

Spring 5-10-2017

COMPUTATIONAL INVESTIGATIONS OF BIOMOLECULAR MOTIONS AND INTERACTIONS IN GENOMIC MAINTENANCE AND REGULATION

Bradley R. Kossmann
Georgia State University

Follow this and additional works at: http://scholarworks.gsu.edu/chemistry_diss

Recommended Citation

Kossmann, Bradley R., "COMPUTATIONAL INVESTIGATIONS OF BIOMOLECULAR MOTIONS AND INTERACTIONS IN GENOMIC MAINTENANCE AND REGULATION." Dissertation, Georgia State University, 2017.
http://scholarworks.gsu.edu/chemistry_diss/129

This Dissertation is brought to you for free and open access by the Department of Chemistry at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Chemistry Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.

COMPUTATIONAL INVESTIGATIONS OF BIOMOLECULAR MOTIONS AND
INTERACTIONS IN GENOMIC MAINTENANCE AND REGULATION

by

Bradley R Kossmann

Under the Direction of Prof. Ivaylo Ivanov

ABSTRACT

The most critical biochemistry in an organism supports the central dogma of molecular biology: transcription of DNA to RNA and translation of RNA to peptide sequence. Proteins are then responsible for catalyzing, regulating and ensuring the fidelity of transcription and translation. At the heart of these processes lie selective biomolecular interactions and specific dynamics that are necessary for complex formation and catalytic activity. Through advanced biophysical and computational methods, it has become possible to probe these macromolecular dynamics and interactions at the molecular and atomic levels to tease out their underlying physical bases. To the end of a more thorough understanding of these physical bases, we have performed studies to probe the motions and interactions intrinsic to the function of biomolecular

complexes: modeling the dual-base flipping strategy of alkylpurine glycosylase D, dynamically tracing evolution and epistasis in the 3-ketosteroid family of nuclear receptors, discovering the allosteric and conformational aspects of transcription regulation in liver receptor homologue 1, leveraging specific contacts in tyrosyl-DNA phosphodiesterase 2 for the development of novel inhibitor scaffolds, and detailing the experimentally observed connection between solvation and sequence-specific binding affinity in PU.1-DNA complexes at the atomic level. While each study seeks to solve system-specific problems, the collection outlines a general and broadly applicable description of the biophysical motivations of biochemical processes.

INDEX WORDS: computational biophysics, computational chemistry, molecular dynamics, DNA repair, transcription regulation

COMPUTATIONAL INVESTIGATIONS OF BIOMOLECULAR MOTIONS AND
INTERACTIONS IN GENOMIC MAINTENANCE AND REGULATION

by

Bradley R Kossmann

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of

Philosophy

In the College of Arts and Sciences

Georgia State University

2017

Copyright by
Bradley R Kossmann
2017

COMPUTATIONAL INVESTIGATIONS OF BIOMOLECULAR MOTIONS AND
INTERACTIONS IN GENOMIC MAINTENANCE AND REGULATION

by

Bradley R Kossmann

Committee Chair: Iwaylo Ivanov

Committee: Donald Hamelberg

Giovanni Gadda

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

March 2017

ACKNOWLEDGEMENTS

My advisor, Prof. Ivaylo Ivanov, has served as a perfect mentor to me and has provided inspiration to me as a scientist. Throughout my graduate tenure I have had ample interesting projects and problems to work on and ample resources to work with. I have been afforded incredible flexibility and freedom to pursue my projects. Most importantly, Prof. Ivanov has spent countless hours of his valuable time patiently teaching and directing me. For all of this, he has earned my respect and gratitude.

I have also been blessed with fantastic collaborative opportunities and fantastic collaborators: Eric Ortlund and his lab (most notably Paul Musille and William Hudson), whose work is prominently featured in this dissertation; Christophe Marchand for his help and patience throughout the TDP2 project; and Gregory Poon for his conception of and enthusiasm throughout the ETS/PU.1 project.

Thank you, to all of the Ivanov Group members that I have had the pleasure of working with: Buddhaev Maiti and Chunli Yan for helping me develop as a researcher, and my fellow graduate students Kathleen Carter, Xiaojun (Max) Xu, Tom Dodd, Shih-Wei (Shawn) Chuo, Bernard Scott and Stephanie Kofsky-Wofford, for your support and camaraderie.

I would also like to thank the other members of my committee: Donald Hamelberg and Giovanni Gadda. Thank you for your time, wisdom and insights throughout graduate school and my dissertation preparation and defense. You have been phenomenal mentors and I appreciate your guidance.

I owe thanks to the Molecular Basis of Disease program at Georgia State for funding my stipend and providing intriguing and valuable seminars and conferences.

Finally, a major thanks to my family. Your support has been the most meaningful motivation for me throughout this and will certainly continue to be in all of my future endeavors.

Thank you.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
LIST OF TABLES	xi
LIST OF FIGURES.....	xii
CHAPTER 1. DNA REPAIR, GENE REGULATION AND SPECIFIC AIMS.....	1
1.1 Glycosylases and base excision repair	1
1.2 Topoisomerase-induced double strand break repair.....	3
1.3 Nuclear receptor mediated gene regulation.....	6
1.4 Specific aims.....	7
CHAPTER 2. METHODS.....	11
2.1 Molecular Dynamics	11
2.1.1 The Molecular Dynamics Forcefield.....	12
2.1.2 Force Integration and Trajectory Accumulation	13
2.1.3 Handling of Nonbonded Interactions	15
2.1.4 Solvent and the Simulation box.....	16
2.2 Partial Nudged Elastic Band	17
2.3 Umbrella Sampling	18
2.3.2 Weighted Histogram Analysis Method	19
2.4 Dynamical Network Analysis	20
2.4.1 Suboptimal Paths.....	21
2.4.2 Community Analysis	22
2.5 Molecular Mechanics-Poisson Boltzmann surface area.....	23
2.6 Grid Inhomogeneous Solvation Theory.....	25

CHAPTER 3. ALKYL PURINE GLYCOSYLASE D EMPLOYS DNA SCULPTING AS A STRATEGY TO EXTRUDE AND EXCISE DAMAGED BASES	27
3.1 Abstract	27
3.2 Author Summary	27
3.3 Introduction	28
3.4 Results and Discussion	30
3.4.1 Pathway and energetics of base pair opening and lesion extrusion by AlkD	30
3.4.2 Structural determinants for DNA bending, double base flipping and catalysis by AlkD....	35
3.5 Conclusions	41
3.6 Methods	41
3.6.1 Model Construction	41
3.6.2 Equilibration Protocol.....	42
3.6.3 Path Optimization	42
3.6.4 Umbrella Sampling Protocol.....	43
 CHAPTER 4. DISTAL SUBSTITUTIONS DRIVE DIVERGENT DNA SPECIFICITY AMONG PARALOGOUS TRANSCRIPTION FACTORS THROUGH SUBDIVISION OF CONFORMATIONAL SPACE	 45
4.1 Abstract	45
4.2 Significance Statement	46
4.3 Introduction	46
4.4 Results.....	48
4.4.1 DNA substrates dictate conformation of the GR DBD	48
4.4.2 GR is the only SR capable of binding nGREs	50
4.4.3 DBD-nGRE binding and subsequent repression is a feature of the ancestral 3-keto SR....	51

4.4.4 GR evolved enhanced inter-monomer allostery at nGREs	54
4.4.5 Subtle, irreversible structural changes enhanced nGRE binding.....	56
4.5 Discussion.....	60
4.6 Materials and Methods.....	62
4.6.1 Protein expression and purification	62
4.6.2 Protein – DNA binding assays.....	63
4.6.3 Crystallization and structure determination.....	63
4.6.4 Molecular dynamics simulations and network analysis	64
4.6.5 Cellular activation and repression assays	66
4.6.6 NMR.....	67
4.6.7 Phylogenetics and ancestral sequence reconstruction.....	67
 CHAPTER 5. UNEXPECTED ALLOSTERIC NETWORK CONTRIBUTES TO LRH-1	
COREGULATOR SPECIFICITY	69
5.1 Summary	69
5.2 Introduction.....	70
5.3 Experimental Procedures.....	72
5.3.1 Reagents	72
5.3.2 Protein expression and purification	72
5.3.3 Structure determination	72
5.3.4 Local conformational analysis.....	75
5.3.5 Synthesis of NBD-DLPE	75
5.3.6 Phospholipid binding assays.....	75
5.3.7 Reporter gene assays.....	76
5.3.8 Model construction for molecular dynamics.....	77
5.3.9 Molecular dynamics	78

5.3.10 Analysis methodology	79
5.4 Results.....	80
5.4.1 Structure of the apo LRH-1 LBD – TIF complex	80
5.4.2 Improved structure of the LRH-1 LBD – E. coli PL – TIF2 complex.....	82
5.4.3 Co-regulator binding interactions are altered by ligand status	88
5.4.4 Ligand and coregulator drive differential effects on local residue environment.....	91
5.4.5 The activated LRH-1 complex exhibits coordinated motions.....	94
5.4.6 MD simulations demonstrate communication between β -sheet-H6 and the AF-H through helices 3, 4, and 5.....	96
5.4.7 Structural and dynamical rationale for lipid and co-regulator agreement.....	99
5.4.8 Modest disruption of interhelical interactions along the allosteric pathway reduces, but does not eliminate, LRH-1 activity.....	103
5.5 Discussion.....	106
5.5.1 Lipid mediated allosteric control of a protein-protein binding interface	106
 CHAPTER 6. DISCOVERY OF SELECTIVE INHIBITORS OF TYROSYL-DNA	
PHOSPHODIESTERASE 2 BY TARGETING THE ENZYME DNA-BINDING CLEFT	
6.1 Abstract	110
6.2 Results.....	110
6.3 Methods	126
6.3.1 Computational methods	126
6.3.1.1 Molecular dynamics	126
6.3.1.2 Ligand optimization and parameterization.....	127
6.3.2 Experimental methods.....	127
6.3.2.1 Recombinant TDP2 assay	127
6.3.2.2 Whole cell extract TDP2 assay	128

6.3.2.3 Recombinant TDP1 assay	128
6.3.2.4 Kinetics experiments	128
CHAPTER 7. PU.1 AND ETS-1 SEQUENCE SPECIFICITY DIVERGENCE THROUGH DIFFERENTIAL ETS-DNA COMPLEX HYDRATION	130
7.1 Abstract	130
7.2 Introduction	131
7.3 Materials and Methods	133
7.3.1 Molecular dynamics setup and simulation	133
7.3.2 Molecular dynamics trajectory analysis.....	134
7.4 Results and Discussion	134
7.4.1 GIST solvation free energies correlate with experimental binding free energies	137
7.4.2 Network analysis reveals sequence-dependent shift in binding mode	141
7.4.3 Ordered solvent increases site-specific complex cohesiveness	143
7.5 Future work	145
CHAPTER 8. PERSPECTIVE	146
8.1 Biomolecular interactions	146
8.2 Motions in biochemical systems	146
8.3 Conclusion	147
REFERENCES	148

LIST OF TABLES

Table 5.1. Data collection and refinement statistics (molecular replacement) 74

Table 5.2 Modes chosen for PC1 and PC2 and the dot products between these modes. 101

Table 6.1. IC₅₀ values against human (*H. sapiens*, hTDP2), mouse (*M. musculus*, mTDP2) and zebrafish (*D. rerio*, zTDP2) TDP2 enzymes and against human TDP1 (hTDP1). 121

Table 7.1. Thermodynamic data for GIST analysis and experimental binding free energies. 136

LIST OF FIGURES

Figure 1.1.1 The base excision repair process.	2
Figure 1.2.1. Topoisomerase 2 relieves superhelical tension.	5
Figure 3.4.1.1. AlkD binding flattens the free energy landscape for lesion extrusion from DNA.	32
Figure 3.4.1.2. AlkD's sculpting of the DNA substrate results in three stable conformations along the flipping pathway. The three stable states in the AlkD/3 mA-DNA PMF. A) initial state; B) kinked intermediate state; and C) fully extruded (final) state. DNA bending is shown schematically in black.....	34
Figure 3.4.2.1. AlkD recognizes DNA through HEAT repeat motifs.	36
Figure 3.4.2.2. AlkD provides a scaffold to accommodate multiple DNA conformations with different degrees of bending.	38
Figure 3.4.2.3. Persistent hydrogen bonding contact observed between the 3 mA base and the phosphate in position -2 along the lesion strand.....	40
Figure 4.4.1.1 The glucocorticoid receptor (GR) DNA binding domain (DBD) adopts distinct conformations to activate and repress transcription.....	49
Figure 4.4.3.1 The GR lineage improved upon an ancestral cellular repressive function that was lost in MR.	52
Figure 4.4.4.1 Although nGRE binding orientation, stoichiometry, and sequence specificity originated at the common ancestor of all 3-keto SRs, GR is capable of enhanced DNA-mediated allosteric communication at nGREs.....	55

Figure 4.4.5.1 A single amino acid substitution far from the DNA binding interface – Ser425Gly – led to an improvement in nGRE binding through subtle effects in SR backbone conformation.	58
Figure 5.4.1.1 Structure of the apo LRH-1 LBD–TIF complex.	81
Figure 5.4.2.1 Structure of the LRH-1 LBD–E.coli PL–TIF2 complex.	83
Figure 5.4.2.2 LRH-1 in vitro lipid binding profile.	86
Figure 5.4.3.1 AF-2 charge clamp engagement is dictated by ligand-coregulator combination. .	90
Figure 5.4.4.1 ProSMART Procrustes central residue analysis of LRH-1 complexes.	93
Figure 5.4.5.1 Correlated motion in LRH-1–PL–coregulator systems.	95
Figure 5.4.6.1 Allosteric paths from binding pocket to co-regulator.	98
Figure 5.4.7.1 Biologically relevant principal modes identified from the projections of the MD trajectories on PC1 vs. PC2.	100
Figure 5.4.8.1 Subtle disruption of residues on or near the allosteric pathway reduces LRH-1 activation.	105
Figure 6.2.1. Flowchart overview of our TDP2 inhibitor discovery process	114
Figure 6.2.2. Inhibition of human recombinant TDP2 (hTDP2) and endogenous human TDP2 from whole cell extracts (hTDP2 WCE) by NSC379576, NSC114532, and NSC3198.	117
Figure 6.2.3. Inhibition of human (<i>H. sapiens</i> , hTDP2), mouse (<i>M. musculus</i> , mTDP2) and zebrafish (<i>D. rerio</i> , zTDP2) TDP2 enzymes by NSC379576, NSC114532, and NSC3198.	119
Figure 6.2.4. Binding poses and per-residue energy decomposition for residues important to binding.	123
Figure 7.4.1.1. GIST solvation energies vs. experimental binding free energies.	138
Figure 7.4.2.1. PU.1 networks for sequences.	140

Figure 7.4.2.2. Summed interface edge weights vs. experimental binding free energies.....	142
Figure 7.4.3.1. Overlay of networks and solvation isosurfaces.....	144

CHAPTER 1. DNA REPAIR, GENE REGULATION AND SPECIFIC AIMS

1.1 Glycosylases and base excision repair

Erroneous chemical modification of DNA by endogenous or exogenous factors is a frequent event, occurring up to 10,000 times per cell per day[1]. These lesions, if left unchecked, can have deleterious effects, causing mutation and strand breaks[1]. The most common type of damage occurs to single bases through, e.g. oxidation and alkylation. Single-nucleotide damage is repaired through the base excision repair (BER) pathway (Figure 1.1.1). In BER, a damage-specific glycosylase first identifies a damage site, rotates the damaged nucleotide out of the base stack and into an active site competent to hydrolyze the N-glycosidic linkage[2, 3]. Upon hydrolysis and product release, the newly formed apurinic/apyrimidinic (AP) site is acted upon by endonuclease, lyase, polymerase and ligase activities to restore the proper Watson-Crick (WC) nucleotide.

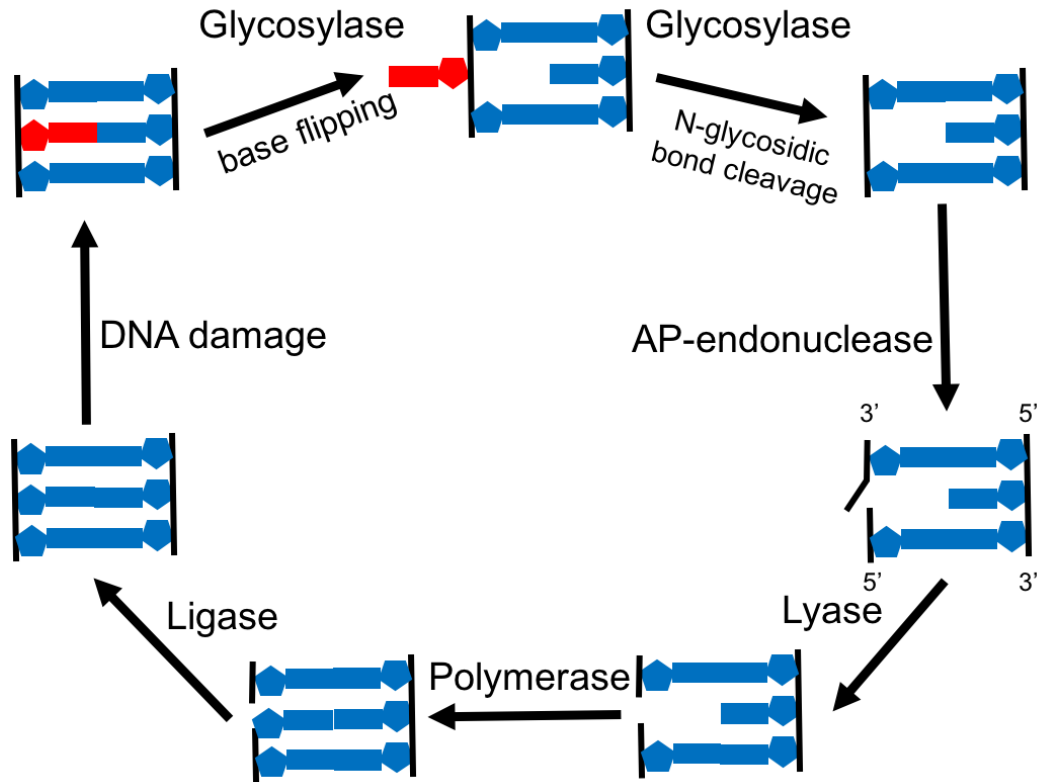


Figure 1.1.1 The base excision repair process.

In base excision repair, glycosylases first identify, extrude and excise a DNA lesion. AP-endonuclease, lyase, polymerase and ligase activity follow to restore the proper Watson-Crick sequence.

Glycosylases' fundamental role in BER, the identification and removal of damaged nucleobases, is generally carried out in three stages. The first stage is the search process, characterized by both one-dimensional sliding along DNA and three-dimensional hopping, with the glycosylase probing base pairs for thermodynamic instability indicative of nucleobase damage. In the second stage, the glycosylase rotates the damaged nucleoside out of the base stack in a process referred to as "base flipping." By exploiting lesion-specific properties to promote eversion of damaged nucleobases over canonical WC base pairs, glycosylases may employ base flipping as part of the lesion recognition process. After base flipping, the glycosylase hydrolyzes the nucleobase from the deoxyribose ring, leaving an AP site to be processed by the remaining BER machinery[4].

1.2 Topoisomerase-induced double strand break repair

The bulk of DNA in eukaryotes exists in tightly compacted nucleosome structures in the cellular nucleus[5]. Prior to replication or transcription, the DNA must be unpacked to be accessible to the replication and transcription-related enzymatic complexes. Due to its tight packing, nucleosomal DNA necessarily has much higher twist and writhe numbers relative to canonical, B-form DNA. To maintain that degree of geometric distortion, the topology of the DNA, referred to as the link number, must be altered. The mathematical relationship between twist (T), writhe (W) and link (L) are given by the very simple equation

$$T + W = L \quad (1.1)$$

Unpacking superhelical DNA for accessibility requires that the linking number be changed, so that the DNA can relax and expand from the highly compact twist and writhe nucleosomal state. DNA topology is altered by DNA topoisomerases [6, 7]. There are six topoisomerases encoded in the human genome, acting on a broad range of nucleic acid substrates

and are grouped into three families; Top1, Top2 and Top3, based on the mechanism of link number alteration[8].

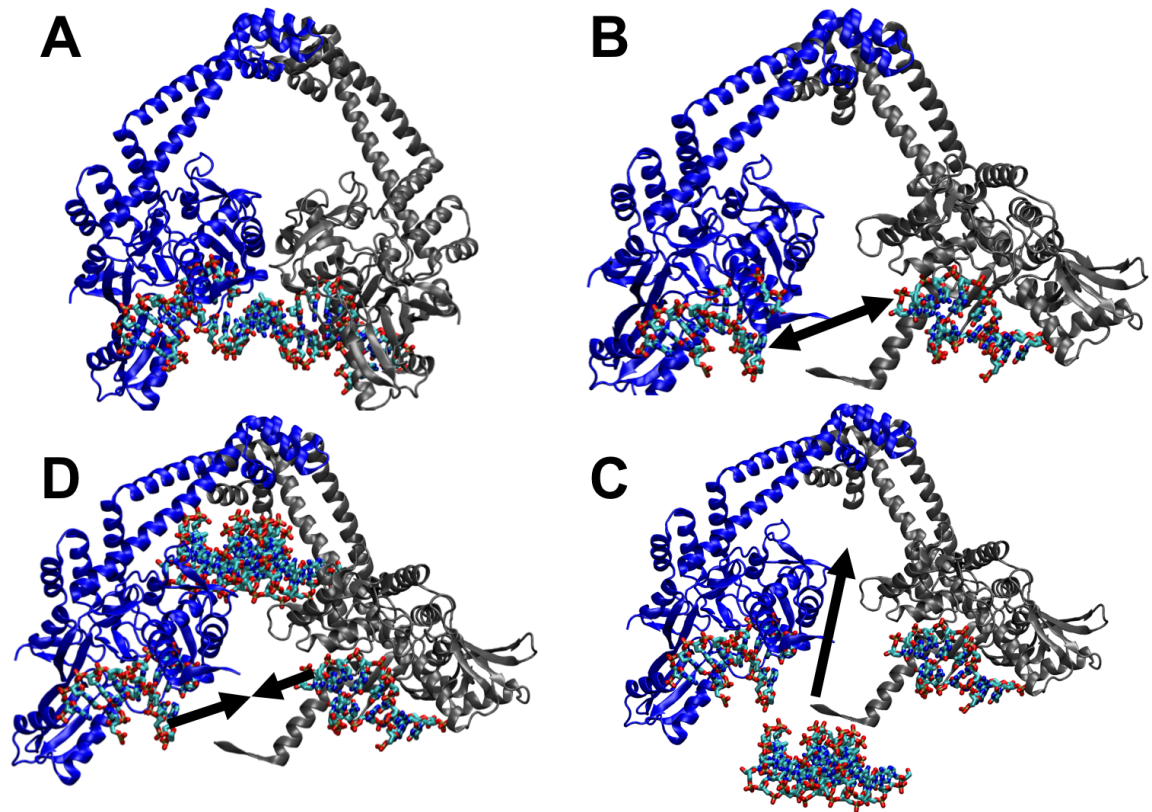


Figure 1.2.1. Topoisomerase 2 relieves superhelical tension.

Topo2 first cleaves duplex DNA (panel A-B) to allow for downstream DNA to pass through (C), before finally religating the strands (D).

Top2 first creates tyrosyl-DNA phosphodiester linkages and a double-strand break (dsb) in the DNA, allowing the the duplex to pass through the dsb to change the linking number (Figure 1.2.1). After this process is complete, Top2 religates the DNA. Although necessary, inducing double-strand breaks in DNA carries a high degree of risk, potentially resulting in cell death if the religation step of topoisomerase activity is not carried out to completion. Top2 can become incompetent to relieve its tyrosyl-phosphodiester linkages and must be catalytically removed from the DNA by tyrosyl-DNA phosphodiesterase 2 (TDP2)[9]. Double strand break repair enzymes can then religate the DNA duplex, preventing cell death.

Because topoisomerase activity is most prevalent during replication and tumor cells undergo division at a vastly higher rate than adult somatic cells, topoisomerase activity is an attractive chemotherapeutic target. A class of chemotherapeutics, known as Top2 poisons and comprised of popular pharmaceuticals such as etoposide, acts to trap Top2 to the substrate DNA, rendering the usually transient Top2-induced double-strand break permanent and therefore halting replication and tumor growth[10]. Unfortunately, TDP2 activity reduces the efficacy of Top2 poisons by reversing the Top2-DNA linkage. Therefore, TDP2 is a highly valuable target for inhibitor discovery efforts aimed at producing a compound that exhibits synergism with Top2 poisons.

1.3 Nuclear receptor mediated gene regulation

While epigenetic alterations serve to promote chromatin packing, a long-term gene regulation strategy, nuclear receptors (NRs) carry out the crucial task of signaling specific genes that require transcription in the short term[11, 12]. NRs generally consist of a ligand binding domain (LBD) and a DNA binding domain (DBD), connected by a long, disordered peptide loop. The DBD serves to anchor the NR to a gene-specific recognition sequence[13]. Signaling

molecules, e.g. hormones and metabolites, bind specific NR LBDs, signaling for the activation or repression of NR-specific genes. Once activation or repression has been signaled to the LBD, the LBD binds to a coregulator, triggering the coregulator to chemically modify either the histone, or other regulatory proteins in a signaling cascade to induce activation or repression[11, 14].

NR LBDs share highly homologous tertiary structure, typically consisting of a 3-layer alpha-helical sandwich[15]. LBDs generally possess two binding sites: firstly, small-molecule binding pockets that accommodate specific signaling molecules (e.g. hormones) and secondly, a coregulator peptide binding site capable of interacting with co-activators and co-repressors, depending on the signaling molecule. Whether the LBD binds a co-activator or co-repressor is dictated by whether the bound small molecule signals for activation or repression[16]. The physical mechanism through which this activation/repression preference is communicated has not been conclusively determined, although some form of allostery has been theorized to play a role[17].

DBDs are responsible for anchoring the NR to recognition sequences associated with specific genes[18]. Sequence specificity for an NR is of prime importance, but its basis is a very convoluted puzzle involving electrostatics, sequence-dependent conformational deformation energies, protein-DNA contacts and solvation energies. Further complicating matters, the members of some NR families (such as the ubiquitous glucocorticoid receptor family) are highly homologous, sharing a common evolutionary ancestor, yet exhibit very different sequence preferences and binding affinities[19].

1.4 Specific aims

The variety of biological systems and methodologies discussed in this dissertation is broad. However, the overarching theme can be summarized as the study of a subset of physical rationale

for the behavior of biomolecular systems: how dynamical description of how single biomolecules and small biomolecular complexes behave and analyses of the physical interactions between biomolecules. The specific aims of this research are to:

- 1) Study the biological process of base flipping at the atomic level, using AlkD as a case-study, tying the thermodynamics of the process to specific conformational events and contact rearrangements. A common enzymatic strategy is to greatly shift chemical equilibrium via small adjustments in the environment. AlkD is an example of this; while remaining an essentially rigid scaffold, AlkD lowers the energetic barrier to base flipping on the order of 10 kcal/mol by the strategic placement and adjustment of protein-DNA contacts.
- 2) Develop a descriptive model of molecular evolution wherein the impact of evolutionary mutations on the function of a protein can be observed using network theory and community analysis. Dynamical network analysis techniques are uniquely suited for mapping the minute concerted motions underlying a biomolecule's function. The DBDs of the 3-ketosteroid family of nuclear receptors maintain extremely high sequence identity, with relatively few and small mutations throughout evolution having altered each member's DNA binding affinity. Although the mutations are subtle, network analysis is able to uncover significant changes in the dynamical topology of 3KS DBDs, resulting in their altered behaviors.
- 3) Identify and trace an allosteric path using dynamical networks, through the LRH-1 LBD. Allostery is well understood in many systems as it pertains to enzymatic activity and biomolecular interactions. The exact atomic-scale phenomena that allostery is rooted in are less clear. In all examples of allostery, there must be a mechanism by which

interactions at one site on a biomolecule are transmitted to another site, whether by contact rearrangements or changes in the dynamics of the system. The LRH-1 LBD coordinates transcription by binding coregulator peptides in one site and exhibits coregulatory specificity depending on the identity of ligand bound in a spatially separate binding site. Dynamical networks are again employed to trace the subtle allosteric motions between the ligand-binding and coregulator-binding sites.

- 4) Identify inhibitor candidate scaffolds by exploiting contacts in an enzyme-substrate binding cleft. Small-molecule inhibitors frequently disrupt enzymatic activity by binding in an active or binding site to preclude activity. Because enzymes have evolved substrate-specific contacts to optimize substrate binding affinity, one successful inhibitor design strategy is to structurally mimic the substrate, taking advantage of the same contacts. TDP2 exhibits a ssDNA binding cleft with many potential target contacts. Beginning with a lead compound known to bind and inhibit TDP2, docking and virtual screening methodologies are used to discover new scaffolds that benefit from exploiting specific substrate-binding residues, and can be experimentally shown to inhibit TDP2 function.
- 5) Illustrate the utility of solvation and solvent-mediated contacts in biomolecular complex formation. While direct contacts between biomolecules in complex are more heavily studied, solvation and solvent-mediated contacts play a major role in modulating binding affinity, contributing in both enthalpy and entropy. The very structurally similar, but sequence-diverged ETS-family DBDs, Ets-1 and PU.1 have experimentally been shown to have very different DNA-binding affinities in different solvent environments, with PU.1-DNA binding affinities apparently relying on multiple water-mediated contacts, while Ets-1-DNA binding affinities do not. By monitoring solvent placement and

electrostatic environment, relative entropic and enthalpic contributions from solvent can be spatially mapped onto a biomolecular complex. Integration of these quantities with respect to volume yields direct thermodynamic evidence of the importance of solvation in ETS-DNA binding, and the energetic isosurfaces provide detailed qualitative information about the localization of water in the complex.

CHAPTER 2. METHODS

The power of modern computational resources has allowed for a proliferation of methods capable of simulating and analyzing large molecular systems over biologically relevant timescales. For biomolecular systems, simulation protocols are generally geared towards generating large ensembles of realistic conformations and take advantage of the inherent stochastic and ergodic nature of molecular systems. From these ensembles, statistical mechanical treatments yield thermodynamic and kinetic data while standard statistical techniques such as principle component analysis and graph analysis can yield important structural information and highlight important motions and correlations relevant to conformational transitions.

2.1 Molecular Dynamics

Molecular dynamics (MD) has become one of the most-utilized methods for molecular simulations in the literature[20]. While there are many variations of MD, the core underlying procedure is to calculate interatomic forces, integrate Newton's equations of motion with respect to a discrete timestep and finally update atomic coordinates. This sequence of events is iterated many times until a sufficiently large ensemble of structures or a sufficiently long time series is generated for analysis.

To initiate an MD simulation, atomic coordinates for the system of interest must be known. Sources for such structures include x-ray crystallography, NMR and homology modeling. Solvent and salt are then added to the system, thus completing the initial setup. The system is then usually subjected to minimization, whereby the atomic coordinates are slowly moved down the potential energy gradient defined by the MD forcefield. Finally, initial atomic velocities are randomly selected from the Boltzmann distribution at the desired temperature, thus initiating the MD simulation.

2.1.1 The Molecular Dynamics Forcefield

The accurate calculation of interatomic forces is the foundation of MD. However, due to the analytical complexity of atomic interactions, approximations are made to make MD computationally tractable: atomic bonds and angles are treated with Hooke's Law, torsions are treated sinusoidally and van der Waals interactions are calculated with an empirically derived 6-12 Lennard-Jones potential. Electrostatic potential is represented, without approximation, by Coulomb's Law. Finally, improper torsions can be used to enforce planarity in aromatic rings. This set of interactions, generally represented in the form of potential energy dependent on the 3-dimensional atomic coordinates of the system (R), is referred to as the MD forcefield[21]:

$$V(R) = \sum_{bonds} k_{bond} (r - r_{0,bond})^2 + \sum_{angles} k_{angle} (\theta - \theta_{0,angle})^2 + \sum_{torsions} k_{\phi} [1 + \cos(n\phi - \delta)] + \sum_{impropers} k_{\phi} (\varphi - \varphi_{0,improper})^2 + \frac{1}{2} \sum_i \sum_{j \neq i} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \epsilon_{ij} \left(\frac{D_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{D_{ij}}{r_{ij}} \right)^6 \quad (2.1)$$

where k represents a force constant, naught-subscripts indicate equilibrium values, q represents atomic charge, r_{ij} is the interatomic distance between the atoms with indices i and j , and D_{ij} is the optimum interatomic distance for van der Waals interactions. Bonds, angles, torsions and impropers are referred to as bonded interactions, whereas coulombic and van der Waals interactions are termed nonbonded interactions. The halving of the nonbonded terms double summation eliminates for double-counting and nonbonded self-interactions are ignored.

The parameters in eq. 2.1 are fitted from a number of sources; bond and angle force constants derived from spectroscopic studies, equilibrium bond lengths, triatomic angles and torsional angles are calculated using quantum mechanical (QM) optimizations, torsional angle barriers and Lennard-Jones parameters are fitted to QM potential energy surface (PES) scans, and atomic charges are fitted using the RESP method[22].

Many research groups have independently developed forcefields, for example AMBER, GROMACS and CHARMM[23-25]. Functional differences between the forcefields are minimal, with the choice of a forcefield for a study being left primarily to the researcher's preference. While these forcefields have roots dating back to the 1960s[26], most are frequently updated to improve fidelity with experimental results. As MD simulation lengths increase, previously unobserved flaws in the forcefield become apparent and forcefield parameters are adjusted accordingly.

MD forcefields explicitly ignore electrons, representing atoms as point charges surrounded by a soft van der Waals sphere. This approximation also necessitates the permanence of atomic bonds throughout an MD simulation, meaning that atomic bonds cannot be formed nor broken, restricting MD simulations to the exploration of conformational transitions and disallowing chemical reactions from taking place.

2.1.2 Force Integration and Trajectory Accumulation

Instantaneous per-atom force vectors can be calculated as the gradient of the potential energy with respect to the atomic coordinates:

$$F = -\nabla V(R) \quad (2.2)$$

Force can then be related to acceleration via

$$F = ma = m \frac{d^2R}{dt^2} \quad (2.3)$$

Acceleration can be integrated with respect to time, yielding

$$R = \frac{1}{2}at^2 + vt + R_0 \quad (2.4)$$

where R is the updated atomic coordinates, a is the instantaneous acceleration, v is the instantaneous velocity and R_0 is the initial atomic coordinates before force integration. In practice, the differential dt in eq. 2.3 is a discrete timestep, denoted from here on as Δt . The

choice of this timestep is critical to capturing physically realistic molecular motions in a practical period of computational time; overly short timesteps will require too many steps and an unrealistic number of calculations to be completed while overly long timesteps will allow atoms to overshoot realistic positions, e.g. two atoms may become perfectly superimposed. The timestep is usually chosen to be 2fs, $\sim 1/5$ of the timescale of hydrogen bond-angle vibrations.

In principle, there are a number of algorithmic schemes for carrying out force integration, with variations of the Verlet algorithm being most popular in practice[27]. The logic of the Verlet algorithm proceeds as follows. First, the two 4th order Taylor series about $R(t)$ for forward and reverse system evolution are:

$$R_{t+1} = R_t + v_t \Delta t + \frac{F_t}{2m} \Delta t^2 + \frac{b}{6} \Delta t^3 + O\Delta t^4 \quad (2.5)$$

$$R_{t-1} = R_t - v_t \Delta t + \frac{F_t}{2m} \Delta t^2 - \frac{b}{6} \Delta t^3 + O\Delta t^4 \quad (2.6)$$

where b is the third derivative of position with respect to time and O is the order of magnitude for higher terms in the Taylor series. Slight rearrangement and summing of equations 2.5 and 2.6 yields

$$R_{t+1} = 2R_t - R_{t-1} + \frac{F_t}{m} \Delta t^2 + O\Delta t^4 \quad (2.7)$$

Equation 2.7 provides a way to propagate atomic positions, conveniently avoiding direct calculation of atomic velocities, and with an error on the order of Δt^4 .

