K275A-ATP        | <2.2e-16  | Significant     | WT vs. E290D-ATP        | <2.2e-16  | Significant     |
| WT vs. E290K-ATP        | <2.2e-16  | Significant     | WT vs. I364L-ATP        | <2.2e-16  | Significant     | WT vs. I364N-ATP        | <2.2e-16  | Significant     |
| K275A vs. E290D         | <2.2e-16  | Significant     | K275A vs. E290K         | 0.004714  | Significant     | K275A vs. I364L         | <2.2e-16  | Significant     |
| K275A vs. I364N         | 2.696e-06 | Significant     | K275A vs. WT-DAS        | <2.2e-16  | Significant     | K275A vs. WT-ATP        | <2.2e-16  | Significant     |
| K275A vs. K275A-ATP     | 0.2347    | Non-significant | K275A vs. E290D-ATP     | 1.24e-13  | Significant     | K275A vs. E290K-ATP     | 0.001487  | Significant     |
| K275A vs. I364L-ATP     | <2.2e-16  | Significant     | K275A vs. I364N-ATP     | 1.601e-15 | Significant     | E290D vs. E290K         | <2.2e-16  | Significant     |
| E290D vs. I364L         | 1.517e-06 | Significant     | E290D vs. I364N         | <2.2e-16  | Significant     | E290D vs. WT-DAS        | <2.2e-16  | Significant     |
| E290D vs. WT-ATP        | <2.2e-16  | Significant     | E290D vs. K275A-ATP     | <2.2e-16  | Significant     | E290D vs. E290D-ATP     | <2.2e-16  | Significant     |
| E290D vs. E290K-ATP     | <2.2e-16  | Significant     | E290D vs. I364L-ATP     | <2.2e-16  | Significant     | E290D vs. I364N-ATP     | <2.2e-16  | Significant     |
| E290K vs. I364L         | <2.2e-16  | Significant     | E290K vs. I364N         | <2.2e-16  | Significant     | E290K vs. WT-DAS        | 0.8891    | Non-significant |
| E290K vs. WT-ATP        | <2.2e-16  | Significant     | E290K vs. K275A-ATP     | <2.2e-16  | Significant     | E290K vs. E290D-ATP     | <2.2e-16  | Significant     |
| E290K vs. E290K-ATP     | <2.2e-16  | Significant     | E290K vs. I364L-ATP     | <2.2e-16  | Significant     | E290K vs. I364N-ATP     | <2.2e-16  | Significant     |
| I364L vs. I364N         | <2.2e-16  | Significant     | I364L vs. WT-DAS        | <2.2e-16  | Significant     | I364L vs. WT-ATP        | <2.2e-16  | Significant     |
| I364L vs. K275A-ATP     | <2.2e-16  | Significant     | I364L vs. E290D-ATP     | <2.2e-16  | Significant     | I364L vs. E290K-ATP     | <2.2e-16  | Significant     |
| I364L vs. I364L-ATP     | <2.2e-16  | Significant     | I364L vs. I364N-ATP     | <2.2e-16  | Significant     | I364N vs. WT-DAS        | <2.2e-16  | Significant     |
| I364N vs. WT-ATP        | 1.001e-10 | Significant     | I364N vs. K275A-ATP     | 0.0001416 | Significant     | I364N vs. E290D-ATP     | 4.883e-09 | Significant     |
| I364N vs. E290K-ATP     | 7.369e-12 | Significant     | I364N vs. I364L-ATP     | 0.001316  | Significant     | I364N vs. I364N-ATP     | 1.253e-13 | Significant     |
| WT-DAS vs. WT-ATP       | 0.01429   | Significant     | WT-DAS vs. K275A-ATP    | <2.2e-16  | Significant     | WT-DAS vs. E290D-ATP    | <2.2e-16  | Significant     |
| WT-DAS vs. E290K-ATP    | <2.2e-16  | Significant     | WT-DAS vs. I364L-ATP    | <2.2e-16  | Significant     | WT-DAS vs. I364N-ATP    | <2.2e-16  | Significant     |
| WT-ATP vs. K275A-ATP    | <2.2e-16  | Significant     | WT-ATP vs. E290D-ATP    | 0.6291    | Non-significant | WT-ATP vs. E290K-ATP    | <2.2e-16  | Significant     |
| WT-ATP vs. I364L-ATP    | <2.2e-16  | Significant     | WT-ATP vs. I364N-ATP    | <2.2e-16  | Significant     | K275A-ATP vs. E290D-ATP | <2.2e-16  | Significant     |
| K275A-ATP vs. E290K-ATP | 0.6015    | Non-significant | K275A-ATP vs. I364L-ATP | <2.2e-16  | Significant     | K275A-ATP vs. I364N-ATP | 2.678e-07 | Significant     |
| E290D-ATP vs. E290K-ATP | <2.2e-16  | Significant     | E290D-ATP vs. I364L-ATP | <2.2e-16  | Significant     | E290D-ATP vs. I364N-ATP | <2.2e-16  | Significant     |
| E290K-ATP vs. I364L-ATP | <2.2e-16  | Significant     | E290K-ATP vs. I364N-ATP | <2.2e-16  | Significant     | I364L-ATP vs. I364N-ATP | <2.2e-16  | Significant     |

Table A5: Wilcoxon signed-rank test results for pairwise RMSD comparisons within the SH2 domain of Lyn kinase. The p-values indicate the statistical significance of differences in structural fluctuations between WT, mutant, and ligand-bound systems, with p-values less than 0.05 considered significant.

| Comparison              | p-Value   | Significance    | Comparison              | p-Value   | Significance    | Comparison              | p-Value  | Significance |
|-------------------------|-----------|-----------------|-------------------------|-----------|-----------------|-------------------------|----------|--------------|
| WT vs. K275A            | <2.2e−16  | Significant     | WT vs. E290D            | <2.2e−16  | Significant     | WT vs. E290K            | <2.2e−16 | Significant  |
| WT vs. I364L            | <2.2e−16  | Significant     | WT vs. I364N            | <2.2e−16  | Significant     | WT vs. WT-DAS           | <2.2e−16 | Significant  |
| WT vs. WT-ATP           | <2.2e−16  | Significant     | WT vs. K275A-ATP        | <2.2e−16  | Significant     | WT vs. E290D-ATP        | <2.2e−16 | Significant  |
| WT vs. E290K-ATP        | <2.2e−16  | Significant     | WT vs. I364L-ATP        | <2.2e−16  | Significant     | WT vs. I364N-ATP        | <2.2e−16 | Significant  |
| K275A vs. E290D         | <2.2e−16  | Significant     | K275A vs. E290K         | 0.1173    | Non-significant | K275A vs. I364L         | <2.2e−16 | Significant  |
| K275A vs. I364N         | 3.679e−10 | Significant     | K275A vs. WT-DAS        | <2.2e−16  | Significant     | K275A vs. WT-ATP        | <2.2e−16 | Significant  |
| K275A vs. K275A-ATP     | <2.2e−16  | Significant     | K275A vs. E290D-ATP     | <2.2e−16  | Significant     | K275A vs. E290K-ATP     | <2.2e−16 | Significant  |
| K275A vs. I364L-ATP     | <2.2e−16  | Significant     | K275A vs. I364N-ATP     | <2.2e−16  | Significant     | E290D vs. E290K         | <2.2e−16 | Significant  |
| E290D vs. I364L         | 0.6364    | Non-significant | E290D vs. I364N         | <2.2e−16  | Significant     | E290D vs. WT-DAS        | <2.2e−16 | Significant  |
| E290D vs. WT-ATP        | <2.2e−16  | Significant     | E290D vs. K275A-ATP     | <2.2e−16  | Significant     | E290D vs. E290D-ATP     | <2.2e−16 | Significant  |
| E290D vs. E290K-ATP     | <2.2e−16  | Significant     | E290D vs. I364L-ATP     | <2.2e−16  | Significant     | E290D vs. I364N-ATP     | <2.2e−16 | Significant  |
| E290K vs. I364L         | <2.2e−16  | Significant     | E290K vs. I364N         | 0.0001829 | Significant     | E290K vs. WT-DAS        | 3.07e−06 | Significant  |
| E290K vs. WT-ATP        | <2.2e−16  | Significant     | E290K vs. K275A-ATP     | <2.2e−16  | Significant     | E290K vs. E290D-ATP     | <2.2e−16 | Significant  |
| E290K vs. E290K-ATP     | <2.2e−16  | Significant     | E290K vs. I364L-ATP     | <2.2e−16  | Significant     | E290K vs. I364N-ATP     | <2.2e−16 | Significant  |
| I364L vs. I364N         | <2.2e−16  | Significant     | I364L vs. WT-DAS        | <2.2e−16  | Significant     | I364L vs. WT-ATP        | <2.2e−16 | Significant  |
| I364L vs. K275A-ATP     | <2.2e−16  | Significant     | I364L vs. E290D-ATP     | <2.2e−16  | Significant     | I364L vs. E290K-ATP     | <2.2e−16 | Significant  |
| I364L vs. I364L-ATP     | <2.2e−16  | Significant     | I364L vs. I364N-ATP     | <2.2e−16  | Significant     | I364N vs. WT-DAS        | <2.2e−16 | Significant  |
| I364N vs. WT-ATP        | <2.2e−16  | Significant     | I364N vs. K275A-ATP     | 0.000179  | Significant     | I364N vs. E290D-ATP     | <2.2e−16 | Significant  |
| I364N vs. E290K-ATP     | 9.852e−13 | Significant     | I364N vs. I364L-ATP     | 0.7164    | Non-significant | I364N vs. I364N-ATP     | <2.2e−16 | Significant  |
| WT-DAS vs. WT-ATP       | 6.339e−15 | Significant     | WT-DAS vs. K275A-ATP    | <2.2e−16  | Significant     | WT-DAS vs. E290D-ATP    | <2.2e−16 | Significant  |
| WT-DAS vs. E290K-ATP    | <2.2e−16  | Significant     | WT-DAS vs. I364L-ATP    | <2.2e−16  | Significant     | WT-DAS vs. I364N-ATP    | <2.2e−16 | Significant  |
| WT-ATP vs. K275A-ATP    | <2.2e−16  | Significant     | WT-ATP vs. E290D-ATP    | <2.2e−16  | Significant     | WT-ATP vs. E290K-ATP    | <2.2e−16 | Significant  |
| WT-ATP vs. I364L-ATP    | <2.2e−16  | Significant     | WT-ATP vs. I364N-ATP    | <2.2e−16  | Significant     | K275A-ATP vs. E290D-ATP | <2.2e−16 | Significant  |
| K275A-ATP vs. E290K-ATP | <2.2e−16  | Significant     | K275A-ATP vs. I364L-ATP | 0.5784    | Non-significant | K275A-ATP vs. I364N-ATP | <2.2e−16 | Significant  |
| E290D-ATP vs. E290K-ATP | 4.382e−09 | Significant     | E290D-ATP vs. I364L-ATP | 8.142e−13 | Significant     | E290D-ATP vs. I364N-ATP | <2.2e−16 | Significant  |
| E290K-ATP vs. I364L-ATP | <2.2e−16  | Significant     | E290K-ATP vs. I364N-ATP | <2.2e−16  | Significant     | I364L-ATP vs. I364N-ATP | <2.2e−16 | Significant  |



Table A6: Wilcoxon signed-rank test results for pairwise RMSD comparisons within the SH3 domain of Lyn kinase. The p-values indicate the statistical significance of differences in structural fluctuations between WT, mutant, and ligand-bound systems, with p-values less than 0.05 considered significant.