After each round of force integration, atomic coordinates can be saved as a snapshot of the system at the current step. The accumulation of these snapshots is a physically realistic time-evolution trajectory that can be analyzed for thermodynamic, conformational and dynamical characteristics.

2.1.3 Handling of Nonbonded Interactions

Nonbonded interactions pose a more daunting computational task than do bonded interactions, because there are potentially $O(n^2)$ nonbonded interactions in a system of n atoms. To reduce the cost of calculating nonbonded interactions, a distance cutoff is set around each atom. Within the cutoff, all nonbonded interactions are computed directly. Due to the rapid spatial decay of the Lennard-Jones potential's r^{-6} dependence, van der Waals interactions outside of a carefully chosen cutoff can be safely ignored. Coulombic interactions do not decay as rapidly and are generally treated with the particle mesh Ewald (PME) method[28].

PME considers all electrostatic interactions between particles in the unit cell, given by the Coulombic component of eq. 2.1

$$V = \frac{1}{2} \sum_i \sum_{j \neq i} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (2.8)$$

Extending eq. 2.8 to include electrostatic interactions in neighboring periodic cells yields

$$V = \frac{1}{2} \sum_N \sum_i \sum_{j \neq i} \frac{q_i q_j}{4\pi\epsilon_0 |r_{ij} + NL|} \quad (2.9)$$

where N is the desired number of periodic units and L is the unit cell length in a given direction. Equation 2.9 converges very slowly and is very expensive to calculate directly. The Ewald method then breaks the distance dependence of eq. 2.9 potential into short-range and long-range terms

$$\frac{1}{r} = \frac{f(r)}{r} + \frac{1-f(r)}{r} \quad (2.10)$$

The first term on the right hand side is the short-range component, while the second is the long-range component. $f(r)$ is taken to be $erfc(r)$. Each atom is then represented as a point charge with local neutralizing charge in the form of a Gaussian distribution

$$\rho_i(r) = \frac{q_i \alpha^3}{\pi^{3/2}} e^{-\alpha^2 r^2} \quad (2.11)$$

resulting the in the short-range term of the Ewald sum taking the form

$$V = \frac{1}{2} \sum_N \sum_i \sum_{j \neq i} \frac{q_i q_j}{4\pi\epsilon_0 |r_{ij} + NL|} \frac{erfc(\alpha|r_{ij} + LN|)}{|r_{ij} + LN|} \quad (2.12)$$

The neutralizing Gaussian charges in eq. 2.11 is corrected by adding a background charge distribution

$$V = \frac{1}{2} \sum_N \sum_i \sum_{j \neq i} \frac{1}{\pi L^3} \frac{q_i q_j}{4\pi\epsilon_0} \frac{4\pi^2}{k^2} \exp\left(-\frac{k^2}{4\alpha^2}\right) \cos(k \cdot r_{ij}) \quad (2.13)$$

where k are reciprocal vectors. Eq. 2.13 is the long-range component of the Ewald summation.

An optimized selection of α is critical to computational efficiency; small values of α lead to faster convergence of eq. 2.13, while larger values of α cause eq. 2.12 to converge more quickly.

The PME method has been further optimized by treating 2.13 as a Fourier series and applying the ubiquitous discrete fast Fourier transform for summation of the long-range terms. Furthermore, some MD engines, such as NAMD, smooth the reciprocal-space discretized point charges over multiple grid points with an Euler spline[29].

2.1.4 Solvent and the Simulation box

While MD can be performed *in vacuo*[30], solvent interactions and explicit electrostatic screening play central roles in the functioning of biomolecules, motivating the use of explicit solvent for all studies in this dissertation. Numerous computationally efficient water models exist for MD[31], with one of the most popular being the TIP3P model used throughout this dissertation. TIP3P water sets rigid bonds between all three atom pairs in each water molecule at equilibrium lengths, eliminating multiple degrees of freedom while maintaining minimal deviations from experimentally measured bulk properties.

MD simulations are typically performed under the canonical (NVT), microcanonical (NVE) or isothermal-isobaric (NPT) ensembles; with N, V, T, P and E indicating constant

number of particles, volume, temperature, pressure and total energy respectively. There are numerous schemes for enforcing constant temperature and pressure in MD[32-34]. While the NVE and NVT ensemble possess computational advantages, NPT most closely replicates laboratory conditions and is used for all of the production simulations herein.

2.2 Partial Nudged Elastic Band

There are often high energetic barriers associated with biochemical transitions and the conformations associated with these transition states are often of great interest. Crossing energetic barriers is a rare phenomenon and artificial forces must frequently be added to observe transition states in MD. Minimum free energy path (MFEP) methods have been developed to trace the conformational transition of a system between endpoints, through saddle points near barriers and valleys elsewhere[35-37]. The MFEP and nearby phase space represent the exponentially most likely states for the system to reside in and contribute most heavily to conformational ensembles during a transition, and are therefore most important to calculating thermodynamic properties and other ensemble statistics. One such MFEP definition method is the partial nudged elastic band (PNEB)[38]. In the PNEB, two endpoints are first defined structurally. Multiple copies, known as beads, of each endpoint are then generated. Each bead is simulated independently, with linear, per-atom forces (F^{\parallel}) imposed to position the conformation of each bead between its neighbors, parallel to the MFEP. This parallel force is defined as

$$F_i^{\parallel} = [k_{i+1}(P_{i+1} - P_i) - k_i(P_i - P_{i-1}) \cdot \tau] \tau \quad (2.15)$$

where i is the bead index, k is a chosen force constant, P_i are the collective Cartesian coordinates of bead i and τ is the tangent vector to the MFEP. Forces are applied on a per-atom basis after rmsd alignment between neighboring beads at every timestep. While the beads are stretched into

a smooth band by the parallel force, the MD forcefield acts perpendicularly to the band (F^\perp) to sample the local energetic terrain, moving beads into local minima:

$$F_i^\perp = -\nabla V(P_i) + (\nabla V(P_i) \cdot \tau)\tau \quad (2.16)$$

where $V(P_i)$ is the potential energy landscape with respect to atomic coordinates. The total forces acting on a bead in PNEB are then:

$$F_i = F_i^\parallel + F_i^\perp \quad (2.17)$$

As the beads are interpolated from the PNEB forces, the MD simulation ensures that they sample in physically reasonable regions of the phase space. PNEB optimization protocols typically consist of multiple stages, changing the system temperature and PNEB force constants to move the band toward the MFEP, in an analogous manner to simulated annealing. After the MFEP is fully optimized, the beads hold conformational clues about the biochemical process and can also be used as starting points for path sampling methods.

2.3 Umbrella Sampling

Due to the timescale limitations of MD imposed by computational hardware, many enhanced sampling methods have been developed to more adequately sample chemical or physical processes and, usually after post-processing, generate realistic equilibrium ensembles suitable for thermodynamic and kinetic analysis[39-41]. One family of enhanced sampling techniques seeks to restrain the system at various points along a reaction coordinate and simulate biased ensembles along the reaction coordinate. The conformational positions along the reaction coordinate chosen for sampling can be generated with an MFEP method, like the PNEB. Unbiasing of the ensembles yields relative probabilities of the system residing in a given state, which can be related to energies from

$$P(R) \propto e^{-E(R)/k_B T} \quad (2.18)$$

where R is the system's coordinates in phase space, E represents free energy, k_B is the Boltzmann constant and T is temperature in Kelvin. The compilation of these energies along the reaction coordinate yields a potential of mean force (PMF) that serves as an effective free energy profile for the process. Most commonly, a PMF is given in terms of either Gibbs free energy or Helmholtz free energy, determined by whether the simulations are carried out in the isothermal-isobaric or canonical ensemble, respectively. An extremely popular example of these methods is umbrella sampling (US). In US, an artificial spring is imposed on the system to restrain it to a region of the reaction coordinate:

$$F_{TOT} = F_{FF} + \frac{1}{2}k(\xi - \xi_0)^2 \quad (2.19)$$

where F_{TOT} is the total instantaneous force vector acting on the system, F_{FF} is the force on the system due to the forcefield and the final term describes a harmonic potential applied to the system at position ξ along the reaction coordinate, about a window center at ξ_0 . There are many options for reaction coordinates, with the only general rules being that the chosen coordinate is well-behaved and varies along with the progress of the chemical process being studied. US can be performed along multiple reaction coordinates simultaneously by adding more harmonic terms to eq. 2.19, and is then capable of producing multi-dimensional PMFs that can be analyzed for interplay between the chosen reaction coordinates. However, sampling time increases proportional to the power of the number of dimensions, resulting in much more computational effort being required to adequately sample higher-dimension reaction spaces.

2.3.2 Weighted Histogram Analysis Method

Biasing simulations to restrict sampling to a small volume of the phase space, as in US, is conceptually and mathematically simple. However, the ensembles generated in US are biased and therefore physically unrealistic. Furthermore, because there are many windows along the

reaction coordinate, all of the biased ensembles must be considered simultaneously to construct a PMF. Unbiasing of the ensemble is critical to the efficacy of US and is accomplished through a post-processing calculation known as the weighted histogram analysis method (WHAM)[42]. In WHAM, the positions from every snapshot of every window are first binned along the reaction coordinate, followed by the iterative solution of the following two equations:

$$P(\xi) = \frac{\sum_{i=1}^N n_i(\xi)}{\sum_{i=1}^N N_i \exp([F_i - U_i(\xi)]/k_B T)} \quad (2.20)$$

$$F_i = -k_B T \ln[\sum_{bins} P(\xi) \exp\left(-\frac{U(\xi)}{k_B T}\right)] \quad (2.21)$$

where i indexes the individual windows, with N being the total number of windows, n represents bin population, U is the biasing potential, F is the change in free energy, P is the unbiased probability and $k_B T$ is defined as above. The resulting set of F_i are the PMF for the process. The primary assumption in applying eq. 2.20 and eq. 2.21 is that neighboring windows sample an overlapping portion of the reaction coordinate, to ensure that the F_i are continuous. While simpler methods exist that rely on fixing system replicas along the reaction coordinate, measuring the force gradients applied through the constraints and integrating the forces along the reaction coordinate to obtain a PMF, these methods suffer from a lack of certainty of the continuous nature of the reaction coordinate. The requirement that neighboring windows have sampling overlap in the US/WHAM method ensures a continuous PMF.

2.4 Dynamical Network Analysis

Network analysis (sometimes known as graph analysis) encompasses a very widely adopted set of methods for studying the dynamics of systems comprising many individual components, such as computer networks, social networks, biomes and physical systems[43]. Each individual in the system is represented by a node in a network, with edges connecting

interacting individuals and the weight of each edge representing an arbitrary measure of the strength of interaction between the individuals it connects.

The definition of nodes in biomolecular systems is arbitrary, but most commonly taken to be each residue in a protein[44], while nucleobases and ribophosphate groups are split into separate nodes in nucleic acids. Edges are then placed nearby nodes and weighted according to the Cartesian covariance between the nodes during an MD simulation:

$$w_{i,j} = E[(x_i - \bar{x}_i)(x_j - \bar{x}_j)] \quad (2.22)$$

where w is the edge weight between nodes i and j , x represents the coordinates of the node, \bar{x} represents the average coordinates of the node throughout the simulation and E expresses the expectation value of the argument. The network holds key, often subtle, dynamical information about the correlated motions within the system. These motions can be linked to conformational phenomena relevant to biological processes. There exists a multitude of network analysis methodologies to address or highlight various aspects of a network.

2.4.1 Suboptimal Paths

Communication between two nodes, here termed the source and sink, separated by multiple edges can occur through many different paths in a well-connected network. In the case of dynamical network analysis for MD simulations, the network is undirected, rendering the terms source and sink interchangeable. The shortest path between a source and a sink is referred to as the optimal path. However, there are a number of slightly longer, nearly-optimal paths that will also contribute significantly to communication between the source and sink. These are known as suboptimal paths and they must be taken into account to adequately map the flow of communication between source and sink[44]. For MD system networks, these communication pathways carry the bulk of allosteric transmissions between source and sink.

When determining the length of a suboptimal path, the edge lengths comprising the path are summed. In the case of molecular dynamics systems, these lengths are taken to be the Cartesian covariance between the nodes joined by an edge. The calculation for suboptimal path length is

$$L_p = \sum_e -\log(w_e) \quad (2.23)$$

where L_p is the length of the path, summed over all edge lengths belonging to said path. Edge lengths are defined as $-\log(w_e)$ because stronger correlations correspond to shorter distances in network space. All suboptimal paths under an arbitrary length cutoff are assumed to contribute significantly to allosteric signaling and combine to form an allosteric tether between source and sink.

2.4.2 Community Analysis

While network structures can be informative, they are often far too complex to analyze visually for large biomolecules. Community analysis is a tool for segregating groups of nodes into larger, semi-autonomous communities[45]. In a biomolecule, communities indicate heavily self-interacting groups of residues and can be used to visualize general dynamical topology and highlight general functional sites.

Many methods exist to separate networks into communities[45]. One popular method, Girvan-Newman (GN)[46], is especially well-suited for analysis of MD-based networks due to the absence of free parameters in the algorithm, leading to simpler implementation and use. The GN algorithm is executed as follows: edges are assigned betweenness values, where betweenness is calculated as the number of shortest paths between any two nodes that traverse the edge; edges are removed from the network individually, in order of decreasing betweenness, calculating modularity at each iteration and identifying communities (groups of nodes not connected to the

rest of the network), until no edges remain; finally, the highest modularity community structure is returned. The key metric in GN, modularity, is defined as

$$Q = \frac{1}{2m} \sum_{vw} [A_{vw} - \frac{k_v k_w}{2m}] - \frac{s_v s_w + 1}{2} \quad (2.24)$$

where Q is modularity[47], k_v is the degree of node v , A_{vw} is the adjacency matrix, s_v is the membership of node v and m is the total number of edges in the graph. The adjacency matrix contains a 1 at position v,w if there is an edge between nodes v and w , and 0 elsewhere. The degree of node v is the number of edges connected to node v . Community membership, s , is set as 1 or -1, depending on which community a node belongs to. It should be noted that that eq. 2.24 is defined for two communities, leading to hierarchical community splitting in practice.

While the modularity optimization method rigorously defines the optimal community structure, the number of communities generated from an MD simulation, especially in relatively rigid complexes where thermal noise contributes significantly to the overall dynamics, can be excessive or errant. A minor adaptation of the GN algorithm has been developed for use in this dissertation, wherein a modularity cutoff is set, as a percentage of the optimal modularity, with the final community structure being that which possesses the fewest communities while maintaining a modularity value that lies within the cutoff. Very small reductions in modularity can lead to large reductions in the number of communities GN produces, yielding more interpretable community structures with similarly high modularity scores to the optimal structure.

2.5 Molecular Mechanics-Poisson Boltzmann surface area

Binding energetics are of special importance in biophysics, determining equilibrium constants for complex formation, inhibitor efficacy and various transition rates. While nonbonded energies can be calculated from an MD trajectory according to the forcefield, the

forcefield approximations render such calculations inadequate for the accurate estimation of binding energetics. Molecular mechanics – Poisson-Boltzmann surface area (MM-PBSA) calculations are a more accurate, thermodynamic cycle-based approach to measuring binding enthalpy[48]. The thermodynamic cycle solves for the solvated binding free energy of the complex

$$\Delta G_{bind,solv} = \Delta G_{bind,vacuum} + \Delta G_{solv,complex} - (\Delta G_{solv,ligand} + \Delta G_{solv,receptor}) \quad (2.25)$$

where

$$\Delta G_{solv} = G_{electrostatic,\epsilon=80} - G_{electrostatic,\epsilon=1} + \Delta G_{hydrophobic} \quad (2.26)$$

and

$$\Delta G_{vacuum} = \Delta E_{MM} - T\Delta S_{NMA} \quad (2.27)$$

$\Delta G_{hydrophobic}$ is an empirical term proportional to the solvent accessible surface area of the moiety, ΔE_{MM} is the change in intramolecular energies from the MD forcefield and is generally negligible and therefore ignored, ΔS_{NMA} is a normal mode analysis-based entropic estimate that is frequently dropped when comparing similar systems, and $G_{electrostatic}$ is the energy attributed to electrostatic interactions, given as

$$G_{electrostatic} = -\frac{1}{2} \int \rho(r)\phi(r)dr \quad (2.28)$$

where the electrostatic potential, $\phi(r)$, is solved for from the Poisson-Boltzmann equation

$$\nabla\epsilon(r)\nabla\phi(r) - \epsilon(r)\lambda(r)\kappa^2 \frac{k_B T}{q} \sinh\left(\frac{q\phi(r)}{k_B T}\right) = -4\pi\rho(r) \quad (2.29)$$

$\rho(r)$ is the free charge density, q is the charge of an electron, $\lambda(r)$ is the Debye length at the simulation ionic strength, and κ^2 is a switching function set to one in electrolyte-accessible regions and zero elsewhere. The difference in electrostatics terms in eq. 2.26 represents the polarization work associated with solvating the complex, by varying the permittivity parameter.

MM-PBSA binding energies can be decomposed on per-residue and pairwise interaction bases by considering only interactions involving a residue or residue pair. While the solvation terms in eq. 2.25 are not pairwise-decomposable, there are schemes to satisfactorily approximate pairwise interaction energies[49]. Pairwise interaction energies can be instructive in designing mutational studies and designing inhibitors to exploit specific contacts.

2.6 Grid Inhomogeneous Solvation Theory

Solvent-solute interactions are important across a spectrum of biological processes, from catalysis to binding interface recognition. Grid inhomogeneous solvation theory (GIST) provides a convenient mathematical framework and computational tool for analyzing solvation energetics[50]. In GIST, solvent energies are decomposed into enthalpic and entropic terms

$$\Delta G = \Delta E_{sw} + \Delta E_{ww} + \Delta S_{trans} + \Delta S_{orient} \quad (2.30)$$

where the subscripts s and w denote solute and water, respectively; ΔS_{trans} is the translational entropy of the solvent and ΔS_{orient} is the orientational entropy of the solvent. ΔE is calculated from the forcefield nonbonded energies. ΔS_{trans} is given as

$$\Delta S_{trans} = -k_B \rho^0 \int g(r) \ln g(r) dr \quad (2.31)$$

where $g(r) \equiv \rho(r)/\rho^0$, ρ^0 is the number density of bulk solvent and $\rho(r)$ is the number density function of the system. ΔS_{orient} is given as

$$\Delta S_{trans} = -k_B \rho^0 \int g(r) S^\omega(r) dr \quad (2.32)$$

with

$$S^\omega(r) = \frac{-k_B}{8\pi^2} \int g(\omega|r) \ln g(\omega|r) d\omega \quad (2.33)$$

and $g(\omega|r) \equiv \rho(\omega|r)/\rho^0$. In GIST, the simulation box is divided into discrete voxels, with eq. 2.30 solved for the time-averaged properties in each voxel. The resulting volumetric energetic information can be integrated with respect to volume to yield the thermodynamic components of

solvation, restricted to specific regions of interest or across the entire system. Visualization of energetic isosurfaces is a useful diagnostic for localizing solvation hotspots.

CHAPTER 3. ALKYL PURINE GLYCOSYLASE D EMPLOYS DNA SCULPTING AS A STRATEGY TO EXTRUDE AND EXCISE DAMAGED BASES

3.1 Abstract

Alkylpurine glycosylase D (AlkD) exhibits a unique base excision strategy. Instead of interacting directly with the lesion, the enzyme engages the non-lesion DNA strand. AlkD induces flipping of the alkylated and opposing base accompanied by DNA stack compression. Since this strategy leaves the alkylated base solvent exposed, the means to achieve enzymatic cleavage had remained unclear. We determined a minimum energy path for flipping out a 3-methyl adenine by AlkD and computed a potential of mean force along this path to delineate the energetics of base extrusion. We show that AlkD acts as a scaffold to stabilize three distinct DNA conformations, including the final extruded state. These states are almost equivalent in free energy and separated by low barriers. Thus, AlkD acts by sculpting the global DNA conformation to achieve lesion expulsion from DNA. *N*-glycosidic bond scission is then facilitated by a backbone phosphate group proximal to the alkylated base.

3.2 Author Summary

DNA repair efficiency is critically dependent on the function of DNA glycosylases. These versatile enzymes perform a remarkably discriminating search for DNA lesions, followed by damage-specific base extrusion into to the enzyme's active site and removal of the damaged bases. Our work elucidates the mechanism of *Bacillus cereus* AlkD, representative of a superfamily of alkylpurine glycosylase enzymes that function differently from all other known glycosylases. AlkD does not employ any direct contacts to the alkylated lesion. Instead, it relies on DNA backbone contacts to extrude the lesion's base-pairing partner. The alkylated base is flipped into solvent allowing *N*-glycosidic bond hydrolysis to occur with no apparent assistance

from any protein side chains. Our work contributes to understanding of this unique base extrusion and excision strategy. We determined a minimum energy path for flipping out a 3-methyl adenine base by AlkD and computed an effective free energy profile for this transition. We show that lesion extrusion relies on DNA sculpting to break up the process into two steps characterized by low free energy barriers and a stable intermediate. AlkD provides a rigid scaffold to accommodate the three distinct DNA conformations and positions a phosphate group to facilitate scission of the alkylated base.

3.3 Introduction

Despite its remarkable stability, DNA is subject to a variety of reactions. Left unchecked, these processes could impair the transmission of vital genetic information and threaten the integrity of the genome. To cope with genomic instability cells have evolved elaborate DNA repair mechanisms. Many cancer therapies are directly impacted by the efficiency of DNA repair. Antitumor drugs often act by inducing DNA lesions, thus, blocking replication in rapidly dividing cancer cells[51]. Upregulating DNA repair is a common mechanism in tumors to develop resistance to chemotherapy. Conversely, suppression of repair activity sensitizes cancer tissues to chemotherapies targeting DNA[51-54]. Specifically, alkylating agents can give rise to alkylpurine lesions[55] such as 3-methyl adenine (3mA) and 7-methyl guanine (7mG). Alkylpurine lesions carry a positive formal charge on the base, resulting in a sheared base pairing orientation and a comparatively labile *N*-glycosidic linkage that is prone to spontaneous hydrolysis. Not only are these lesions cytotoxic themselves, their propensity for spontaneous depurination could result in other, more deleterious forms of damage (*e.g.* single-strand or double-strand DNA breaks)[56, 57]. Alkylation lesions are processed by the cell's base excision

repair (BER) machinery[58-60] to replace the damaged base with its correct Watson-Crick analog.

In BER, DNA N-glycosylases are the first line of defense against damage in genomic DNA. These enzymes efficiently and specifically recognize and excise single-base lesions[2, 61]. Structures of glycosylase enzymes reveal that DNA binding is accompanied by a multitude of conformational changes preceding active site chemistry. Specifically, nucleotide flipping (base extrusion) is a commonly employed strategy wherein a deoxynucleotide swings out of the DNA helix and is accommodated in the enzyme's catalytic pocket. This process has been the object of intense experimental focus for over two decades. Nonetheless, consensus has not been achieved regarding the pathways and molecular events that accompany base extrusion. Indeed, controversy has persisted regarding the precise role of the glycosylase (active or passive) in dislodging lesions from DNA. A number of alkylpurine-specific glycosylases have been characterized[62-65]. *Bacillus cereus* AlkD belongs to a unique superfamily of N3- and N7-alkylpurine glycosylases (present in all three domains of life) that function differently from all other known glycosylases[66-68]. Commonly, DNA glycosylases must flip damaged nucleotides out of the DNA base stack into damage specific pockets and must also accommodate the resulting DNA distortion by intercalating a side chain into the stack to replace the extrahelical nucleotide. To achieve efficient N-glycosidic bond scission, glycosylases must also provide side chains suitable to act as a general base in catalysis. By contrast, AlkD does not employ any direct contacts to the alkylated lesion. Instead, it relies on DNA backbone contacts to extrude the lesion's base-pairing partner. The alkylated base is flipped into the cytosol, allowing for hydrolysis to occur with no apparent assistance from any protein side chains[67].

3.4 Results and Discussion

3.4.1 Pathway and energetics of base pair opening and lesion extrusion by AlkD

Previous studies of DNA flipping[69-73] have relied primarily on intuitive reaction coordinates such as a pseudo torsional angle. However, AlkD works by sculpting the DNA backbone. A local reaction coordinate such as a pseudo dihedral is, in this case, inadequate. To describe such complex conformational transitions of DNA it is advantageous to employ path optimization methods such as the partial nudged elastic band (PNEB)[38, 74]. A key requirement is that the initial and final states in the transition be known. In this respect, crystal structures of AlkD with DNA containing a 3-deaza-3mA or tetrahydrofuran (THF) are available to represent the pre-extrusion and post-excision complexes[67]. From these structures we constructed and equilibrated models for the initial and final AlkD/3mA-DNA states and then computed a minimum energy path (MEP) for base extrusion of 3mA by AlkD using PNEB (Figure S1). Umbrella sampling was then performed to describe the free energy profile of this conformational transition using molecular dynamics[75]. The reaction coordinate ξ for the transition was defined as $\xi = rmsd_i - rmsd_f$ where $rmsd_{i(f)}$ denotes root-mean-square deviation from the initial and final state. For clarity ξ was further normalized to vary from 0 to 1 (corresponding to the initial and final state, respectively). Among the advantages of ξ as a reaction coordinate is the ability to adequately describe global DNA bending induced by AlkD and to distinguish between concerted and sequential base flipping events. While the PNEB optimization involved all atoms of the AlkD/DNA complex, the definition of ξ involved rmsd over the nucleic acid heavy atoms alone. Additionally, we modeled B-form DNA with identical sequence and applied steered molecular dynamics to rotate the base opposite the 3mA out of the base stack. Umbrella sampling was performed with the same protocol as for the AlkD/DNA complex, except ξ involved rmsd over

the lesion pair and the base pairs immediately above or below in the stack. The obtained PMF profiles are shown in Figure 3.4.4.1. The reference profile suggests that base flipping in canonical B-DNA in the absence of AlkD proceeds with a steep initial rise in free energy as soon as the base departs from the stack. At ξ value of 0.4, ΔG reaches ~ 10 kcal/mol and continues to increase to ~ 14 kcal/mol albeit with a lesser slope afterward. The extruded state is thus represented by a broad plateau region with no apparent stabilization of the nucleotide anywhere outside the initial stacked conformation. Previous work on base flipping[20] in DNA is fully consistent with this view of the extrusion process.

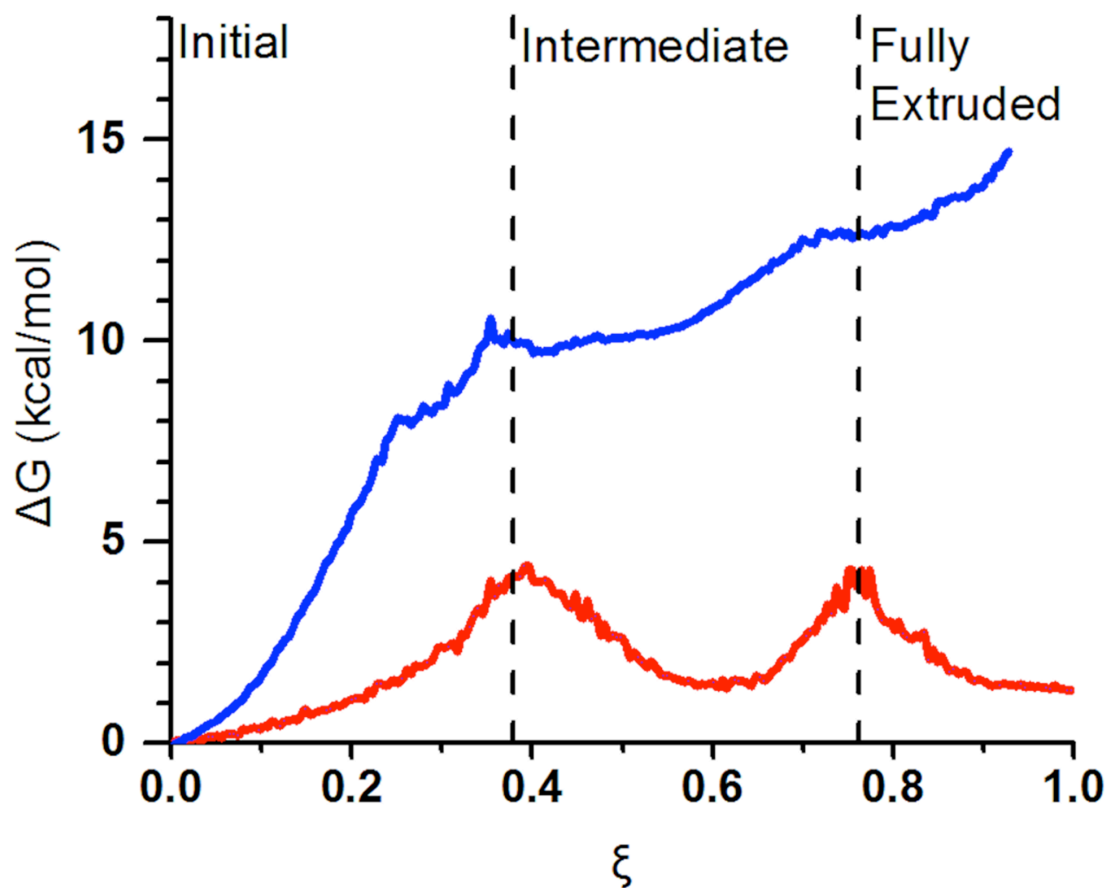


Figure 3.4.1.1. AlkD binding flattens the free energy landscape for lesion extrusion from DNA. Effective free energy profiles for base flipping in the presence (red line) and absence (blue line) of AlkD. The rmsd-based reaction coordinate, ξ was normalized to vary from 0 to 1.

Here we show that AlkD association to DNA substantially lowers the energy barrier for base flipping and provides a relatively flat free energy landscape characterized by three stable states (Figure 3.4.1.2) denoted as initial, intermediate and final (fully extruded). Notably, the barriers that separate these states are ~ 3.9 and 3.0 kcal/mol, respectively. Two separate flipping events are resolved in the PNEB path and the PMF. The opposing base is extruded first and accommodated in a shallow pocket on the surface of AlkD. In this orientation the nucleotide is stabilized primarily through residue contacts to the DNA backbone, while the base itself remains solvent exposed. Two factors contribute to the observed barrier: (i) strain accompanying the opposing base rotation around the DNA backbone; and (ii) water penetration into the space previously occupied by the base in the DNA stack. The way AlkD relieves both of these factors after the initial barrier is to severely kink the DNA substrate while still preserving the 3mA position in the base stack. Analysis of the umbrella sampling windows with the program Curves+[76] reveals that AlkD moderately bends the DNA in the initial state by 12.7° ; severely kinks the DNA near the lesion in the intermediate state by 26.9° ; and straightens the DNA to a negligible bend of 3.6° in the final state. The origin of the second barrier in the PMF is the rotation of the 3mA lesion out of the DNA stack. Collapse of the water filled cavity left by the base and repositioning of two contacts to the lesion strand (Thr39 and Arg43) leads to the final fully extruded state. Base stack compression restores stacking interactions and removes the DNA kink. We note that the three stable states are almost equivalent in terms of free energy with initial and final differing by just $\sim 2 k_B T$ units of thermal energy. Thus, AlkD specifically stabilizes the extruded state allowing sufficient lifetime of 3mA in the cytosol to accomplish hydrolysis. The low barriers among the three states could ensure frequent transitions on the ms timescale associated with base flipping.

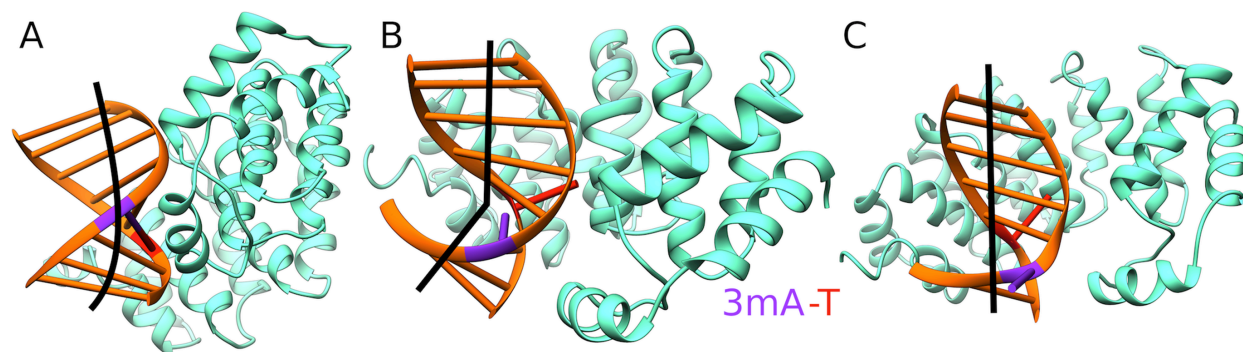


Figure 3.4.1.2. AlkD's sculpting of the DNA substrate results in three stable conformations along the flipping pathway. The three stable states in the AlkD/3 mA-DNA PMF. A) initial state; B) kinked intermediate state; and C) fully extruded (final) state. DNA bending is shown schematically in black.

3.4.2 Structural determinants for DNA bending, double base flipping and catalysis by AlkD

As a complement to its unique strategy, AlkD is structurally comprised almost entirely of HEAT-repeat motifs, more commonly known to mediate protein-protein rather than protein-nucleic acid interactions[68]. In AlkD, repeats 2 through 6 are comprised of two antiparallel helices H1 and H2 that are oriented with a minor right-handed twist. The carboxy-terminal helix (H2) lines the concave surface of the DNA binding cleft and provides positively charged residues to recognize the non-lesion DNA strand (Figure 3.4.2.1a). The nucleotide opposite to the 3mA is extruded into a shallow pocket on the protein surface with no specific contacts to the base (Figure 3.4.2.1b). The only major polar interaction involves the Arg148 residue, which doubly hydrogen bonds to the 3' phosphate group of the extruded base. Two bulky tryptophan residues, W109 and W187, flank the DNA backbone to the 3' and 5' ends of the extruded base, sterically hindering rotations about the DNA backbone. These contacts are formed during the transition from the initial to the intermediate state and persist in the final state.