| Comparison              | p-Value   | Significance    | Comparison              | p-Value   | Significance    | Comparison              | p-Value   | Significance |
|-------------------------|-----------|-----------------|-------------------------|-----------|-----------------|-------------------------|-----------|--------------|
| WT vs. K275A            | <2.2e-16  | Significant     | WT vs. E290D            | <2.2e-16  | Significant     | WT vs. E290K            | <2.2e-16  | Significant  |
| WT vs. I364L            | <2.2e-16  | Significant     | WT vs. I364N            | <2.2e-16  | Significant     | WT vs. WT-DAS           | <2.2e-16  | Significant  |
| WT vs. WT-ATP           | <2.2e-16  | Significant     | WT vs. K275A-ATP        | <2.2e-16  | Significant     | WT vs. E290D-ATP        | <2.2e-16  | Significant  |
| WT vs. E290K-ATP        | <2.2e-16  | Significant     | WT vs. I364L-ATP        | <2.2e-16  | Significant     | WT vs. I364N-ATP        | <2.2e-16  | Significant  |
| K275A vs. E290D         | <2.2e-16  | Significant     | K275A vs. E290K         | <2.2e-16  | Significant     | K275A vs. I364L         | <2.2e-16  | Significant  |
| K275A vs. I364N         | <2.2e-16  | Significant     | K275A vs. WT-DAS        | 9.034e-16 | Significant     | K275A vs. WT-ATP        | <2.2e-16  | Significant  |
| K275A vs. K275A-ATP     | 1.186e-05 | Significant     | K275A vs. E290D-ATP     | <2.2e-16  | Significant     | K275A vs. E290K-ATP     | <2.2e-16  | Significant  |
| K275A vs. I364L-ATP     | 2.136e-05 | Significant     | K275A vs. I364N-ATP     | <2.2e-16  | Significant     | E290D vs. E290K         | <2.2e-16  | Significant  |
| E290D vs. I364L         | 0.799     | Non-significant | E290D vs. I364N         | <2.2e-16  | Significant     | E290D vs. WT-DAS        | <2.2e-16  | Significant  |
| E290D vs. WT-ATP        | <2.2e-16  | Significant     | E290D vs. K275A-ATP     | <2.2e-16  | Significant     | E290D vs. E290D-ATP     | <2.2e-16  | Significant  |
| E290D vs. E290K-ATP     | <2.2e-16  | Significant     | E290D vs. I364L-ATP     | <2.2e-16  | Significant     | E290D vs. I364N-ATP     | <2.2e-16  | Significant  |
| E290K vs. I364L         | <2.2e-16  | Significant     | E290K vs. I364N         | 0.3174    | Non-significant | E290K vs. WT-DAS        | <2.2e-16  | Significant  |
| E290K vs. WT-ATP        | <2.2e-16  | Significant     | E290K vs. K275A-ATP     | <2.2e-16  | Significant     | E290K vs. E290D-ATP     | <2.2e-16  | Significant  |
| E290K vs. E290K-ATP     | <2.2e-16  | Significant     | E290K vs. I364L-ATP     | <2.2e-16  | Significant     | E290K vs. I364N-ATP     | <2.2e-16  | Significant  |
| I364L vs. I364N         | <2.2e-16  | Significant     | I364L vs. WT-DAS        | <2.2e-16  | Significant     | I364L vs. WT-ATP        | <2.2e-16  | Significant  |
| I364L vs. K275A-ATP     | <2.2e-16  | Significant     | I364L vs. E290D-ATP     | <2.2e-16  | Significant     | I364L vs. E290K-ATP     | <2.2e-16  | Significant  |
| I364L vs. I364L-ATP     | <2.2e-16  | Significant     | I364L vs. I364N-ATP     | <2.2e-16  | Significant     | I364N vs. WT-DAS        | 1.32e-10  | Significant  |
| I364N vs. WT-ATP        | <2.2e-16  | Significant     | I364N vs. K275A-ATP     | 0.1015    | Non-significant | I364N vs. E290D-ATP     | <2.2e-16  | Significant  |
| I364N vs. E290K-ATP     | <2.2e-16  | Significant     | I364N vs. I364L-ATP     | <2.2e-16  | Significant     | I364N vs. I364N-ATP     | <2.2e-16  | Significant  |
| WT-DAS vs. WT-ATP       | <2.2e-16  | Significant     | WT-DAS vs. K275A-ATP    | <2.2e-16  | Significant     | WT-DAS vs. E290D-ATP    | <2.2e-16  | Significant  |
| WT-DAS vs. E290K-ATP    | <2.2e-16  | Significant     | WT-DAS vs. I364L-ATP    | <2.2e-16  | Significant     | WT-DAS vs. I364N-ATP    | <2.2e-16  | Significant  |
| WT-ATP vs. K275A-ATP    | <2.2e-16  | Significant     | WT-ATP vs. E290D-ATP    | <2.2e-16  | Significant     | WT-ATP vs. E290K-ATP    | <2.2e-16  | Significant  |
| WT-ATP vs. I364L-ATP    | <2.2e-16  | Significant     | WT-ATP vs. I364N-ATP    | <2.2e-16  | Significant     | K275A-ATP vs. E290D-ATP | <2.2e-16  | Significant  |
| K275A-ATP vs. E290K-ATP | <2.2e-16  | Significant     | K275A-ATP vs. I364L-ATP | <2.2e-16  | Significant     | K275A-ATP vs. I364N-ATP | <2.2e-16  | Significant  |
| E290D-ATP vs. E290K-ATP | 0.005952  | Significant     | E290D-ATP vs. I364L-ATP | <2.2e-16  | Significant     | E290D-ATP vs. I364N-ATP | 1.251e-07 | Significant  |
| E290K-ATP vs. I364L-ATP | <2.2e-16  | Significant     | E290K-ATP vs. I364N-ATP | 0.2965    | Non-significant | I364L-ATP vs. I364N-ATP | <2.2e-16  | Significant  |



Table A7: RMSD-based clustering on the MD trajectories. The table lists the number of frames per cluster, maximum and average intra-cluster dissimilarity, cluster diameter, and inter-cluster separation.

| System    | Cluster | Size | Max Dissim. | Avg. Dissim. | Diameter | Separation |
|-----------|---------|------|-------------|--------------|----------|------------|
| WT        | 1       | 2179 | 382.19      | 98.68        | 542.10   | 17.91      |
| WT        | 2       | 1936 | 388.46      | 94.82        | 449.49   | 10.38      |
| WT        | 3       | 1886 | 398.54      | 101.07       | 438.12   | 10.38      |
| E290D     | 1       | 1641 | 115.03      | 38.75        | 137.17   | 6.96       |
| E290D     | 2       | 359  | 126.41      | 42.15        | 170.20   | 6.96       |
| E290K     | 1       | 4233 | 393.06      | 72.81        | 425.76   | 7.96       |
| E290K     | 2       | 1768 | 439.87      | 147.69       | 603.38   | 7.96       |
| I364L     | 1       | 3118 | 345.74      | 82.17        | 435.37   | 9.65       |
| I364L     | 2       | 1738 | 347.05      | 124.72       | 496.50   | 9.65       |
| I364L     | 3       | 1145 | 310.48      | 102.34       | 391.70   | 11.25      |
| I364N     | 1       | 4875 | 382.76      | 51.91        | 409.35   | 11.34      |
| I364N     | 2       | 1126 | 383.63      | 79.19        | 525.17   | 11.34      |
| K275A     | 1       | 4637 | 220.04      | 54.96        | 271.47   | 13.40      |
| K275A     | 2       | 336  | 243.79      | 108.13       | 456.15   | 12.53      |
| K275A     | 3       | 387  | 192.40      | 67.92        | 369.86   | 12.53      |
| K275A     | 4       | 287  | 188.68      | 45.48        | 241.30   | 15.03      |
| K275A     | 5       | 354  | 247.52      | 48.93        | 330.38   | 32.43      |
| WT-DAS    | 1       | 4136 | 331.26      | 76.31        | 374.73   | 11.68      |
| WT-DAS    | 2       | 1865 | 391.20      | 77.83        | 423.45   | 11.68      |
| WT-ATP    | 1       | 4144 | 310.65      | 64.38        | 371.13   | 7.12       |
| WT-ATP    | 2       | 1857 | 313.23      | 142.47       | 565.95   | 7.12       |
| E290D-ATP | 1       | 5501 | 410.90      | 65.01        | 449.78   | 7.55       |
| E290D-ATP | 2       | 500  | 415.25      | 133.81       | 597.88   | 7.55       |
| E290K-ATP | 1       | 3913 | 140.26      | 46.01        | 170.99   | 8.53       |
| E290K-ATP | 2       | 1931 | 173.59      | 48.63        | 222.09   | 8.53       |
| E290K-ATP | 3       | 157  | 311.33      | 55.64        | 447.85   | 18.70      |
| I364L-ATP | 1       | 5302 | 177.08      | 52.53        | 218.18   | 7.36       |
| I364L-ATP | 2       | 699  | 277.88      | 88.93        | 351.47   | 7.36       |
| I364N-ATP | 1       | 2520 | 67.95       | 39.76        | 103.92   | 6.44       |
| I364N-ATP | 2       | 1134 | 99.78       | 36.33        | 143.77   | 6.55       |
| I364N-ATP | 3       | 1282 | 204.19      | 42.18        | 246.18   | 6.83       |
| I364N-ATP | 4       | 1065 | 247.64      | 58.57        | 289.62   | 6.44       |
| K275A-ATP | 1       | 4533 | 360.82      | 66.98        | 405.05   | 7.52       |
| K275A-ATP | 2       | 1468 | 405.32      | 67.63        | 444.86   | 7.52       |

Table A8: Wilcoxon rank-sum test results for pairwise comparisons of ligand RMSD distributions across systems. A  $p$ -value threshold of 0.05 was used to determine statistical significance.

| Comparison              | $p$ -Value              | Significance    |
|-------------------------|-------------------------|-----------------|
| WT-ATP vs. WT-DAS       | $< 2.2 \times 10^{-16}$ | Significant     |
| WT-ATP vs. K275A-ATP    | $< 2.2 \times 10^{-16}$ | Significant     |
| WT-ATP vs. E290D-ATP    | $< 2.2 \times 10^{-16}$ | Significant     |
| WT-ATP vs. E290K-ATP    | $< 2.2 \times 10^{-16}$ | Significant     |
| WT-ATP vs. I364L-ATP    | $< 2.2 \times 10^{-16}$ | Significant     |
| WT-ATP vs. I364N-ATP    | $< 2.2 \times 10^{-16}$ | Significant     |
| WT-DAS vs. K275A-ATP    | 0.6218                  | Not Significant |
| WT-DAS vs. E290D-ATP    | $< 2.2 \times 10^{-16}$ | Significant     |
| WT-DAS vs. E290K-ATP    | $< 2.2 \times 10^{-16}$ | Significant     |
| WT-DAS vs. I364L-ATP    | $< 2.2 \times 10^{-16}$ | Significant     |
| WT-DAS vs. I364N-ATP    | $1.36 \times 10^{-4}$   | Significant     |
| K275A-ATP vs. E290D-ATP | $< 2.2 \times 10^{-16}$ | Significant     |
| K275A-ATP vs. E290K-ATP | $< 2.2 \times 10^{-16}$ | Significant     |
| K275A-ATP vs. I364L-ATP | $< 2.2 \times 10^{-16}$ | Significant     |
| K275A-ATP vs. I364N-ATP | $6.11 \times 10^{-7}$   | Significant     |
| E290D-ATP vs. E290K-ATP | $< 2.2 \times 10^{-16}$ | Significant     |
| E290D-ATP vs. I364L-ATP | $< 2.2 \times 10^{-16}$ | Significant     |
| E290D-ATP vs. I364N-ATP | $< 2.2 \times 10^{-16}$ | Significant     |
| E290K-ATP vs. I364L-ATP | $< 2.2 \times 10^{-16}$ | Significant     |
| E290K-ATP vs. I364N-ATP | $< 2.2 \times 10^{-16}$ | Significant     |
| I364L-ATP vs. I364N-ATP | $< 2.2 \times 10^{-16}$ | Significant     |

Table A9:  $p$ -values for Pearson correlation significance between systems. Values less than 0.05 are considered statistically significant.

| Systems   | WT       | WT-ATP   | WT-DAS   | K275A   | K275A-ATP | E290D    | E290D-ATP | E290K    | E290K-ATP | I364L    | I364L-ATP | I364N    | I364N-ATP |
|-----------|----------|----------|----------|---------|-----------|----------|-----------|----------|-----------|----------|-----------|----------|-----------|
| WT        | 0        | 7.26E-11 | 9.01E-05 | 0.2425  | 0.7947    | 0.00894  | 0.7327    | 0.01716  | 0.06322   | 0.01496  | 0.33319   | 0.00764  | 0.86731   |
| WT-ATP    | 7.26E-11 | 0        | 7.71E-05 | 0.23909 | 0.86908   | 0.01210  | 0.81578   | 0.02273  | 0.06595   | 0.01995  | 0.32193   | 0.01011  | 0.88367   |
| WT-DAS    | 9.01E-05 | 7.71E-05 | 0        | 0.17824 | 0.44469   | 0.02251  | 0.78695   | 0.04018  | 0.15660   | 0.03720  | 0.35211   | 0.01133  | 0.83684   |
| K275A     | 0.24254  | 0.23909  | 0.17824  | 0       | 0.09287   | 0.02769  | 0.74644   | 0.04101  | 0.04088   | 0.02494  | 0.69187   | 0.03955  | 0.36018   |
| K275A-ATP | 0.79471  | 0.86908  | 0.44469  | 0.09287 | 0         | 0.73697  | 0.14552   | 0.70846  | 0.53654   | 0.68448  | 0.09850   | 0.64746  | 0.04494   |
| E290D     | 0.00894  | 0.01210  | 0.02251  | 0.02769 | 0.73697   | 0        | 0.82337   | 7.10E-09 | 8.73E-05  | 2.50E-11 | 0.61881   | 4.25E-10 | 0.97466   |
| E290D-ATP | 0.73273  | 0.81578  | 0.78695  | 0.74644 | 0.14552   | 0.82337  | 0         | 0.82981  | 0.32898   | 0.76866  | 0.00127   | 0.82321  | 0.00136   |
| E290K     | 0.01716  | 0.02273  | 0.04018  | 0.04101 | 0.70846   | 7.10E-09 | 0.82981   | 0        | 0.00017   | 1.03E-10 | 0.65681   | 3.51E-08 | 0.98624   |
| E290K-ATP | 0.06322  | 0.06595  | 0.15660  | 0.04088 | 0.53654   | 8.73E-05 | 0.32898   | 0.00017  | 0         | 7.90E-05 | 0.62453   | 9.50E-05 | 0.75966   |
| I364L     | 0.01496  | 0.01995  | 0.03720  | 0.02494 | 0.68448   | 2.50E-11 | 0.76866   | 1.03E-10 | 7.90E-05  | 0        | 0.62108   | 4.60E-09 | 0.95895   |
| I364L-ATP | 0.33319  | 0.32193  | 0.35211  | 0.69187 | 0.09850   | 0.61881  | 0.00127   | 0.65681  | 0.62453   | 0.62108  | 0         | 0.67099  | 0.39058   |
| I364N     | 0.00764  | 0.01011  | 0.01133  | 0.03955 | 0.64746   | 4.25E-10 | 0.82321   | 3.51E-08 | 9.50E-05  | 4.60E-09 | 0.67099   | 0        | 0.99493   |
| I364N-ATP | 0.86731  | 0.88367  | 0.83684  | 0.36018 | 0.04494   | 0.97466  | 0.00136   | 0.98624  | 0.75966   | 0.95895  | 0.39058   | 0.99493  | 0         |