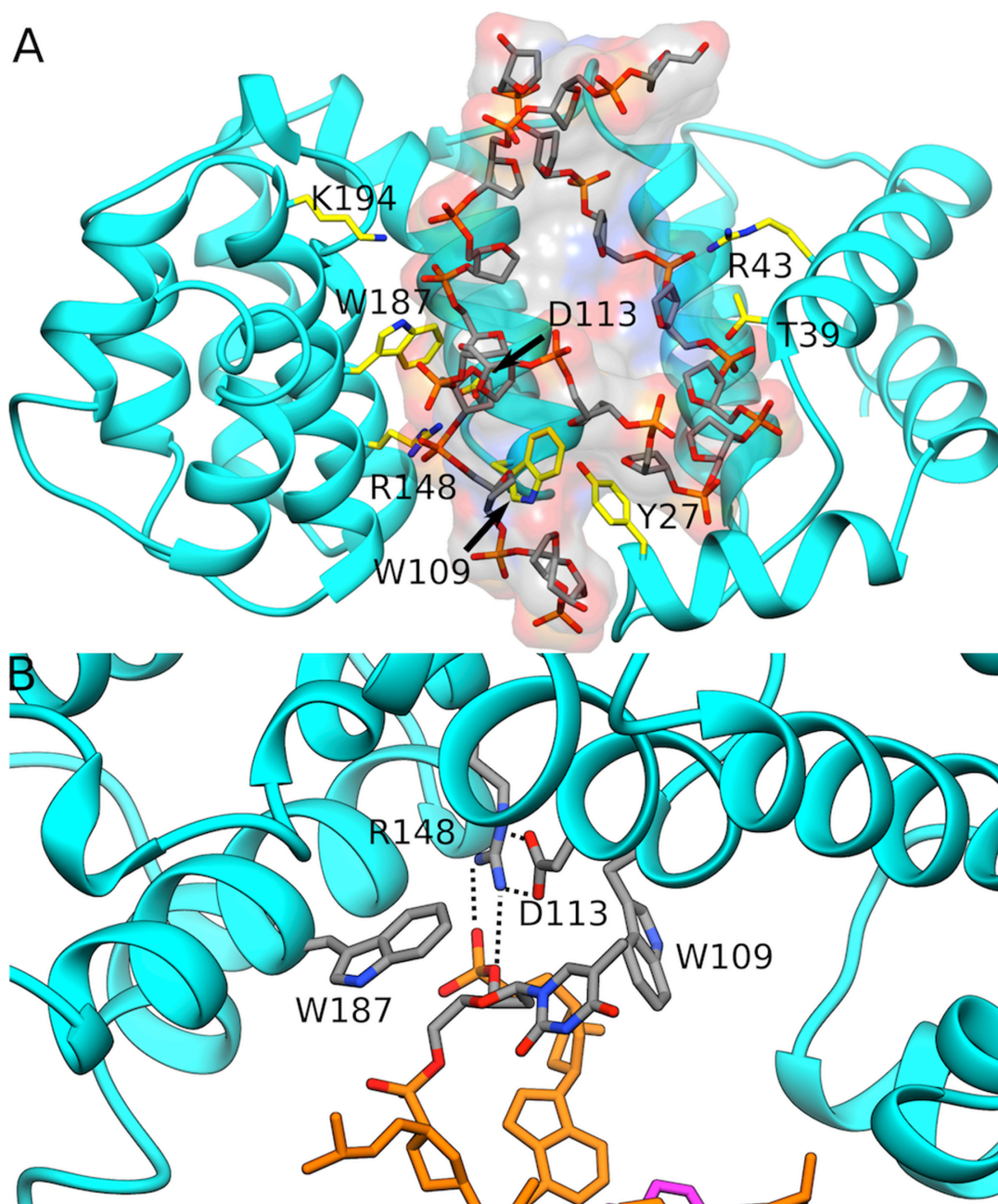


Figure 3.4.2.1. AlkD recognizes DNA through HEAT repeat motifs.

A) Overall architecture of the AlkD-DNA complex with residues contacting the DNA backbone shown explicitly and labelled; B) The mode of recognition of the extrahelical thymine base opposite to the 3 mA lesion. DNA is shown in stick representation and as a transparent colored surface. AlkD is shown in cartoon representation.

Surprisingly, the large conformational transitions of the DNA are accompanied by only minor changes in the AlkD conformation (Figure 3.4.2.2). Energy decomposition with the NAMDEnergy plugin of VMD[77] shows a rise in DNA conformational energy (primarily from the torsional component) and a concomitant increase in favorable protein-DNA interactions as base extrusion proceeds from initial to intermediate to final state. This corresponds to side chain adjustment of the residues contacting the DNA in these state (Figure 3.4.2.2a). At the same time we found that the change in rmsd from initial to final (computed over all heavy atoms of AlkD) was only ~ 2 Å. Thus, AlkD requires no significant motion of the protein itself to flip and expose the 3mA lesion. Instead, it provides a concave positively charged groove that is wide enough to accommodate multiple DNA conformations with different degrees of bending. Indeed, the only significant switch in AlkD-DNA contacts corresponding to the second barrier in the PMF was the observed repositioning of residues Thr39 and Arg43 with respect to the lesion strand (Figure 3.4.2.2b). These two residues shifted their hydrogen bonds by one phosphate group in the 3' direction along the DNA backbone. Repositioning Thr39 and Arg43 has the dual effect of energetically stabilizing the final extruded 3mA conformation and discouraging 3mA reinsertion into the DNA stack.

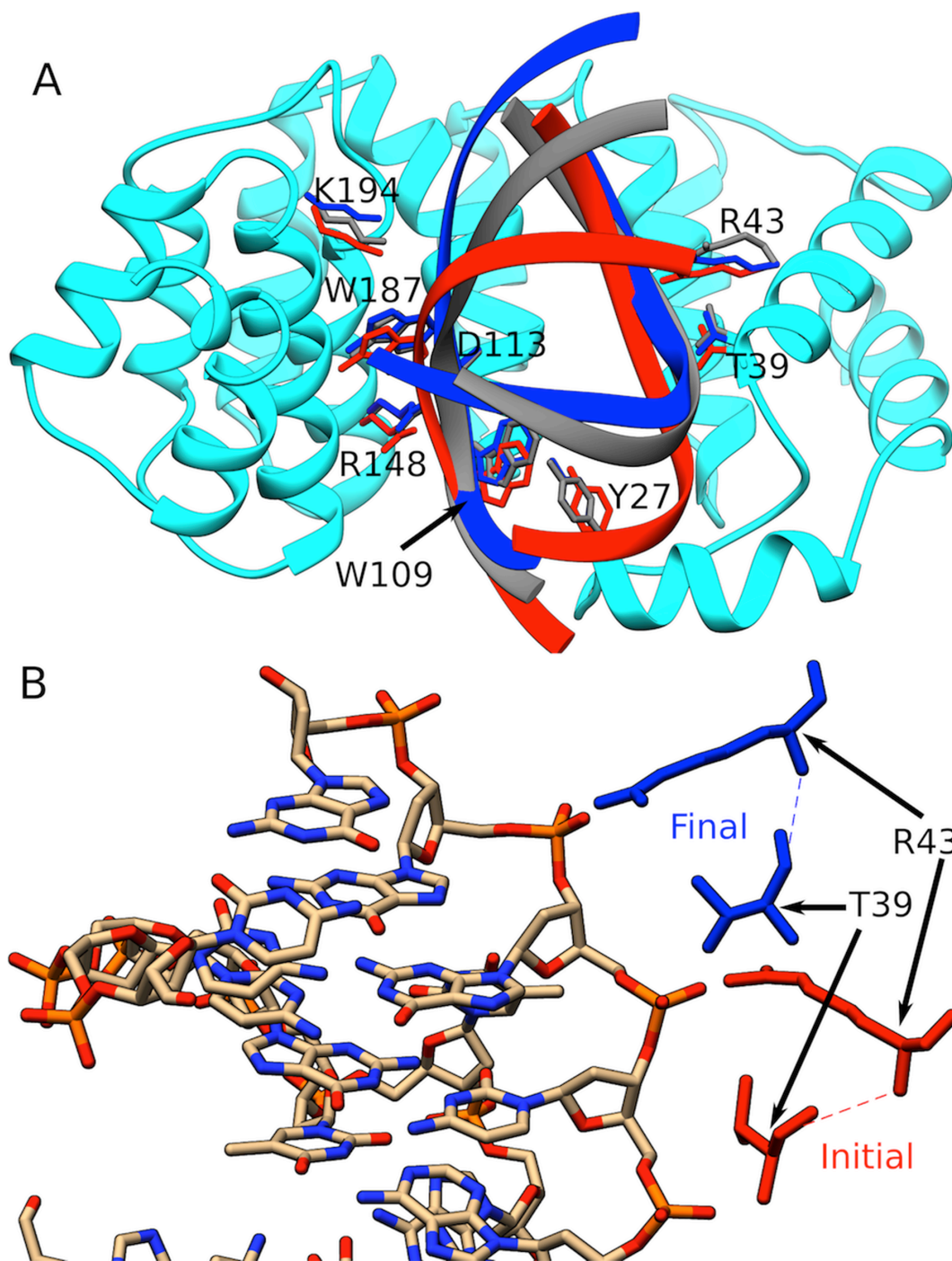


Figure 3.4.2.2. AlkD provides a scaffold to accommodate multiple DNA conformations with different degrees of bending.

A) DNA substrate in the initial (red), intermediate (gray) and final (blue) states of AlkD base extrusion. AlkD is represented in cyan. Minimal side chain movement is sufficient to accommodate the DNA conformational transitions; B) Shift in hydrogen bonding contacts from the intermediate to the final state. Two AlkD residues, Arg43 and Thr39, making direct contacts to the lesion strand shift their binding positions along the DNA backbone.

Sculpting the DNA substrate to promote 3mA base eversion is obviously necessary for removal of the lesion. However, solvent exposure is not sufficient to explain the 230-fold catalytic rate enhancement (over the spontaneous rate of hydrolysis) offered by AlkD[68]. Recent biochemical evidence has pointed to AlkD's role in stabilizing a catalytically competent conformation by positioning a phosphate group in proximity to the lesion ribose. The mechanistic proposal is that the phosphate would serve a role analogous to a protein carboxylate group in stabilizing the developing positive charge on the lesion ribose in the transition state (TS)[78]. In our MD simulation of the extruded state we observe persistent direct interaction of the lesion with the phosphate in position -2 (Figure 3.4.2.3). However, the interaction occurs through hydrogen bonding to the 3mA base rather than the ribose ring. This is reasonable as the 3mA base carries a formal positive charge. Thus, it is possible AlkD employs an alternative strategy to stabilize the TS by differential hydrogen bonding to the 3mA lesion. Altering hydrogen bonding to the base in the TS is not unprecedented and has also been proposed to contribute to catalysis by the prototypical glycosylase UDG[79]. This does not preclude a role for the phosphate stabilizing the ribose charge if the distance to the ribose decreases further in the TS complex.

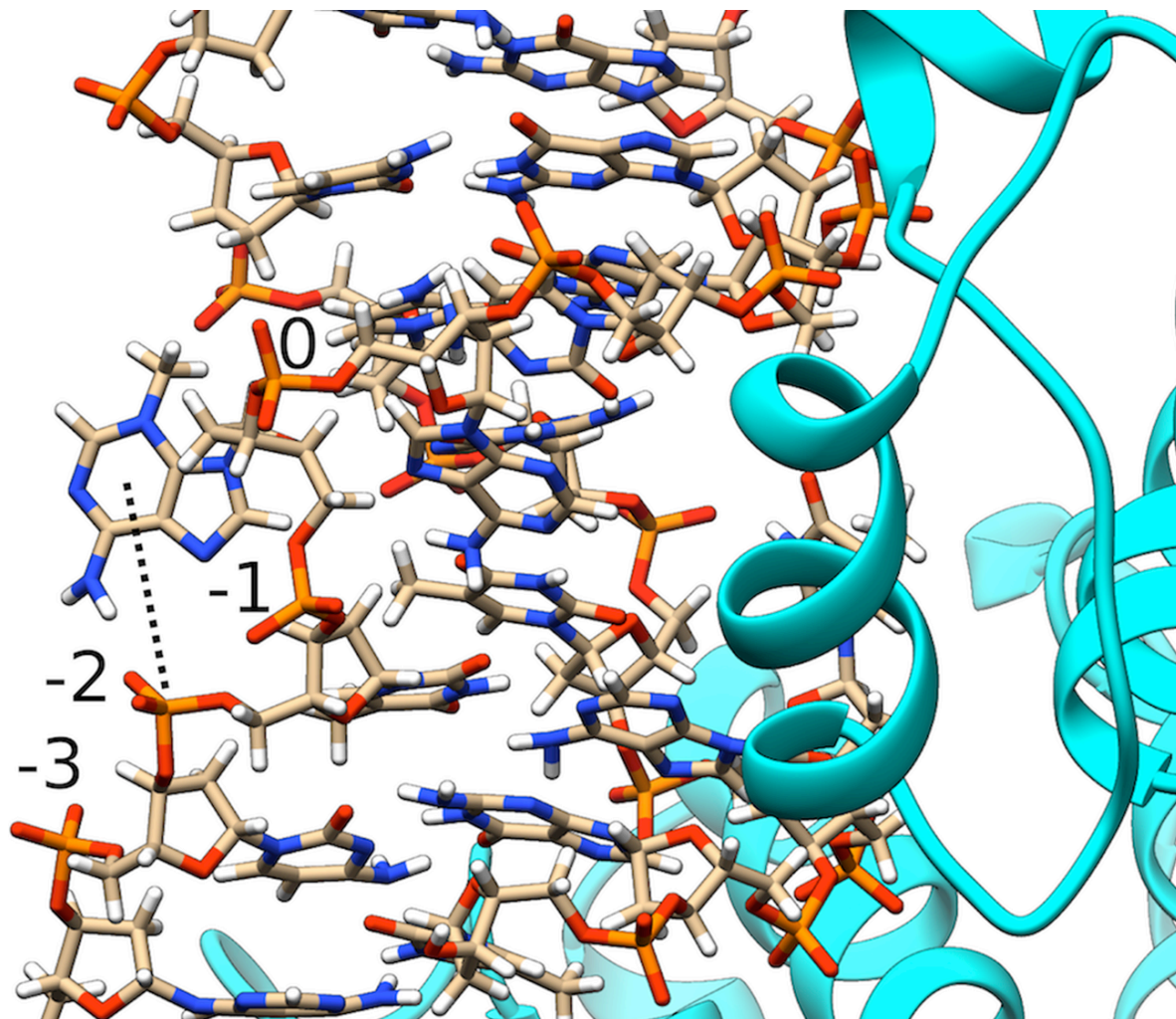


Figure 3.4.2.3. Persistent hydrogen bonding contact observed between the 3 mA base and the phosphate in position -2 along the lesion strand.
Phosphate groups are numbered starting from the 3 mA lesion at position 0.

3.5 Conclusions

In summary, lesion extrusion by AlkD relies on DNA sculpting to break up the process into two steps, which are characterized by low free energy barriers and a stable intermediate. The end result is a flattened free energy landscape along the path from the initial to the fully extruded state. The rigid arrangement of HEAT-repeat helices results in a C-shaped, positively charged cleft providing a scaffold to accommodate three distinct DNA conformations with different degrees of bending. Finally, excision of the 3mA base itself is dependent on the natural chemical instability of the alkylpurine *N*-glycosidic linkage and on a phosphate group suitably positioned to interact with the lesion in the extruded state. In this respect, the AlkD/DNA complex acts much like a DNAzyme, using the DNA backbone for catalysis. However, binding to AlkD's C-shaped cleft is required to achieve a catalytically competent conformation.

3.6 Methods

3.6.1 Model Construction

Models for the pre- and post-extrusion states (denoted initial and final) were constructed from two AlkD/DNA crystal structures[67] (Protein Data Bank accession codes 3JX7 and 3JXZ, respectively). The partial nudged elastic band method (PNEB)[38, 74] requires an identical number of atoms in each replica of the band. Therefore, the DNA sequence in the final model was changed to match the DNA construct for the initial model. This construct comprised a 9 base-pair DNA duplex with lesion strand sequence 5'-ACT(3mA)ACGGG-3'. The protein-DNA complexes were solvated with 9,977 TIP3P water molecules[80] in a box with dimensions 73.9 x 64.0 x 72.9 Å. Hydrogen atoms, Na⁺ counterions and solvent were introduced using the Xleap module of AMBER11[81] with the AMBER Parm99SB parameter set[82] and refined parameters for nucleic acids dihedrals (BSC0)[83]. 3mA force field parameters were determined

with the Antechamber module[84, 85] of AMBER. Partial charges for 3mA were obtained by RESP fitting after DFT calculations performed at the BLYP/6-31G*[86] level with the Gaussian03[87] program.

3.6.2 Equilibration Protocol

The systems were equilibrated using the NAMD 2.8 code[75, 88] and minimized for 10,000 steps with harmonic restraints on the protein and nucleic acid atoms to remove unfavorable contacts. The systems were then gradually brought up to 300 K in the NVT ensemble while keeping the protein and nucleic acid atoms restrained. The equilibration was continued for another 2 ns in the NPT ensemble and the harmonic restraints were gradually released. Next, the simulations were continued for additional 11 ns of unrestrained molecular dynamics to ensure fully equilibrated initial and final states for PNEB.

3.6.3 Path Optimization

To determine a MEP connecting the pre- and post-extrusion AlkD configurations we employed the PNEB method - a chain-of-replicas method that involves concurrent optimization of a number of copies of the simulated system (denoted as replicas or beads). We chose to represent the path by a total of 30 replicas - 15 copies of the equilibrated initial and final states, respectively. By gradually spreading the replicas from the initial and final states we allow the PNEB optimization process to discover the MEP in a fully unbiased way. All atoms of the AlkD/DNA complex were included in the path optimization. Simulations were carried out with a 1-fs integration step in the NVT ensemble at 300K. The minimum and maximum values for force constants between replicas were varied in from 0 to 4.5 kcal mol⁻¹ Å⁻² (k_{\min}) and from 0.25 to 4.5 kcal mol⁻¹ Å⁻² (k_{\max}). The PNEB protocol involved gradual ramping up of the force constant over

2 ns and subsequent scaling down to $2.0 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ over another 4 ns. The PNEB band was optimized at 300K for 5ns and then gradually brought back to 0 K in the last 1 ns.

Additionally, we modeled canonical B-form DNA with sequence identical to the 9-mer from the AlkD/DNA complex. Equilibration involved 1,000 steps of minimization, 5 ps of NVT dynamics to bring the temperature to 300 K and 200 ps of dynamics in the NPT ensemble. After equilibration, we applied SMD to rotate the thymine base opposite the 3mA lesion out of the base stack. The base was rotated 180° through the minor groove with constant velocity for 4 ns.

3.6.4 Umbrella Sampling Protocol

Umbrella sampling was performed to compute a PMF along the optimized PNEB path using the collective variables module of NAMD 2.8[75, 88] The collective variables module has a predefined RMSD variable (root-mean-square deviation of a group of atoms with respect to a reference structure). The module first calculates the best superposition of the atom group onto the set of reference coordinates before evaluating RMSD. The reaction coordinate (RC) was defined as $\xi = rmsd_i - rmsd_f$ where $rmsd_{i(f)}$ denoted root-mean-square deviation from the initial and final state, respectively. Each PNEB replica provided a configuration that was used to initiate an umbrella sampling window. Difference RMSD from initial and final was computed for this bead configuration and a harmonic umbrella potential ($k = 3.0 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$) was applied and centered at the computed ξ value. The collective variable module internally distributes the applied force onto the selected atoms according to the definition of the RC to maintain small deviation from the center of each window. For the AlkD/DNA complex ξ was defined over the nucleic acid heavy atoms. For the reference DNA system, ξ was defined over the heavy atoms of the lesion pair and the base pairs immediately above or below in the DNA stack. The production runs were performed in the NPT ensemble (1 atm and 300 K) for 10 ns per window with the

smooth particle mesh Ewald algorithm[89], short-range non-bonded cutoff at 10 Å and a switching function applied at 8.5 Å. The r-RESPA multiple timestep method[27] was employed with a 2-fs time step for bonded interactions, 2-fs for short-range non-bonded interactions and 4-fs for electrostatic interactions.

To analyze the results we used the weighted histogram analysis method (WHAM) as implemented in the code by Alan Grossfield[90] The first 2 ns from each window were considered equilibration and only the subsequent 8 ns were used for analysis. In total, 26 windows were sufficient to ensure uninterrupted coverage of the RC for WHAM calculations. Error bars were calculated by repeating the WHAM calculations in 2 ns increments (from 2 to 8 ns over the trajectories) and computing standard deviation. For the canonical B-form DNA control runs we carried out umbrella sampling with a $3.0 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ harmonic restraint imposed on the RC with the same protocol as for the AlkD/DNA complex. Since the B-DNA systems required less time to reach convergence, we ran each window for 5 ns.

CHAPTER 4. DISTAL SUBSTITUTIONS DRIVE DIVERGENT DNA SPECIFICITY AMONG PARALOGOUS TRANSCRIPTION FACTORS THROUGH SUBDIVISION OF CONFORMATIONAL SPACE

4.1 Abstract

Many genomes contain families of paralogs - proteins with divergent function that evolved from a common ancestral gene after a duplication event. To understand how paralogous transcription factors evolve divergent DNA specificities, we examined how the glucocorticoid receptor and its paralogs evolved to bind activating response elements [(+)GREs] and negative glucocorticoid response elements (nGREs). We show that binding to nGREs is a property of the glucocorticoid receptor (GR) DNA binding domain (DBD) not shared by other members of the steroid receptor family. Using phylogenetic, structural, biochemical, and molecular dynamics techniques, we show that the ancestral DBD from which GR and its paralogs evolved was capable of binding both nGRE and (+)GRE sequences due to its ability to assume multiple DNA-bound conformations. Subsequent amino acid substitutions in duplicated daughter genes selectively restricted protein conformational space, causing this dual DNA-binding specificity to be selectively enhanced in the GR lineage and lost in all others. Key substitutions that determined the receptors' response element binding specificity were far from the proteins' DNA-binding interface and interacted epistatically to change the DBD's function through DNA-induced allosteric mechanisms. These amino acid substitutions subdivided both the conformational and functional space of the ancestral DBD among the present-day receptors, allowing a paralogous family of transcription factors to control disparate transcriptional programs despite high sequence identity.

4.2 Significance Statement

Most organisms contain families of related proteins that evolved from duplication of an ancestral gene. Using the example of DNA binding by the steroid hormone receptors, this work examines the structural mechanisms by which these related proteins evolved separate functions during their history. We show that a functionally promiscuous ancestor was capable of accessing multiple protein conformations to bind disparate DNA sequences. This functional and conformational diversity were divided among daughter genes after gene duplication, allowing evolutionarily related proteins to generate disparate transcriptional outcomes in response to signaling input.

4.3 Introduction

Gene duplication is a key factor in the acquisition of novel protein function over evolution[91]. Most species, including humans, encode numerous families of paralogs – genes with divergent function that evolved from a common ancestral gene after a duplication event. Many models have been proposed to explain how new functions arise and are subdivided among paralogous proteins[92], and promiscuity of ancestral genes is hypothesized to be a major factor in the evolution of novel functions in protein families[[93, 94] but see [95]]. Understanding these processes is critical for a deeper understanding of molecular evolution as well as for developing therapies to target specific proteins contained within larger gene families. In this article, we consider the example of dual DNA-binding specificity by the glucocorticoid receptor (GR), a member of the steroid hormone receptor (SR) family. Composed of six members, the SR family is responsible for mediating the intracellular effects of steroid hormones, which are cholesterol-derived molecules that effect long-range, long-lasting physiological effects in target tissue[96, 97].

GR is capable of both activating[98, 99] and repressing[100-102] transcription. To activate transcription, GR's DNA binding domain (DBD) cooperatively dimerizes on inverted repeat activating glucocorticoid response elements, or (+)GREs (**Fig. 1A**)[103]. All 3-keto SRs – comprising of the glucocorticoid, mineralocorticoid, and androgen and progesterone receptors – bind to identical DNA sequences termed (+)GREs. The structural basis of GR-mediated transcriptional repression has remained less clear; however, recent work has shown that negative glucocorticoid response elements, or nGREs, play a role in GR-mediated transcriptional repression[102]. At nGREs, monomeric GR DBD binds to an everted repeat with negative cooperativity [104](Figure 4.4.1.1A). GR down-regulates the transcription of many anti-inflammatory genes, making GR agonists a mainstay of treatment of diseases such as asthma and arthritis[105]. However, GR's closest paralog, the mineralocorticoid receptor (MR), causes a pro-inflammatory transcriptional state when activated[106]. Despite their opposing effects on inflammation, GR agonists prolong the lifespan of MR^{-/-} mice, indicating that GR and MR share some overlapping transcriptional effects[107]. Likewise, in prostate cancer cells, GR activation can compensate for the pharmacological blockade of the androgen receptor [108]. Thus, separate members of the SR family can exhibit both overlapping and distinct transcriptional effects, yet the evolutionary mechanisms driving this phenomenon are unknown.

In this work, we demonstrate that despite their shared affinity for (+)GREs, GR is the only 3-keto SR capable of nGRE binding and subsequent transcriptional repression. Surprisingly, we find that nGRE-binding was a feature of the ancestral 3-keto SR DBD, and this feature was selectively retained and improved in the GR lineage throughout SR evolution. We show that a small number of amino acid substitutions differentially affected nGRE- and (+)GRE-

binding among SR DBDs, leading the SRs to become a family of transcription factors with specific and diverse responses to steroid hormone signaling input.

4.4 Results

4.4.1 DNA substrates dictate conformation of the GR DBD

As a dual DNA/RNA binding protein, GR binds a diverse number of nucleic acid substrates - including (+)GRE DNA, nGRE DNA, mRNA, and the lincRNA Gas5 – through its DBD[109]. To determine if GR's nucleic acid substrate alters the conformation of its DBD, we performed 2D [^1H , ^{15}N] HSQC on ^{15}N -labeled human GR (hGR) DBD free in solution as well as bound to both a consensus (+)GRE and an nGRE from the *TSLP* promoter (Fig. 1A-B). As expected, GR binding to both the (+)GRE and nGRE caused marked chemical shift perturbations within the GR DBD. Strikingly, the GR DBD adopted distinct conformations when bound to an nGRE *versus* an (+)GRE, in agreement with the crystal structures solved of GR DBD – (+)GRE and GR DBD – nGRE complexes[110, 111] (Figure 4.4.1.1A). Residues comprising the GR DBD's dimerization loop, such as Ala458, Gly459, and Arg460, showed significant chemical shift perturbations when bound to (+)GRE DNA, but not nGRE DNA, supporting the notion that the GR DBD binds to the *TSLP* nGRE as two monomers[104]. Additionally, NMR peaks for arginine side chains appeared when the DBD was bound to nGRE, indicating altered rates of exchange of DBD-(+)GRE and DBD-nGRE complexes.

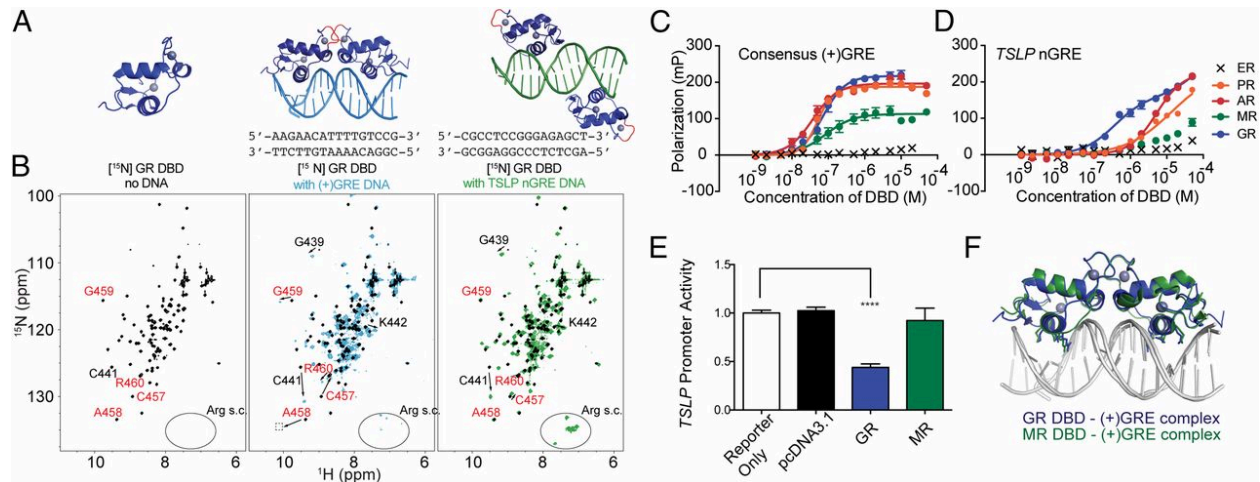


Figure 4.4.1.1 The glucocorticoid receptor (GR) DNA binding domain (DBD) adopts distinct conformations to activate and repress transcription.

(A) At left, the NMR structure of free GR DBD in solution[112] is shown. The GR DBD can dimerize on an inverted repeat (+)GRE element[111] to activate transcription (center), or can bind as two monomers on an nGRE (right) to repress transcription[104]. (B) NMR spectra of ^{15}N -labeled GR DBD in the absence of DNA (left), or bound to (+)GRE DNA (center) or nGRE DNA (right). DNA contacting residues (e.g. G439 or K442) are shifted upon nGRE or (+)GRE binding, but residues of the dimerization loop (red) are substantially shifted upon (+)GRE binding but not binding to the *TSLP* nGRE. Differences between (+)GRE- and nGRE-bound GR DBD occur at other residues, including the appearance of NMR peaks for arginine side chains when bound to nGRE (lower right), indicating altered rates of exchange of DBD-(+)GRE and DBD-nGRE complexes. Together, these results indicate that GR DBD adopts two distinct conformations in order to activate or repress transcription when bound to DNA (See also SI Appendix, Fig. S1). (C) All 3-keto SR DBDs bind to a consensus (+)GRE with nanomolar affinity, but only hGR binds to nGREs with nanomolar affinities (D). (E) While full-length hGR is capable of repressing a constitutively active nGRE-containing reporter, MR is not, despite high sequence and structural (F) similarity within the DBDs[110, 111].

Further 2D [^1H , ^{15}N] HSQC experiments with differing DBD:DNA molar ratios indicate that GR DBD – nGRE binding is characterized by two non-identical, monomeric binding events. In particular, GR DBD residues at the protein-DNA interface are affected by higher concentrations of GR DBD relative to nGRE DNA, whereas residues at the dimerization interface are not. This demonstrates that two dimerization-independent GR-DNA binding events occur between the GR DBD and the *TSLP* nGRE.

4.4.2 GR is the only SR capable of binding nGREs

The 3-keto SRs exhibit extremely high sequence conservation within the DBD. This sequence conservation is mirrored by functional conservation at (+)GREs: all 3-keto SRs are capable of binding to a consensus (+)GRE with high affinity (Figure 4.4.1.1C). However, *in vitro* binding experiments demonstrate that GR is the only SR capable of binding with nanomolar affinity to the *TSLP* nGRE (Figure 4.4.1.1D).

To test whether the inability of the non-GR 3-keto SRs to bind nGREs extends to cellular repression of nGRE-containing promoters, we tested the ability of MR - GR's closest paralog - to repress a constitutively active promoter containing the *TSLP* nGRE. In line with *in vitro* binding results, MR was unable to repress transcription from this element (Figure 4.4.1.1E), confirming that GR and MR exhibit a divergence in function at nGREs, which is quite remarkable considering their high sequence identity within the DBDs as well as the small size of the domain (75 amino acids). Moreover, the x-ray crystal structures of the GR[111] and MR[110] DBDs bound to (+)GRE DNA are superimposable, suggesting that subtle structural and evolutionary mechanisms underlie the functional differences among the 3-keto SRs (Fig. 1F).

4.4.3 DBD-nGRE binding and subsequent repression is a feature of the ancestral 3-keto SR

To trace the evolutionary history of divergent response element specificity among the SRs from their well-established phylogeny[19], we reconstructed sequences of ancestral DBDs from key nodes within the 3-keto SR evolutionary lineage. All ancestral sequences were strongly supported (mean posterior probabilities between 0.96 and 0.99 across sites) with very few ambiguously reconstructed residues.

We used overlap-extension PCR to insert these DBDs into the full-length hGR and tested their ability to both activate a simple (+)GRE reporter and repress the *TSLP* nGRE under a constitutively-active promoter. All extant and ancestral proteins within both the GR and MR lineage activated a (+)GRE reporter (Figure 4.4.3.1A-B). Surprisingly, the AncSR2 DBD, the common ancestor of all 3-keto SRs, was able to repress the *TSLP* nGRE from baseline, although this activity was not significant when corrected for multiple comparisons (Figure 4.4.3.1C). This slight repressive ability was retained in the AncCR DBD, the common ancestor of both GR and MR, lost in the ancestral MR (AncMR), and enhanced in the GR lineage (Figure 4.4.3.1C-D). In particular, a dramatic increase in the ability of GR to repress nGREs occurred between AncGR to AncGR2, which represents the GR protein in the common ancestor of all jawed vertebrates and all bony vertebrates, respectively [113] (Figure 4.4.3.1C).

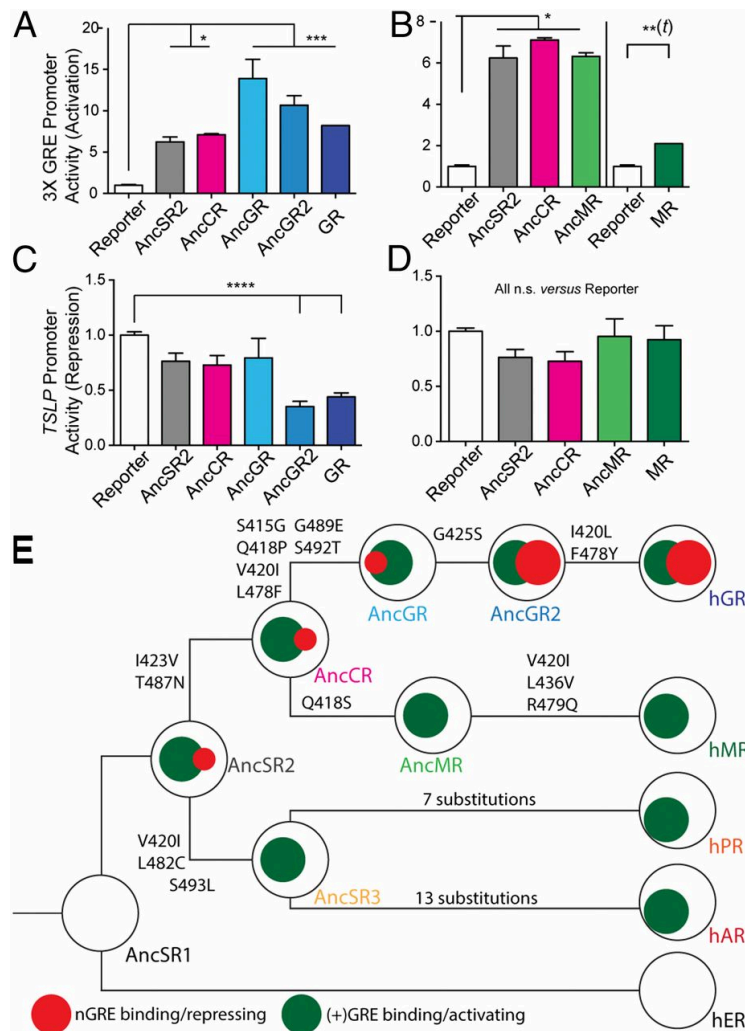


Figure 4.4.3.1 The GR lineage improved upon an ancestral cellular repressive function that was lost in MR.

(A) Reconstructed ancestral DBDs were inserted into full-length human GR using overlap-extension PCR and transfected into HeLa cells with the indicated reporter and treated with 1 μ M dexamethasone (or aldosterone, for human MR only). All DBDs in the GR lineage, including AncSR2, are capable of activating a (+)GRE reporter (B); similar results are found in the MR lineage (C). (D) The AncSR2 DBD is capable of repressing transcription from an nGRE-containing promoter, and this ability was retained in the GR lineage and further enhanced at the AncGR2 node. (E) The ability to repress at the *TSLP* nGRE was lost in the MR lineage at AncMR, consistent with *in vitro* binding data (See also SI Appendix, Fig. S3). (F) To summarize, (+)GRE and nGRE-binding and repressive ability of SR DBDs are mapped onto their phylogeny. (+)GRE binding was derived at the AncSR2 node [114], along with a moonlighting nGRE binding and *transrepressive* function. While (+)GRE binding was preserved throughout the clade, nGRE binding was lost at AncCR and AncSR3 and preserved (and enhanced) in the GR lineage. Green circles represent the ability of a given DBD to bind to and activate from a consensus (+)GREs. Red circles indicate the ability of a DBD to repress the *TSLP* nGRE in cells more or less than 50% (large and small circles, respectively). Above each branch are the amino acid substitutions between each node, using hGR numbering.