Table A10: Dynamic, interface, and evolutionary properties of Lyn allosteric connectors across Src family kinases. For each connector, the table lists its region, correlation with activity classification, module mixing, interface involvement, and interface-weighted score. Residue identities across SFKs are shown alongside Shannon entropy and Lyn-specific conservation scores, offering an integrated view of dynamic importance and evolutionary variation.

| Connector | Region | Cor-with-act | p-val   | Mean-mix | mix-weighted-corr | Interface | Interface-freq | Interface-score | Lyn | Hck | Lck | Blk | Fgr | Fyn | Src | Yes | Shannon score | Identity score |
|-----------|--------|--------------|---------|----------|-------------------|-----------|----------------|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|---------------|----------------|
| K343      | SH1C   | 0.90         | 3.5e-05 | 0.78     | 0.70              | SH2       | 1.00           | 1.39            | K   | K   | K   | R   | Q   | N   | Q   | Q   | 0.59          | 38             |
| D66       | SH3    | 0.75         | 0.00    | 0.77     | 0.58              | SH2       | 1.00           | 1.15            | D   | I   | N   | H   | T   | T   | T   | T   | 0.54          | 13             |
| S228      | Linker | 0.60         | 0.03    | 0.90     | 0.54              | SH3       | 1.00           | 1.08            | S   | S   | T   | R   | I   | K   | T   | T   | 0.51          | 25             |
| C226      | SH2    | 0.65         | 0.02    | 0.77     | 0.50              | SH3       | 1.00           | 0.99            | C   | C   | C   | C   | C   | C   | C   | C   | 1.00          | 100            |
| R304      | SH1N   | 0.57         | 0.04    | 0.83     | 0.48              | SH1C      | 1.00           | 0.96            | R   | K   | R   | R   | Q   | Q   | Q   | P   | 0.59          | 38             |
| Y266      | SH1N   | 0.51         | 0.08    | 0.86     | 0.44              | SH3       | 1.00           | 0.87            | Y   | Y   | Y   | Y   | W   | W   | W   | W   | 0.77          | 50             |
| A307      | SH1N   | 0.56         | 0.05    | 0.78     | 0.43              | Linker    | 1.00           | 0.86            | A   | A   | A   | A   | A   | A   | A   | A   | 1.00          | 100            |
| V250      | SH1N   | 1.00         | 2.5e-18 | 0.79     | 0.79              | none      | 0.00           | 0.79            | V   | E   | V   | V   | E   | I   | E   | E   | 0.68          | 38             |
| E245      | Linker | 0.49         | 0.09    | 0.78     | 0.38              | SH1N      | 1.00           | 0.76            | E   | E   | E   | Q   | S   | E   | E   | E   | 0.76          | 75             |
| P488      | SH1C   | 0.86         | 1.4e-04 | 0.86     | 0.74              | none      | 0.00           | 0.74            | P   | P   | P   | P   | P   | P   | P   | P   | 1.00          | 100            |
| P73       | SH3    | 0.98         | 3.5e-09 | 0.76     | 0.74              | none      | 0.00           | 0.74            | P   | D   | S   | D   | D   | D   | D   | D   | 0.76          | 13             |
| E312      | SH1N   | 0.45         | 0.12    | 0.82     | 0.37              | SH3       | 1.00           | 0.74            | E   | K   | Q   | K   | E   | E   | E   | E   | 0.70          | 63             |
| I227      | Linker | 0.48         | 0.10    | 0.77     | 0.37              | SH2       | 1.00           | 0.74            | I   | M   | Q   | V   | T   | H   | P   | P   | 0.37          | 13             |
| G65       | SH3    | 0.96         | 3.2e-07 | 0.76     | 0.73              | none      | 0.00           | 0.73            | G   | D   | D   | K   | V   | V   | V   | V   | 0.60          | 13             |
| G325      | SH1N   | 0.90         | 3.1e-05 | 0.78     | 0.70              | none      | 0.00           | 0.70            | G   | G   | G   | G   | G   | G   | G   | G   | 1.00          | 100            |
| S326      | SH1N   | 0.85         | 2.6e-04 | 0.82     | 0.70              | none      | 0.00           | 0.70            | S   | S   | S   | C   | S   | S   | S   | S   | 0.88          | 88             |
| Y265      | SH1N   | 0.44         | 0.14    | 0.86     | 0.38              | Linker    | 0.85           | 0.69            | Y   | T   | Y   | Y   | T   | T   | T   | T   | 0.78          | 38             |
| A255      | SH1N   | 0.87         | 1.0e-04 | 0.79     | 0.69              | none      | 0.00           | 0.69            | A   | A   | A   | S   | T   | N   | Q   | Q   | 0.50          | 38             |
| R244      | Linker | 0.43         | 0.14    | 0.79     | 0.34              | SH1N      | 1.00           | 0.68            | R   | R   | R   | R   | R   | R   | R   | R   | 1.00          | 100            |
| K481      | SH1C   | 0.88         | 6.4e-05 | 0.76     | 0.67              | none      | 0.00           | 0.67            | K   | K   | K   | R   | R   | K   | R   | K   | 0.78          | 63             |
| F347      | SH1C   | 0.83         | 4.0e-04 | 0.78     | 0.65              | none      | 0.00           | 0.65            | F   | F   | M   | M   | M   | M   | M   | M   | 0.81          | 25             |
| N116      | SH3    | 0.81         | 8.2e-04 | 0.79     | 0.64              | none      | 0.00           | 0.64            | N   | N   | N   | N   | N   | N   | N   | N   | 1.00          | 100            |
| M263      | SH1N   | 0.74         | 0.00    | 0.86     | 0.64              | none      | 0.00           | 0.64            | M   | M   | M   | M   | L   | M   | M   | M   | 0.88          | 88             |
| K324      | SH1N   | 0.82         | 6.1e-04 | 0.78     | 0.64              | none      | 0.00           | 0.64            | K   | K   | N   | R   | H   | K   | K   | K   | 0.65          | 63             |
| V373      | SH1C   | 0.79         | 0.00    | 0.79     | 0.63              | none      | 0.00           | 0.63            | V   | I   | I   | I   | I   | I   | I   | I   | 0.88          | 13             |
| D346      | SH1C   | 0.40         | 0.17    | 0.78     | 0.31              | SH2       | 1.00           | 0.62            | D   | D   | D   | D   | D   | D   | D   | D   | 1.00          | 100            |
| F258      | SH1N   | 0.82         | 5.5e-04 | 0.76     | 0.62              | none      | 0.00           | 0.62            | F   | F   | F   | F   | F   | F   | F   | F   | 1.00          | 100            |
| V308      | SH1N   | 0.39         | 0.19    | 0.78     | 0.30              | SH3       | 1.00           | 0.61            | V   | V   | V   | V   | V   | V   | V   | V   | 1.00          | 100            |
| I351      | SH1C   | 0.78         | 0.00    | 0.78     | 0.61              | none      | 0.00           | 0.61            | I   | I   | I   | I   | V   | V   | I   | I   | 0.81          | 75             |
| E260      | SH1N   | 0.80         | 0.00    | 0.76     | 0.61              | none      | 0.00           | 0.61            | E   | E   | E   | E   | D   | E   | E   | E   | 0.88          | 88             |
| A384      | SH1C   | 0.77         | 0.00    | 0.78     | 0.60              | none      | 0.00           | 0.60            | A   | A   | A   | A   | A   | A   | A   | A   | 1.00          | 100            |
| P146      | SH2    | 0.37         | 0.21    | 0.79     | 0.29              | SH1C      | 1.00           | 0.59            | P   | P   | P   | P   | P   | F   | A   | P   | 0.76          | 75             |
| G264      | SH1N   | 0.67         | 0.01    | 0.86     | 0.58              | none      | 0.00           | 0.58            | G   | A   | G   | G   | G   | G   | G   | G   | 0.88          | 88             |
| M322      | SH1N   | 0.74         | 0.00    | 0.78     | 0.58              | none      | 0.00           | 0.58            | M   | M   | M   | M   | M   | M   | M   | M   | 1.00          | 100            |
| M294      | SH1N   | 0.55         | 0.05    | 0.82     | 0.45              | SH1C      | 0.23           | 0.55            | M   | M   | M   | M   | M   | M   | M   | M   | 1.00          | 100            |
| I475      | SH1C   | 0.70         | 0.01    | 0.77     | 0.54              | none      | 0.00           | 0.54            | I   | I   | L   | V   | A   | L   | L   | L   | 0.60          | 25             |
| G256      | SH1N   | 0.68         | 0.01    | 0.79     | 0.54              | none      | 0.00           | 0.54            | G   | G   | G   | G   | G   | G   | G   | G   | 1.00          | 100            |
| A323      | SH1N   | 0.69         | 0.01    | 0.78     | 0.54              | none      | 0.00           | 0.54            | A   | A   | E   | A   | C   | N   | S   | S   | 0.51          | 38             |
| N372      | SH1C   | 0.68         | 0.01    | 0.79     | 0.53              | none      | 0.00           | 0.53            | N   | N   | N   | N   | N   | N   | N   | N   | 1.00          | 100            |
| A286      | SH1N   | 0.63         | 0.02    | 0.83     | 0.52              | none      | 0.00           | 0.52            | A   | A   | A   | A   | A   | S   | A   | A   | 0.88          | 88             |
| S83       | SH3    | 0.67         | 0.01    | 0.77     | 0.51              | none      | 0.00           | 0.51            | S   | S   | G   | Q   | T   | S   | S   | S   | 0.65          | 63             |
| V69       | SH3    | 0.66         | 0.01    | 0.77     | 0.51              | none      | 0.00           | 0.51            | V   | V   | I   | V   | I   | V   | V   | V   | 0.81          | 75             |
| P342      | SH1C   | 0.66         | 0.01    | 0.77     | 0.51              | none      | 0.00           | 0.51            | P   | P   | N   | P   | P   | P   | P   | P   | 0.88          | 88             |
| F420      | SH1C   | 0.61         | 0.03    | 0.83     | 0.50              | none      | 0.00           | 0.50            | F   | F   | F   | F   | F   | F   | F   | F   | 1.00          | 100            |

Table A11: Src family kinase inhibitors, their clinical development phase, inhibitor type, available PDB structures, and corresponding Src family targets.