Remarkably few mutations led to divergence of function within this small protein domain – allowing a unique opportunity to pursue the detailed mechanisms by which functions are altered and distributed among paralogous proteins. To illuminate the biochemical and structural mechanisms by which repression at nGREs was selectively retained and enhanced in the GR lineage, we recombinantly expressed and tested all ancestral DBDs for *in vitro* binding to both (+)GRE and nGRE DNA. All ancestral DBDs bound to a consensus (+)GRE, as expected given the ability of all extant 3-keto SR DBDs to activate from these sequences (Figure 4.4.1.1E and ref. [115]). The ancestral 3-keto SR, AncSR2, bound nGREs with affinity in the high nanomolar range, implying that nGRE-binding originated at the ancestor of all 3-keto SRs and was then retained in GR. Unlike hGR, the AncSR2 DBD displayed no negative cooperativity on nGREs; instead, negative cooperativity of DBD-nGRE binding emerged gradually along the lineage from AncSR2 to hGR [104].

The ancestral capacity to bind the *TSLP* nGRE was lost along the lineage leading to AncSR3, the common ancestor of the androgen and progesterone receptors (Figure 4.4.3.1E). Only three historical amino acid substitutions occurred along this branch. We found that none of the individual changes affected the affinity of AncSR2 for nGREs or (+)GREs. When combined, however, the three mutations ablated nGRE binding, pointing to a strong epistatic interaction among historical substitutions. Along the other lineage - leading towards the GR and MR - the ancestral capacity to bind nGRE was retained in AncCR, ancestor of MR and GR, despite two amino acid changes in the DBD sequence. After the duplication of AncCR to produce separate GR and MR genes, nGRE binding was lost in the lineage leading to MR, consistent with cellular repression data (Figure 4.4.3.1C). A single historical substitution mediated this loss in function,

and our experiments indicate that this mutation interacted epistatically with the two earlier amino acid substitutions during the AncSR2-AncCR interval to abolish nGRE binding.

4.4.4 GR evolved enhanced inter-monomer allostery at nGREs

After establishing that GR alone retained an ancestral ability to bind to nGREs, we sought to obtain the structural mechanisms by which this expanded DNA specificity was uniquely retained and enhanced in the GR lineage. The crystal structure of the AncSR2 DBD – *TSLP* nGRE complex reveals that the GR-nGRE binding orientation and sequence specificity originated at the ancestor of all 3-keto SRs, before the emergence of vertebrates[116] (Figure 4.4.4.1A-B), despite the superior ability of GR DBD to repress nGRE-mediated transcription in cells (Figure 4.4.3.1C). NMR data from (+)GRE- and nGRE-bound GR (Figure 4.4.1.1) indicate that DBD conformation differs when bound to distinct response elements, and such allosteric changes have been shown to affect transcriptional output[111, 117]. Molecular dynamics (MD) trajectories followed by community and sub-optimal path analysis of the SR – DNA complexes show that AncSR2's daughter proteins in the GR and MR lineage contain more complex community organizations, resulting in diverging allosteric communication upon DNA binding among AncSR2-derived paralogs: at nGREs, the community connecting DBD monomers is much larger in hGR versus AncSR2, indicating that DNA-mediated inter-monomer communication and resulting negative cooperativity at nGREs[104] is enhanced in hGR compared to AncSR2 (Figure 4.4.4.1C-F). At (+)GREs, the community structure of AncSR2 bound to a (+)GRE is relatively simple compared to human MR (hMR) and hGR, with each monomer consisting of one large community (α and β) in direct allosteric communication, with some communication diverted through residues in the protein dimerization loops and through the DNA response element itself.

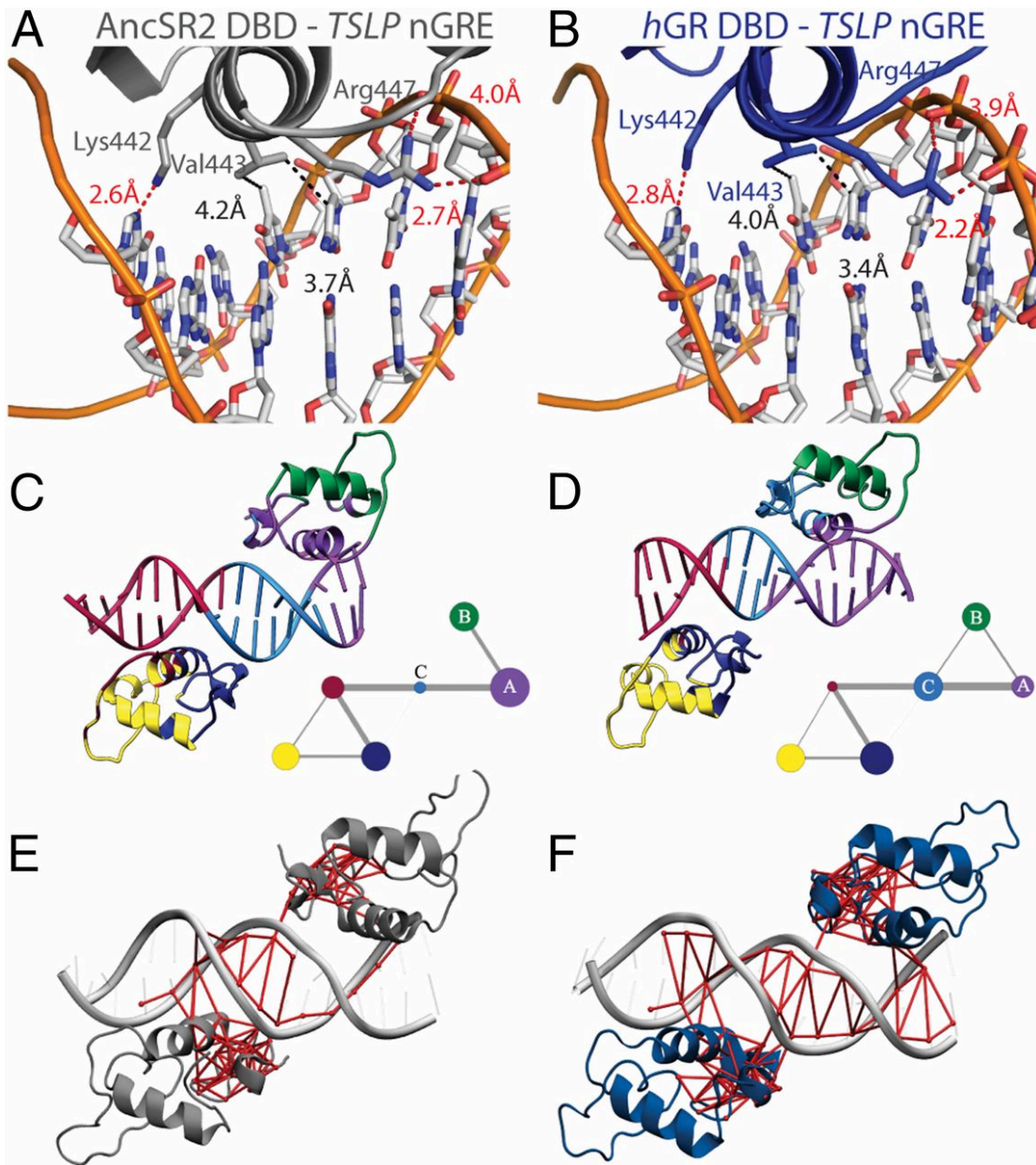


Figure 4.4.4.1 Although nGRE binding orientation, stoichiometry, and sequence specificity originated at the common ancestor of all 3-keto SRs, GR is capable of enhanced DNA-mediated allosteric communication at nGREs.

(A,B) hGR retains ancient amino acid-DNA contacts to ensure sequence specificity when binding nGREs, as shown by the AncSR2 – nGRE (c) and human GR – nGRE crystal structures [104]. (C,D) Community analysis of nGRE-bound AncSR2 and human GR DBDs[104]. At nGREs, each DBD monomer communicates through a central community (community C). Community C largely consists of DNA in the AncSR2 DBD – DNA complex, but this community is expanded to include part of the GR protein in the human GR DBD – DNA complex. (E, F) The larger community in hGR, relative to AncSR2, enhances communication between DBD monomers at nGREs, as also shown by suboptimal path analysis between the AncSR2 monomers (panel E; 4,158 pathways) and human GR monomers (panel F; 26,165 pathways) when nGRE-bound. This enhanced communication correlates to an increase in negative cooperativity at nGREs observed throughout the GR lineage.

In contrast to both AncSR2 and hGR, such extensive DNA-mediated communication between DBD monomers is not a feature of the MR branch of the SR phylogeny. MD trajectories and resulting community analysis of the AncMR – (+)GRE and MR – (+)GRE crystal structures[110], reveal a qualitatively different community organization and routes of communication, compared to their GR paralogs. A greater fraction of the total communication between AncMR monomers occurs directly through the dimerization interface, as opposed to through the DNA, relative to the GR lineage – a trend more pronounced in hMR. Given the link between transcriptional output and allosteric communication[111, 117], the fracturing of the community organization and weakening of allosteric cohesiveness leads to a network in AncMR and hMR with decreased intra-DNA allosteric communication. This correlates with weaker binding to nGRE sequences, where protein dimerization does not occur. It is noteworthy that all amino acid substitutions driving these allosteric changes occurred far from the DNA binding surface, which is likely necessary to maintain (+)GRE.

4.4.5 Subtle, irreversible structural changes enhanced nGRE binding

To understand the effects of the key historical substitutions on nGRE binding, we determined the X-ray crystal structures of several ancestral DBDs in complex with a nGRE or (+)GRE. We found that Gly425Ser - the key substitution in the GR lineage that enhanced repression at nGREs (Figure 4.4.3.1) - is involved in a change in conformational freedom of the DBD. In the AncGR2 – DNA crystal structure, the sidechain of Ser425 is solvent exposed and makes no contacts with DNA or the remainder of the GR DBD (Figure 4.4.5.1A-B). Due to the conformational freedom granted by glycine residues (Figure 4.4.5.1C), we hypothesized that subtle changes in backbone conformation may underlie the large effect caused by the Gly425Ser substitution. All extant and ancestral 3-keto SRs (other than GRs) contain the ancestral glycine

and occupy glycine-only backbone conformations when bound to (+)GREs, as visualized by a Ramachandran plot (Figure 4.4.5.1D-F). Intriguingly, the AncSR2 – (+)GRE crystal structure[114] contains two dimers; Gly425 of one dimer occupies glycine-only Ramachandran space and Gly425 of the second dimer occupies general Ramachandran space (Figure 4.4.5.1D), indicating that AncSR2 may be more dynamic and able to occupy a wider range of conformational ensembles. The Gly425Ser substitution at the AncGR2 node locked the GR lineage into a restricted subset of Ramachandran space, as compared to the MR and AncSR3 lineages (Figure 4.4.5.1E-F). MD trajectories show that position 425 of AncGR2 occupies a separate subset of Ramachandran space from AncGR, eliminating any artifacts from crystal packing (Figure 4.4.5.1G). As a result of these conformational changes, community analysis reveals that the Gly425Ser substitution decreases direct communication of the two DBD monomers via the dimerization interface and instead increases inter-protein communication via DNA on activating elements, a trend that is preserved in hGR.

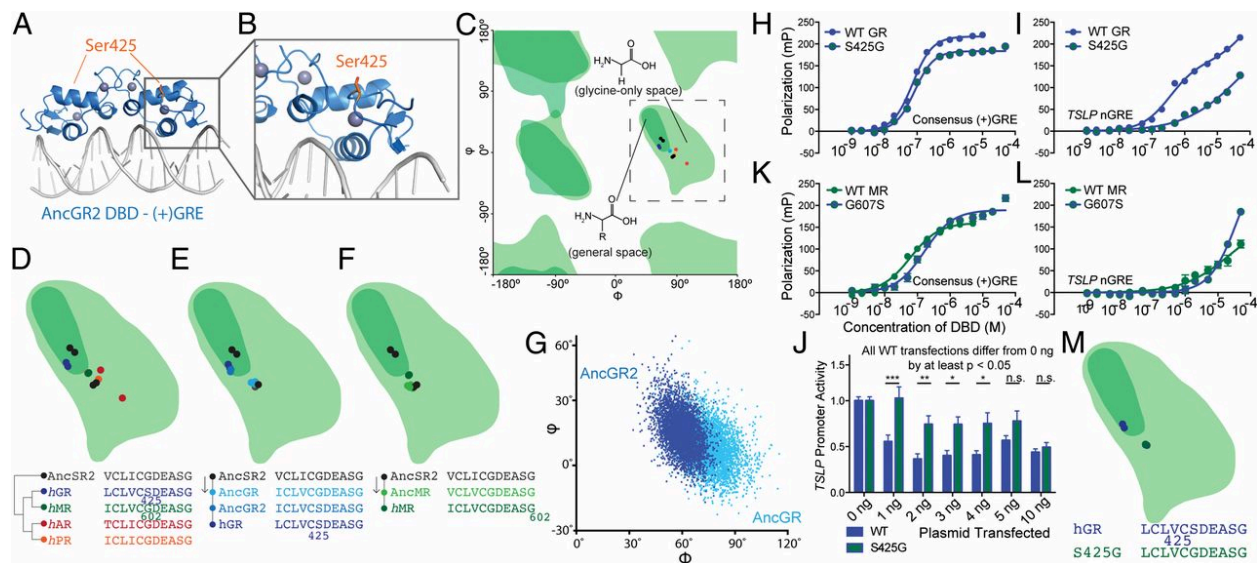


Figure 4.4.5.1 A single amino acid substitution far from the DNA binding interface – Ser425Gly – led to an improvement in nGRE binding through subtle effects in SR backbone conformation.

(A) Location of Ser425 within the overall crystal structure of the AncGR2 DBD – (+)GRE complex; Ser425 is solvent exposed and contacts no other side chains in AncGR2 (B). (C) Ramachandran plot of the AncSR2 and extant human DBDs. Non-glycine (“general”) amino acids in a protein polymer are restricted in their backbone conformations (phi/psi angles) as compared to glycine residues (dark vs. light green, respectively). (D); AncSR2 and hGR DBDs occupy general Ramachandran space at position 425, in contrast to the other three extant human DBDs. (E) hGR’s conformational space was decided at the AncGR2 node, where the Gly425Ser substitution occurred. (F) MR retains a glycine at the homologous position, and assumes a backbone conformation only accessible to glycine residues. (G) Backbone conformation of AncGR- and AncGR2-(+)GRE complexes differs over the course of 200 ns MD trajectories. (H,I) Comparison of binding between WT and S425G GR DBD at a consensus (+)GREs and the *TSLP* nGRE. (J) Reporter gene assay of a constitutively active reporter gene controlled by the *TSLP* nGRE; full-length WT or Ser425Gly GR were transfected at indicated amounts into HeLa cells co-transfected treated with dexamethasone. (K,L) Comparison of binding between WT and Gly607Ser MR DBD at a consensus (+)GRE and the *TSLP* nGRE. (M) The Ser425Gly mutation reverts GR to an MR-like backbone conformation.

Reversal of this substitution (Ser425Gly) in hGR causes the protein to retain high-affinity binding to a (+)GRE but to lose much of its affinity for nGREs (Figure 4.4.5.1H-I); similarly, in cells, full-length GR with the Ser425Gly mutation is a much less potent repressor of the *TSLP* nGRE (Figure 4.4.5.1J). Together, these data indicate the two amino acid changes between AncGR2 and hGR, Ile420Leu and Phe478Tyr, locked hGR into dependence on the Gly425Ser substitution for even low-affinity binding to nGREs, a hypothesis supported by sequence alignments and similar to the evolutionary ‘ratchet’ observed in SR ligand binding domains[118, 119]. The crystal structure of the GR DBD Ser425Gly mutant bound to DNA reveals that residue 425 reverts to a glycine-only conformation (Figure 4.4.5.1M), confirming that backbone conformation is the likely mechanism for the large-effect seen by the Gly425Ser substitution. Even when not bound to DNA, 2D [¹H,¹⁵N] HSQC NMR experiments reveal that reversal of the Gly425Ser substitution in the hGR DBD results in large conformational changes in residues comprising the DNA-binding interface, consistent with the historical substitution’s effects on DNA-mediated inter-protein communication. Mutation of residue 425 to a non-glycine amino acid, such as alanine, does not cause a loss of nGRE binding, confirming that backbone dynamics are responsible for the effects of the Gly425Ser substitution. Intriguingly, the Ser425Gly mutation has been widely studied in the context of hGR and shown to have deleterious effects on the repression of T-bet, AP-1, and NF-κB by hGR[120-123], suggesting DNA binding-mediated effects play a larger role in pro-inflammatory transcriptional repression by the GR than has been previously assumed.

Although the historical Gly425Ser substitution conferred nGRE-mediated repression at the AncGR-AncGR2 transition, it is not sufficient to improve affinity for nGREs when introduced into the human MR (Figure 4.4.5.1K-L). This result indicates an additional role for

epistatic mutations during steroid receptor DBD evolution: either mutations occurred in the lineage leading to AncGR - but not that to MR - which were necessary for Gly425Ser to yield repression at nGREs, or restrictive mutations occurred in the lineage leading to MR, which prevented Gly425Ser from having that same effect. In either case, amino acid substitutions that do not affect function on (+)GREs modified the capacity of other mutations to affect function nGREs, changing the evolutionary potential of the protein in a lineage-specific fashion. Together, the epistatic and function-switching mutations caused the diverging paralogs to follow evolutionary paths that yielded similar transcription factors that control distinct transcriptional programs and effect specific responses to signaling input.

4.5 Discussion

These results illuminate the mechanism by which closely related transcription factors with similar function and high sequence identity can evolve distinct specificity to alternate response elements. In this case, a moonlighting ancestor, the AncSR2 DBD, developed a weak affinity for and repressive ability at nGREs, likely as it evolved toward a (+)GRE-binding phenotype[114]. This nGRE-binding ability was subsequently lost at two independent time points during 3-keto SR evolution, but maintained and enhanced in the lineage leading to modern-day hGR (Figure 4.4.3.1). These findings strongly support the hypothesis that subdivision of ancestral promiscuous functions is an important mechanism of divergent function among paralogs[124].

The detailed biochemical, structural, and molecular dynamics studies described here present a unified view of the molecular basis for alternate response element specificity among paralogous transcription factors. In this case, enhancement of nGRE-binding ability in the GR lineage and its loss in the MR and AncSR3 lineages were critical for development of divergent

DNA specificity. Both computational and directed evolution studies have implicated epistasis as a primary factor in molecular evolution[125, 126], and studies of historical protein divergence have established a major role for epistasis in specific cases[119, 127-132].

In the two loss-of-function events described here, three amino acid substitutions combine to accomplish a loss of nGRE binding with no effects on (+)GRE binding. The temporal order in which these substitutions occurred can be partially resolved on the phylogeny: the two earliest-occurring changes did not change binding to either element, but once they were in place, addition of the third and last substitution abolished nGRE binding and regulation. These findings demonstrate that initially ‘neutral’ amino acid substitutions lay the necessary groundwork for later function-changing mutations with which they interact epistatically. The mutations that drove functional divergence among the 3-keto SRs were far from the DNA-reading α -helix, presumably because binding to both nGREs and (+)GREs requires the same DNA-binding interface of the SR proteins, likely imposing strong selective constraint on this region of the protein in all DBDs in the family.

Functional promiscuity and conformational dynamism have been hypothesized to contribute to protein evolvability [124, 133], but there have been very few direct studies of the historical evolution of a protein’s occupancy of the ensemble of potential conformers[134, 135]. Our findings reinforce the importance of these factors. We found that the ancestral 3-keto SR, AncSR2, was capable of binding both nGREs and (+)GREs (Figure 4.4.3.1). To bind both response elements, our NMR experiments show that the hGR DBD accesses two conformational states (Figure 4.4.1.1A-B), and it is likely that other members of the family that exhibit nGRE binding - including AncSR2 - utilize a similar mechanism for their promiscuous functions. Our observations suggest that the ancestral promiscuity and conformational dynamism was retained

and refined in hGR, and is likely a feature of the many DNA/RNA binding proteins encoded within the human genome[109]. Our results are consistent with previous hypotheses that ‘evolvable’ proteins may exist in closely-related but functionally distinct conformers whose distribution may be easily perturbed by mutation[136]. In the case of the SRs, the Gly425Ser substitution at the AncGR to AncGR2 transition led to a discrete partitioning of Ramachandran space between the GR and MR lineage (Figure 4.4.5.1C-F). Despite the significant distance of residue 425 from the DBD’s DNA binding interface (Figure 4.4.5.1A-B), its alteration was sufficient to change the evolving proteins’ occupancy of conformational space and confer alternate DNA binding specificity on MR and GR. In this way, two paralogous DNA binding domains evolved specific control over distinct transcriptional programs, despite their high primary sequence identity and structural similarity.

4.6 Materials and Methods

4.6.1 Protein expression and purification

Proteins were expressed as described previously[104, 110]. Proteins were cloned into pMCSG7 vector, which contains an N-terminal 6X His tag. The expression vector was transformed into BL-21(DE3)pLysS *E. coli*, which were grown in TB media and induced with 300 μ M IPTG at an OD of \sim 0.8 for 4 hours at 30 $^{\circ}$ C. Cells were lysed via sonication in 20 mM Tris 7.4, 1 M NaCl, 25 mM imidazole, and 5% glycerol. SR DBDs were purified from the supernatant using a nickel affinity chromatography column (HisTrap) with fast protein liquid chromatography (FPLC). Gel filtration via FPLC was used to further purify the DBDs into a buffer of 20 mM Tris 7.4, 100 mM NaCl, and 5% glycerol. Protein was concentrated to \sim 4 mg/ml and flash frozen in liquid N₂ until further use.

4.6.2 Protein – DNA binding assays

All DNA oligos for this study were purchased from Integrated DNA Technologies. For DNA binding assays, increasing amounts of indicated protein were added to 10 nM of 5' carboxyfluorescein-labeled DNA oligos and fluorescence polarization values were measured using a Biotek Synergy plate reader. All binding experiments were performed in 100 mM NaCl, 20 mM Tris-HCl pH 7.4, and 5% glycerol. Sequences used for binding assays were: (+)GRE: 5'-(FAM)CCAGAACAGAGTGTCTGA-3' and 5'-TCAGAACACTCTGTTCTGG-3'; *TSLP* nGRE: 5'-(FAM)CCGCCTCCGGGAGAGCTG and 5' - CAGCTCTCCCGGAGGCGG - 3', where (FAM) indicates the position of the carboxyfluorescein dye. hGR-*TSLP* nGRE and hMR – (+)GRE binding data were also reported in refs [104, 110]. GraphPad Prism was used for fitting of data, and cooperativity was calculated using the one-site specific binding model with hill slope. In all panels, points or bars indicate mean, and error bars represent s.e.m.

4.6.3 Crystallization and structure determination

All structures were crystallized via hanging drop vapor diffusion. The AncSR2 – *TSLP* nGRE complex was crystallized in 0.1 M HEPES (pH 7.5), 10% PEG 20000, 5% glycerol, and 5% ethanol. The AncGR – (+)GRE complex was crystallized in 0.1 M HEPES (pH 7.5), 12% PEG 20000, and 5% glycerol. The AncGR2 – (+)GRE complex was crystallized in 0.1 M HEPES (pH 7.5) and 15% PEG 8000. The AncMR – (+)GRE complex was crystallized in 0.1 M MES (pH 6.5), 20% PEG 6000, and 5% glycerol. The GR Ser425Gly – (+)GRE complex was crystallized in 0.1 M HEPES (pH 7.5), 20% PEG 8000, and 4% ethylene glycol. Proteins were flash-cooled in liquid N₂ after soaking in cryoprotectant consisting of the crystallization condition plus additional PEG and glycerol. DNA constructs used for crystallization were 5' - CGCCTCCGGGAGAGCT - 3' and 5' – AGCTCTCCCGGAGGCG - 3' for the AncSR2 - *TSLP*

nGRE complex; and 5'- TCAGAACACTCTGTTCTG -3', and 5'- CCAGAACAGAGTGTCTG -3' for the AncGR-, AncGR2- AncMR- and hGR Ser425Gly-(+)GRE complexes.

Data were collected at Southeast Regional Collaborative Access Team (SER-CAT) 22-ID (or 22-BM) beamline at the Advanced Photon Source (APS), Argonne National Laboratory, Argonne, IL, USA. Supporting institutions may be found at www.ser-cat.org/members.html. Data were processed using HKL-2000 and phased with molecular replacement using PHASER in the PHENIX suite, and refined using phenix.REFINE[137, 138]. COOT was used for model building[139], and both PyMOL (Schrodinger, LLC) and Chimera[140] (Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco) were used for structure visualization and figure generation. The PDB_REDO server was also used for structure optimization and validation[141]. Sample electron density for all structures reported here is located in Fig. S12.

4.6.4 Molecular dynamics simulations and network analysis

Eight systems were prepared for molecular dynamics (MD) simulation: (+)GRE-AncSR2 (PDB 4OOR), AncGR (PDB 5CBX), AncGR2 (PDB 5CBY), hGR (PDB 3G6R), AncMR2 (PDB 5CBZ), hMR (PDB 4TNT); nGRE- AncSR2 (PDB 5CC0), and hGR (PDB 4HN5). Each protein monomer was capped with acetyl and N-methyl groups at the N- and C-termini, respectively. The complexes were placed in a rectilinear simulation box extending no less than 12 Å from the protein-DNA along each dimension. The systems were solvated with TIP3P[142] water molecules. Na⁺ and Cl⁻ ions were added to maintain neutral charge with excess 0.15M NaCl concentration to mimic physiological conditions. Systems were set up with

the xLeap module of the AmberTools11[85] package with the parm99-bsc0 forcefield[143, 144]. Cys-Zn²⁺ tetrahedral geometries in the zinc fingers were maintained using dummy atoms.

All minimization and MD was performed with the NAMD2.9 simulation package[145] using a 2-fs timestep, the r-RESPA method[146] for force integration, and the SHAKE[147] algorithm to restrain bonds all bonds involving hydrogen atoms. Non-bonded interactions were calculated within a 12 Å cutoff, with a switching function applied between 10 Å and 12 Å. Long-range electrostatics were treated with the smooth particle mesh Ewald (PME) scheme[148], with full electrostatics calculated every 2 steps and 1-4 interaction scaling set at 0.83. Each system was subjected to 10,000 steps of conjugate gradient minimization, followed by a 200 ps MD simulation in the NVT ensemble, with 5 kcal/mol×Å² harmonic restraints on all protein and DNA heavy atoms, while smoothly heating from 0-300 K. Seven stages of restraint release were then carried out in the NPT ensemble in 1 ns segments, incrementally releasing first protein sidechains, followed by protein backbone, DNA nucleosides, and finally DNA backbone heavy atoms. An unrestrained, 20 ns NPT simulation was performed to fully equilibrate the system. Each system was then sampled for 200 ns in the NPT ensemble. 10,000 evenly spaced frames were taken from each trajectory for network analysis in the NetworkView[149] plugin of the VMD[150] visualization and analysis program.

Network theory was employed to highlight evolutionary alterations in the allosteric network of 3-KS complexes. Networks are constructed by defining all protein alpha-carbon and DNA C1' atoms in a system as nodes and using Cartesian covariance as a measure of communication within the network. Any pair of nodes that reside within a 4.5 Å cutoff for more than 75% of the MD trajectory are connected via an edge, with the weight of the edge being proportional to the covariance between the nodes. From this raw data, networks were resolved

into communities, which are groups of nodes with correlated motions, using the Girvan-Newman algorithm. While constructing the community graphs, the minimum number of communities possible were generated while maintaining at least 92.3% of the maximum modularity[47] to prevent excessive community subdivision. The magnitude of communication flow between communities was quantified by the total betweenness[151] of all edges that transition between communities. In the case of the nGRE complexes, suboptimal paths between monomers and through the DNA were identified using the Floyd-Warshall algorithm[152].

4.6.5 Cellular activation and repression assays

HeLa cells were transfected with 10 ng of the indicated receptor, 50 ng of the indicated firefly luciferase reporter, and 10 ng of *Renilla* luciferase under the control of the constitutively active pRL-TK promoter. For transfection, OptiMEM media without antibiotics was used with FuGene HD, according to the manufacturer's protocols. Otherwise, HeLa cells were passaged in MEM α (Life Technologies) supplemented with 10% stripped FBS (Atlanta Biologicals) and penicillin/streptomycin (Life Technologies). Twenty-four hours after transfection, cells were treated with 1 μ M dexamethasone, except for hMR conditions in which aldosterone was used. Twenty-four hours after ligand treatment, firefly and *Renilla* luciferase activities were measured using the DualGlo kit (Promega) according to the manufacturer's protocol on a Biotek Synergy plate reader.

In all figures, *, **, ***, and **** indicate $P < 0.05$, 0.01, 0.001, and 0.0001, respectively and points or bars indicate mean, and error bars represent s.e.m. In Figure 4.4.1.1F and Figure 4.4.3.1, one-way ANOVA followed by Tukey's post-hoc test was performed, except for hMR vs reporter in Figure 4.4.3.1C (Student's *t* test), where the full-length MR protein was used. For Figure 4.4.5.1J, two-way ANOVA with Sidak's multiple comparisons test was used.

4.6.6 NMR

¹⁵N-labeled GR DBD and S425G were expressed in *Escherichia coli* BL21(DE3) pLysS cells as TEV-cleavable hexahistidine-tagged fusion proteins using M9 media, with ¹⁵NH₄Cl as the sole nitrogen source. Proteins were purified through a Ni-NTA column against a 500 mM imidazole gradient, using wash buffer containing 50 mM Tris (pH 8.0), 100 mM NaCl, 15 mM imidazole, 1 mM TCEP. The hexahistidine tag was subsequently cleaved with TEV protease overnight at 4 °C. The protein solutions were passed through the Ni NTA column anew, and the flow through containing purified protein was collected. Proteins were verified to be >99% pure by SDS-PAGE.

NMR data were collected on a 700 MHz Bruker NMR instrument equipped with a QCI cryoprobe. For DNA experiments, the 19-nt GRE and TSLP nGRE DNA duplexes were reconstituted in 20 mM phosphate (pH 6.7), 100 mM NaCl, 1 mM TCEP, 10% D₂O buffer to 1.7-2.0 mM, subsequently annealed by denaturing at 95 °C for 3 minutes and equilibrated to room temperature (20-23°C) overnight. For protein experiments, 2D [¹H,¹⁵N]-HSQC spectra were collected at 25 °C for ~250 uM of free ¹⁵N-labelled GR DBD protein and complexed with 1.5:1 of GRE or 0.44:1/2.1:1 of TSLP nGRE DNA duplex in the same NMR buffer. Chemical shift perturbations were assigned using previously published GR DBD NMR chemical shifts[117] and calculated using the minimum chemical shift perturbation procedure. Data were processed using Bruker Topspin and analysed with NMRViewJ (OneMoon Scientific, Inc).

4.6.7 Phylogenetics and ancestral sequence reconstruction

Annotated protein sequences for nuclear receptors were downloaded from UniPROTKB/TrEMBL, GenBank, the JGI genome browser, and Ensembl. Australian lungfish (*Neoceratodus forsteri*) receptor sequences were amplified by degenerate PCR from frozen

tissue samples (kind gift of Angel Amores). Steroid receptor sequences were amplified by degenerate PCR, and full-length sequences were obtained by 5' and 3' nested rapid amplification of cDNA ends (RACE) using the SMART-RACE kit (BD Biosciences, Palo Alto, CA). A total of 234 steroid and related receptor sequences containing both DNA binding and ligand binding domains were aligned using the Multiple Sequence Alignment by Log-Expectation (MUSCLE) program[153]. The alignment was checked to ensure alignment of the nuclear receptor AF-2 domain and manually edited to remove lineage-specific indels. The N-terminal variable region and hinge region were removed from the alignment file, as these areas could not be aligned reliably among sequences.

Phylogenies were inferred from these alignments using PHYML v2.4.5[154] and the Jones-Taylor-Thornton model with gamma-distributed among-site rate variation and empirical state frequencies, which was the best-fit evolutionary model selected using the Akaike Information Criterion implemented in PROTTEST software. Statistical support for each node was evaluated by the chi-squared confidence statistic derived from approximate likelihood ratio calculations[155].

The ancestral DBDs were reconstructed by the maximum likelihood method[156] on a two-branch rearrangement of the ML phylogeny that requires fewer gene duplications and losses to explain the distribution of SRs in agnathans and jawed vertebrates using the Codeml module of PAML v3.14[157] and Lazarus software[158] assuming a free eight-category gamma distribution of among-site rate variation and the Jones-Taylor-Thornton protein model. Average probabilities were calculated across all DBD sites except those containing indels.

CHAPTER 5. UNEXPECTED ALLOSTERIC NETWORK CONTRIBUTES TO LRH-1 COREGULATOR SPECIFICITY

5.1 Summary

Phospholipids (PLs) are unusual signaling hormones sensed by the nuclear receptor liver receptor homolog-1 (LRH-1), which has evolved a novel allosteric pathway to support appropriate interaction with coregulators depending on ligand status. LRH-1 plays an important role in controlling lipid and cholesterol homeostasis and is a potential target for the treatment of metabolic and neoplastic diseases. While the prospect of modulating LRH-1 via small molecules is exciting, the molecular mechanism linking PL structure to transcriptional coregulator preference is unknown. Previous studies showed that binding to an activating PL-ligand, such as dilauroylphosphatidylcholine (DLPC), favors LRH-1's interaction with transcriptional coactivators to upregulate gene expression. Both crystallographic and solution-based structural studies showed that DLPC binding drives unanticipated structural fluctuations outside of the canonical activation surface in an alternate activation function (AF) region, encompassing the β -sheet-H6 region of the protein. However, the mechanism by which dynamics in the alternate AF influences coregulator selectivity remains elusive. Here we pair x-ray crystallography with molecular modeling to identify an unexpected allosteric network that traverses the protein ligand binding pocket and links these two elements to dictate selectivity. We show that communication between the alternate AF region and classical AF2 is correlated with the strength of the coregulator interaction. This work offers the first glimpse into the conformational dynamics that drive this unusual PL-mediated nuclear hormone receptor activation.

5.2 Introduction

Phospholipids (PLs) are best known for their structural role in membranes and as synthesis material for potent signaling molecules, such as diacylglycerol, leukotrienes, and inositol phosphates. Recent evidence, however, suggests intact PLs are able to directly modulate the activity of transcription factors involved in lipid homeostasis, such as sterol regulatory element-binding protein 1 (SREBP-1), and some members of the nuclear receptor (NR) family of ligand-regulated transcription factors, including peroxisome proliferator activated receptor α (PPAR α ; NR1C1), steroidogenic factor 1 (SF-1; NR5A1) and human liver receptor homologue-1 (LRH-1; NR5A2) [159-162]. LRH-1 regulates the expression of genes central to embryonic development, cell cycle progression, steroid synthesis, lipid and glucose homeostasis, and local immune function [163-170]. Thus, LRH-1 is an enticing pharmaceutical target for the treatment of metabolic and neoplastic diseases [164].

Although the endogenous ligand for hLRH-1 is currently unknown, oral treatment with the exogenous PL agonist dilauroylphosphatidylcholine (PC 12:0-12:0; DLPC) lowers serum lipid levels, reduces liver fat accumulation, and improves glucose tolerance in a LRH-1 dependent manner in a diabetic mouse model [171]. Activation of LRH-1 by DLPC drives increased glucose uptake by muscle and increases the rate of both glycolysis and glycogen synthesis with a concomitant reduction in fatty acid metabolism [172]. These observations suggest LRH-1 agonists may resolve glucose homeostasis related-diseases. New evidence suggests that LRH-1 may also be targeted to relieve chronic ER stress. Activation of LRH-1 by synthetic DLPC or the small molecule RJW100 induces *Plk3*, which is required for the activation of ATF2 and the induction of its target genes, which play a key role in resolving ER stress [173]. Given its potential therapeutic value, LRH-1 has been the subject of multiple attempts to identify

small molecule modulators [174-177]. These attempts have been met with mixed success due in part to our limited understanding of LRH-1's mechanism of activation.