| Inhibitor   | Phase | Type | PDB IDs                                                                                  | Target | Family |
|-------------|-------|------|------------------------------------------------------------------------------------------|--------|--------|
| Bafetinib   | 2     | 2.0  | 2e2b                                                                                     | LYN    | Src    |
| Bosutinib   | 4     | 1.0  | 3soa, 3ue4, 4mxo, 4mxx, 4mxy, 4mxz, 4qmn, 5ajq, 5i9x, 5vc3, 5vcb, 5vcc, 5wgl, 5wgm, 5wgn | LYN    | Src    |
| Bosutinib   | 4     | 1.0  | 3soa, 3ue4, 4mxo, 4mxx, 4mxy, 4mxz, 4qmn, 5ajq, 5i9x, 5vc3, 5vcb, 5vcc, 5wgl, 5wgm, 5wgn | HCK    | Src    |
| Bosutinib   | 4     | 1.0  | 3soa, 3ue4, 4mxo, 4mxx, 4mxy, 4mxz, 4qmn, 5ajq, 5i9x, 5vc3, 5vcb, 5vcc, 5wgl, 5wgm, 5wgn | SRC    | Src    |
| Dasatinib   | 4     | 1.0  | 2gqg, 2y60, 2zva, 3g5d, 3k54, 3lfa, 3oct, 3oht, 3qlg, 3sxx, 4qms, 4xv2, 5hbp, 5ixc       | SRC    | Src    |
| Dasatinib   | 4     | 1.0  | 2gqg, 2y60, 2zva, 3g5d, 3k54, 3lfa, 3oct, 3oht, 3qlg, 3sxx, 4qms, 4xv2, 5hbp, 5ixc       | LCK    | Src    |
| Dasatinib   | 4     | 1.0  | 2gqg, 2y60, 2zva, 3g5d, 3k54, 3lfa, 3oct, 3oht, 3qlg, 3sxx, 4qms, 4xv2, 5hbp, 5ixc       | YES    | Src    |
| Dasatinib   | 4     | 1.0  | 2gqg, 2y60, 2zva, 3g5d, 3k54, 3lfa, 3oct, 3oht, 3qlg, 3sxx, 4qms, 4xv2, 5hbp, 5ixc       | FYN    | Src    |
| Nintedanib  | 4     | 1.0  | 3c7q, 5maf, 5te0, 6nec                                                                   | LCK    | Src    |
| Nintedanib  | 4     | 1.0  | 3c7q, 5maf, 5te0, 6nec                                                                   | LYN    | Src    |
| Nintedanib  | 4     | 1.0  | 3c7q, 5maf, 5te0, 6nec                                                                   | SRC    | Src    |
| Ponatinib   | 4     | 2.0  | 3ik3, 3oxz, 3zos, 4c8b, 4qrc, 4tyj, 4u0i, 4uxq, 4v01, 4v04, 6eg9, 6p9m                   | LCK    | Src    |
| Ponatinib   | 4     | 2.0  | 3ik3, 3oxz, 3zos, 4c8b, 4qrc, 4tyj, 4u0i, 4uxq, 4v01, 4v04, 6eg9, 6p9m                   | SRC    | Src    |
| Ponatinib   | 4     | 2.0  | 3ik3, 3oxz, 3zos, 4c8b, 4qrc, 4tyj, 4u0i, 4uxq, 4v01, 4v04, 6eg9, 6p9m                   | LYN    | Src    |
| Rebastinib  | 2     | 2.0  | 3qri, 3qrj, 5g6v, 6cnh, 6mwe, 8ub8                                                       | LYN    | Src    |
| Regorafenib | 4     | 2.0  | –                                                                                        | FRK    | Src    |
| Saracatinib | 3     | 1.0  | 2h8h, 4qmx, 5vcx, 5vd3, 6zgc                                                             | SRC    | Src    |
| Vandetanib  | 4     | 1.5  | 2ivu                                                                                     | BRK    | Src    |

Table A12: Similarity of full-length Src family kinases to Lyn based on BLAST sequence alignment. Blue text highlights Lyn-related kinases; red indicates Src-related kinases.

| Protein    | Organism     | Max Score | Total Score | Query Cover | E value | Identity% | Length | UniProt ID |
|------------|--------------|-----------|-------------|-------------|---------|-----------|--------|------------|
| <b>Lyn</b> | Homo sapiens | 1066      | 1066        | 100%        | 0       | 100.00%   | 512    | P07948     |
| <b>Hck</b> | Homo sapiens | 754       | 754         | 100%        | 0       | 70.54%    | 526    | P08631     |
| <b>Lck</b> | Homo sapiens | 685       | 685         | 94%         | 0       | 67.29%    | 509    | P06239     |
| <b>Fyn</b> | Homo sapiens | 590       | 590         | 87%         | 0       | 62.72%    | 509    | P06241     |
| <b>Yes</b> | Homo sapiens | 592       | 592         | 87%         | 0       | 62.25%    | 543    | P07947     |
| <b>Fgr</b> | Homo sapiens | 564       | 564         | 87%         | 0       | 61.12%    | 509    | P09769     |
| <b>Blk</b> | Homo sapiens | 645       | 645         | 100%        | 0       | 60.35%    | 505    | P51451     |
| <b>Src</b> | Homo sapiens | 572       | 572         | 86%         | 0       | 60.27%    | 536    | P12931     |

Table A13: Similarity of Src family kinase SH1 domains to Lyn based on BLAST sequence alignment. Blue text highlights Lyn-related kinases; red indicates Src-related kinases.

| Protein    | Organism     | Max Score | Total Score | Query Cover | E value | Identity%      | Length | UniProt ID |
|------------|--------------|-----------|-------------|-------------|---------|----------------|--------|------------|
| <b>Lyn</b> | Homo sapiens | 530       | 530         | 100%        | 0.0     | <b>100.00%</b> | 512    | P07948     |
| <b>Hck</b> | Homo sapiens | 441       | 441         | 100%        | 6e-155  | 81.57%         | 526    | P08631     |
| <b>Lck</b> | Homo sapiens | 407       | 407         | 100%        | 8e-142  | 74.90%         | 509    | P06239     |
| <b>Blk</b> | Homo sapiens | 395       | 395         | 100%        | 2e-137  | 71.88%         | 505    | P51451     |
| <b>Fyn</b> | Homo sapiens | 378       | 378         | 100%        | 5e-130  | 68.63%         | 537    | P06241     |
| <b>Src</b> | Homo sapiens | 370       | 370         | 100%        | 7e-127  | 65.49%         | 536    | P12931     |
| <b>Yes</b> | Homo sapiens | 367       | 367         | 100%        | 6e-126  | 66.27%         | 543    | P07947     |
| <b>Fgr</b> | Homo sapiens | 355       | 355         | 100%        | 5e-121  | 65.88%         | 529    | P09769     |

Table A14: Structural features selected for supervised classification of active-like and inactive-like ensembles in full-length Lyn kinase. Each feature is grouped by functional category, describes a specific geometric relationship between residues or domains, and exhibits differential values across conformational ensembles.

| Feature   | Category                 | Geometric Relationship                    | Ensemble                |
|-----------|--------------------------|-------------------------------------------|-------------------------|
| K275–E290 | Catalytic region         | Catalytic– $\alpha$ C helix contact       | Lower in active-like    |
| D385–K275 | Catalytic region         | Catalytic site contact                    | Lower in active-like    |
| E290–D385 | Catalytic region         | Catalytic site contact                    | Lower in active-like    |
| K275–M294 | Catalytic region         | Catalytic– $\alpha$ C helix contact       | Lower in active-like    |
| Y397–R366 | Activation segment       | Activation loop–SH1C contact              | Lower in inactive-like  |
| Y397–E289 | Activation segment       | Activation loop– $\alpha$ C helix contact | Lower in active-like    |
| E157–K162 | Intradomain architecture | SH2 internal contact                      | Lower in inactive-like  |
| R390–K361 | Intradomain architecture | SH1C internal contact                     | Lower in active-like    |
| SH1 Lobes | Intradomain architecture | SH1N–SH1C lobes                           | Lower in inactive-like  |
| Rspine    | Intradomain architecture | F386–M294–L305 angle                      | Higher in active-like   |
| D498–K137 | Interdomain coupling     | SH1C–SH2 domain proximity                 | Lower in inactive-like  |
| D206–K89  | Interdomain coupling     | SH2–SH3 domain proximity                  | Higher in inactive-like |
| I77–N449  | Interdomain coupling     | SH3–SH1C lobe proximity                   | Lower in inactive-like  |
| SH3–Nlobe | Interdomain coupling     | Center-of-mass proximity                  | Lower in inactive-like  |
| SH2–Clobe | Interdomain coupling     | Center-of-mass proximity                  | Lower in inactive-like  |
| SH2–SH3   | Interdomain coupling     | Center-of-mass proximity                  | Lower in inactive-like  |

# Appendix C: Mathematical Framework

## Allosteric Connector Scoring

The quantitative scoring procedure used to identify allosteric connectors was based on three components: centrality–state correlation, module-mixing behavior, and interface frequency.

### Centrality–State Correlation

For each residue  $r$ , a betweenness centrality vector across all  $N$  systems was defined as:

$$\mathbf{b}_r = (b_{r,1}, \dots, b_{r,N}). \quad (\text{A1})$$

System activity was encoded as a binary vector:

$$\mathbf{y} = (y_1, \dots, y_N), \quad (\text{A2})$$

with  $y_s = 1$  for active-like systems and  $y_s = 0$  otherwise. The Pearson correlation coefficient between these vectors was computed as:

$$\rho_r = \text{corr}(\mathbf{b}_r, \mathbf{y}). \quad (\text{A3})$$

### Module-Mixing Index

For each residue  $r$  in system  $s$ , the module-mixing score was defined as  $M_{r,s}$ . The average mixing across all systems was given by:

$$\mu_r = \frac{1}{N} \sum_{s=1}^N M_{r,s}. \quad (\text{A4})$$

### Interface Frequency

The number of systems in which residue  $r$  appeared at an interdomain interface was counted as  $k_r$ . The normalized interface frequency was:

$$f_r = \frac{k_r}{N}. \quad (\text{A5})$$

## Mixing-Weighted Correlation

The first two features were combined as:

$$\text{mix\_weighted\_corr}_r = \rho_r \cdot \mu_r. \quad (\text{A6})$$

## Final Interface-Weighted Score

The complete interface-weighted score for each residue was:

$$\text{interface\_weighted\_score}_r = \text{mix\_weighted\_corr}_r \cdot (1 + f_r). \quad (\text{A7})$$

Residues with interface-weighted scores above 0.5 were classified as putative allosteric connectors.

## Entropy-Based Conservation Scoring

Evolutionary conservation of Lyn connector positions across the Src family was quantified using a gap-aware normalized Shannon entropy metric.

### Shannon Entropy

Entropy was computed for each alignment position as:

$$H = - \sum_{i=1}^n p_i \log_2 p_i \quad (\text{A8})$$

where  $p_i$  denotes the observed frequency of amino acid  $i$  at the corresponding alignment column, including gaps.

### Normalized Conservation Score

Entropy values were normalized to a 21-symbol distribution (20 amino acids plus gap), yielding the final conservation score:

$$\text{Conservation Score} = 1 - \frac{H}{\log_2(21)} \quad (\text{A9})$$

This normalization assigns a score of 1 to fully conserved positions and values approaching 0 to highly diverse positions.

# Publications

## Publication Resulting from this Thesis:

- [1] **M. Rabipour**, F. Hassenrück, E. Pallaske, F. Röhrig, M. Hallek, J. R. Alvarez-Idaboy, and R. Rebolledo-Rios. Allosteric Coupling in Full-Length Lyn Kinase Revealed by Molecular Dynamics and Network Analysis. *International Journal of Molecular Sciences*, 26:5835, 2025. doi:10.3390/ijms26125835

## Other Publications:

- [2] F. Hassenrück, M. Farina-Morillas, L. Neumann, F. Landini, S. J. Blakemore, **M. Rabipour**, J. R. Alvarez-Idaboy, C. P. Pallasch, M. Hallek, R. Rebolledo-Rios, and G. Krause. Functional Impact and Molecular Binding Modes of Drugs That Target the PI3K Isoform p110 $\delta$ . *Communications Biology*, 6:603, 2023. doi:10.1038/s42003-023-04921-z
- [3] **M. Rabipour**, Z. Sekhavat Pour, R. Sahraei, M. Ghaemy, M. Erfani Jazi, and T. E. Mlsna. pH-Sensitive Nanocomposite Hydrogels Based on Poly(Vinyl Alcohol) Macromonomer and Graphene Oxide for Removal of Cationic Dyes from Aqueous Solutions. *Journal of Polymers and the Environment*, 28:2488–2500, 2020. doi:10.1007/s10924-019-01625-6



# Bibliography

- [1] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298:1912–1934, 2002.
- [2] S. H. Seok. Structural insights into protein regulation by phosphorylation and substrate recognition of protein kinases/phosphatases. *Life*, 11(9):957, 2021.
- [3] X. Wu, Z. Yang, J. Zou, H. Gao, Z. Shao, C. Li, and P. Lei. Protein kinases in neurodegenerative diseases: current understandings and implications for drug discovery. *Signal Transduction and Targeted Therapy*, 10(1):146, 2025.
- [4] M. E. M. Noble, J. A. Endicott, and L. N. Johnson. Protein kinase inhibitors: insights into drug design from structure. *Science*, 303(5665):1800–1805, 2004.
- [5] R. Reinhardt and T. A. Leonard. A critical evaluation of protein kinase regulation by activation loop autophosphorylation. *eLife*, 12:e88210, 2023.
- [6] H. Zhang, X. Cao, M. Tang, G. Zhong, Y. Si, H. Li, F. Zhu, Q. Liao, L. Li, J. Zhao, J. Feng, C. Kaulich, M. Li, S. Wang, F. Wang, L. Chen, L. Li, Z. Xia, H. Liang, X. H. Liang, T. Lu, and B. Zhao. A subcellular map of the human kinome. *eLife*, 10:e64943, 2021.
- [7] S. Soleymani, N. Gravel, L. C. Huang, W. Yeung, E. Bozorgi, N. G. Bendzunas, P. Saeedi, K. Yin, P. Moreno, K. S. Pawelczak, K. Trachana, T. Barber, U. Farooq, A. Nasir, G. L. Johnson, A. Goldberg, and N. Kannan. Dark kinase annotation, mining, and visualization using the protein kinase ontology. *PeerJ*, 11:e16087, 2023.
- [8] J. R. Johnson, S. D. Santos, T. Johnson, U. Pieper, M. Strumillo, O. Wagih, and P. Beltrao. Prediction of functionally important phospho-regulatory events in *Xenopus laevis* oocytes. *PLoS Computational Biology*, 11(8):e1004362, 2015.
- [9] P. Cohen. The origins of protein phosphorylation. *Nature Cell Biology*, 4(5):E127–E130, 2002.
- [10] L. J. Wilson, A. Linley, D. E. Hammond, F. E. Hood, J. M. Coulson, D. J. MacEwan, and I. A. Prior. New perspectives, opportunities, and challenges in exploring the human protein kinome. *Cancer Research*, 78(1):15–29, 2018.
- [11] T. Hunter. Signaling—2000 and beyond. *Cell*, 100(1):113–127, 2000.

- [12] K. S. Metz, E. M. Deoudes, M. E. Berginski, I. Jimenez-Ruiz, B. A. Aksoy, J. Hammerbacher, and D. H. Phanstiel. Coral: clear and customizable visualization of human kinome data. *Cell Systems*, 7(3):347–350, 2018.
- [13] S. K. Hanks and T. Hunter. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *The FASEB Journal*, 9(8):576–596, 1995.
- [14] P. Creixell, A. Palmeri, C. J. Miller, H. J. Lou, C. C. Santini, L. N. Chan, and R. Linding. Hierarchical organization endows the kinase domain with regulatory plasticity. *Cell Reports*, 25(5):1300–1314, 2018.
- [15] L. M. Iyer, L. Aravind, S. L. Coon, D. C. Klein, and E. V. Koonin. Evolution of cell–cell signaling in animals: did late horizontal gene transfer from bacteria have a role? *Trends in Genetics*, 20(7):292–299, 2004.
- [16] Z. A. Knight, H. Lin, and K. M. Shokat. Targeting the cancer kinome through polypharmacology. *Nature Reviews Cancer*, 10(2):130–137, 2010.
- [17] N. Sugiyama, H. Imamura, and Y. Ishihama. Large-scale discovery of substrates of the human kinome. *Scientific Reports*, 9(1):10503, 2019.
- [18] S. S. Taylor and A. P. Kornev. Protein kinases: evolution of dynamic regulatory proteins. *Trends in Biochemical Sciences*, 36(2):65–77, 2011.
- [19] V. Modi and R. L. Jr Dunbrack. Defining a new nomenclature for the structures of active and inactive kinases. *Proceedings of the National Academy of Sciences*, 116(14):6818–6827, 2019.
- [20] S. S. Taylor, M. M. Keshwani, J. M. Steichen, and A. P. Kornev. Evolution of the eukaryotic protein kinases as dynamic molecular switches. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1602):2517–2528, 2012.
- [21] A. P. Kornev, N. M. Haste, S. S. Taylor, and L. F. Ten Eyck. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proceedings of the National Academy of Sciences*, 103(47):17783–17788, 2006.
- [22] J. C. Burton, F. Royer, and N. J. Grimsey. Spatiotemporal control of kinases and the biomolecular tools to trace activity. *Journal of Biological Chemistry*, 300(11):107846, 2024.
- [23] M. Huber and T. Brummer. Enzyme is the name—adapter is the game. *Cells*, 13(15):1249, 2024.

- [24] C. J. DiRusso, M. Dashtiahanger, and T. D. Gilmore. Scaffold proteins as dynamic integrators of biological processes. *Journal of Biological Chemistry*, 298(12):102628, 2022.
- [25] J. Zhang, P. L. Yang, and N. S. Gray. Targeting cancer with small molecule kinase inhibitors. *Nature Reviews Cancer*, 9(1):28–39, 2009.
- [26] S. Baltschukat, B. S. Engstler, A. Huang, H. X. Hao, et al. Capmatinib (inc280) is active against models of non-small cell lung cancer and other cancer types with defined mechanisms of met activation. *Clinical Cancer Research*, 25(10):3164–3175, 2019.
- [27] R. K. Murthy, S. Loi, A. Okines, E. Paplomata, et al. Tucatinib, trastuzumab, and capecitabine for HER2-positive metastatic breast cancer. *New England Journal of Medicine*, 382(7):597–609, 2020.
- [28] J. J. O’Shea, D. M. Schwartz, A. V. Villarino, M. Gadina, et al. The jak-stat pathway: impact on human disease and therapeutic intervention. *Annual Review of Medicine*, 66(1):311–328, 2015.
- [29] C. Harrison, J. J. Kiladjian, H. K. Al-Ali, H. Gisslinger, et al. Jak inhibition with ruxolitinib versus best available therapy for myelofibrosis. *New England Journal of Medicine*, 366(9):787–798, 2012.
- [30] M. L. Wynn, A. C. Ventura, J. A. Sepulchre, H. J. García, et al. Kinase inhibitors can produce off-target effects and activate linked pathways by retroactivity. *BMC Systems Biology*, 5(1):156, 2011.
- [31] M. E. Noble, J. A. Endicott, and L. N. Johnson. Protein kinase inhibitors: insights into drug design from structure. *Science*, 303(5665):1800–1805, 2004.
- [32] C. Arter, L. Trask, S. Ward, S. Yeoh, and R. Bayliss. Structural features of the protein kinase domain and targeted binding by small-molecule inhibitors. *Journal of Biological Chemistry*, 298(8):102050, 2022.
- [33] S. Chakraborty, T. Inukai, L. Fang, M. Golkowski, and D. J. Maly. Targeting dynamic atp-binding site features allows discrimination between highly homologous protein kinases. *ACS Chemical Biology*, 14(6):1249–1259, 2019.
- [34] M. Zhang, Y. Liu, H. Jang, and R. Nussinov. Strategy toward kinase-selective drug discovery. *Journal of Chemical Theory and Computation*, 19(5):1615–1628, 2023.
- [35] J. Li, C. Gong, H. Zhou, J. Liu, X. Xia, W. Ha, et al. Kinase inhibitors and kinase-targeted cancer therapies: recent advances and future perspectives. *International Journal of Molecular Sciences*, 25(10):5489, 2024.

- [36] G. L. Gonzalez-Del Pino, K. Li, E. Park, A. M. Schmoker, B. H. Ha, and M. J. Eck. Allosteric MEK inhibitors act on BRAF/MEK complexes to block MEK activation. *Proceedings of the National Academy of Sciences*, 118(36):e2107207118, 2021.
- [37] J. Trepel, M. Mollapour, G. Giaccone, and L. Neckers. Targeting the dynamic HSP90 complex in cancer. *Nature Reviews Cancer*, 10(8):537–549, 2010.
- [38] A. M. Gross, B. Glassberg, P. L. Wolters, E. Dombi, et al. Selumetinib in children with neurofibromatosis type 1 and asymptomatic inoperable plexiform neurofibroma at risk for developing tumor-related morbidity. *Neuro-Oncology*, 24(11):1978–1988, 2022.
- [39] M. Rask-Andersen, J. Zhang, D. Fabbro, and H. B. Schiöth. Advances in kinase targeting: current clinical use and clinical trials. *Trends in Pharmacological Sciences*, 35(11):604–620, 2014.
- [40] U.S. Food and Drug Administration. FDA approves selumetinib for adults with neurofibromatosis type 1 with symptomatic, inoperable plexiform neurofibromas. Available online. Content current as of: November 19, 2025.
- [41] M. A. Lemmon and J. Schlessinger. Cell signaling by receptor tyrosine kinases. *Cell*, 141(7):1117–1134, 2010.
- [42] D. R. Robinson, Y. M. Wu, and S. F. Lin. The protein tyrosine kinase family of the human genome. *Oncogene*, 19(49):5548–5557, 2000.
- [43] M. K. Paul and A. K. Mukhopadhyay. Tyrosine kinase—role and significance in cancer. *International Journal of Medical Sciences*, 1(2):101, 2004.
- [44] I. Amata, M. Maffei, and M. Pons. Phosphorylation of unique domains of src family kinases. *Frontiers in Genetics*, 5:181, 2014.
- [45] C. L. Abram and S. A. Courtneidge. Src family tyrosine kinases and growth factor signaling. *Experimental Cell Research*, 254(1):1–13, 2000.
- [46] A. M. Eshaq, T. W. Flanagan, S. Y. Hassan, S. A. Al Asheikh, et al. Non-receptor tyrosine kinases: their structure and mechanistic role in tumor progression and resistance. *Cancers*, 16(15):2754, 2024.
- [47] M. A. Ortiz, T. Mikhailova, X. Li, B. A. Porter, A. Bah, and L. Kotula. Src family kinases, adaptor proteins and the actin cytoskeleton in epithelial-to-mesenchymal transition. *Cell Communication and Signaling*, 19(1):67, 2021.
- [48] E. L’Estrange-Stranieri, T. A. Gottschalk, M. D. Wright, and M. L. Hibbs. The dualistic role of lyn tyrosine kinase in immune cell signaling: implications for systemic lupus erythematosus. *Frontiers in Immunology*, 15:1395427, 2024.

- [49] E. Ingley. Functions of the lyn tyrosine kinase in health and disease. *Cell Communication and Signaling*, 10(1):21, 2012.
- [50] C. A. Lowell. Src-family kinases: rheostats of immune cell signaling. *Molecular Immunology*, 41(6–7):631–643, 2004.
- [51] S. M. Thomas and J. S. Brugge. Cellular functions regulated by src family kinases. *Annual Review of Cell and Developmental Biology*, 13(1):513–609, 1997.
- [52] J. Meade, C. Fernandez, and M. Turner. The tyrosine kinase lyn is required for B cell development beyond the t1 stage in the spleen: rescue by over-expression of bcl-2. *European Journal of Immunology*, 32(4):1029–1034, 2002.
- [53] E. J. Brodie, S. Infantino, M. S. Low, and D. M. Tarlinton. Lyn, lupus, and (B) lymphocytes, a lesson on the critical balance of kinase signaling in immunity. *Frontiers in Immunology*, 9:401, 2018.
- [54] P. Scapini, S. Pereira, H. Zhang, and C. A. Lowell. Multiple roles of lyn kinase in myeloid cell signaling and function. *Immunological Reviews*, 228(1):23–40, 2009.
- [55] W. Lu, K. M. Skrzypczynska, and A. Weiss. Acute csk inhibition hinders B cell activation by constraining the PI3 kinase pathway. *Proceedings of the National Academy of Sciences*, 118(43):e2108957118, 2021.
- [56] B. F. Brian 4th and T. S. Freedman. The Src-family kinase lyn in immunoreceptor signaling. *Endocrinology*, 162(10):bqab152, 2021.
- [57] N. Koutras, V. Morfos, K. Konnaris, A. Kouvela, et al. Integrated signaling and transcriptome analysis reveals src family kinase individualities and novel pathways controlled by their constitutive activity. *Frontiers in Immunology*, 14:1224520, 2023.
- [58] P. M. Weerawarna and T. I. Richardson. Lyn kinase structure, regulation, and involvement in neurodegenerative diseases: A mini review. *Kinases and Phosphatases*, 1(1):23–38, 2023.
- [59] Y. Xu, K. W. Harder, N. D. Huntington, M. L. Hibbs, and D. M. Tarlinton. Lyn tyrosine kinase: accentuating the positive and the negative. *Immunity*, 22(1):9–18, 2005.
- [60] S. Chakraborty, E. Ahler, J. J. Simon, L. Fang, et al. Profiling of drug resistance in Src kinase at scale uncovers a regulatory network coupling autoinhibition and catalytic domain dynamics. *Cell Chemical Biology*, 31(2):207–220, 2024.
- [61] J. Pei and Q. Cong. Computational analysis of regulatory regions in human protein kinases. *Protein Science*, 32(10):e4764, 2023.