We have shown that DLPC is able to bind directly to the LRH-1 ligand binding domain (LBD) and activate the receptor by affecting receptor dynamics in an alternate activation function (AF) region, encompassing the β -sheet-H6 region of the protein, to alter co-regulator binding preference [178]. Importantly, it seems that DLPC may promote activation by relieving LRH-1 from repression by the non-canonical co-repressor NR SHP, which mimics a co-activator using the canonical Leu-X-X-Leu-Leu (where X is any amino acid) nuclear coactivator interaction motif [179, 180]. In the absence of ligand, the alternate AF is highly dynamic and mutations that restrict motion in this region ablate transactivation [178]. SHP is a robust corepressor of LRH-1-mediated transactivation in the liver can recognize both apo LRH-1 and LRH-1 when bound to a non-ideal ligand such as bacterial PLs *in vitro* [179, 181, 182]. It is unclear how LRH-1 discriminates between SHP and coactivators such as TIF2 that bind using a similar LxxLL motif to recognize the active NR orientation. Further, how does human LRH-1 recognize coactivators in the absence of ligand? How do PLs varying only in their acyl tail composition show differing abilities to drive transactivation? Which ligand/coregulator states are appropriate for *in silico* ligand design?

This incomplete understanding of what dictates LRH-1's PL and coregulator selectivity limits our ability to guide the design of robust small molecule modulators for this intriguing pharmacological target. To address these questions, we have generated a novel crystal structure of the LRH-1-TIF2 complex in an apo state, as well as a higher resolution structure of LRH-1 bound to *E. coli* PLs. These crystal structures, in combination with novel lipid binding assays, molecular dynamics simulations and principle component analysis (PCA) have allowed us to

identify an unexpected allosteric network that may contribute to PL-mediated NR signaling and coregulator selectivity.

5.3 Experimental Procedures

5.3.1 Reagents

Chemicals were purchased from Sigma, Fisher or Avanti PLs. pMALCH10T and the vector for His tagged TEV were a gift from John Tesmer (UT Austin). pLIC_MBP and pLIC_HIS were gifts from John Sondek (UNC, Chapel Hill). Peptides were synthesized by RS Synthesis (Louisville, KY). DNA oligonucleotide primers were synthesized by IDT (Coralville, IA USA).

5.3.2 Protein expression and purification

The human LRH-1 LBD (residues 291–541) was purified as described previously [183]. Purified protein was dialyzed against 60 mM NaCl, 100 mM ammonium acetate (pH 7.4), 1 mM DTT, 1 mM EDTA and 2 mM CHAPS and concentrated using centrifugal filters with a 10-kDa cutoff to 5–7 mg ml⁻¹. For apo LRH-1 crystallization, purified LRH-1 LBD was incubated with 1,2- ditetracosanoyl-sn-glycero-3-phosphocholine (PC 24:0–24:0) (Avanti Polar Lipids) and GSK8470, a weak and labile agonist, at a final PC24:ligand:protein ratio of 20:3:1 [175]. The receptor was purified away from unbound PC 24:0–24:0 and the weakly bound agonist by size exclusion chromatography, dialyzed against 60 mM NaCl, 100 mM ammonium acetate, pH 7.4, 1 mM DTT, 1 mM EDTA and 2 mM CHAPS and concentrated to 5–7 mg ml⁻¹.

5.3.3 Structure determination

Both the apo LRH-1 LBD–TIF2 complex and the LRH-1 LBD–*E. coli* PL–TIF2 complex crystals were generated by hanging-drop vapor diffusion at 20 °C from solutions containing 1 µl of protein at 6.5 mg ml⁻¹ in complex with a peptide derived human TIF2 NR box 3 (+H3N-

KENALLRYLLDKDD-CO2-) at a 1:4 molar ratio and 1 μ l of the following crystal mixture: 0.7-1 M di-Sodium Malonate, 0.1 M HEPES pH 7.4, 0.5% Jeffamine ED-2001. Crystals were cryoprotected in crystallant containing 20% (v/v) glycerol and flash-frozen in liquid N₂. Data for the apo LRH-1 LBD-TIF2 NRBox3 complex were collected to 1.75 Å resolution at 100 K using a wavelength of 0.9999 Å at 22-BM at the Southeast Regional Collaborative Access Team (SER-CAT) at the Advanced Photon Source and were processed and scaled with HKL2000 [184]. Data for the LRH-1 LBD-*E. coli* PL-TIF2 complex were collected to 1.75 Å resolution at 100 K using a wavelength of 0.9999 Å at 22-ID at the Southeast Regional Collaborative Access Team (SER-CAT) at the Advanced Photon Source and were processed and scaled with HKL2000 [184]. Initial phases for both structures were determined using LRH-1 PDB 1YOK as a molecular replacement search model. The structures were refined using the PHENIX suite of programs, and model building was carried out in COOT [185, 186]. The final model for the LRH-1-TIF2 complex contains LRH-1 residues 300-538 and TIF2 residues 742-752; it shows good geometry, with 98.4% and 1.6% of the residues in the favored and allowed regions of the Ramachandran plot, respectively. The final model for the LRH-1-*E. coli* PL-TIF2 NRbox3 complex contains LRH-1 residues 298-538 and TIF2 residues 743-750; it shows good geometry, with 98.7% and 1.3% of the residues in the favored and allowed regions of the Ramachandran plot, respectively. Data collection and refinement statistics are listed in Table 1. Coordinates and structure factors have been deposited with the Protein Data Bank under accession codes 4PLD and 4PLE.

Table 5.1. Data collection and refinement statistics (molecular replacement)

	LRH-1 – TIF2 NRbox3	LRH-1 – <i>E. coli</i> PL – TIF2 NRbox3
Data collection		
Space group	P2 ₁ 2 ₁ 2 ₁	P2 ₁
Cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	45.8, 65.7, 83.5	65.9, 76.9, 100.8
(°)	90.0, 90.0, 90.0	90.0, 95.5, 90.0
Resolution (Å)	1.75 (1.81 – 1.75)*	1.75 (1.81 – 1.75)*
<i>R</i> _{merge}	6.6 (30.6)	6.6 (30.9)
<i>I</i> / <i>sI</i>	18.99 (2.8)	12.8 (3.2)
Completeness (%)	99.4 (96.22)	92.6 (63.8)
Redundancy	3.9 (3.3)	3.6 (3.2)
Refinement		
Resolution (Å)	1.75	1.75
No. reflections	25933	6751
<i>R</i> _{work} / <i>R</i> _{free}	18.7 / 22.4	20.67 / 23.4
No. atoms		
Protein	2026	8117
Ligand/ion	42	493
Water	137	378
<i>B</i> -factors		
Protein	23.9	27.0
Ligand/ion	29.2	37.5
Water	29.4	32.7
R.m.s. deviations		
Bond lengths	0.008	0.006
(Å)		
Bond angles (°)	1.41	1.03
PDB	4PLD	4PLE

*Data collected from a single crystal. Values in parentheses are for highest-resolution shell.

5.3.4 Local conformational analysis

ProSMART is an alignment tool that provides a conformation-independent structural comparison of two proteins based upon the alignment of corresponding overlapping fragments of the protein chains [187]. We performed ProSMART analyses among five LRH-1 structures with different bound ligands and coregulator peptides, representing different activation states: apo-SHP (fully repressed; PDB: 4DOR), apo-TIF2, *E. coli* PL-SHP (PDB: 1YUC), *E. coli* PL-TIF2, and DLPC-TIF2 (fully activated; PDB: 4DOS). This allowed for a detailed analysis of the local structural dissimilarities between two proteins independently of their global conformations. The local backbone conformation of available LRH-1 crystal structures were compared to generate the Procrustes score, which is the r.m.s.d. of the central residue of two corresponding structural fragments of length n , where n is an odd number of amino acids

5.3.5 Synthesis of NBD-DLPE

DLPE (dilauroylphosphatidylethanolamine; 50 mg, 90 μ mol), NBD-Cl (4-chloro-7-nitrobenzofuran; 50 mg, 250 μ mol), and triethylamine (17.5 μ L) were dissolved in 5 mL 1:1 CHCl_3 :MeOH and stirred for 2 h at room temperature. The reaction mixture was dried, reconstituted in a minimal volume of CHCl_3 , and purified by TLC on silica in 9:1 CHCl_3 :MeOH ($R_f = 0.36$). The product was extracted with CHCl_3 , filtered, and evaporated to yield 37 mg (50 μ mol, 56% yield) NBD-DLPE. Product identity and purity was verified by mass spectrometry, with a single peak corresponding to NBD-DLPE at m/z 741.38671.

5.3.6 Phospholipid binding assays

To characterize PL-binding, we developed an equilibrium based FRET assay using DCIA-labeled LRH-1 LBD as the donor and NBD-DLPE as the acceptor. Recombinant LRH-1 from *E. coli* was fluorescently labeled with DCIA (7-diethylamino-3-((4'-

(iodoacetyl)amino)phenyl)-4-methylcoumarin; Molecular Probes, Inc.; Eugene, Oregon USA) according to manufacturer instructions, and further purified by gel filtration chromatography to remove excess dye. All experiments were performed in assay buffer containing 150 mM NaCl, 10 mM Tris HCl (pH 7.4), 5% glycerol, and 0.1% N-octyl- β -D-glucopyranoside. All PL stocks were prepared as small unilamellar vesicles via sonication from evaporated chloroform stocks reconstituted in assay buffer. The binding affinity of NBD-DLPE to LRH-1 was measured using a constant concentration of 150 nM unlabeled or DCIA-LRH-1, and 0 – 100 μ M NBD-DLPE. Competition assays were performed with constant concentrations of 150 nM DCIA-LRH-1 and 5 μ M NBD-DLPE, with 0 – 100 μ M competing PL. Fluorescence intensity was measured on a Synergy 4 plate reader (Biotek; Winooski, VT USA) equipped with 380/20 nm excitation and 460/40 nm emission filters. All assays were performed in triplicate on black 384-well plates in a total volume of 50 μ L. Data for unlabeled LRH-1 were subtracted from corresponding DCIA-LRH-1 data to remove background fluorescence, and all background-corrected data were expressed as percent fluorescence intensity of fully unbound DCIA-LRH-1 (i.e. 0 M NBD-DLPE). Data were processed with GraphPad Prism 5 (GraphPad Software, Inc.).

5.3.7 Reporter gene assays

Transactivation of wild type and mutant LRH-1 was measured via luciferase-based reporter gene assay. HEK 293T cells were seeded into 24-well plates and incubated at 37°C in complete media (DMEM supplemented with 10% charcoal/dextran-stripped FBS and 1% penicillin-streptomycin) until approximately 90% confluent. Each well was then transiently transfected in OptiMem using Lipofectamine 3000 with plasmids encoding firefly luciferase under control of the *shp* promoter (SHP-luc; 500 ng/well), renilla luciferase under constitutive activation via the CMV reporter (pRLCMV; 10 ng/well), and wild-type or mutant LRH-1 in the

pCI mammalian expression vector (100 ng/well). Transfection was ended after 4h incubation at 37°C via the replacement of transfection mixture with complete media, and cells were incubated overnight. Luciferase activity was measured using the Dual-Glo luciferase assay system (Promega; Madison, WI USA). Statistical analyses were performed in GraphPad Prism 5 (GraphPad Software, Inc.; La Jolla, CA USA), via one-factor ANOVA followed by Dunnett's multiple comparison test using wild-type LRH-1 as a control. Data are the results of five independent experiments. All mutations were introduced into the wildtype LRH-1/pCI construct using the QuikChange II Lightning Multi site-directed mutagenesis kit (Agilent Technologies, Inc.; Santa Clara, CA USA).

5.3.8 Model construction for molecular dynamics

Five models were constructed to examine the structural and allosteric impacts of ligand/co-regulator agreement: 1) apoLRH-1 – TIF2 NRBox3, 2) LRH-1 – DLPC – TIF2 NRBox3, 3) LRH-1 – E. coli PLs – TIF2 NRBox3, 4) apoLRH-1 – SHP NRBox2, 5) LRH-1 – E. coli PLs – SHP NRBox2. Agreement is defined here by simultaneous binding of an activating lipid and coactivator or by the binding of a corepressor in the absence of ligand. In every case, residues 297-540 from the LRH-1 LBD form the core of the complex, with additions of 2-3 residues at either terminus as necessary to maintain consistent sequences between models, using the program xLeap, part of the AmberTools11 suite [85]. All five systems were solvated with TIP3P water in a rectangular box with equilibrated dimensions of 67 Å X 70 Å X 72 Å and neutralized with sodium and chloride ions to a salt concentration of 0.15 M.

Briefly, the first model containing the LRH-1 LBD, in the apo state, bound to a TIF2 co-activator peptide was modeled directly from the novel apoLRH-1 – TIF2 NRBox3 crystal structure. The second model, containing DLPC in the binding pocket, bound to a TIF2 co-

activator peptide was modeled directly from PDB ID 4DOS [178]. The third system, comprised of the LRH-1 LBD with the *E. coli* PL in the binding pocket, bound to a TIF2 peptide was modeled from the novel LRH-1 – *E. coli* PLs – TIF2 NRBox3 crystal structure. While electron density in the crystal structure is insufficient to identify the head group of the bound lipid, mass spectrometry results suggest phosphatidylglycerol and phosphatidylethanolamine to be the predominant PL isoforms [178]. Thus, we modeled a bacterial phosphatidylethanolamine with 16 and 18 carbons on the sn1 and sn2 position, respectively, derived from PDB ligand EPH, which is herein referred to as *E. coli* PL. The fourth model consists of the LRH-1 LBD, in the apo state, bound to a SHP co-repressor peptide, constructed from the LRH-1 LBD (derived from the apoLRH-1 LBD – TIF2 NRBox3 structure), with the SHP peptide (PDB ID: 4DOR) [178] modeled in place of TIF2 via superposition of LRH-1 LBD residues 340-382 and 533-538. The charge clamp specific contacts between LRH-1 residues Arg361 and Glu534 and the SHP peptide were enforced with harmonic restraints during the equilibration phase of the molecular dynamics simulation and released before the production runs. The final model, LRH-1 containing DLPC in the binding pocket and bound to a SHP co-repressor peptide was constructed from PDB: 4DOS [178] with the SHP co-repressor modeled in place of TIF2 as described in the previous model.

5.3.9 Molecular dynamics

The CHARMM27 [188] force field for lipids and proteins was employed for all simulations. All systems were subjected to 10,000 steps of steepest-descent minimization, heated to 300 K under the canonical ensemble for 100 ps. Finally, positional restraints were incrementally released first on the protein sidechains, followed by the backbone, under the isobaric-isothermal ensemble. Production runs were performed under constant pressure and

temperature, totaling 212 ns of unrestrained molecular dynamics for each system, with 12 ns discarded as equilibration, resulting in 200 ns of production simulation time per system. All simulations were performed with NAMD 2.9 [75], employing the r-RESPA [27] multiple time step method, with bonded and short-range interactions evaluated every 2 fs and long-range electrostatics evaluated every 4 fs with the smooth Particle Mesh Ewald method [29]. The short-range non-bonded interactions were calculated used a cutoff of 10 Å with a switching function at 8.5 Å. The integration time step was 2 fs and the SHAKE algorithm was applied to fix the bonds between the hydrogens and the heavy atoms. Parameters and topology for the *E. coli* PL ligand were obtained from the general lipid parameters available in CHARMM27.

5.3.10 Analysis methodology

For all analyses, 10,000 evenly spaced frames were taken from the 200 ns production runs to allow for sufficient statistical sampling. Covariance matrices were constructed using the program Carma [189] over all alpha-carbons to produce per-residue statistics. The NetworkView plugin [149] in VMD [150], along with the programs subopt, included in the NetworkView package, and Carma were employed to produce dynamical networks for each system, along with suboptimal path analyses. The ptraj module of AmberTools11 was used for structural averaging as well as Cartesian principal component analysis over protein backbone atoms and, over the same 10,000-frame trajectories used for the covariance analyses. Principal components were projected onto the molecular dynamics trajectories, with snapshots binned according to their displacements along the components. Temporal correlations between modes are lost in this approach but heavily sampled regions of the conformational subspace are more easily identified.

5.4 Results

5.4.1 Structure of the apo LRH-1 LBD – TIF complex

LRH-1 is able to bind to both coactivators and corepressor proteins in the absence of a ligand. To visualize the structural perturbations necessary to bind to coactivators in its apo form, we crystallized apo LRH-1 LBD bound to a fragment of the coactivator TIF2 and determined its structure to 1.75 Å (Figure 5.4.1.1A). There is no visible electron density to support modeling a bound ligand. The opening to the ligand binding pocket is constricted by 2 Å, which reduces the volume of the ligand binding pocket from 1554 Å³ in the LRH-1 – TIF2 – DLPC complex to roughly 940 Å³. (Figure 5.4.3.1A vs B and D). This is in stark contrast to the apo LRH-1 LBD – SHP NRBox2 structure reported previously, which lacks electron density for the entirety of the alternate AF (Figure 5.4.1.1A vs C). Unlike the ligand binding pocket of rodent LRH-1, the ligand binding pocket constriction is not stabilized by any intramolecular interactions [190]. However, it is possible that the alternate AF, which comprises nearly one third of the binding pocket, may be visible due to fortuitous interactions with a crystallographic symmetry mate. Regardless, this shows remarkable flexibility of the ligand binding pocket.

The structure contains a single CHAPS detergent molecule that docks on H10 and H12 via hydrophobic interactions and two hydrogen bonds between the CHAPS 7-OH and Glu-514, and the CHAPS phosphate and Tyr-518. CHAPS also makes extensive contact with a crystallographic symmetry mate (Figure 5.4.1.1C). Thus, two molecules within the crystal create a cleft for CHAPS binding.

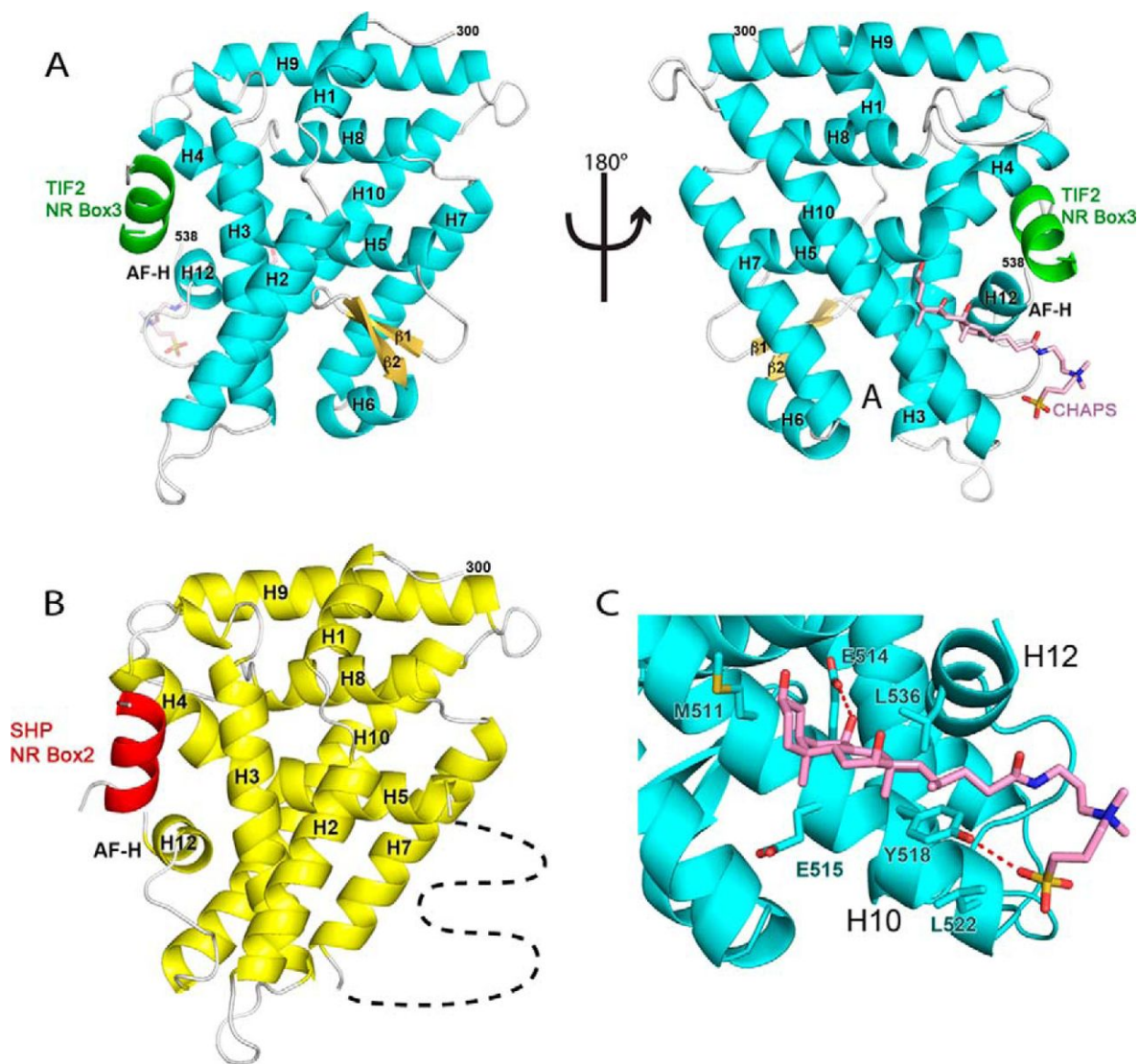


Figure 5.4.1.1 Structure of the apo LRH-1 LBD-TIF complex.

A, Ribbon diagram of apo LRH-LBD (α -helices, teal; β -strands, yellow) with the TIF2 NR box 3 peptide (orange). The surface bound CHAPS is depicted as sticks (C, pink; O, red; S, yellow; N, blue). The AF-2 surface is defined by H3, H4 and H12. *B*, Ribbon diagram of apo LRH-SHP NRBox2 complex (PDB ID 4DOR) with the unobserved alternate AF region (defined by $\beta 1$ -2 and H6) represented by a dashed line. *C*, Close up view of the bound CHAPS molecules included in the crystallization buffer.

5.4.2 Improved structure of the LRH-1 LBD – *E. coli* PL – TIF2 complex

To generate a more accurate model for molecular dynamics studies, and as a control in our crystallization experiments, we crystallized the LRH-1 LBD – *E. coli* PL – TIF2 complex and determined its structure to 1.75 Å (Figure 5.4.2.1A). This represents an improved resolution over the existing LRH-1 LBD – *E. coli* PL – TIF2 structures, which were both solved to 2.5 Å [191, 192]. The structure is highly similar to the previous structures with an r.m.s.d. of 0.6 Å over main chain atoms and maintains *E. coli* PLs in the binding pocket (Figure 5.4.2.1B). The lipid acyl tails show a decrease in electron density near their termini, which is similar to previous observations for the bound *E. coli* PLs and the LRH-1 – DLPC complex [178] (Figure 5.4.2.1B). This observation further supports the hypothesis that LRH-1 specifically recognizes its PL ligands near the glycerolphosphate backbone and the exact position of the acyl tails is less important than the amount of space they occupy in the deeper portion of the ligand binding pocket.

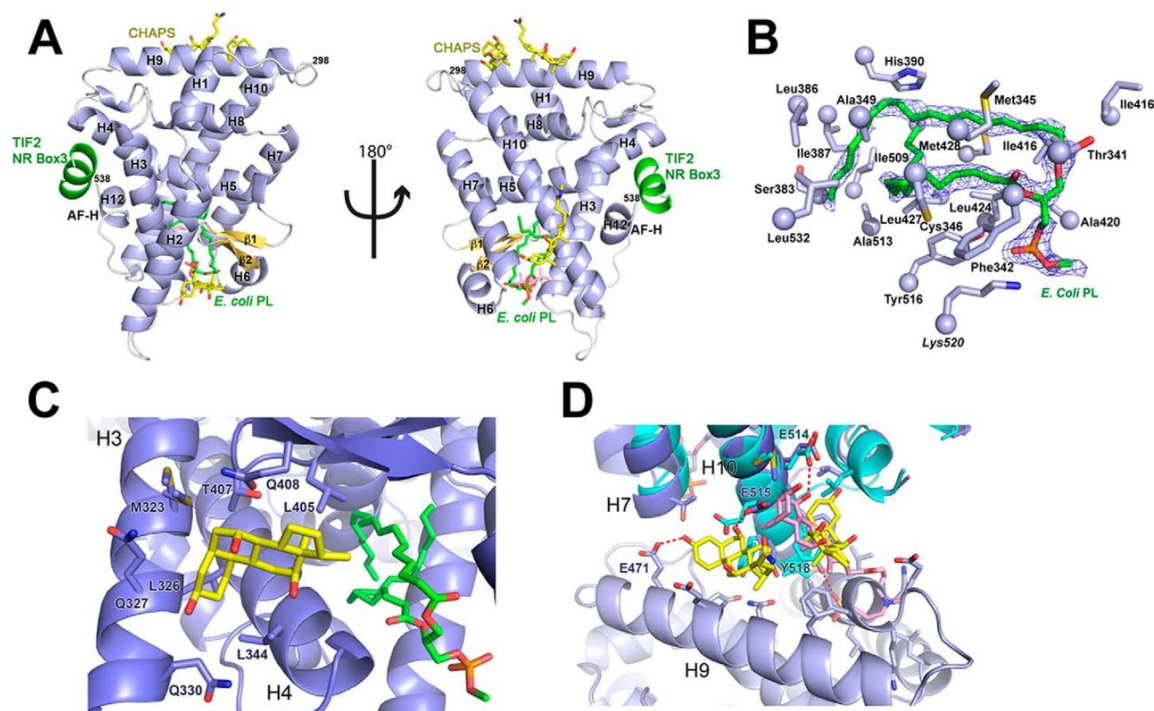


Figure 5.4.2.1 Structure of the LRH-1 LBD–*E.coli* PL–TIF2 complex.

A, Ribbon diagram of *E. coli* PL bound LRH-LBD (α -helices, teal; β -strands, yellow) with the TIF2 NR box 3 peptide (green). The bound *E. coli* PL is depicted as sticks (C, green; O, red; P, magenta) The surface bound CHAPS is depicted as sticks (C, yellow; O, red; S, yellow; N, blue). *B*, 2Fo – Fc electron density (contoured at 1 σ) for the bound *E. coli* PL observed in this structure, along with side chains lining the ligand-binding pocket of hLRH-1 that contact this ligand. *C*, Close up view of the bound CHAPS molecules included in the crystallization buffer along H3 and H4 in close proximity to the bound PL. Residues within 4.2 Å are depicted as sticks. *D*, Close up view of the bound CHAPS along H9 which interact with a crystallographic symmetry mate and in a position overlapping the CHAPS site in the apo LRH-1 – TIF2 complex. Residues within 4.2 Å of CHAPS are depicted as sticks.

The structure contains three CHAPS detergent molecules that dock onto the surface of the protein and make interactions with crystallographic symmetry mates. One CHAPS molecule is secured in the cleft between H3 and the β -sheets via hydrophobic interactions. A second CHAPS molecule mediates contact between two copies of the LRH-1 monomer and is secured by hydrophobic interactions along H10 and a hydrogen bond with Glu-515 of one monomer, and hydrophobic interactions along H9 and a hydrogen bond with Glu-471 of the second monomer. The third CHAPS is adjacent to the second, and also mediates contact between two LRH-1 monomers via hydrophobic interactions with H10 of the first monomer and H9 of the second, but does not make any hydrogen bonds with either monomer. The CHAPS molecules contacting two LRH-1 monomers are unique relative to the apo LRH-1 LBD – TIF complex, while the CHAPS occupying the site near H10 shows a partial overlap with the well ordered CHAPS in the apo structure (Figure 5.4.2.1D). In contrast to the excellent electron density for the CHAPS bound in the apo LRH-1 LBD – TIF complex, the CHAPS bound at this site in the E. coli lipid bound complex shows electron density for only the sterane ring. This is likely due to greater thermal motion or reduced CHAPS occupancy at this site in the crystal. Interestingly, CHAPS is docked at regions within LRH-1 that show most exchange in HDX studies and the most conformational fluctuations in crystal structures. It is possible that these are sites for protein-protein or protein-lipid interaction in the cell.

LRH-1 can bind to several PLs [160, 183, 192, 193], yet only PCs have been shown to drive transactivation [171, 178, 192, 193]. It is unclear why LRH-1 responds only to PCs in cells; this may be intrinsic to the receptor or due to uncharacterized PL transporters capable of delivering PC ligands. In order to elucidate the mechanisms by which PLs differentially activate LRH-1, it is critical to determine the effects of head group and tail variation on binding. To

characterize differential PL binding, we developed a FRET-based PL-binding assay monitoring the ability of NBD-labeled DLPE to bind to DCIA-labeled LRH-1 (Figure 5.4.2.2A). This binding event quenches the DCIA fluorescence, which can be recovered upon the competitive binding of unlabeled lipids (Figure 5.4.2.2B, E).

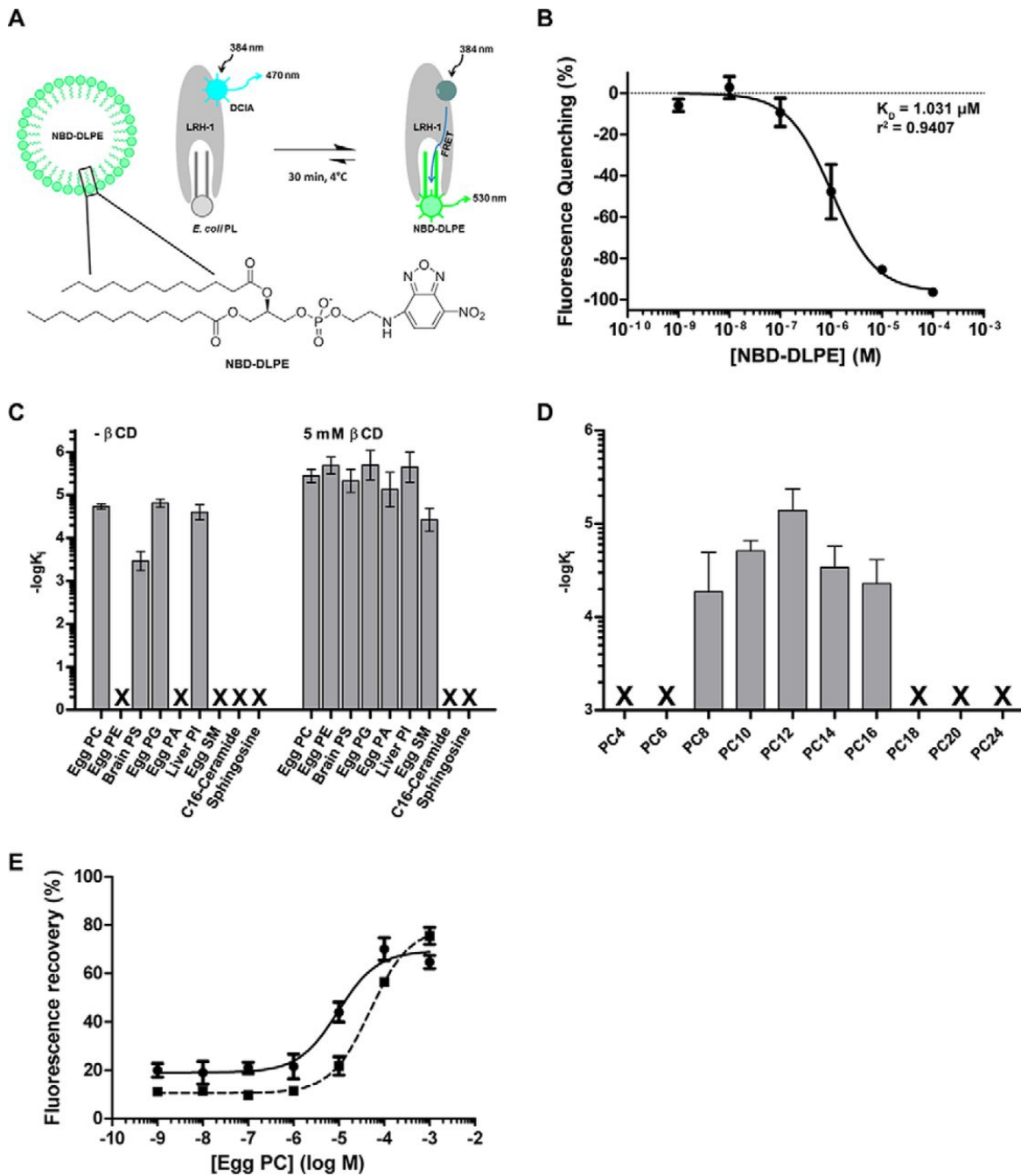


Figure 5.4.2.2 LRH-1 in vitro lipid binding profile.

Binding affinities of LRH-1 to PLs of differing head group and tail compositions. *A*, PL binding was measured relative to probe ligand NBD-DLPE via FRET quenching of DCIA-labeled LRH-1. *B*, Binding affinity of LRH-1 to NBD-DLPE probe. *C*, Relative binding affinities of competing PLs of differing head groups; 5 mM β -cyclodextrin added as indicated. *D*, Relative binding affinities of competing saturated PCs of differing tail lengths. Data are reported as the means + S.E.M. of three independent experiments. The presence of an X instead of a bar indicates that no binding was observed. *E*, Example of an individual competitive binding curves for NBD-DLPE displacement. Solid line represents the inclusion of 5 mM β -cyclodextrin while the dashed line is without 5 mM β -cyclodextrin as described in the methods.

Prior to engulfing PLs, LRH-1 must extract PL from the lipid membrane – a step typically conducted by PL transport proteins that contain amphipathic structural elements to facilitate partitioning in membranes [194-196]. In the absence of lipid chaperones, we find that LRH-1 extracts and binds PC, PG, and PI with micromolar affinity, but cannot extract PE, PS, PA, SM, ceramide, or sphingosine (Figure 5.4.2.2C). Thus, LRH-1's ability to bind PLs from vesicles is sensitive to the nature of the head group. However, addition of 5 mM β -cyclodextrin, a small molecule chaperone widely used for the delivery of hydrophobic small molecules, enables the binding of PC, PE, PS, PG, PA, with low micromolar affinity (Figure 5.4.2.2C). LRH-1 is unable to bind sphingosine and ceramide despite the presence of β -cyclodextrin suggesting that extraction from vesicles is not a limiting factor; rather, these lipids do not fit well within the ligand binding pocket. These extracts contain a range of PL isoforms and the PC mixture showed the highest maximum displacement of bound NBD-DLPE (Figure 5.4.2.2E and data not shown).

We then investigated LRH-1's intrinsic selectivity for PL tail composition by testing a range of saturated PCs. Surprisingly, only PCs with mid-length chains of 8-16 carbons bind to LRH-1, with DLPC showing the strongest affinity. We observed no change in binding with the inclusion of 5 mM β -cyclodextrin (data not shown). These findings mirror previously published activation data, which demonstrate that LRH-1 is most strongly activated by the 11- and 12-carbon saturated PCs, DUPC and DLPC [171]. Thus, PL selectivity is driven by the length of the fatty acid tails *in vitro* suggesting that the amount of space filled by the acyl tails is a critical determinant of binding.