- [62] J. A. Moroco, J. K. Craigo, R. E. Iacob, T. E. Wales, et al. Differential sensitivity of Src-family kinases to activation by SH3 domain displacement. *PloS One*, 9(8):e105629, 2014.
- [63] J. J. Alvarado, L. Betts, J. A. Moroco, T. E. Smithgall, and J. I. Yeh. Crystal structure of the Src family kinase Hck SH3-SH2 linker regulatory region supports an SH3-dominant activation mechanism. *Journal of Biological Chemistry*, 285(46):35455–35461, 2010.
- [64] The UniProt Consortium. UniProtKB – P07948 (LYN) – human, 2024. <https://www.uniprot.org/uniprotkb/P07948/entry>.
- [65] A. Contri, A. M. Brunati, L. Trentin, A. Cabrelle, et al. Chronic lymphocytic leukemia B cells contain anomalous Lyn tyrosine kinase, a putative contribution to defective apoptosis. *The Journal of Clinical Investigation*, 115(2):369–378, 2005.
- [66] F. Zonta, M. A. Pagano, L. Trentin, E. Tibaldi, et al. Lyn sustains oncogenic signaling in chronic lymphocytic leukemia by strengthening SET-mediated inhibition of PP2A. *Blood*, 125(24):3747–3756, 2015.
- [67] P. H. Nguyen, O. Fedorchenko, N. Rosen, M. Koch, et al. LYN kinase in the tumor microenvironment is essential for the progression of chronic lymphocytic leukemia. *Cancer Cell*, 30(4):610–622, 2016.
- [68] A. F. Vom Stein, R. Rebollido-Rios, A. Lukas, M. Koch, et al. LYN kinase programs stromal fibroblasts to facilitate leukemic survival via regulation of c-JUN and THBS1. *Nature Communications*, 14(1):1330, 2023.
- [69] J. L. Chen, P. Y. Chu, C. T. Huang, W. L. Wang, et al. Targeting SET to counteract LYN activation demonstrates anti-cancer potential in diffuse large B-cell lymphoma. *Clinical and Experimental Medicine*, 25(1):345, 2025.
- [70] P. L. McCormack and S. J. Keam. Dasatinib: a review of its use in the treatment of chronic myeloid leukaemia and philadelphia chromosome-positive acute lymphoblastic leukaemia. *Drugs*, 71(13):1771–1795, 2011.
- [71] P. Negi, R. S. Cheke, and V. M. Patil. Recent advances in pharmacological diversification of Src family kinase inhibitors. *Egyptian Journal of Medical Human Genetics*, 22(1):52, 2021.
- [72] N. K. Williams, I. S. Lucet, S. P. Klinken, E. Ingley, and J. Rossjohn. Crystal structures of the Lyn protein tyrosine kinase domain in its apo- and inhibitor-bound state. *Journal of Biological Chemistry*, 284(1):284–291, 2009.

- [73] K. M. Choi, E. Cho, G. Bang, S. J. Lee, et al. Activity-based protein profiling reveals potential dasatinib targets in gastric cancer. *International Journal of Molecular Sciences*, 21(23):9276, 2020.
- [74] R. Mshaik, J. Simonet, A. Georgievski, L. Jamal, and others. HSP90 inhibitor NVP-BEP800 affects stability of SRC kinases and growth of T-cell and B-cell acute lymphoblastic leukemias. *Blood Cancer Journal*, 11(3):61, 2021.
- [75] R. Kohal, Bhavana, P. Kumari, A. K. Sharma, G. D. Gupta, et al. Fyn, Blk, and Lyn kinase inhibitors: a mini-review on medicinal attributes, research progress, and future insights. *Bioorganic & Medicinal Chemistry Letters*, 102:129674, 2024.
- [76] J. P. Changeux and S. J. Edelstein. Allosteric mechanisms of signal transduction. *Science*, 308(5727):1424–1428, 2005.
- [77] R. Nussinov and C. J. Tsai. Allostery in disease and in drug discovery. *Cell*, 153(2):293–305, 2013.
- [78] S. S. Yadav and W. T. Miller. Cooperative activation of Src family kinases by SH3 and SH2 ligands. *Cancer Letters*, 257(1):116–123, 2007.
- [79] N. R. Gough and C. G. Kalodimos. Exploring the conformational landscape of protein kinases. *Current Opinion in Structural Biology*, 88:102890, 2024.
- [80] V. R. Mingione, Y. Paung, I. R. Outhwaite, and M. A. Seeliger. Allosteric regulation and inhibition of protein kinases. *Biochemical Society Transactions*, 51(1):373–385, 2023.
- [81] Z. M. Khan, A. M. Real, W. M. Marsiglia, A. Chow, et al. Structural basis for the action of the drug trametinib at KSR-bound MEK. *Nature*, 588(7838):509–514, 2020.
- [82] A. A. Wylie, J. Schoepfer, W. Jahnke, S. W. Cowan-Jacob, et al. The allosteric inhibitor ABL001 enables dual targeting of BCR–ABL1. *Nature*, 543(7647):733–737, 2017.
- [83] S. Bowerman and J. Wereszczynski. Detecting allosteric networks using molecular dynamics simulation. 578:429–447, 2016.
- [84] C. L. McClendon, G. Friedland, and M. P. Jacobson. Quantifying correlations between allosteric sites in thermodynamic ensembles. *Biophysical Journal*, 98(3):385a–386a, 2010.
- [85] F. Noé, G. De Fabritiis, and C. Clementi. Machine learning for protein folding and dynamics. *Current Opinion in Structural Biology*, 60:77–84, 2020.

- [86] M. W. Iqbal, M. Shahab, Z. Ullah, G. Zheng, et al. Integrating machine learning and structure-based approaches for repurposing potent tyrosine protein kinase Src inhibitors to treat inflammatory disorders. *Scientific Reports*, 15(1):1836, 2025.
- [87] V. R. Mingione, Y. Paung, I. R. Outhwaite, and M. A. Seeliger. Validation of an allosteric binding site of Src kinase identified by unbiased ligand binding simulations. *Journal of Molecular Biology*, 434(19):167763, 2022.
- [88] V. Gapsys, W. Kopec, D. Matthes, and B. L. de Groot. Biomolecular simulations at the exascale: from drug design to organelles and beyond. *Current Opinion in Structural Biology*, 88:102887, 2024.
- [89] O. T. Unke, M. Stöhr, S. Ganscha, T. Unterthiner, et al. Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments. *Science Advances*, 10(14):eadn4397, 2024.
- [90] Q. Cui, T. Pal, and L. Xie. Biomolecular QM/MM simulations: What are some of the “burning issues”? *The Journal of Physical Chemistry B*, 125(3):689–702, 2021.
- [91] S. K. Niazi. Quantum mechanics in drug discovery: a comprehensive review of methods, applications, and future directions. *International Journal of Molecular Sciences*, 26(13):6325, 2025.
- [92] A. O. Dohn. Multiscale electrostatic embedding simulations for modeling structure and dynamics of molecules in solution: a tutorial review. *International Journal of Quantum Chemistry*, 120(21):e26343, 2020.
- [93] D. F. Hahn, V. Gapsys, B. L. de Groot, D. L. Mobley, and G. Tresadern. Current state of open source force fields in protein–ligand binding affinity predictions. *Journal of Chemical Information and Modeling*, 64(13):5063–5076, 2024.
- [94] J. Huang and A. D. MacKerell. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *Journal of Computational Chemistry*, 34(25):2135–2145, 2013.
- [95] P. S. Nerenberg and T. Head-Gordon. New developments in force fields for biomolecular simulations. *Current Opinion in Structural Biology*, 49:129–138, 2018.
- [96] A. C. Beatrice. Advancements and future directions in molecular dynamics (MD) simulations. *IDOSR Journal of Applied Sciences*, 9(1):21–26, 2024.
- [97] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958, 2010.



- [98] M. Li. Molecular dynamics simulations: a systematic review of techniques and applications in biochemistry. *Computational Molecular Biology*, 14, 2024.
- [99] A. Romero-Rivera, M. Garcia-Borràs, and S. Osuna. Computational tools for the evaluation of laboratory-engineered biocatalysts. *Chemical Communications*, 53(2):284–297, 2017.
- [100] P. Bauer, B. Hess, and E. Lindahl. GROMACS 2022 manual. Zenodo, 2022. <https://doi.org/10.5281/zenodo.6103568>.
- [101] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, et al. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, 2005.
- [102] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.
- [103] L. D. Schuler, X. Daura, and W. F. Van Gunsteren. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *Journal of Computational Chemistry*, 22(11):1205–1218, 2001.
- [104] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, et al. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone  $\phi$ ,  $\psi$  and side-chain  $\chi_1$  and  $\chi_2$  dihedral angles. *Journal of Chemical Theory and Computation*, 8(9):3257–3273, 2012.
- [105] P. Li, B. P. Roberts, D. K. Chakravorty, and K. M. Merz. Rational design of particle mesh Ewald compatible Lennard-Jones parameters for +2 metal cations in explicit solvent. *Journal of Chemical Theory and Computation*, 9(6):2733–2748, 2013.
- [106] J. Yoo and A. Aksimentiev. Improved parametrization of  $\text{Li}^+$ ,  $\text{Na}^+$ ,  $\text{K}^+$ , and  $\text{Mg}^{2+}$  ions for all-atom molecular dynamics simulations of nucleic acid systems. *The Journal of Physical Chemistry Letters*, 3(1):45–50, 2012.
- [107] O. Allnér, L. Nilsson, and A. Villa. Magnesium ion–water coordination and exchange in biomolecular simulations. *Journal of Chemical Theory and Computation*, 8(4):1493–1502, 2012.
- [108] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, et al. A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, 103(19):8577–8593, 1995.
- [109] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, et al. OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology*, 13(7):e1005659, 2017.

- [110] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008.
- [111] P. Mark and L. Nilsson. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *The Journal of Physical Chemistry A*, 105(43):9954–9960, 2001.
- [112] Y. Wu, H. L. Tepper, and G. A. Voth. Flexible simple point-charge water model with improved liquid-state properties. *The Journal of Chemical Physics*, 124(2), 2006.
- [113] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *The Journal of Physical Chemistry B*, 119(16):5113–5123, 2015.
- [114] J. L. Abascal and C. Vega. A general purpose model for the condensed phases of water: TIP4P/2005. *The Journal of Chemical Physics*, 123(23), 2005.
- [115] M. W. Mahoney and W. L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potentials. *The Journal of Chemical Physics*, 112(20):8910–8922, 2000.
- [116] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.
- [117] B. Hess, H. Bekker, H. J. Berendsen, and J. G. Fraaije. LINCS: a linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.
- [118] V. Kräutler, W. F. Van Gunsteren, and P. H. Hünenberger. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of Computational Chemistry*, 22(5):501–508, 2001.
- [119] J. A. Lemkul. Introductory tutorials for simulating protein dynamics with GRO-MACS. *The Journal of Physical Chemistry B*, 128(39):9418–9435, 2024.
- [120] GROMACS Development Team. Molecular dynamics – GROMACS 2025.4 reference manual, 2025. <https://manual.gromacs.org/documentation/current/reference-manual/algorithms/molecular-dynamics.html>.
- [121] GROMACS Tutorial Authors. Position-restrained equilibration – GROMACS tutorial, 2023. <https://tutorials.gromacs.org/>.

- [122] R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, 2021.
- [123] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2025. R package version 2.1.8.1.
- [124] C. Hennig. *fpc: Flexible Procedures for Clustering*, 2024. R package version 2.2-13.
- [125] R. Rebollido-Rios and O. Kramer. Sequence-based protein pocket prediction via Prott5 embeddings and spatial sampling. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 1312–1315, 2025.
- [126] M. AlQuraishi. Machine learning in protein structure prediction. *Current Opinion in Chemical Biology*, 65:1–8, 2021.
- [127] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.
- [128] F. Noé, A. Tkatchenko, K. R. Müller, and C. Clementi. Machine learning for molecular simulation. *Annual Review of Physical Chemistry*, 71(1):361–390, 2020.
- [129] Y. Wang, J. M. L. Ribeiro, and P. Tiwary. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Current Opinion in Structural Biology*, 61:139–145, 2020.
- [130] P. Campitelli, T. Modi, S. Kumar, and S. B. Ozkan. The role of conformational dynamics and allostery in modulating protein evolution. *Annual Review of Biophysics*, 49(1):267–288, 2020.
- [131] O. Schueler-Furman and S. J. Wodak. Computational approaches to investigating allostery. *Current Opinion in Structural Biology*, 41:159–171, 2016.
- [132] D. Medina-Ortiz, S. Contreras, C. Quiroz, and Á. Olivera-Nappa. Development of supervised learning predictive models for highly non-linear biological, biomedical, and general datasets. *Frontiers in Molecular Biosciences*, 7:13, 2020.
- [133] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio. Un-supervised learning methods for molecular simulation data. *Chemical Reviews*, 121(16):9722–9758, 2021.
- [134] C. Wang, Y. Chen, Y. Zhang, K. Li, et al. A reinforcement learning approach for protein–ligand binding pose prediction. *BMC Bioinformatics*, 23(1):368, 2022.
- [135] C. Cortes and V. Vapnik. Support–vector networks. *Machine Learning*, 20(3):273–297, 1995.