5.4.3 Co-regulator binding interactions are altered by ligand status

The canonical model of NR activation revolves primarily around a mobile ligand-sensing helix (H12), termed the AF-H. When a receptor is bound to an agonist the AF-H packs against helices 3 and 4 of the LBD forming a surface, termed activation function 2 (AF-2), which enables interaction with coactivator proteins containing a LxxLL motif [197]. This helical peptide inserts its leucines into a groove on the AF-2 surface and is further stabilized by a charge clamp interactions with Arg 361 on H3 and Glu 534 on the AF-H. An equivalent charge clamp is conserved across NRs and represents a general mechanism for activation [198]. LRH-1, like some other orphan NRs, is able to form a productive AF-2 in the absence and presence of ligands in available crystal structures. This makes inferences regarding ligand potency based on backbone positioning within the AF-2 alone challenging. Nevertheless, we compared coregulator binding at the AF-2 across all available crystals structures and observed that regardless of the ligand state, Arg 361 on H3 forms the expected charge clamp interaction. In contrast, we were surprised to find that Glu 534, on the AF-H, does not make the expected charge clamp interaction with coregulator peptide under all circumstances (Figure 5.4.3.1). This does not appear to be an artifact of crystal packing. Instead, the conformation of Glu 534 correlates with an agreement between the ligand and the bound coregulator peptide. Agreement is defined here by simultaneous binding of an activating lipid and coactivator or by the binding of a corepressor in the absence of ligand. When apo or bound to a poorly activating ligand, Glu 534 is rotated out of hydrogen bond distance with the coactivator TIF2 peptide (Figure 5.4.3.1A-C). In contrast, when LRH-1 is bound to a strong agonist such as DLPC, Glu 534 makes the expected hydrogen bond with a backbone amide of the TIF2 peptide (Figure 5.4.3.1D). This charge clamp interaction is also observed in apo or *E. coli* PL bound LRH-1 when complexed to a peptide

derived from the corepressor SHP (Figure 5.4.3.1E-F). These observations suggest that LRH-1 has an extensive allosteric network that appropriately tunes the receptors ability to stabilize very similar LxxLL motifs present in coactivators and corepressors.

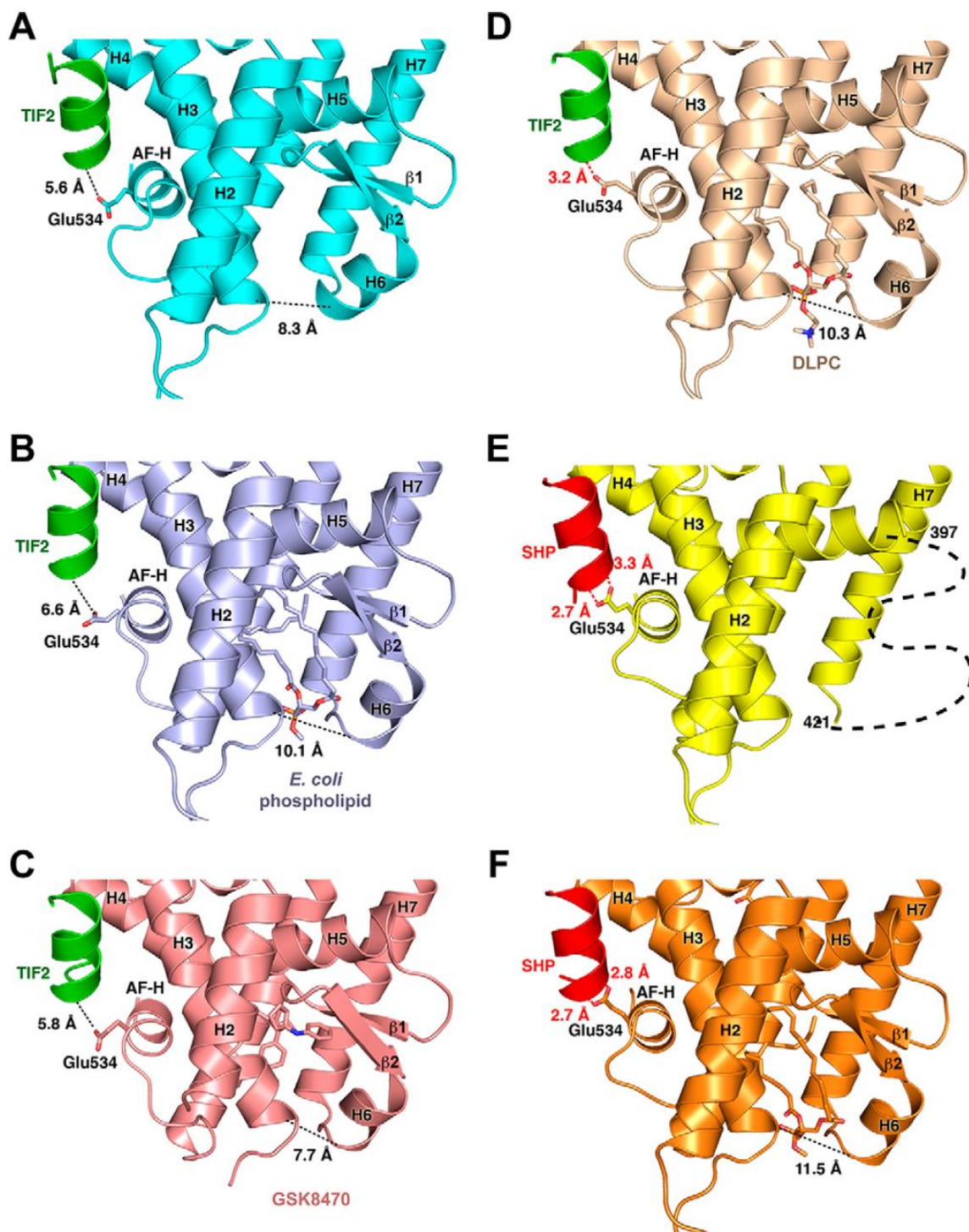


Figure 5.4.3.1 AF-2 charge clamp engagement is dictated by ligand-coregulator combination.

Ligand binding pocket entrance measurements and analysis of Glu 534 – peptide charge clamp engagement for the *A*, apo LRH-1-TIF2 complex, *B*, LRH-1-*E. coli* PL-TIF2 complex, *C*, LRH-1-GSK8470-TIF2 complex (PDB ID: 3PLZ), *D*, LRH-1-DLPC-TIF2 complex (PDB ID: 4DOS), *E*, apo LRH-1-SHP complex (PDB ID: 4DOR), *F*, LRH-1-*E. coli* PL-SHP complex (PDB ID: 1YUC).

5.4.4 Ligand and coregulator drive differential effects on local residue environment

Supposition of multiple LRH-1 – ligand structures revealed only subtle differences in the coregulator binding surface. We therefore used ProSMART to compare the local residue environment to identify how differential ligand and coregulator peptide binding affects local structure [187]. Caution of course must be taken with the interpretation of these results since the crystal structures included in this analysis are derived from multiple crystal forms.

LRH-1 shows the greatest conformational similarity between structures where both ligand and coregulator status are in agreement within the structural complex (Figure 5.4.4.1). Greater conformational dissimilarity is seen when one or both complexes are not in ligand-coregulator agreement, indicating that such agreement is crucial in maintaining a stable complex, regardless of whether that complex is activated or repressed. In all coregulator states, the addition of a ligand stabilizes the alternate AF region compared to apo, as demonstrated by the high structural dissimilarity seen in this region compared to the apo-TIF2 structure (Figure 5.4.4.1E, F, H), and the fact that this area could not be modeled in the apo-SHP complex. As expected, the highest structural dissimilarity is seen in the AF-2 and alternate AF (β -sheets/ H6), the respective interaction sites for the coregulator peptide and PL head group. SHP poorly discriminates between apo and bacterial PL – bound receptor, and shows high structural similarity throughout the ligand binding domain (Figure 5.4.4.1A). In contrast, the LRH-1 TIF2 complexes show strong differences with LRH-1 SHP complexes regardless of ligand status, even in cases where the ligand is the same or nonexistent (Figure 5.4.4.1B-E, G, I). Thus, unlike SHP, TIF2-bound conformations are sensitive to the nature of the bound ligand. All LRH-1 – TIF2 complexes exhibit moderate or high structural dissimilarity in both the AF-H and the preceding loop, and the alternate AF region (Figure 5.4.4.1F, H, J). The greatest agreement among the

LRH-1 – TIF2 complexes is seen between the *E. coli* PL and DLPC bound structures (Figure 5.4.4.1J), indicating that while TIF2 is sensitive to the presence or absence of a ligand, it does not strongly discriminate between ligands so long as one is present. This is consistent with previous coregulator recruitment studies, which show only a 3-fold difference in binding affinities between TIF2 and *E. coli* PL or DLPC bound LRH-1 (20.1 μ M and 6.5 μ M, respectively) [178]. Taken together, the ProSMART analyses suggest that ligand-coregulator agreement promotes the stabilization of LRH-1 into either an active or repressed conformation, with detectable but subtle structural differences between these conformations. These conformational variations are also in line with the prior HDX data showing conformational variation between the same structural elements (i.e. the alternate AF, the AF2 and in H8 and 9) [159].

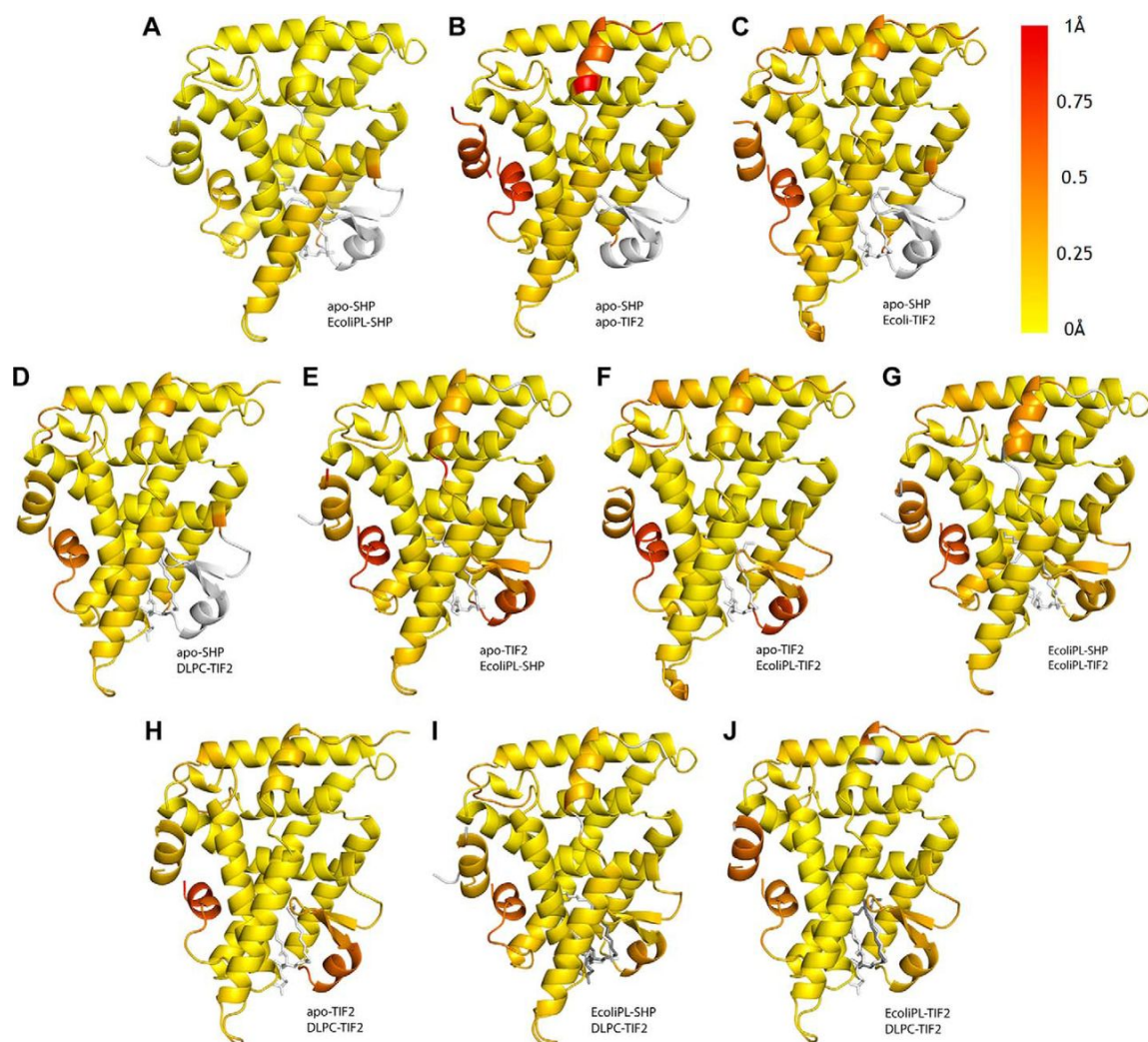


Figure 5.4.4.1 ProSMART Procrustes central residue analysis of LRH-1 complexes.

ProSMART analysis of LRH-1 with differentially bound ligands and coregulator peptides. Models were colored by the Procrustes score of the central residue of an aligned fragment pair according to the legend at top right. Areas colored white were omitted from the analysis. The following pairwise comparisons were made: (A) apo-SHP (PDB ID: 4DOR) vs *E. coli* PL-SHP (PDB ID: 1YUC); (B) apo-SHP vs apo-TIF2; (C) apo-SHP vs *E. coli* PL-TIF2; (D) apo-SHP vs DLPC-TIF2 (PDB ID: 4DOS); (E) apo-TIF2 vs *E. coli* PL-SHP; (F) apo-TIF2 vs *E. coli* PL-TIF2; (G) *E. coli* PL-SHP vs *E. coli* PL-TIF2; (H) apo-TIF2 vs DLPC-TIF2; (I) *E. coli* PL-SHP vs DLPC-TIF2; (J) *E. coli* PL-TIF2 vs DLPC-TIF2.

5.4.5 The activated LRH-1 complex exhibits coordinated motions

To analyze the dynamic coupling of structural elements in LRH-1, we computed cross-correlation (normalized covariance) matrices for the C- α atoms in each of the five systems with the program Carma [189]. A covariance matrix contains a great deal of information regarding the dynamics within a system, in this case describing the correlation of motions \mathbf{r}_i and \mathbf{r}_j for residues i and j , taken from their respective α Carbons. Element (i, j) of the covariance matrix is calculated as $\langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \rangle$. In essence, this type of covariance matrix provides a way of visualizing whether the motions of two residues within a complex are correlated, anti-correlated or non-correlated (i.e. whether the motion over the course of the MD trajectory is related by 180°) (i.e. whether the motion vectors of the residues are parallel or anti-parallel). A cross-correlation matrix is simply a covariance matrix that is normalized to vary between -1 (perfectly anti-correlated) and 1 (perfectly correlated) (Figure 5.4.5.1).

The motions in residues within helices 4 through 9 of the LRH-1 LBD become correlated upon lipid binding in the presence of a co-activator (Figure 5.4.5.1C). The correlation of these motions in the lipid and co-activator bound systems is muted in the apo states as well as the DLPC-bound LBD in complex with the SHP peptide (Figure 5.4.5.1A, D). This suggests that both lipid and co-regulator binding impact an allosteric network through the LRH-1 core, requiring the lipid pocket and AF-H elements to be in agreement to yield an active complex. Lipid may therefore allow correlated motions in LRH-1 to favor TIF2 binding while in the apo state these motions are eliminated, thereby favoring SHP binding.

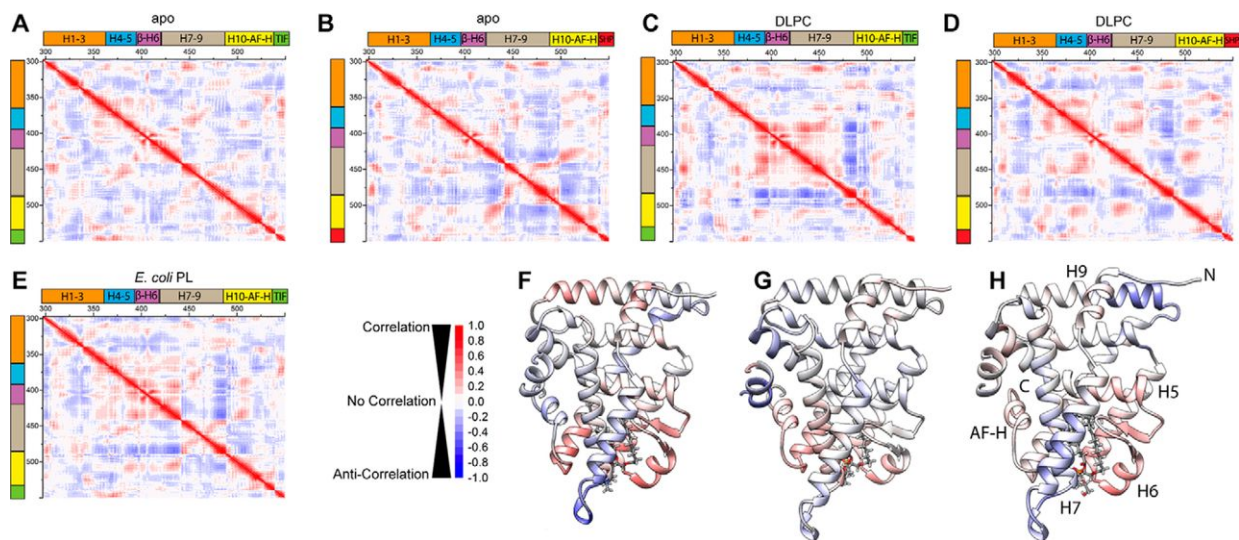


Figure 5.4.5.1 Correlated motion in LRH-1-PL-coregulator systems.

Cross-correlation matrices showing correlated and anti-correlated motion over the 200 ns MD simulation for *A*, apo LRH-1-TIF2 complex, *B*, apo LRH-1-SHP complex, *C*, LRH-1-DLPC-TIF2 complex, *D*, LRH-1-DLPC-SHP complex, *E*, LRH-1-*E. coli* PL-TIF2 complex.

We have projected cross-correlation between the lipid head group phosphorus atom and all protein residues, for each lipid-bound system studied, onto the LRH-1 structure (Figure 5.4.5.1F-H). We find that the lipid displays some positive covariance with the β -H6 region of the complex, and some negative covariance with H9 and H2. The DLPC – LRH-1 – SHP system shows similar behavior, but with smaller magnitudes, likely due to its disagreement status.

5.4.6 MD simulations demonstrate communication between β -sheet-H6 and the AF-H through helices 3, 4, and 5

We have previously discovered that LRH-1 contains an alternate activation function region that encompasses the β -sheet-H6 portion the ligand-binding pocket. Our data suggested that the dynamics of this region are coupled to ligand binding and receptor activation [178]. To identify the relevant communication pathways contributing to these observations, we constructed dynamical networks to identify the most prevalent communication pathways between the β -sheet-H6 region and the bound co-regulator (Figure 5.4.5.1). Dynamical networks, as defined in the field of network theory, describe the communication pathways between components of a system. In a dynamical network, every component is taken to be a “node” and a communication between two nodes defines an “edge.” In the methodology employed here, each protein residue’s α carbon is a node and any two nodes must be within a distance cutoff of 4.5 Å for 75% of the MD trajectory and the strength of communication between two nodes, or the “edge weight,” is determined from the covariance between the two nodes. A communication path between two distant nodes is then a chain of edges that connect them and the optimal path transmits communication between two nodes through the fewest number of edges possible and is likely to carry more communication than any other single path. The optimal path and a relatively small set of slightly longer suboptimal paths are expected to carry the majority of communications

between two edges. Monitoring the strength and number of suboptimal paths between two distant nodes can yield detailed insight into the strength of communication, or in macromolecular systems, allostery.

These pathways show much stronger communication when the lipid pocket and AF-H domain states are in agreement than otherwise (Figure 5.4.6.1A-E). The number of communication pathways increases greatly upon lipid pocket – AF-H state agreement, especially expanding outward from the β -sheet–H6 region and into the co-regulator itself. This strongly supports our previous hypothesis that the β -sheet–H6 and AF-H regions communicate to control LRH-1 activation. Furthermore, the vast majority of communication paths proceed through helices 3, 4, and 5. These same helices showed the most protection from deuterium exchange in prior HDX studies suggesting that their rigidity may facilitate the flow of information through the receptor [178]. Therefore, the allosteric pathway between the β -sheet–H6 region and AF-H like traverse through helices 3, 4, and 5 (Figure 5.4.5.1B, D). These helices present an optimal tether between the allosteric switches. It is interesting to note that many of the mutations that affect LRH-1 PL binding, coregulator sensitivity, and overall activation lie directly on or immediately adjacent to this pathway (Figure 5.4.6.1F)[178, 183, 199].

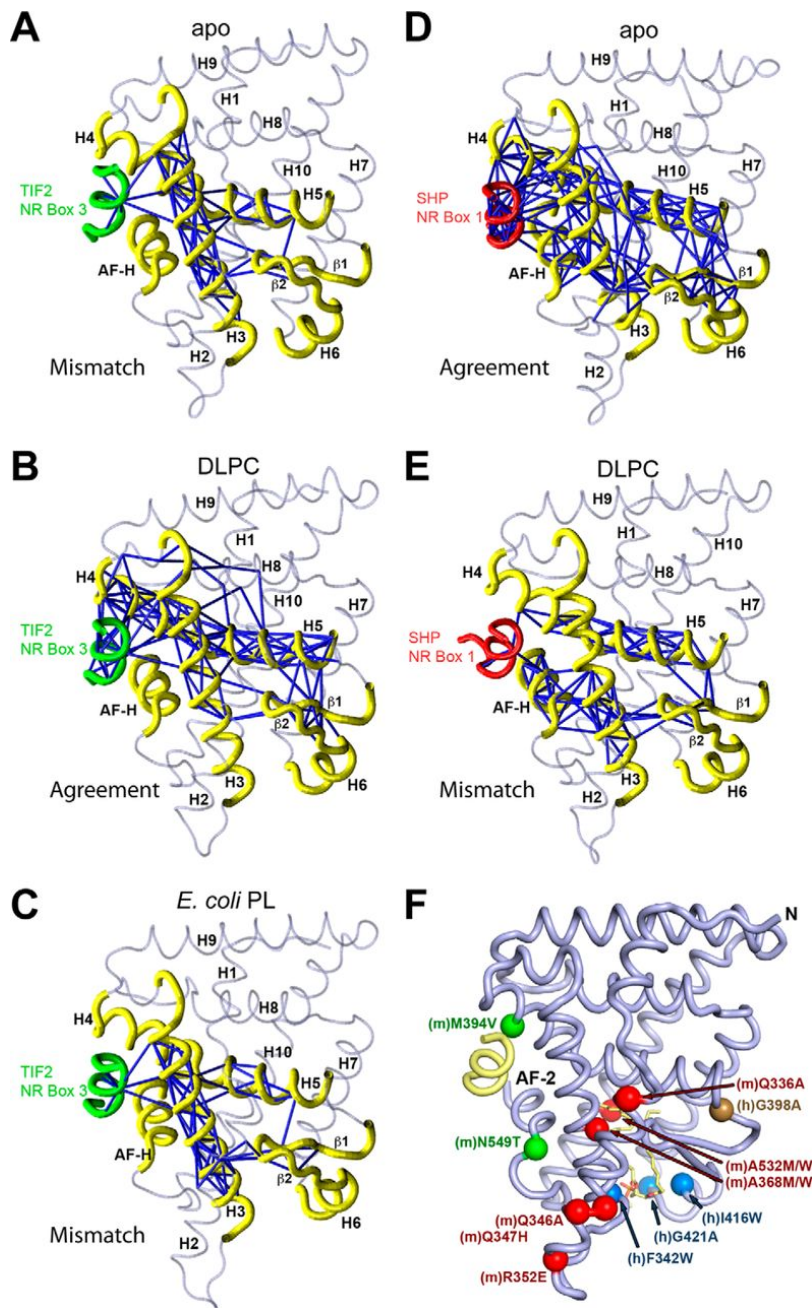


Figure 5.4.6.1 Allosteric paths from binding pocket to co-regulator.

Allosteric communication pathways between the β -sheet-H6 and co-regulator binding regions of the LRH-1 LBD in the *A*, apo LRH-1-TIF2, *B*, LRH-1-DLPC-TIF2, *C*, LRH-1-*E. coli* PL-TIF2, *D*, apo LRH-1-SHP and *E*, LRH-1-DLPC-SHP complexes. Cartoon loop view of LRH-1 showing thick loops (yellow, LRH-1; green, TIF2; red, SHP) for regions of the protein identified along the allosteric path. *F*, LRH-1 mutations that alter PL binding or coregulator recruitment lie on or adjacent to the allosteric pathway. Known mutations of mouse (m) or human (h) LRH-1 LBD are shown as C- α spheres on the LRH-1 protein backbone. Mutations shown in green enhance the degree of LRH-1 activation in response to coactivator binding; mutations shown in red selectively decrease LRH-1 sensitivity to SHP without affecting overall activation; mutations shown in brown decrease overall LRH-1 activity without affecting PL binding; mutations shown in blue decrease PL binding and overall activity.

5.4.7 Structural and dynamical rationale for lipid and co-regulator agreement

To identify and functionally significant collective motions of the residues forming the allosteric network within LRH-1, we employed principal component analysis (PCA) [200]. In PCA, the C- α covariance matrix is diagonalized to yield eigenvectors, denoted as principal modes, and eigenvalues, representing the mean square fluctuation along each principal mode. Projections of the MD trajectory onto the principal modes are called the principal components. By reducing the dimensionality of the data, PCA recapitulates the most important dynamical features from the MD trajectories. Thus, the first few principal modes, known as the essential dynamics, are likely to describe the collective, global motions of LRH-1 involved in the allosteric response to ligand and coregulator binding.

We have identified two modes that are indicative of the lipid-binding pocket's state and the bound coregulator, named PC1 and PC2 and have projected snapshots from the molecular dynamics trajectories onto these modes in Figure 5.4.7.1. To ensure comparability and uniformity of the modes studied, we optimized the total root mean square inner product (r.m.s.i.p.) across all systems' essential dynamics. The r.m.s.i.p. method for optimizing subspace overlap does not guarantee that the same mode number will be selected from each system, as some variation in the ordering of principal modes is expected, even for highly similar systems[201]. A table of the modes chosen for PC1 and PC2 and the dot products between these modes are included in table 5.4.7.1.

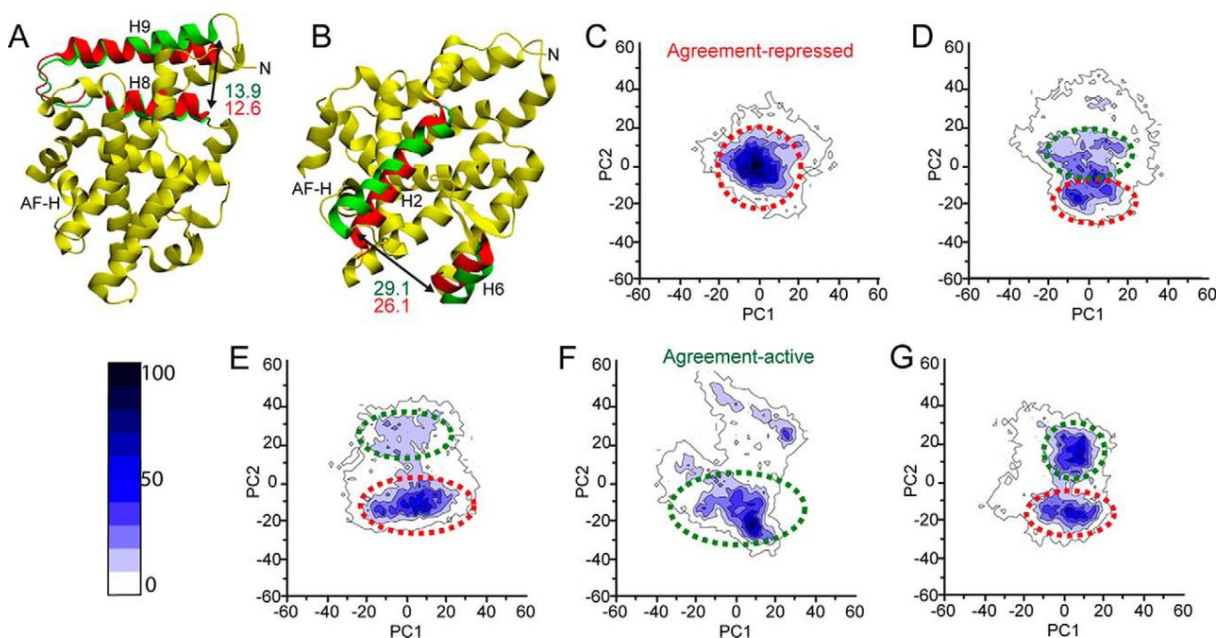


Figure 5.4.7.1 Biologically relevant principal modes identified from the projections of the MD trajectories on PC1 vs. PC2.

An outward swing of helix 9, *A*, contributes to PC1, while opening motions at the mouth of the lipid binding pocket, *B*, result in translation along PC2. Projections of snapshots taken from MD onto PC1 and PC2 in *C*, apo LRH-1–SHP and *D*, LRH-1–DLPC–SHP, *E*, apo LRH-1–TIF2, *F*, LRH-1–DLPC–TIF2, *G*, LRH-1–*E. coli* PL–TIF2 complexes. Higher densities indicate more populated regions of the conformational subspace. Scale bar indicates how many snapshots (out of 10,000) were collected within a contour. Green and red rings indicate activating and repressing regions of the subspace, respectively.

Table 5.2 Modes chosen for PC1 and PC2 and the dot products between these modes.

<i>PC1</i>	apo SHP :					Average	+/-
	2	dlpc SHP : 3	apo TIF : 3	dlpc TIF : 3	pl TIF : 3		
apo SHP : 2	1.0000	0.0756	0.0626	0.4221	0.4954	0.4111	0.3834
dlpc SHP : 3	0.0756	1.0000	0.1587	0.4538	0.2609	0.3898	0.3692
apo TIF : 3	0.0626	0.1587	1.0000	0.2138	0.2337	0.3338	0.3783
dlpc TIF : 3	0.4221	0.4538	0.2138	1.0000	0.6050	0.5389	0.2931
pl TIF : 3	0.4954	0.2609	0.2337	0.6050	1.0000	0.5190	0.3112

<i>PC2</i>	apo SHP :					Average	+/-
	4	dlpc SHP : 1	apo TIF : 1	dlpc TIF : 1	pl TIF : 1		
apo SHP : 4	1.0000	0.3522	0.0943	0.5088	0.0987	0.4108	0.3734
dlpc SHP : 1	0.3522	1.0000	0.1388	0.5782	0.2195	0.4578	0.3457
apo TIF : 1	0.0943	0.1388	1.0000	0.1446	0.0122	0.2780	0.4071
dlpc TIF : 1	0.5088	0.5782	0.1446	1.0000	0.3446	0.5153	0.3184
pl TIF : 1	0.0987	0.2195	0.0122	0.3446	1.0000	0.3350	0.3923

* Modes and dot products selected for PC1 and PC2 via RMSIP. Each system label is formatted as “binding-pocket-state coregulator : mode”. Averages of dot products and standard deviations for each system with respect to each other system are included. In each PC, the LRH-1—DLPC—TIF2 system (highlighted in green) was the most central eigenvector, having the highest average inner product with the other systems.

In the projections (Figure 5.4.7.1C-G), areas of high density indicate regions of high conformational probability. Snapshots from the most densely populated regions of each system's conformational subspace were collected and averaged to obtain representative structures for comparison (Figure 5.4.7.1A, B). PC1 is characterized by an outward motion of helix 9 relative to helix 8 and the core of the LBD, with the distance from N332 to T422 measuring 29.1 Å in the DLPC – LRH-1 – TIF2 model and 26.1 Å in the apo-LRH-1 – SHP model (Figure 5.4.7.1A). PC2 consists primarily of an opening motion near the mouth of the lipid-binding pocket, with the distance from Q444 to N487 measuring 13.9 Å in the most prevalent conformation in DLPC – LRH-1 – TIF2 and just 12.6 Å for the dominant apo-LRH-1-SHP conformation (Figure 5.4.7.1B).

Projections of the MD trajectories onto these principal modes (Figure 5.4.7.1C-G) illustrates that DLPC binding promotes conformations with high values of PC2, while apo- and bacterial long-tail lipid bound states tend to exhibit conformations of lower PC2 magnitude. Coregulator binding influences the dominant conformation's placement along PC1, with all TIF2-bound complexes exhibiting primary centroids near +10 and SHP-bound complexes exhibiting centroids near -10. It is worth noting that the long-tail *E. coli* lipid and apo-TIF2 complexes both share two common clusterings, with the former maintaining nearly equal populations near each center and the latter undergoing a population shift toward the repression-promoting region of the subspace. These results show that lipid binding and coregulator binding both impact motions in the LRH-1 LBD and that combinations of those motions result in only two distinct and stable conformations: repressing and activating. We also observed that “disagreement” complexes exist in mixed populations between the two states.

R.m.s.d. alignment of the resultant repressing and activating structures (Figure 5.4.7.1A-B), respectively, reveals an upward shift in helices 2 and 3 in the activated structure perturbs the

AF-H backbone by an r.m.s.d. of 1.2 Å. This alters the binding position of the coregulator and provides a mechanism by which binding pocket status directly impacts coregulator choice through PC2. Similarly, overlaying the average repressing structure from both the apo-LRH-1 – TIF and apo-LRH-1 – SHP, the differing binding position of the coregulator presses outward on helix 4, causing a slight rearrangement in helices 8 and 9, leading to the motion observed in PC1. Interestingly, the large motions identified in PC1 and PC2 encompass the same regions showing the highest conformational movement in previous HDX studies [178]. In these prior studies, apo LRH-1 shows rapid exchange in helices 2, 3 and 6 (PC1) and helices 8 and 9 (PC2) with complete exchange of these elements occurring in 60 seconds. These same elements are the most sensitive to ligand status showing the strongest projection in the LRH-1 – phospholipid complex.

5.4.8 Modest disruption of interhelical interactions along the allosteric pathway reduces, but does not eliminate, LRH-1 activity

In order to verify that the allosteric pathway we identified plays a role in LRH-1 transcriptional activity, we generated mutant forms of LRH-1 designed to perturb the communication network between phospholipid and coregulator. We took care to avoid residues that make direct contact with the ligand, the AF-2 (coregulator binding) surface or the β -sheet–H6 (alternate AF) region since these would all be expected to reduce LRH-1 transactivation. We instead sought to disrupt LRH-1's allostery one shell of residues away from these surfaces. Helix 5 was identified as a central feature of the pathway (Figure 5.4.6.1D,B), as its junction with helix 10 creates the cleft against which helix 12 docks to establish the AF-2 coregulator binding surface (Figure 5.4.8.1A). Moreover, the acyl tails of long-chain PLs dock against these helices. We hypothesized that differences in tail length or unsaturation may exert variable amounts of pressure against these helices, which is then transmitted along the allosteric network to the AF-2

site, affecting coregulator binding. The junction between helices 5 and 10 displays hydrogen bonding between residues S383 and E514, and electrostatic interactions between residues E384 and R507 (Figure 5.4.8.1B). To disrupt these interactions, we generated mutant forms of LRH-1 (S383A, E384Q + R507H, and S383A + E384Q + R507H) and measured their transcriptional activity in HEK 293T cells via luciferase reporter gene assay.

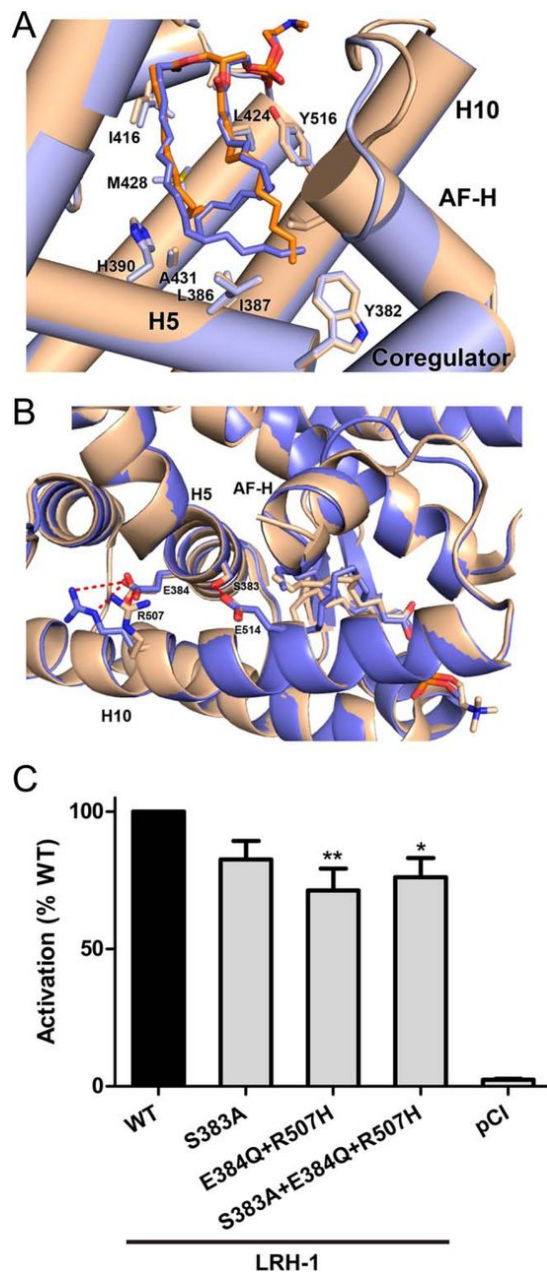


Figure 5.4.8.1 Subtle disruption of residues on or near the allosteric pathway reduces LRH-1 activation.

A, Close up view of the PL binding pocket of DLPC- (beige/orange) and *E. coli* PL-bound LRH-1 (blue). Helices are shown as cylinders and helix 3 has been hidden. Residues within 4.2 Å of the phospholipid are depicted as sticks. *B*, The junction of helices 5 and 10 displays hydrogen bonding (red dashes) between S383 and E514, and electrostatic interactions between E384 and R507. In the active conformation, helix 12 docks against this junction to support the AF-2 coregulator binding surface, driving gene transactivation and transrepression. *C*, Abolition of the electrostatic interaction between helices 5 and 10 via at E384Q/R507H mutation causes a subtle but significant reduction in LRH-1 transcriptional activity. LRH-1 activity was measured via luciferase reporter gene assay in transiently transfected HEK 293T cells. Data are the combined results of 5 independent experiments. Statistical significance is represented as *: $p < 0.05$; **: $p < 0.01$.

These mutant forms of LRH-1 all showed a slight decrease in basal activity, which achieved statistical significance specifically upon disruption of the electrostatic interaction between E384 and R507 (Figure 5.4.8.1C). Importantly, none of these mutations fully abrogated LRH-1 activity, indicating that these mutations did not fatally inactivate the receptor.

5.5 Discussion

Robust signaling pathways must not only respond to activating ligands but must discriminate against the wrong ones to reduce noise [202]. For LRH-1, this challenge is amplified since its ligands include highly abundant intact PLs that comprise a large fraction of cell membranes. It is possible that LRH-1 displays an intrinsic set of selection criteria for PL isoforms, that PL delivery to the receptor is facilitated by a soluble lipid transport proteins, or a combination of the two. Our results show that LRH-1 is able to bind a wide range of PLs *in vitro*, but can extract only PCs, PGs, and PIs from a membrane/ vesicle without assistance from a molecular chaperone. Inclusion of a non-specific lipid chaperone, β -cyclodextrin, permits the binding of all glycerophospholipids tested. This is in line with structural studies since the majority of recognition occurs via contacts with the lipid tails and phosphoglycerol backbone. Thus, LRH-1 lipid preference is driven more so by the composition of the PL tails than by the head group, which protrudes from the receptor surface. Remarkably, while LRH-1 can readily accommodate a range of medium-chain saturated PLs, affinity is highest for the 11- and 12-carbon PCs shown to selectively drive receptor activation in cells [171].

5.5.1 Lipid mediated allosteric control of a protein-protein binding interface

Intact PLs are unusual ligands and LRH-1 has evolved to respond to them, via a novel allosteric pathway to support appropriate interaction with coregulators depending on the ligand status. The idea that ligand binding can drive the selective recruitment of different coregulators

has been hypothesized before; previous MD studies have indicated that the SHP – LRH-1 interaction is weakened upon the binding of phosphatidylserine (PS) to the apo receptor, while binding of DAX-1 and PROX1 is strengthened [203], suggesting that an avenue exists for communication between the LBP and the AF-2 cleft. While no studies have demonstrated a role for PS in the regulation of LRH-1's target genes, recent HDX studies that compared LRH-1 bound to *E. coli* PLs and DLPC demonstrated increased flexibility in both the mouth of the LBP and the AF-2 region in DLPC-bound LRH-1 [178]. Furthermore, stabilizing the mouth of the LBP in apo hLRH-1 by replacing residues 419-424 with the corresponding mouse LRH-1 sequence enhances binding of the coactivators TIF-2 and PGC1 α [204]. In the absence of PLs, the receptor accesses a greater amount of conformational space and readily interacts with corepressors. Medium-chain PLs appear to promote productive motions that favor coactivator interaction and disfavor SHP interaction, perhaps by suppressing non-activating (non-productive) motions to drive selective interaction with coregulators. LRH-1's allosteric network connecting the β -sheet–H6 region may be an evolutionary adaptation that allowed LRH-1 to sense these unusually large ligands and discriminate against fatty acids and cholesterol-derived ligands which would also fit in the receptor's large hydrophobic pocket.

Ideally, structure-function work should be performed and interpreted in the context of the full-length protein. Obtaining a structure of the intact receptor has been challenging, likely due to the large amount of disorder in the linker region connecting the DNA and ligand binding domains. Thus, we modeled systems for which there was empirical structural and biochemical data. In addition, LRH-1 transactivation has been shown to be affected by posttranslational modifications located on the hinge (i.e. phosphorylation, acetylation and SUMOylation) [170]. Phosphorylation of the serine residues S238 and S243 in the hinge region of the human LRH-1

by the mitogen-activated protein kinase ERK1/2 enhances its activity [205]. LRH-1 also been shown to be acetylated in the basal state and is bound by the small heterodimer partner (SHP)-sirtuin 1 (SIRT1) transrepressive complex. Surprisingly SIRT1 does not modulate LRH-1 directly, thus what is driving the acetylation and deacetylation of LRH-1 is not established [206]. LRH-1 transactivation is also controlled by SUMO conjugation to lysine 289 [207]. SUMOylation was shown to drive LRH-1 localization in nuclear bodies, whereby SUMO-conjugated LRH-1 is preferentially sequestered in these bodies preventing it from binding to DNA [207]. Recently, Dr. Kristina Schoonjans's lab showed that SUMOylated LRH-1 interacts with PROX-1, a corepressor, to control 25% of LRH-1 gene targets in the liver. Mutation of lysine 289 to an arginine specifically ablates PROX-1 interaction, without affecting other canonical coregulator interactions.

Emerging evidence suggests that NR activation does not occur via the classically described "mouse trap" model, whereby the AF-H swings from an inactive to active state upon agonist binding. Both experimental and modeling studies are inconsistent with radical repositioning of H12 away from the AF-2 in apoNRs [208-211]. Rather, subtle local conformational adaptations are observed in H12 as well as other regions within the LBD such as the H11-H12 loop, H3 and H5 [211]. These subtle conformational differences between structures may be functionally important, representing a shift between conformational ensembles, but are difficult to identify via inspection of superimposed crystal structures. Previous work with both steroid receptors and fatty acid sensing NRs, have also revealed remarkable flexibility in this region comprising bottom half of the ligand binding pocket including H3, H6-H7 and H11 [212, 213]. In the absence of ligand, NRs are partially unfolded. Recent NMR studies focused on PPAR γ show that in the apo state only half of the expected peaks appear on the intermediate

exchange timescale (milliseconds-to-microseconds). NMR supports a model whereby NRs sample a range of conformations in the apo state. Full-agonists drive this equilibrium towards a more classically active conformation by protecting residues comprising the ligand binding pocket and AF-2 from intermediate exchange, while partial agonists only partially stabilize the regions of the receptor [214]. The β -sheet region may also play an important role in mediating PPAR γ 's response to ligands [208]. While the dynamics in this region are important for mediating ligand action, activation by partial agonists is mediated by the ability of a solvent inaccessible serine residue in this region to be phosphorylated [208].

Given LRH-1's limited selectivity criteria *in vitro*, it is possible that access to endogenous ligands are controlled both temporally and spatially by phospholipid transfer proteins. For example, phospholipid transfer proteins such as phosphatidylinositol transfer protein α and phosphatidylcholine transfer protein are both capable of transporting intact PLs into the nucleus [215, 216]. The effect of tail unsaturation has also not yet been studied, but it is likely that the bends introduced by *cis* unsaturation would allow the LRH-1 ligand binding pocket to accommodate longer-chain acyl tails promoting potent receptor activation. Given the diverse composition of PL tails *in vivo*, these studies are best guided by lipodimics-based identification of endogenous PL ligands. Current limitations in the ability to isolate LRH-1 from mammalian tissue has limited the field's ability to identify endogenous ligands, though these studies are underway.

CHAPTER 6. DISCOVERY OF SELECTIVE INHIBITORS OF TYROSYL-DNA PHOSPHODIESTERASE 2 BY TARGETING THE ENZYME DNA-BINDING CLEFT

6.1 Abstract

Tyrosyl-DNA phosphodiesterase 2 (TDP2) processes protein/DNA adducts resulting from abortive DNA topoisomerase II (Top2) activity. TDP2 inhibition could provide synergism with the Top2 poison class of chemotherapeutics. By virtual screening of the NCI diversity small molecule database, we identified selective TDP2 inhibitors and experimentally verified their selective inhibitory activity. Three inhibitors exhibited low-micromolar IC₅₀ values. Molecular dynamics simulations revealed a common binding mode for these inhibitors, involving association to the TDP2 DNA-binding cleft. MM-PBSA per-residue energy decomposition identified important interactions of the compounds with specific TDP2 residues. These interactions could provide new avenues for synthetic optimization of these scaffolds.

6.2 Results

Tyrosyl-DNA phosphodiesterase (TDP) activity is necessary to cleave the tyrosyl-DNA linkage between a trapped topoisomerase and its substrate DNA[217]. In humans, there are two known TDPs: TDP1 cleaves 3' type-IB topoisomerase-DNA linkages[218], while TDP2 cleaves 5' type-II topoisomerase-DNA linkages[219][220].

Topoisomerase II (Top2) poisons act by trapping Top2 on its DNA substrate, causing the normally transient Top2-mediated double strand breaks to become permanent, resulting in cell death[221]. For this reason, Top2 poisons are widely used cancer therapeutics. TDP2 activity reduces the efficacy of Top2 poisons and is therefore an attractive anticancer drug target, with TDP2 deficient cells exhibiting extreme sensitivity to Top2 poisons[219][222]. The viability of TDP2 knockout mice indicates that TDP2 inhibition is theoretically possible without

unacceptable side effects[222]. A TDP2 inhibitor could have great potential for synergistic effects when used in combination with Top2 poisons and could greatly increase the efficacy of such treatments.

Small-angle X-ray scattering analysis shows that TDP2 consists of a ~110-residue, disordered N-terminal domain and a 255-residue, globular catalytic domain[223, 224]. Only the catalytic domain is necessary for phosphodiesterase activity, while the N-terminal tail is thought to interact with cellular signaling machinery. Although no structures exist for human TDP2 (hTDP2), multiple crystal structures of TDP2 have been solved including *C. elegans* (cTDP2)[224], *D. rerio* (zTDP2)[224] and *M. musculus* (mTDP2)[223]. mTDP2 serves as an excellent structural homologue to hTDP2, with the variants' catalytic domains sharing 78% sequence identity and nearly 100% sequence. A structure of mTDP2* bound to a substrate analog (PDB accession code 4GZ1) shows a short DNA-binding cleft, contacting 3 DNA phosphates, leading directly into the active site. Kinetic studies indicate that mammalian TDP2 is highly specific for 5'-tyrosine overhangs, as opposed to 3' overhangs, blunt ends or other adducts, with the exception of p-nitrophenol, a compound frequently used as a DNA adduct in screening assays[223][225]. Interestingly, TDP2 shares very similar active site geometry and catalytic mechanism with the base excision repair enzyme apurinic/apyrimidinic endonuclease 1 (APE1), along with 14% sequence identity and 30% sequence similarity[224]. Despite these similarities, TDP2 does not show endonuclease activity. From mutational and structural data, Schellenberg, et al. proposed that hydrophobic contacts made to W307 and F325 by the substrate DNA backbone serve as the basis for the specificity toward 5'-tyrosine overhangs by forming favorable Van der Waals interactions with the substrate deoxyribose ring[223]. Crystal structures

of the *D. rerio* and *C. elegans* TDP2 homologues in complex with substrate DNA provide further evidence that hydrophobic contacts are primarily responsible for substrate binding[224]. Binding of a 3'-tyrosine substrate would juxtapose a phosphate group in place of a ribose ring, unfavorably forcing a negatively charged group to contact hydrophobic sidechains. Moreover, the mutation of certain residues lining the DNA-binding cleft greatly alters catalytic activity in hTDP2. Of note are R231, R266, W297 and F315, all of which are important to the activity of TDP2 on 4-nucleotide overhang 5'-tyrosine substrate DNA, with mutations being very deleterious to catalysis[223].

In contrast to TDP2, TDP1 cleaves 3' adducts and is capable of acting on a relatively broad range of substrates[226]. TDP1 also displays some activity against 5'-tyrosine overhangs, causing specificity overlap with TDP2[227, 228]. Designing a TDP2 inhibitor is complicated by the similar catalytic activity and substrate characteristics of TDP1. A potential inhibitor must be strongly selective for TDP2 to limit binding competition by TDP1. Because the TDP2 binding cleft may hold the key for TDP2 substrate specificity, it presents a promising region for targeting inhibitors that do not have activity against TDP1.

The discovery of selective TDP2 inhibitors based on toxoflavin and deazaflavin scaffolds has recently been reported, in which the authors began the inhibitor search with a high-throughput *in vitro* screening of 100,000 compounds[229]. Small molecule docking of this scaffold shows that inhibitors of this type are likely to bind in the active site of TDP2, directly blocking catalysis. While these compounds show promise, they exhibit slight inhibition of TDP1 at 100 μ M. In addition, toxoflavins and deazaflavins also have undesirable characteristics for drug scaffolds. Toxoflavins are susceptible to redox activity and deazaflavins have poor cell permeability.

The difficulty in predicting the binding selectivity of a compound necessitates a high-throughput lead screening and optimization protocol. To this end, we have carried out an inhibitor discovery protocol (Figure 6.2.1) to identify selective TDP2 inhibitors. Our protocol exploits the large scale docking of 11,000 compounds from the 250,000-compound Open National Cancer Institute (NCI) Database [230] followed by virtual screening (VS) and *in vitro* assays using whole cell extract (WCE). With this protocol, we have discovered three potent and selective small molecule inhibitors of TDP2. Results from molecular docking, molecular dynamics (MD), molecular mechanics – Poisson-Boltzmann surface area (MM-PBSA) calculations, biochemical and kinetics experiments provide evidence that our inhibitors bind in the DNA-binding cleft of TDP2, effectively blocking substrate binding and catalysis. The scaffolds of these inhibitors have the potential to be further optimized by targeting specific contacts made to residues in the TDP2 binding cleft, increasing both potency and selectivity for TDP2.

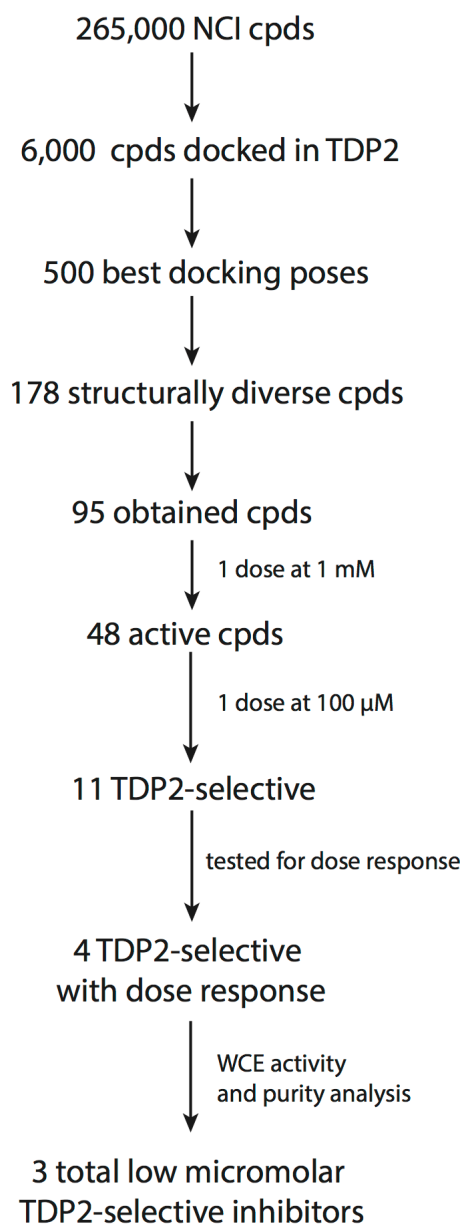


Figure 6.2.1. Flowchart overview of our TDP2 inhibitor discovery process

Compound 21, an APE1 inhibitor provided by Dr. Neamati, was chosen to initiate our screening protocol to identify hTDP2-selective inhibitors because it inhibited hTDP2 with an IC_{50} of 12, 20 μ M ($n=2$) without affecting hTDP1 up to 200 μ M (data not shown). To predict how a ligand will bind to hTDP2, a homology model of the catalytic domain was constructed from the substrate analog-bound mTDP2 structure (PDB accession code 4GZ1) using the program Modeller[231]. Although the homologues have extraordinary sequence similarity, and therefore nearly identical secondary structure, minor alterations to the surface of the enzyme can influence preferred ligand binding positions. Compound 21 was first docked into the hTDP2 homology model. Docking was performed with the program AutoDock4[232], after preparing a ligand/protein atom-pair interaction grid with AutoGrid4 within the search area. The search area was comprised of both the DNA binding cleft and active site, the two regions of the enzyme known to directly impact catalytic activity. The three best-scoring docking poses that were distinct from the others with respect to ligand conformation and ligand-protein contacts were chosen for further analysis. Each pose was simulated for 30ns in MD, using the NAMD2.9 package[75]. The MD trajectory which resulted in the most stable binding pose was clustered with respect to the root mean square deviation (rmsd) of the atomic coordinates of the small molecule using the hierarchical clustering algorithm ($\epsilon=10.0$), as implemented in the ptraj[233] MD analysis program. The representative frame from the most populated cluster was then used for VS. VS against this dominant conformation was performed as follows: sets of fifty random conformers for each compound in the NCI database were first constructed using the Openeye Software program Omega[234] and screened for structural and electrostatic similarity to the dominant MD conformer using Openeye's ROCS[235] and EON[236] programs, respectively. The 11,000 top hits from the VS, as determined by averaging structural and electrostatic

Tanimoto overlaps with the lead compound, were then docked using the AutoDock4 program into our hTDP2 homology model.

From the top 500 docking hits, a set of 178 was selected for optimal structural diversity. From these 178 compounds, 95 were available and were tested against recombinant human TDP2 (hTDP2) at a single concentration of 1 mM (Figure 6.2.1). 48 active compounds were further tested at a single concentration of 100 μ M against hTDP2. 11 TDP2-selective inhibitors were further evaluated in dose response against both recombinant human TDP2 and TDP1 (hTDP2 and hTDP1) enzymes (Figure 6.2.1). Four TDP2-selective low micromolar inhibitors i.e. not inhibiting hTDP1 up to 111 μ M, were checked by LC/MS to assess for their presence in the vial and evaluated for their ability to inhibit endogenous hTDP2 within whole cell extracts from human TDP2-complemented DT40 chicken cells (hTDP2 WCE). The TDP2-selective inhibitors, NSC375976, NSC114532 and NSC3198 were found to inhibit hTDP2 with IC_{50} values of 3.5 ± 1.4 , 4.1 ± 0.9 and 9.3μ M, respectively while all being inactive against recombinant hTDP1 up to 111 μ M (Figure 6.2.2 and Table 6.2.1). All 3 compounds show robustness while being exposed to an increased complexity reaction mixture because they retain low micromolar activity against whole cell extracts (WCE) containing endogenous hTDP2 (Figure 6.2.2).

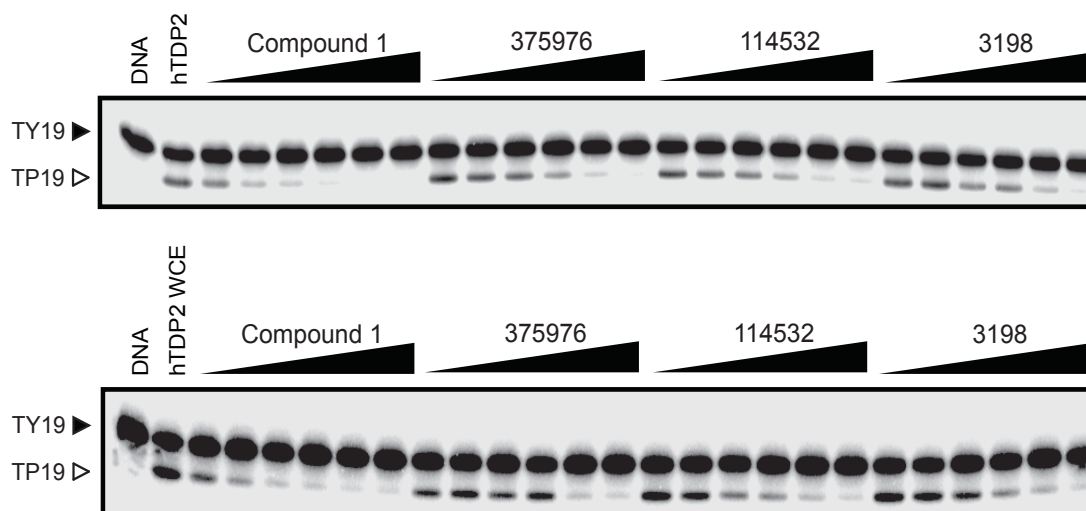


Figure 6.2.2. Inhibition of human recombinant TDP2 (hTDP2) and endogenous human TDP2 from whole cell extracts (hTDP2 WCE) by NSC379576, NSC114532, and NSC3198.

Concentrations of compounds are 0.5, 1.4, 4.1, 12.3, 37 and 111 μM . Concentrations of the positive control Compound 1²² are 0.005, 0.017, 0.05, 0.15, 0.46, 1.4 μM .

The inhibition of TDP2 by these 3 compounds was then assessed across species by comparing their potencies against human (*H. sapiens*, hTDP2), mouse (*M. musculus*, mTDP2) and zebrafish (*D. rerio*, zTDP2) TDP2 enzymes (Figure 6.2.3). Only NSC379576 inhibited efficiently all three enzymes with IC₅₀ values of 15 and 5.2 μM for mTdp2 and zTDP2, respectively (Figure 6.2.3 and Table 6.2.1). NSC114532 and NSC3198 both inhibited zTDP2 with IC₅₀ values of 15 and 20 μM, respectively but interestingly, mTDP2 was totally resistant to both compounds up to 111 μM (Figure 6.2.3 and Table 6.2.1). These results suggest that the TDP2 binding site for both NSC114532 and NSC3198 is not conserved in the mouse enzyme similarly to what was recently observed for the first reported selective TDP2 inhibitor Compound 1 (Figure 6.2.2)[237].

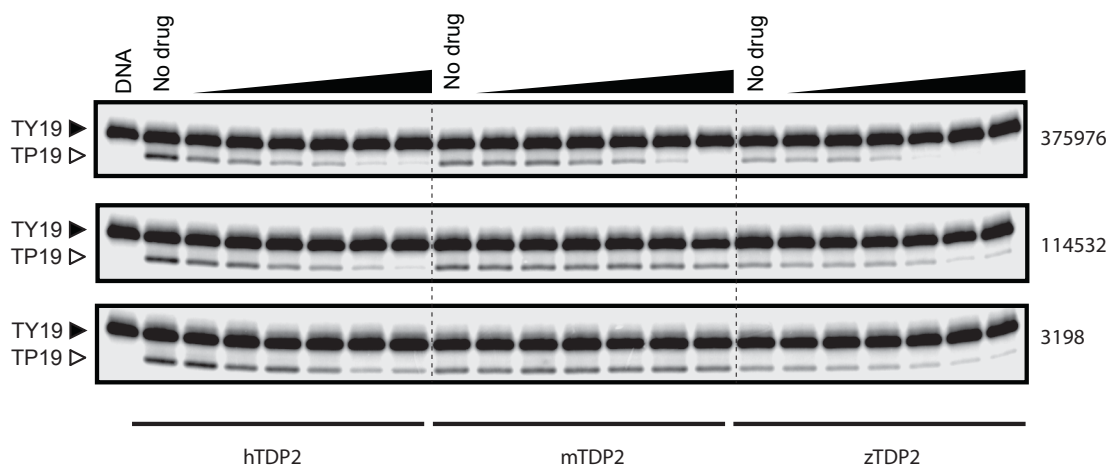
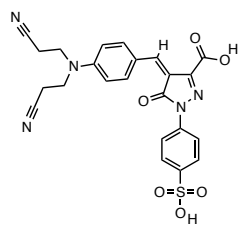
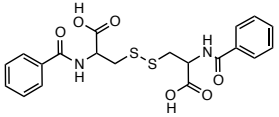
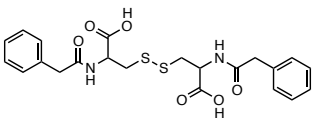


Figure 6.2.3. Inhibition of human (*H. sapiens*, hTDP2), mouse (*M. musculus*, mTDP2) and zebrafish (*D. rerio*, zTDP2) TDP2 enzymes by NSC379576, NSC114532, and NSC3198. Concentrations of compounds are 0.5, 1.4, 4.1, 12.3, 37 and 111 μ M.

A potential concern with the chemotype of NSC114532 and NSC3198 is the presence of a disulfide bond in the center of the structure (Table 6.2.1). The presence of a high concentration of DTT (1mM) in the reaction buffer could therefore potentially cleave the molecule in two. To study the potential impact of the presence of a reducing agent on the stability of NSC114532 and NSC3198, we tested their ability to inhibit hTDP2 in the presence or absence of DTT. We did not observe any difference in the inhibition of hTDP2 by these two compounds under these conditions, suggesting that both compounds are not affected by the composition of the reaction buffer.

Table 6.1. IC₅₀ values against human (*H. sapiens*, hTDP2), mouse (*M. musculus*, mTDP2) and zebrafish (*D. rerio*, zTDP2) TDP2 enzymes and against human TDP1 (hTDP1).

Compound	Structure	IC ₅₀ (μM)			
		hTDP2	mTDP2	zTDP2	hTDP1
NSC375976		3.5 ± 1.4 (n=4)	15 (n=1)	5.2 (n=1)	>111
NSC114532		4.1 ± 0.9 (n=4)	>111	15 (n=1)	>111
NSC3198		9.0, 9.5 (n=2)	>111	20 (n=1)	>111

Molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) calculations[238], as implemented in the AmberTools14 analysis suite[85], were performed on NSC375976, NSC114532 and NSC3198. MM-PBSA is a thermodynamic cycle based method, whereby the solvation enthalpies of the complex ($\Delta H_{complex}$), receptor ($\Delta H_{receptor}$) and ligand (ΔH_{ligand}) are estimated independently, then subtracted to obtain the binding enthalpy ($\Delta H_{binding}$) via:

$$\Delta H_{binding} = \Delta H_{complex} - \Delta H_{receptor} - \Delta H_{ligand} \quad (6.1)$$

From MM-PBSA, NSC114532 and NSC3198 were determined to have binding enthalpies of -18 ± 12 and -30 ± 7.9 kcal/mol, respectively. NSC114532 and NSC3198 have essentially similar binding enthalpies, as expected from their structural similarity and similar IC50 values. NSC375976 has a higher binding enthalpy, 0.1 ± 4.7 kcal/mol, than both NSC114532 and NSC3198, although it possesses the lowest IC50 of the group. This result is less surprising given the dissimilarity of the ligands. It is also possible that the IC50 values do not correlate perfectly with the binding enthalpies due to ligand interactions with other species in WCE. To compare the binding modes of the three inhibitors, we have performed MM-PBSA per-residue energy decomposition[239]. This method calculates the enthalpy of each residue's non-bonded interactions with the ligand, highlighting residue-ligand contacts that are essential to ligand binding, as well as those that are detrimental to it.

NSC114532 and NSC3198 are symmetrical molecules around the disulfide bond. Because these two compounds are very close structural analogs, their modes of binding are strikingly similar, with the majority of the binding enthalpy coming from hydrogen bonding and electrostatic interactions between the ligands' carboxy groups and R231 and R266 (Figure 6.2.4A and 6.2.4B).

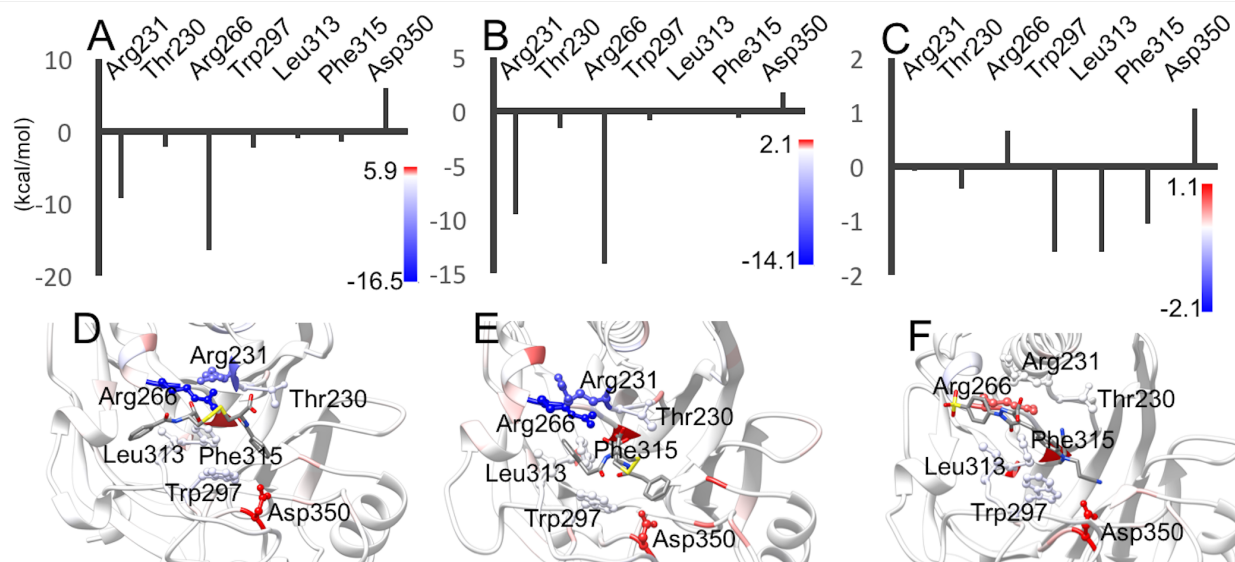


Figure 6.2.4. Binding poses and per-residue energy decomposition for residues important to binding.

Per-residue energy decomposition results for (A) NSC114532, (B) NSC3198 and (C) NSC375976. Dominant binding poses of (D) NSC114532, (E) NSC3198 and (F) NSC375976.

NSC114532 and NSC3198 each also form favorable hydrophobic interactions with residues W297, L313 and F315 that lie in the DNA binding cleft (Figure 6.2.4D and 6.2.4E). NSC375976 does not share the basic chemical scaffold of NSC114532 and NSC3198, yet binds in nearly the same position (Figure 6.2.4C and 6.2.4F). In contrast to NSC114532 and NSC3198, NSC375976 binds predominantly to hydrophobic residues, with W297, L313 and F315 each contributing \sim 1 kcal/mol in enthalpy. The NSC375976 sulfonate group is not able to form favorable electrostatic interactions with R231 and R266, with the adjacent phenyl group clashing slightly with R266. All three inhibitors form favorable van der Waals interactions with the T230 methyl group. The most unfavorable interaction between these inhibitors and hTDP2 is with D350. Each inhibitor places hydrophobic functional groups near the charged D350 side chain, incurring an enthalpic penalty relative to a solvent-exposed D350.

Residues R231, R266, W297, L313 and F315 have all been shown via mutational studies to interact favorably with the DNA substrate, with W297 and F315 having been proposed to form the basis of TDP2 selectivity towards 5'-tyrosine adducts[223]. Clearly, ligand binding in this region of the protein precludes substrate recognition by TDP2, thus inhibiting catalytic activity. Forming favorable contacts with the residues that select for 5' adduct binding is likely to be the cause of these inhibitors' selectivity towards TDP2, as 5' adducts have been shown to be very poor substrates for TDP1. It should be noted that the two distinct molecular scaffolds represented by NSC114532, NSC3198 and NSC375976 both take advantage of the same structural characteristics of the DNA-binding groove, leaving open the possibility of the discovery of other unique scaffolds that have similar binding modes.

To further probe the basis of selectivity of NSC114532, NSC3198 and NSC375976, we docked each into hTDP1, using the same methodology as with hTDP2, again including both the

active site and DNA-binding groove in the search area. The top docking poses from each compound identify the hTDP1 active site as the primary binding target. These results do not indicate that NSC114532, NSC3198 or NSC375976 bind hTDP1. Rather, they serve simply as an indication that, when forced to interact with hTDP1 *in silico*, these inhibitors preferentially interact with the active site. This provides another line of evidence supporting the hypothesis that DNA-binding cleft interactions dictate substrate specificity between hTDP1 and hTDP2 and that our inhibitors select for hTDP2 over hTDP1 based on many of the same interactions. The docking poses of NSC114532 and NSC3198 in hTDP1 are again similar, with both compounds forming hydrogen bonds with S399 and K495. NSC3198 stacks one phenyl ring against H263, while leaving approximately half of the molecule solvent-exposed with no notable favorable interactions with the enzyme. NSC114532 forms no notable favorable stacking interactions with either phenyl ring and forms an electrostatic clash between a carboxylate group and G538. NSC375976 remains largely solvent-exposed by packing against one wall of the active site while reaching slightly into the active site pocket. The only favorable hydrophobic interaction NSC375976 forms is between the phenylsulfonate ring and the aliphatic portion of the S459 sidechain, while the only hydrogen bond to the enzyme is made between the sulfonate group and H263. The combination of unfavorable interactions and inability to take advantage of nonspecific Van der Waals interactions explains these ligands' inability to bind to hTDP1 and, therefore, their selectivity for hTDP2.

NSC114532, NSC3198 and NSC375976 can each be tailored to increase binding affinity to hTDP2 by improving electrostatic complementarity to the binding cleft to enhance binding enthalpy while leaving the Van der Waals contacts in place to maintain selectivity. It is important to note that NSC3198 and NSC114532 form very favorable electrostatic interactions with R231

and R266, residues that are important to the catalytic function of hTDP2. In the future development of these scaffolds, care must be taken to preserve these interactions, meaning that the carboxy groups, as well as the 4-atom linker that separates them, appear to be absolutely essential to effective binding. NSC375976 can likely form very favorable interactions between its sulfonate group and R231 and R266, if the sulfonate group is brought more proximal to the center of the compound. NSC375976 holds the lowest IC_{50} of the compounds studied, taking advantage of this potentially very favorable interaction may greatly improve binding characteristics.