- [136] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.
- [137] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [138] J. Hu and S. Szymczak. A review on longitudinal data analysis with random forest. *Briefings in Bioinformatics*, 24(2):bbad002, 2023.
- [139] Y. Kim, B. A. Kwon, and K. Kang. Assessing the impact of data impurity measures on decision tree algorithms in spam classification. *Journal of Artificial Intelligence Research and Applications*, 1(2):33–46, 2024.
- [140] G. Louppe, L. Wehenkel, A. Suter, and P. Geurts. Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems*, volume 26, 2013.
- [141] O. Rainio, J. Teuvo, and R. Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086, 2024.
- [142] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
- [143] P. Thölke, Y. J. Mantilla-Ramos, H. Abdelhedi, C. Maschke, et al. Class imbalance should not throw you off balance: choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, 277:120253, 2023.
- [144] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.
- [145] B. Nolen, S. Taylor, and G. Ghosh. Regulation of protein kinases: controlling activity through activation segment conformation. *Molecular Cell*, 15(5):661–675, 2004.
- [146] N. Amatya, D. Y. W. Lin, and A. H. Andreotti. Dynamic regulatory features of the protein tyrosine kinases. *Biochemical Society Transactions*, 47(4):1101–1116, 2019.
- [147] J. A. Endicott, M. E. Noble, and L. N. Johnson. The structural basis for control of eukaryotic protein kinases. *Annual Review of Biochemistry*, 81(1):587–613, 2012.
- [148] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.
- [149] G. P. Lisi and J. P. Loria. Allostery in enzyme catalysis. *Current Opinion in Structural Biology*, 47:123–130, 2017.

- [150] D. Shukla, Y. Meng, B. Roux, and V. S. Pande. Activation pathway of src kinase reveals intermediate states as targets for drug design. *Nature Communications*, 5:3397, 2014.
- [151] J. S. Tokarski, J. A. Newitt, C. Y. J. Chang, J. D. Cheng, et al. The structure of dasatinib (BMS-354825) bound to activated ABL kinase domain elucidates its inhibitory activity against imatinib-resistant ABL mutants. *Cancer Research*, 66(11):5790–5797, 2006.
- [152] L. L. Remsing Rix, U. Rix, J. Colinge, O. Hantschel, et al. Global target profile of the kinase inhibitor bosutinib in primary chronic myeloid leukemia cells. *Leukemia*, 23(3):477–485, 2009.
- [153] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, et al. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.
- [154] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, et al. Cancer genome landscapes. *Science*, 339(6127):1546–1558, 2013.
- [155] L. Sutto and F. L. Gervasio. Effects of oncogenic mutations on the conformational free-energy landscape of EGFR kinase. *Proceedings of the National Academy of Sciences of the USA*, 110(26):10616–10621, 2013.
- [156] B. Webb and A. Sali. Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics*, 54(1):5.6.1–5.6.37, 2016.
- [157] S. Berndt, V. V. Gurevich, and T. M. Iverson. Crystal structure of the SH3 domain of human lyn non-receptor tyrosine kinase. *PLoS ONE*, 14(4):e0215140, 2019.
- [158] L. L. Jin, L. E. Wybenga-Groot, J. Tong, P. Taylor, et al. Tyrosine phosphorylation of the Lyn Src homology 2 (SH2) domain modulates its binding affinity and specificity. *Molecular & Cellular Proteomics*, 14(3):695–706, 2015.
- [159] H. Yuki, K. Kikuzato, Y. Koda, J. Mikuni, et al. Activity cliff for 7-substituted pyrrolo-pyrimidine inhibitors of HCK explained in terms of predicted basicity of the amine nitrogen. *Bioorganic & Medicinal Chemistry*, 25(16):4259–4264, 2017.
- [160] C. Colovos and T. O. Yeates. Verification of protein structures: patterns of non-bonded atomic interactions. *Protein Science*, 2(9):1511–1519, 1993.
- [161] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269):pl1, 2013.

- [162] J. D. Stephenson, P. Totoo, D. F. Burke, J. Jänes, et al. Protvar: mapping and contextualizing human missense variation. *Nucleic Acids Research*, 52(W1):W140–W147, 2024.
- [163] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, et al. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.
- [164] J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 2019.
- [165] UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 46(5):2699–2699, 2018.
- [166] J. Jumper, R. Evans, A. Pritzel, T. Green, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.
- [167] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, et al. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, 2010.
- [168] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, et al. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894, 2019.
- [169] J. Cheng, G. Novati, J. Pan, C. Bycroft, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492, 2023.
- [170] A. Rives, J. Meier, T. Sercu, S. Goyal, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.
- [171] A. C. Carrera, K. Alexandrov, T. M. Roberts, et al. The conserved lysine of the catalytic domain of protein kinases is actively involved in the phosphotransfer reaction and not required for anchoring atp. *Proceedings of the National Academy of Sciences of the USA*, 90(2):442–446, 1993.
- [172] H. S. Meharena, X. Fan, L. G. Ahuja, M. M. Keshwani, et al. Decoding the interactions regulating the active state mechanics of eukaryotic protein kinases. *PLoS Biology*, 14(11):e2000127, 2016.
- [173] G. G. Krivov, M. V. Shapovalov, R. L. Dunbrack, et al. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics*, 77(4):778–795, 2009.

- [174] J. Zheng, E. A. Trafny, D. R. Knighton, N. H. Xuong, et al. 2.2 Å refined crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MnATP and a peptide inhibitor. *Acta Crystallographica Section D: Biological Crystallography*, 49(3):362–365, 1993.
- [175] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, et al. Gaussian 16. 2016.
- [176] A. V. Marenich, C. J. Cramer, and D. G. Truhlar. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *The Journal of Physical Chemistry B*, 113(18):6378–6396, 2009.
- [177] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1), 2007.
- [178] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.
- [179] R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, et al. MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations. (LA-UR-19-29136), 2019.
- [180] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.
- [181] Y. Cui, R. Ali, M. Clay, P. Rossi, et al. Conformational landscape adaptations enable processive phosphorylation by Src family kinases. *Science*, 390(6779):eadw8310, 2025.
- [182] Y. Meng, M. P. Pond, and B. Roux. Tyrosine kinase activation and conformational flexibility: lessons from Src-family tyrosine kinases. *Accounts of Chemical Research*, 50(5):1193–1201, 2017.
- [183] J. Gizzio, A. Thakur, A. Haldane, and R. M. Levy. Evolutionary divergence in the conformational landscapes of tyrosine vs serine/threonine kinases. *eLife*, 11:e83368, 2022.
- [184] R. Roskoski. Properties of FDA-approved small molecule protein kinase inhibitors: A 2025 update. *Pharmacological Research*, page 107723, 2025.
- [185] T. S. Ulmer, J. M. Werner, and I. D. Campbell. SH3-SH2 domain orientation in Src kinases: NMR studies of Fyn. *Structure*, 10(7):901–911, 2002.
- [186] M. A. Young, S. Gonfloni, G. Superti-Furga, B. Roux, and J. Kuriyan. Dynamic coupling between the SH2 and SH3 domains of c-src and Hck underlies their inactivation by c-terminal tyrosine phosphorylation. *Cell*, 105(1):115–126, 2001.

- [187] H. Machiyama, T. Yamaguchi, Y. Sawada, T. M. Watanabe, and H. Fujita. SH3 domain of c-Src governs its dynamics at focal adhesions and the cell membrane. *The FEBS Journal*, 282(20):4034–4055, 2015.
- [188] C. Greenman, P. Stephens, R. Smith, G. L. Dalglish, et al. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158, 2007.
- [189] C. C. David and D. J. Jacobs. Principal component analysis: a method for determining the essential dynamics of proteins. pages 193–226, 2013.
- [190] L. Q. Yang, P. Sang, Y. Tao, Y. X. Fu, et al. Protein dynamics and motions in relation to their functions: several case studies and the underlying mechanisms. *Journal of Biomolecular Structure and Dynamics*, 32(3):372–393, 2014.
- [191] J. Alba, M. Montagna, and M. D’Abramo. Modelling the activation pathways in full-length Src kinase. *Biophysica*, 1(2):238–248, 2021.
- [192] C. Daday, S. de Buhr, D. Mercadante, and F. Gräter. Mechanical force can enhance c-Src kinase activity by impairing autoinhibition. *Biophysical Journal*, 121(5):684–691, 2022.
- [193] M. A. Strelkova, A. P. Tolstova, V. A. Mitkevich, I. Y. Petrushanko, et al. Structure of full-length Src kinase and its key phosphorylated states: molecular dynamics study. *International Journal of Molecular Sciences*, 25(22):12391, 2024.
- [194] L. N. Johnson, M. E. Noble, and D. J. Owen. Active and inactive protein kinases: structural basis for regulation. *Cell*, 85(2):149–158, 1996.
- [195] V. Modi and R. L. Dunbrack. Kincore: a web resource for structural classification of protein kinases and their inhibitors. *Nucleic Acids Research*, 50(D1):D654–D664, 2022.
- [196] L. M. Wodicka, P. Ciceri, M. I. Davis, J. P. Hunt, et al. Activation state-dependent binding of small molecule kinase inhibitors. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1804(3):543–550, 2010.
- [197] W. Xu, S. C. Harrison, and M. J. Eck. Three-dimensional structure of the tyrosine kinase c-Src. *Nature*, 385(6617):595–602, 1997.
- [198] M. A. Seeliger, M. Young, M. N. Henderson, P. Pellicena, et al. High yield bacterial expression of active c-Abl and c-Src tyrosine kinases. *Protein Science*, 14(12):3135–3139, 2005.
- [199] N. M. Levinson, O. Kuchment, K. Shen, M. A. Young, et al. A Src-like inactive conformation in the Abl tyrosine kinase domain. *PLoS Biology*, 4(5):e144, 2006.



- [200] Z. Q. Bao, D. M. Jacobsen, and M. A. Young. Briefly bound to activate: transient binding of a second catalytic magnesium activates the structure and dynamics of CDK2 kinase for catalysis. *Structure*, 19(6):675–690, 2011.
- [201] C. Kanellopoulou, A. B. George, E. Masutani, J. L. Cannons, et al.  $Mg^{2+}$  regulation of kinase signaling and immune function. *Journal of Experimental Medicine*, 216(8):1828–1842, 2019.
- [202] A. Ogawa, Y. Takayama, H. Sakai, K. T. Chong, et al. Structure of the carboxyl-terminal Src kinase, Csk. *Journal of Biological Chemistry*, 277(17):14351–14354, 2002.
- [203] K. Nam, A. R. A. Thodika, S. Tischlik, C. Phoeurk, et al. Magnesium induced structural reorganization in the active site of adenylate kinase. *Science Advances*, 10(32):eado5504, 2024.
- [204] K. Mukherjee, M. Sharma, H. Urlaub, G. P. Bourenkov, et al. CASK functions as a  $Mg^{2+}$ -independent neurexin kinase. *Cell*, 133(2):328–339, 2008.
- [205] A. Matte, L. W. Tari, and L. T. Delbaere. How do kinases transfer phosphoryl groups? *Structure*, 6(4):413–419, 1998.
- [206] G. Henkelman, M. X. LaBute, C. S. Tung, P. W. Fenimore, et al. Conformational dependence of a protein kinase phosphate transfer reaction. *Proceedings of the National Academy of Sciences*, 102(43):15347–15351, 2005.
- [207] M. Huse and J. Kuriyan. The conformational plasticity of protein kinases. *Cell*, 109(3):275–282, 2002.
- [208] E. Pucheta-Martínez, G. Saladino, M. A. Morando, J. Martinez-Torrecuadrada, et al. An allosteric cross-talk between the activation loop and the ATP binding site regulates the activation of Src kinase. *Scientific Reports*, 6:24235, 2016.
- [209] D. M. Martins, P. O. Fernandes, L. A. Vieira, V. G. Maltarollo, et al. Structure-guided drug design targeting Abl kinase: how structure and regulation can assist in designing new drugs. *ChemBioChem*, 25(23):e202400296, 2024.
- [210] D. L. Mobley and K. A. Dill. Binding of small-molecule ligands to proteins: “what you see” is not always “what you get”. *Structure*, 17(4):489–498, 2009.
- [211] X. Liu, D. C. Speckhard, T. R. Shepherd, Y. J. Sun, et al. Distinct roles for conformational dynamics in protein-ligand interactions. *Structure*, 24(12):2053–2066, 2016.
- [212] J. Wu, N. A. Jonniya, S. P. Hirakis, C. Olivieri, et al. Role of the  $\alpha C$ – $\beta 4$  loop in protein kinase structure and dynamics. *Elife*, 12:RP91980, 2024.

- [213] H. N. Motlagh, J. O. Wrabl, J. Li, and V. J. Hilser. The ensemble nature of allostery. *Nature*, 508(7496):331–339, 2014.
- [214] L. G. Ahuja, P. C. Aoto, A. P. Kornev, G. Veglia, et al. Dynamic allostery-based molecular workings of kinase: peptide complexes. *Proceedings of the National Academy of Sciences*, 116(30):15052–15061, 2019.
- [215] Z. H. Foda, Y. Shan, E. T. Kim, D. E. Shaw, and M. A. Seeliger. A dynamically coupled allosteric network underlies binding cooperativity in Src kinase. *Nature Communications*, 6:5939, 2015.
- [216] R. Roskoski. Src protein-tyrosine kinase structure, mechanism, and small molecule inhibitors. *Pharmacological Research*, 94:9–25, 2015.
- [217] B. J. Grant, A. P. Rodrigues, K. M. ElSawy, J. A. McCammon, et al. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, 22(21):2695–2696, 2006.
- [218] V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, et al. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.
- [219] J. Vertrees. InterfaceResidues script for PyMOL, 2010. <https://pymolwiki.org/index.php/InterfaceResidues>.
- [220] Schrödinger, LLC. *The PyMOL Molecular Graphics System, Version 3.1.0*. Schrödinger, LLC, New York, NY, 2024. <https://pymol.org/>.
- [221] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, et al. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.
- [222] K. Katoh and D. M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.
- [223] J. A. Capra and M. Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882, 2007.
- [224] A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, et al. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 2009.
- [225] F. Carles, S. Bourg, C. Meyer, and P. Bonnet. PKIDB: a curated, annotated and updated database of protein kinase inhibitors in clinical trials. *Molecules*, 23(4):908, 2018.

- [226] H. M. Berman and S. K. Burley. Protein data bank (PDB): Fifty-three years young and having a transformative impact on science and society. *Quarterly Reviews of Biophysics*, 58:e9, 2025.
- [227] E. Papaleo, G. Saladino, M. Lambrughi, K. Lindorff-Larsen, et al. The role of protein loops and linkers in conformational dynamics and allostery. *Chemical Reviews*, 116(11):6391–6423, 2016.
- [228] Rebollido-Rios Lab. Dataset: Allosteric Connectors and Network Analysis of Full-length Lyn Kinase, 2025. <https://doi.org/10.5281/zenodo.16986533>.
- [229] E. Ingley. Src family kinases: regulation of their activities, levels and identification of new pathways. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1784(1):56–65, 2008.
- [230] B. J. Mayer. SH3 domains: complexity in moderation. *Journal of Cell Science*, 114(7):1253–1263, 2001.
- [231] F. Bauer, K. Schweimer, H. Meiselbach, S. Hoffmann, et al. Structural characterization of Lyn-SH3 domain in complex with a herpesviral protein reveals an extended recognition motif that enhances binding affinity. *Protein Science*, 14(10):2487–2498, 2005.
- [232] S. T. Arold, T. S. Ulmer, T. D. Mulhern, J. M. Werner, et al. The role of the SH3–SH2 interface in the regulation of Src kinases. *Journal of Biological Chemistry*, 276(20):17199–17205, 2001.
- [233] R. Roskoski. Classification of small molecule protein kinase inhibitors based upon the structures of their drug–enzyme complexes. *Pharmacological Research*, 103:26–48, 2016.
- [234] A. C. Dar, T. K. Das, K. M. Shokat, and R. L. Cagan. Chemical genetic discovery of targets and anti-targets for cancer polypharmacology. *Nature*, 486(7401):80–84, 2012.
- [235] O. Laufkötter, H. Hu, F. Miljković, and J. Bajorath. Structure- and similarity-based survey of allosteric kinase inhibitors, activators, and closely related compounds. *Journal of Medicinal Chemistry*, 65(2):922–934, 2022.
- [236] S. P. Bhujbal, J. Jun, H. Park, J. Moon, et al. Gaining insights into key structural hotspots within the allosteric binding pockets of protein kinases. *International Journal of Molecular Sciences*, 25(9):4725, 2024.
- [237] A. P. Kornev, S. S. Taylor, and L. F. Ten Eyck. A helix scaffold for the assembly of active protein kinases. *Proceedings of the National Academy of Sciences*, 105(38):14377–14382, 2008.

- [238] Y. Pan and M. M. Mader. Principles of kinase allosteric inhibition and pocket validation. *Journal of Medicinal Chemistry*, 65(7):5288–5299, 2022.
- [239] C. Kim, H. Ludewig, A. Hadzipasic, S. Kutter, et al. A biophysical framework for double-drugging kinases. *Proceedings of the National Academy of Sciences*, 120(34):e2304611120, 2023.
- [240] T. P. Hughes, M. J. Mauro, J. E. Cortes, H. Minami, et al. Asciminib in chronic myeloid leukemia after ABL kinase inhibitor failure. *New England Journal of Medicine*, 381(24):2315–2326, 2019.
- [241] E. J. Choi. Asciminib: the first-in-class allosteric inhibitor of BCR::ABL1 kinase. *Blood Research*, 58(S1):S29–S36, 2023.
- [242] C. J. Tsai and R. Nussinov. A unified view of “how allostery works”. *PLoS Computational Biology*, 10(2):e1003394, 2014.
- [243] K. Gunasekaran, B. Ma, and R. Nussinov. Is allostery an intrinsic property of all dynamic proteins? *Proteins: Structure, Function, and Bioinformatics*, 57(3):433–443, 2004.
- [244] S. Majumdar, F. Di Palma, F. Spyraakis, S. Decherchi, et al. Molecular dynamics and machine learning give insights on the flexibility-activity relationships in the tyrosine kinome. *Journal of Chemical Information and Modeling*, 63:4814–4826, 2023.
- [245] N. Lesgidou and M. Vlassi. Community analysis of large-scale molecular dynamics simulations elucidated dynamics-driven allostery in tyrosine kinase 2. *Proteins: Structure, Function, and Bioinformatics*, 92(4):474–498, 2024.
- [246] A. Dixit and G. M. Verkhivker. Computational modeling of allosteric communication reveals organizing principles of mutation-induced signaling in ABL and EGFR kinases. *PLoS Computational Biology*, 7(10):e1002179, 2011.
- [247] D. Pincus, J. P. Pandey, Z. A. Feder, P. Creixell, et al. Engineering allosteric regulation in protein kinases. *Science Signaling*, 11(555):eaar3250, 2018.
- [248] S. Betzi, R. Alam, M. Martin, S. R. Jakkaraj, et al. Discovery of a potential allosteric ligand binding site in CDK2. *ACS Chemical Biology*, 6(5):492–501, 2011.
- [249] L. Palmieri and G. Rastelli.  $\alpha$  helix displacement as a general approach for allosteric modulation of protein kinases. *Drug Discovery Today*, 18(7–8):407–414, 2013.
- [250] C. Olivieri, Y. Wang, C. Walker, M. V. Subrahmanian, et al. The  $\alpha$ C- $\beta$ 4 loop controls the allosteric cooperativity between nucleotide and substrate in the catalytic subunit of protein kinase A. *eLife*, 12:RP91506, 2024.

- [251] S. Sakthiah, G. P. Cao, S. P. Gupta, and K. W. Lee. Overview of the structure and function of protein kinases. *Current Enzyme Inhibition*, 13(2):81–88, 2017.
- [252] S. H. Schmidt, J. H. Weng, P. C. Aoto, D. Boassa, et al. Conformation and dynamics of the kinase domain drive subcellular location and activation of LRRK2. *Proceedings of the National Academy of Sciences*, 118(23):e2100844118, 2021.
- [253] M. Audagnotto, W. Czechtizky, L. De Maria, H. Käck, et al. Machine learning/molecular dynamic protein structure prediction approach to investigate the protein conformational ensemble. *Scientific Reports*, 12(1):10018, 2022.
- [254] C. L. McClendon, A. P. Kornev, M. K. Gilson, and S. S. Taylor. Dynamic architecture of a protein kinase. *Proceedings of the National Academy of Sciences*, 111(43):E4623–E4631, 2014.
- [255] S. Yang and B. Roux. Src kinase conformational activation: Thermodynamics, pathways, and mechanisms. *PLoS Computational Biology*, 4(3):e1000047, 2008.
- [256] M. Bernetti, S. Bosio, V. Bresciani, F. Falchi, et al. Probing allosteric communication with combined molecular dynamics simulations and network analysis. *Current Opinion in Structural Biology*, 86:1–10, 2024.
- [257] I. Reveguk and T. Simonson. Classifying protein kinase conformations with machine learning. *Protein Science*, 33(4):e4918, 2024.
- [258] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2:19–25, 2015.
- [259] M. Rabipour, F. Hassenrück, E. Pallaske, F. Röhrig, et al. Allosteric coupling in full-length Lyn kinase revealed by molecular dynamics and network analysis. *International Journal of Molecular Sciences*, 26(12):5835, 2025.
- [260] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [261] R. Reinhardt and T. A. Leonard. A critical evaluation of protein kinase regulation by activation loop autophosphorylation. *eLife*, 12:e88210, 2023.
- [262] M. Tong, J. G. Pelton, M. L. Gill, W. Zhang, et al. Survey of solution dynamics in Src kinase reveals allosteric cross talk between the ligand binding and regulatory sites. *Nature Communications*, 8(1):2160, 2017.
- [263] S. H. Chong, H. Oshima, and Y. Sugita. Allosteric changes in the conformational landscape of Src kinase upon substrate binding. *Journal of Molecular Biology*, 437(2):168871, 2025.

- [264] D. Prakaash, C. Fagnen, G. P. Cook, O. Acuto, et al. Molecular dynamics simulations reveal membrane lipid interactions of the full-length lymphocyte specific kinase (Lck). *Scientific Reports*, 12(1):21121, 2022.
- [265] M. P. Pond, R. Eells, B. W. Treece, F. Heinrich, et al. Membrane anchoring of Hck kinase via the intrinsically disordered SH4-U and length scale associated with subcellular localization. *Journal of Molecular Biology*, 432(9):2985–2997, 2020.
- [266] I. Moarefi, M. LaFevre-Bernt, F. Sicheri, M. Huse, et al. Activation of the Src-family tyrosine kinase Hck by SH3 domain displacement. *Nature*, 385(6617):650–653, 1997.
- [267] F. Sicheri, I. Moarefi, and J. Kuriyan. Crystal structure of the Src family tyrosine kinase Hck. *Nature*, 385(6617):602–609, 1997.
- [268] S. J. Parsons and J. T. Parsons. Src family kinases, key regulators of signal transduction. *Oncogene*, 23(48):7906–7909, 2004.
- [269] J. F. Apperley. Part i: Mechanisms of resistance to imatinib in chronic myeloid leukaemia. *The Lancet Oncology*, 8(11):1018–1029, 2007.
- [270] P. O. F. Martins, D. M. Martins, A. de Souza Bozzi, J. P. A. Martins, et al. Molecular insights on ABL kinase activation using tree-based machine learning models and molecular docking. *Molecular Diversity*, 25(3):1301–1314, 2021.
- [271] M. Shahab, G. Zheng, A. Khan, D. Wei, et al. Machine learning-based virtual screening and molecular simulation approaches identified novel potential inhibitors for cancer therapy. *Biomedicines*, 11(8):2251, 2023.
- [272] N. Jura, N. F. Endres, K. Engel, S. Deindl, et al. Mechanism for activation of the EGF receptor catalytic domain by the juxtamembrane segment. *Cell*, 137(7):1293–1307, 2009.
- [273] R. Choudhary, V. Walhekar, A. Muthal, D. Kumar, et al. Machine learning facilitated structural activity relationship approach for the discovery of novel inhibitors targeting EGFR. *Journal of Biomolecular Structure and Dynamics*, 41(22):12445–12463, 2023.
- [274] S. R. Hubbard. Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and atp analog. *The EMBO Journal*, 16(18):5572–5581, 1997.
- [275] F. S. Krebs, C. Britschgi, S. Pradervand, R. Achermann, et al. Structure-based prediction of BRAF mutation classes using machine-learning approaches. *Scientific Reports*, 12(1):12528, 2022.

- [276] L. Piazza, C. Poles, G. Bononi, C. Granchi, et al. Machine learning-based virtual screening for the identification of novel CDK-9 inhibitors. *Biomolecules*, 16(1):12, 2026.

# Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbstständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit -einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen -noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Dr. Rocio Rebolledo-Rios betreut worden.

## Übersicht der Publikationen:

Rabipour, M., Hassenrück, F., Pallaske, E., Röhrig, F., Hallek, M., Alvarez-Idaboy, J. R., Kramer, O., & Rebolledo-Rios, R. (2025). *Allosteric Coupling in Full-Length Lyn Kinase Revealed by Molecular Dynamics and Network Analysis*. *International Journal of Molecular Sciences*, 26(12), 5835. <https://doi.org/10.3390/ijms26125835>

Ich versichere, dass ich alle Angaben wahrheitsgemäß nach bestem Wissen und Gewissen gemacht habe und verpflichte mich, jedmögliche, die obigen Angaben betreffenden Veränderungen, dem Promotionsausschuss unverzüglich mitzuteilen.

Datum: 13. Juli 2026

Unterschrift: Mina Rabipour