6.3 Methods

6.3.1 Computational methods

6.3.1.1 Molecular dynamics

Molecular dynamics simulations were performed with the NAMD 2.9[75] code with the AMBER ff99SB forcefield[240]. The smooth particle mesh Ewald method (SPME)[29] was used to evaluate long-range electrostatics, with a 12 Å cutoff for non-bonded interactions and a switching function between 10 Å and 12 Å. The r-RESP multiple timestep integration scheme[27] was employed for force integration. The use of a 2fs timestep was enabled by applying SHAKE[241] constraints to all bonds between hydrogen and heavy atoms. The hTDP2 homology model – ligand complexes were solvated in TIP3P water[142] in a rectangular box measuring 67 Å X 68 Å X 72 Å using the XLeap program, a part of the AmberTools package[85]. Starting structures were subjected to 10,000 steps of steepest descent minimization, followed by heating to 300K in the NVT ensemble for 50 ps, with 10 kcal/molÅ² restraints on all protein and ligand heavy atoms. The systems were then simulated in the NPT ensemble for 2 ns,

smoothly releasing the positional restraints. Final production simulations were performed for 30ns at 300K in the NPT ensemble.

6.3.1.2 Ligand optimization and parameterization

Atomic charges were computed with the RESP method[242] in AmberTools after geometry minimization in Gaussian09[243]. Forcefield parameters for the ligands were taken from the general AMBER forcefield (GAFF)[84], with parameters determined using the antechamber functionality in AmberTools and additional force field parameters, when necessary, determined using the parmchk program in AmberTools.

6.3.2 Experimental methods

6.3.2.1 Recombinant TDP2 assay

TDP2 reactions were carried out as described previously[244] with the following modifications. The 18-mer single-stranded oligonucleotide DNA substrate (TY18, α 32P-cordycepin-3'-labeled) was incubated at 1 nM with 25 pM recombinant human TDP2 in the absence or presence of inhibitor for 15 min at room temperature in the LMP2 assay buffer containing 50 mM Tris-HCl, pH 7.5, 80 mM KCl, 5 mM MgCl₂, 0.1 mM EDTA, 1 mM DTT, 40 μ g/mL BSA, and 0.01% Tween 20. Reactions were terminated by the addition of 1 volume of gel loading buffer [99.5% (v/v) formamide, 5 mM EDTA, 0.01% (w/v) xylene cyanol, and 0.01% (w/v) bromophenol blue]. Samples were subjected to a 16% denaturing PAGE with multiple loadings at 12-min intervals. Gels were dried and exposed to a PhosphorImager screen (GE Healthcare). Gel images were scanned using a Typhoon 8600 (GE Healthcare), and densitometry analyses were performed using the ImageQuant software (GE Healthcare).

6.3.2.2 Whole cell extract TDP2 assay

DT40 knockout cells (1 x 10⁷) for TDP2 (TDP2^{-/-}) complemented with human TDP2 (hTDP2) were collected, washed, and centrifuged. Cell pellets were then resuspended in 100 μ L of CellLytic M cell lysis reagent (SIGMA-Aldrich C2978). After 15 min on ice, lysates were centrifuged at 12,000 g for 10 min, and supernatants were transferred to a new tube. Protein concentrations were determined using a Nanodrop spectrophotometer (Invitrogen), and whole cell extracts were stored at -80 °C. The TY18 DNA substrate was incubated at 1 nM with 5 μ g/mL of whole cell extracts in the absence or presence of inhibitor for 15 min at room temperature in the LMP2 assay buffer. Reactions were terminated and treated similarly to recombinant TDP2 reactions (see above).

6.3.2.3 Recombinant TDP1 assay

A 5'-[³²P]-labeled single-stranded DNA oligonucleotide containing a 3'-phosphotyrosine (N14Y) was incubated at 1 nM with 10 pM recombinant TDP1 in the absence or presence of inhibitor for 15 min at room temperature in the LMP1 assay buffer containing 50 mM Tris-HCl, pH 7.5, 80 mM KCl, 2 mM EDTA, 1 mM DTT, 40 μ g/mL BSA, and 0.01% Tween 20. Reactions were terminated and treated similarly to recombinant TDP2 reactions (see above).

6.3.2.4 Kinetics experiments

To determine the kinetic parameters for the inhibition of TDP2 by NSC379576 and NSC114532, 10 pM of TDP2 was incubated at room temperature with 5, 10, 20 and 800 nM of cold TY19 substrate in the absence or presence of 2 and 5 μ M of compound in the LMP2 assay buffer containing 50 mM Tris-HCl, pH 7.5, 80 mM KCl, 5 mM MgCl₂, 0.1 mM EDTA, 1 mM DTT, 40 μ g/mL BSA, and 0.01% Tween 20. All reactions were spiked with 1 nM of ³²P-labeled TY19. The extent of reaction progression was followed in a time-dependent manner and terminated at

different times by adding 1 volume of gel loading buffer. Samples were analyzed by 16% denaturing PAGE, and the initial portions of the reaction curves were fitted to a linear equation to approximate the pre-steady-state reaction velocities using the Prism software (Graphpad). A Lineweaver-Burk plot was then generated with the pre-steady-state reaction velocities and the corresponding substrate concentrations.

CHAPTER 7. PU.1 AND ETS-1 SEQUENCE SPECIFICITY DIVERGENCE THROUGH DIFFERENTIAL ETS-DNA COMPLEX HYDRATION

7.1 Abstract

The DNA-binding domains of ETS family transcription factors are highly structurally homologous despite strongly divergent primary sequences. At opposite ends of the ETS sequence homology spectrum lie the PU.1 and Ets-1 ETS family members (~30% sequence homology). This relatively low sequence homology results in high functional divergence, with PU.1 and Ets-1 exhibiting strongly differing preferences for DNA binding site sequence, despite their very high structural homology (~1.4Å RMSD). While experiments have shown that PU.1 and Ets-1 sequence preference differences are driven at least partially by hydration and salt interactions, the dynamical and conformational differences underlying these effects have remained elusive. We have applied Grid Inhomogeneous Solvation Theory (GIST) to molecular dynamics trajectories to monitor hydration energetics of in a set of PU.1-DNA and Ets-1-DNA complexes differing by PU.1 binding affinity, finding that increases in hydration in the GIST analysis correlate well with increases in PU.1 binding affinity. Dynamical network analysis reveals that these differences in hydration are coupled to allosteric community formation in the DNA. Higher PU.1-affinity sequences exhibit more coherent dynamical DNA communities in PU.1.

7.2 Introduction

The ETS family of transcription factors consist of functionally diverse gene regulators that share a structurally conserved DNA-binding domain known as the ETS domain. Typical of eukaryotic gene families, ETS domains are results of gene duplication followed by divergent evolution [245], resulting in low primary sequence homology and a correspondingly low level of functional redundancy [246] on the one hand, but a highly homologous fold with overlapping DNA site preferences on the other hand [247]. ETS domains therefore represent excellent model systems for studying the biophysical mechanisms responsible for conferring functional heterogeneity on structurally homologous systems. To this end, system-level studies have focused on extrinsic mechanisms of specificity such as cellular expression profiles [248], relative positioning of DNA sites [249], and binding partner proteins [250]. Efforts to identify intrinsic determinants of specificity have met more limited success despite a plethora of high-resolution structures of ETS domains alone or in complex with DNA. We recently reported that the backbone trajectories of the ETS domains of murine PU.1 (Spi1) and Ets-1, the two more sequence-divergent ETS paralogs (~30% amino acid homology), aligned to an RMSD of 1.4 Å [251], well below the precision of the crystallographic structures themselves (>2 Å).

Over the past several years, we have carried out extensive biophysical characterizations of the ETS/DNA complex, comparing PU.1 with Ets-1 as model systems, and identified remarkable heterogeneity in their interactions with the DNA backbone, ions, and water molecules. While the two ETS domains bind their respective optimal sequence with indistinguishably high, nM-affinities under physiologic conditions, the underlying thermodynamics, binding kinetics and the underlying coupling of complex formation with hydration changes, electrostatics, conformational dynamics, and DNA curvature are vastly

different, in some cases qualitatively so [251-253]. More recently, we reported that the two domains are differentially sensitive to CpG-methylation and in fact oppositely inhibited by with hemi-methylated DNA, a phenomenon that arises from the specific effects of CpG methylation on DNA backbone structure [254].

One of the most startling aspects of the mechanistic heterogeneity in DNA recognition between PU.1 and Ets-1 is the hydration properties of the two ETS domains in their DNA-bound state. Using a chemically broad range of cosolutes, we have shown by osmotic stress that, while Ets-1/DNA binding leads to a small net release of hydration water, the high-affinity PU.1/DNA binding is associated with substantial water uptake [252]. The result is that the high-affinity PU.1/DNA binding is distinctly osmotically sensitive, and is substantially destabilized at moderate osmolality (<1 osmolal) [253]. Significantly, osmotic sensitivity is lost in binding to low-affinity sites. The thermodynamic data, however, do not provide a structure to the heterogeneity in hydration of the two ETS paralogs. Here, we report a computational analysis of site-specific DNA complexes of the two proteins by molecular dynamics simulation, using their respective co-crystal structures [255, 256] as starting points.

Our computational analyses indicate that PU.1 binding drives sequence-dependent hydration differences along the DNA binding groove and at the PU.1-DNA interface. This effect is far less pronounced in Ets-1-DNA binding. Furthermore, dynamical network community analysis indicates that these hydration effects arise concurrently with differences in the dynamical nature of the PU.1-bound DNA, with increased dynamical cohesion in highly-solvated portions of the binding interface.

7.3 Materials and Methods

7.3.1 Molecular dynamics setup and simulation

Four DNA sequences were selected for optimal PU.1 affinity differences (binding affinities reported as relative to sequence 2): 5'-AAAACCGGAAGTGGG-3' (-2.46±0.82 kJ/mol), 5'-AAAAAAGGAAGTGGG-3' (0 kJ/mol), 5'-AAAAAAGGAAGAGGG-3' (6.02±0.87 kJ/mol) and 5'-AAAAAAGGAATGGGG-3' (11.14±0.81 kJ/mol)[257]. Each sequence was simulated in the absence of protein, in complex with PU.1 and in complex with Ets-1, resulting in a total of 12 simulations. DNA-only systems were built using the NAB program[258] in the AmberTools14 analysis suite[85]. All PU.1 systems were modeled from the crystal structure 1PUE[256]. DNA sequences in the PU.1-DNA complexes were manually changed from the original crystal structures by altering atom names and building missing coordinates using the xLeap module of AmberTools14. Each PU.1-DNA system was placed in a rectangular waterbox of TIP3P[142] molecules, with no side of the box being less than 10Å from the edge of the solute, in xLeap. Counterions and NaCl were then added to reach a salt concentration of 0.15M. Finally, each system was parameterized with the AMBER14SB forcefield. DNA-only systems were solvated in truncated octahedral waterboxes, with all other system preparation details identical to those of the PU.1-DNA systems. All MD simulations were performed using the CUDA implementation of PMEMD in AMBER14[259]. A 2fs timestep was employed for force integration, with nonbonded interactions calculated directly within a 10 Å cutoff and long-range electrostatics calculated via smooth particle mesh Ewald[29]. Each system was first minimized using a conjugate gradient for 5,000 cycles, followed by heating to 300K over 500ps with 5 kcal/mol Å² harmonic positional restraints imposed on all solute heavy atoms in the canonical ensemble. Harmonic restraints were then iteratively released, first on sidechain

and nucleobase atoms, followed by backbone atoms, over 5ns in the isothermal-isobaric ensemble. Finally, production runs of 500 ns were performed, with the first 100ns discarded as equilibration time, leaving 400 ns for analysis.

7.3.2 Molecular dynamics trajectory analysis

Conformational and hydration analyses were performed using cpptraj[233], as implemented in AmberTools14. All analyses were performed over 10,000 evenly spaced frames from the 400ns production trajectories. Hydration parameters and surfaces were calculated using grid inhomogeneous solvation theory (GIST)[50]. In GIST, a spatial grid is analyzed for relative water residence times and dynamics, relative to bulk solvent. This data allows for the direct calculation of energetic parameters and surfaces. In each case, the GIST grid was set to encompass the DNA and DNA-protein interfaces, separated into 0.5 Å voxels.

Network analysis was performed in the NetworkView[44] plugin of VMD 1.9.1[150]. Networks were constructed as follows: a node was assigned to each protein residue, with the representative atom chosen to be the C_{α} , while two nodes were assigned to each nucleotide; one comprising the phosphate and ribose atoms and the other comprising the nucleobase atoms. If two residues remained within a cutoff contact distance of 4.5 Å for >75% of the trajectory, an edge was placed between their respective nodes, with the edge weight set to the nodes' Cartesian covariance.

7.4 Results and Discussion

In previous osmotic stress studies of hydration changes in DNA recognition by ETS domains, PU.1 and Ets-1 exhibited sharply different osmotic sensitivities that depended only on osmolality of the solution, not the identity of the chemically diverse range of osmolytes employed [252]. Even nicotinamide, a hydrotrope that exhibits preferential mutual attraction

(and a *negative* osmotic coefficient) in water, destabilizes the high-affinity PU.1/DNA complex as well as “typical” co-solutes such as betaine and sucrose [253]. These observations offered confidence that the differential effects of the co-solutes between PU.1 and Ets-1 indeed reflected the hydration properties of DNA recognition, not any weak interactions with the co-solutes themselves. The molecular simulations here are aimed at exploring the structural basis of these experimental observations.

The DNA sites were selected based on experimentally determined affinities for the ETS domain of PU.1 [257] in order to examine how hydration changes correlate with DNA sequence preference.

Table 7.1. Thermodynamic data for GIST analysis and experimental binding free energies.

SEQUENCE	H_{SOLV} (KCAL/MOL)	TS_{SOLV} (KCAL/MOL)	G_{SOLV} (KCAL/MOL)	$\Delta G_{\text{BINDING}}$ EXP. (KCAL/MOL)
1	-153	-14	-139	-0.6
2	-164	-18	-145	0
3	-103	-18	-85	1.4
4	-75	-6	-69	2.7

7.4.1 GIST solvation free energies correlate with experimental binding free energies

As shown in table 7.4.1 and figure 7.4.1.1, GIST analysis reveals that binding free energies correlate well with the experimentally determined binding free energies for PU.1. While a loss in solvent entropy negatively impacts the overall solvation free energy, enthalpic contributions dominate PU.1-DNA solvent interactions, accounting for ~80-90% of the total solvation free energy in each case at 300K. The magnitude of the solvation energies and binding free energies do not match perfectly, indicating that other factors contribute to binding energy, most likely sequence-dependent DNA deformation energies [260].

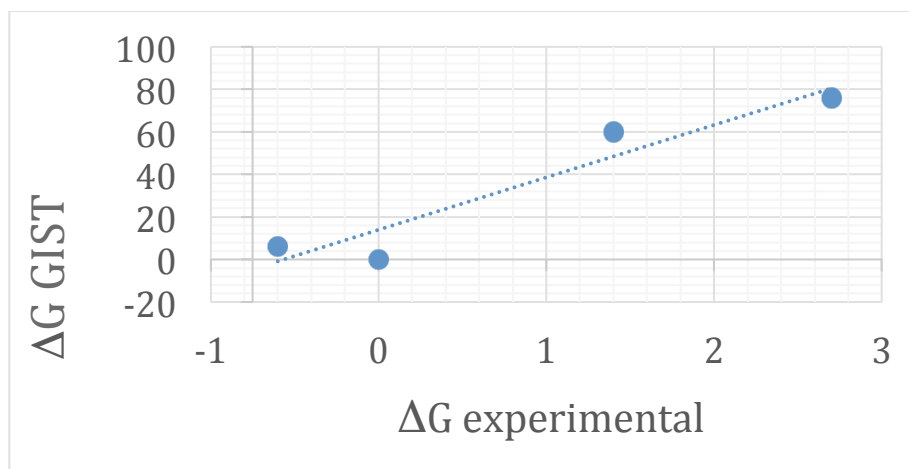


Figure 7.4.1.1. GIST solvation energies vs. experimental binding free energies.

Gist and experimental binding energies correlate well, although despite discrepancy in magnitude.

The solvation energy of sequence 2 does not fit the general trend of the other sequences. Nearly 400ns into the MD simulation, Arg93, an important DNA-binding interface residue capable of direct readout of the consensus portion of the DNA sequence, reorients to form contacts away from the consensus site. This reorientation denotes a qualitative shift in PU.1 binding mode and may represent a statistically significant contribution to the GIST energetics. This transition is observed in the other systems, although it occurs later in the trajectory and likely does not impact the GIST analysis for the other sequences as greatly.

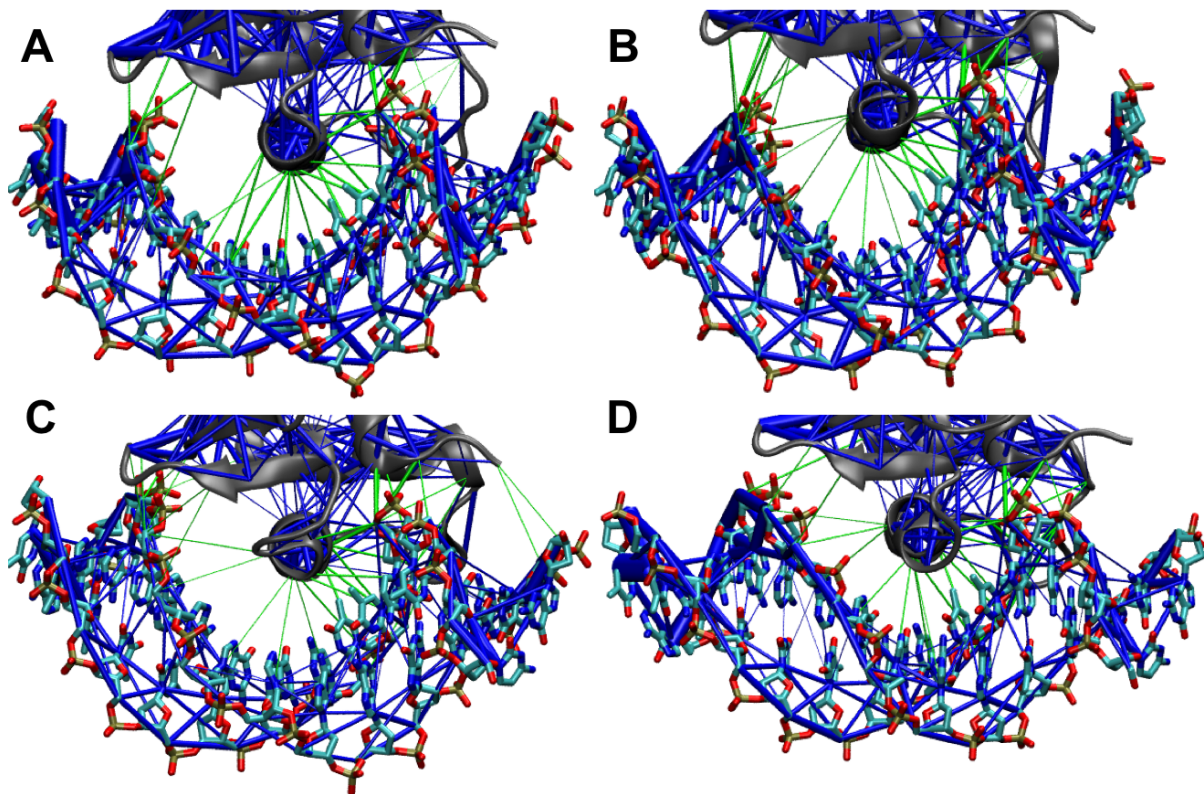


Figure 7.4.2.1. PU.1 networks for sequences.

PU.1-DNA edges (sequences 1-4 in panels A-D, respectively) are green, all other network edges are blue. A qualitative increase and rearrangement in edges is observed in higher-affinity sequences.

7.4.2 Network analysis reveals sequence-dependent shift in binding mode

Network analysis is a direct measure of the cohesiveness of a complex and the strength of specific interactions. As shown in figure 7.4.2.1, PU.1 alters the positioning of the strongest contacts from the 3' end (figure 7.4.2.1, left side of panel A-D left) of the consensus sequence DNA strand to the 5' end, as binding affinity decreases. The rearrangement of edges along the DNA-binding interface indicates preferential binding of the highest-affinity DNA sequences at the DNA-binding interface and to the 5' end of the sequence.

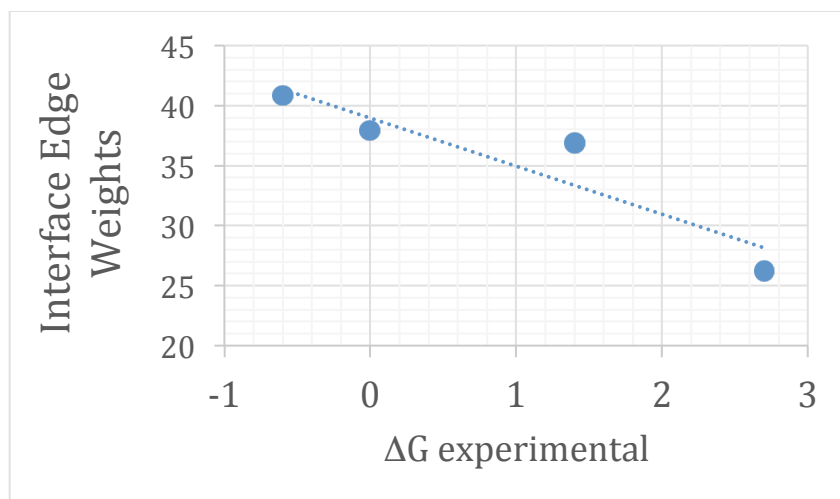


Figure 7.4.2.2. Summed interface edge weights vs. experimental binding free energies. Dynamical coordination is strongly correlated with binding affinity.

Finally, quantification of edge weights for the 28 commonly-held and highest-weight edges between all systems (figure 7.4.2.2) shows a clear linear relationship between binding affinity and dynamical cooperation, as has been observed previously in the case of 3KS nuclear receptor DBD community structures[261].

7.4.3 Ordered solvent increases site-specific complex cohesiveness

While solvent ordering decreases rotational and translational entropy, enthalpic contributions along the PU.1-DNA interface dominate, thus increasing the overall cohesiveness of the complex by mediating electrostatic interactions. Overlaying the GIST free energy isosurfaces with the dynamical networks (figure 7.4.3.1) onto the PU.1-DNA complex structures clearly illustrates that PU.1-DNA motions are most strongly correlated where solvation is greatest, most notably near the consensus binding sequence.

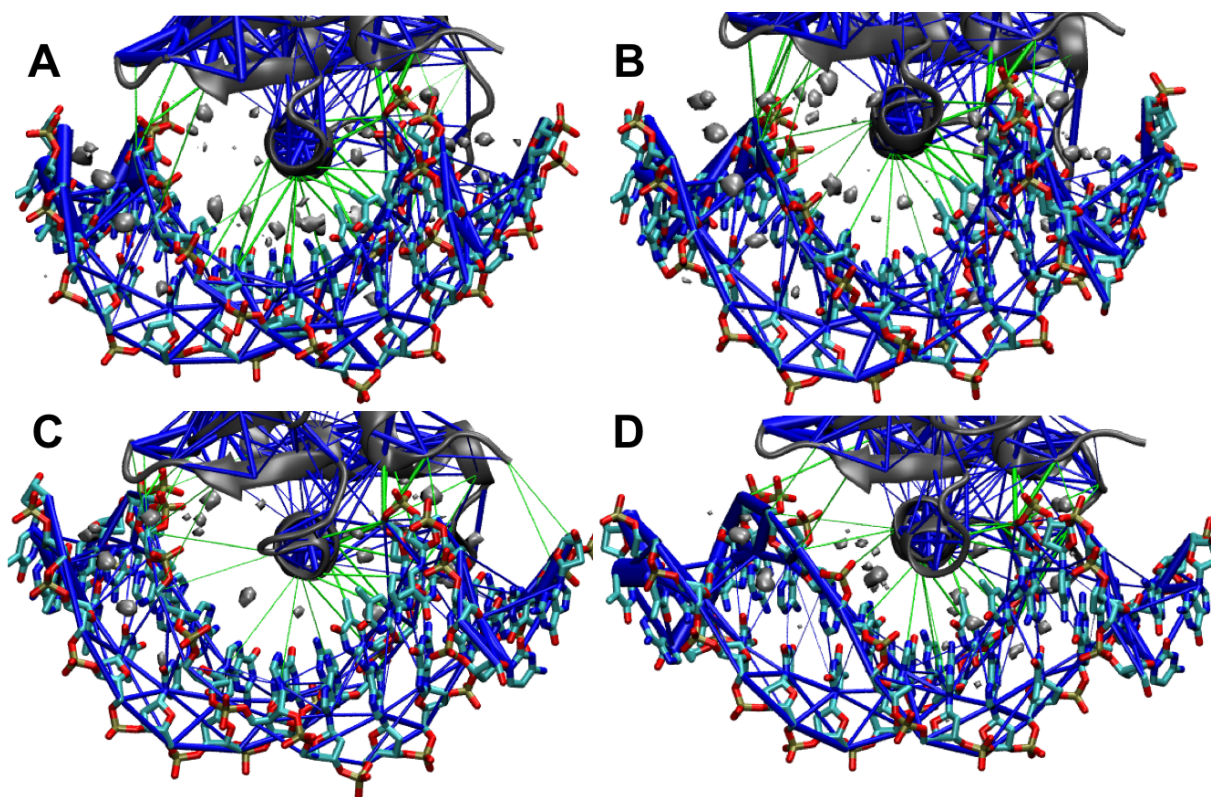


Figure 7.4.3.1. Overlay of networks and solvation isosurfaces.

Network edges collect prominently in regions of the binding interface exhibiting the most solvation.

This qualitative connection between ordered solvent placement and dynamical coordination provides further evidence of the solvent-mediated PU.1 DNA-binding affinity hypothesis proposed from experimental and structural evidence. Interestingly, in all cases network edges are placed in correspondence with interface solvation. Increases in complex-solvent interactions thus drive dynamical cooperation between PU.1 and DNA and enhance binding affinity.

7.5 Future work

The Arg93 repositioning observed in the MD trajectories is likely to have a major impact on both the dynamical networks and solvent ordering in each system. Therefore, the data presented in this chapter is merely preliminary and is subject to change. Nonetheless, the evidence is strong that the interplay between solvent ordering and complex cohesiveness is the driving factor behind PU.1 sequence specificity. As of the writing of this dissertation, the MD simulations are being extended into the microsecond regime to improve the statistical sampling of all PU.1-DNA contacts and conformational states. Given the high binding affinities of NR systems, it is not unrealistic to expect full statistical sampling to require longer timescales.

CHAPTER 8. PERSPECTIVE

8.1 Biomolecular interactions

Specific biomolecular contacts determine the structure of biomolecular complexes. These interactions are critical to inhibitor and drug design, as shown in the discovery of new TDP2 inhibitor scaffolds. While the scaffolds are chemically unique, they interact with TDP2 in very similar ways, making largely the same contacts. In AlkD, rearrangement of the hydrogen bonds formed between Thr39 and Arg43 and the DNA backbone stabilizes the DNA in an unfavorable conformation to shift the chemical equilibrium toward base flipping and stabilize the extrahelical lesion for a sufficient period of time to promote hydrolysis. Solvation and solvent-mediated contacts can also contribute to complex formation. PU.1 modulates sequence-specific binding affinities by maximizing favorable hydration. This favorable solvation lowers the binding free energy of high-affinity DNA sequences although few differences in direct PU.1-DNA contacts are experimentally observed between systems.

8.2 Motions in biochemical systems

Internal motions in biomolecular systems can elucidate and differentiate function. AlkD lowers the energetic barrier of a highly unfavorable dual-base flipping mechanism by splitting it into two distinct single-base flipping steps, while promoting base stack collapse to stabilize the complex at every stage. In LRH-1, PCA reveals two distinct activating and repressing conformational states, as well as the intrinsic motions spanning those states. Subtle motions represented by suboptimal pathways in LRH-1 reveal an experimentally verified allosteric tether linking ligand and co-activator binding sites. Community analysis of the 3-ketosteroid family of NR DBDs highlights a dynamically derived link between evolutionary epistasis and binding affinities for +GRE complexes and details the allosteric signaling path through DNA that favors

GR dimerization on nGRE sequences over SR. Finally, improved solvation energetics in PU.1-DNA complexes contribute to more strongly concerted internal motions, connecting the experimental binding affinities to the solvation dynamics observed in MD.

8.3 Conclusion

The studies comprising this dissertation encompass a broad swath of computational methods and their applications to a selection of diverse biochemical systems. These works share few common threads beyond the biological nature of the questions they have sought to answer. However, the physical phenomena underlying these processes are broadly generalizable. An expanded knowledge of these phenomena enhances not only the potential for leveraging biochemistry in biotechnological pursuits, but also elicits a deeper and more thorough understanding of biology at the most fundamental levels.

REFERENCES

1. Lindahl, T., *Instability and decay of the primary structure of DNA*. Nature, 1993. **362**(6422): p. 709-15.
2. Friedman, J.I. and J.T. Stivers, *Detection of damaged DNA bases by DNA glycosylase enzymes*. Biochemistry, 2010. **49**(24): p. 4957-67.
3. Krokan, H.E., R. Standal, and G. Slupphaug, *DNA glycosylases in the base excision repair of DNA*. Biochem J, 1997. **325 (Pt 1)**: p. 1-16.
4. Seeberg, E., L. Eide, and M. Bjoras, *The base excision repair pathway*. Trends Biochem Sci, 1995. **20**(10): p. 391-7.
5. Venkatesh, S. and J.L. Workman, *Histone exchange, chromatin structure and the regulation of transcription*. Nat Rev Mol Cell Biol, 2015. **16**(3): p. 178-89.
6. Wang, J.C., *Cellular roles of DNA topoisomerases: a molecular perspective*. Nat Rev Mol Cell Biol, 2002. **3**(6): p. 430-40.
7. Champoux, J.J., *DNA topoisomerases: structure, function, and mechanism*. Annu Rev Biochem, 2001. **70**: p. 369-413.
8. Pommier, Y., et al., *Roles of eukaryotic topoisomerases in transcription, replication and genomic stability*. Nat Rev Mol Cell Biol, 2016. **17**(11): p. 703-721.
9. Nitiss, J.L. and K.C. Nitiss, *Tdp2: a means to fixing the ends*. PLoS Genet, 2013. **9**(3): p. e1003370.
10. Nitiss, J.L., *Targeting DNA topoisomerase II in cancer chemotherapy*. Nat Rev Cancer, 2009. **9**(5): p. 338-50.
11. Lonard, D.M. and W. O'Malley B, *Nuclear receptor coregulators: judges, juries, and executioners of cellular regulation*. Mol Cell, 2007. **27**(5): p. 691-700.

12. McKenna, N.J., R.B. Lanz, and B.W. O'Malley, *Nuclear receptor coregulators: cellular and molecular biology*. *Endocr Rev*, 1999. **20**(3): p. 321-44.
13. Kumar, R. and E.B. Thompson, *The structure of the nuclear hormone receptors*. *Steroids*, 1999. **64**(5): p. 310-9.
14. Kuo, M.H. and C.D. Allis, *Roles of histone acetyltransferases and deacetylases in gene regulation*. *Bioessays*, 1998. **20**(8): p. 615-26.
15. Moras, D. and H. Gronemeyer, *The nuclear receptor ligand-binding domain: structure and function*. *Curr Opin Cell Biol*, 1998. **10**(3): p. 384-91.
16. Weatherman, R.V., R.J. Fletterick, and T.S. Scanlan, *Nuclear-receptor ligands and ligand-binding domains*. *Annu Rev Biochem*, 1999. **68**: p. 559-81.
17. Changeux, J.P. and S.J. Edelman, *Allosteric mechanisms of signal transduction*. *Science*, 2005. **308**(5727): p. 1424-8.
18. Khorasanizadeh, S. and F. Rastinejad, *Nuclear-receptor interactions on DNA-response elements*. *Trends Biochem Sci*, 2001. **26**(6): p. 384-90.
19. Thornton, J.W., *Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions*. *Proc Natl Acad Sci U S A*, 2001. **98**(10): p. 5671-6.
20. Karplus, M. and J.A. McCammon, *Molecular dynamics simulations of biomolecules*. *Nat Struct Biol*, 2002. **9**(9): p. 646-52.
21. Ponder, J.W. and D.A. Case, *Force fields for protein simulations*. *Adv Protein Chem*, 2003. **66**: p. 27-85.

22. Wang, J.M., P. Cieplak, and P.A. Kollman, *How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?* Journal of Computational Chemistry, 2000. **21**(12): p. 1049-1074.
23. Vanommeslaeghe, K., et al., *CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields.* J Comput Chem, 2010. **31**(4): p. 671-90.
24. Oostenbrink, C., et al., *A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6.* J Comput Chem, 2004. **25**(13): p. 1656-76.
25. Cornell, W.D., et al., *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995).* Journal of the American Chemical Society, 1996. **118**(9): p. 2309-2309.
26. Monticelli, L. and D.P. Tieleman, *Force fields for classical molecular dynamics.* Methods Mol Biol, 2013. **924**: p. 197-213.
27. Tuckerman, M., B.J. Berne, and G.J. Martyna, *Reversible Multiple Time Scale Molecular-Dynamics.* Journal of Chemical Physics, 1992. **97**(3): p. 1990-2001.
28. Darden, T., D. York, and L. Pedersen, *Particle Mesh Ewald - an $N \cdot \log(N)$ Method for Ewald Sums in Large Systems.* Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
29. Essmann, U., et al., *A Smooth Particle Mesh Ewald Method.* Journal of Chemical Physics, 1995. **103**(19): p. 8577-8593.
30. Roux, B. and T. Simonson, *Implicit solvent models.* Biophys Chem, 1999. **78**(1-2): p. 1-20.

31. Harrach, M.F. and B. Drossel, *Structure and dynamics of TIP3P, TIP4P, and TIP5P water near smooth and atomistic walls of different hydroaffinity*. J Chem Phys, 2014. **140**(17): p. 174501.
32. Lippert, R.A., et al., *Accurate and efficient integration for molecular dynamics simulations at constant temperature and pressure*. Journal of Chemical Physics, 2013. **139**(16).
33. Litniewski, M., *Molecular-Dynamics Method for Simulating Constant Temperature Volume and Temperature Pressure Systems*. Journal of Physical Chemistry, 1993. **97**(15): p. 3842-3848.
34. Toxvaerd, S., *Molecular-Dynamics at Constant Temperature and Pressure*. Physical Review E, 1993. **47**(1): p. 343-350.
35. Maragliano, L., et al., *String method in collective variables: minimum free energy paths and isocommittor surfaces*. J Chem Phys, 2006. **125**(2): p. 24106.
36. West, A.M., R. Elber, and D. Shalloway, *Extending molecular dynamics time scales with milestoning: example of complex kinetics in a solvated peptide*. J Chem Phys, 2007. **126**(14): p. 145104.
37. E, W., W. Ren, and E. Vanden-Eijnden, *Finite temperature string method for the study of rare events*. J Phys Chem B, 2005. **109**(14): p. 6688-93.
38. Bergonzo, C., et al., *A Partial Nudged Elastic Band Implementation for Use with Large or Explicitly Solvated Systems*. Int J Quantum Chem, 2009. **109**(15): p. 3781.
39. Okamoto, Y., *Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations*. J Mol Graph Model, 2004. **22**(5): p. 425-39.

40. Hamelberg, D., J. Mongan, and J.A. McCammon, *Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules*. J Chem Phys, 2004. **120**(24): p. 11919-29.
41. Abrams, C. and G. Bussi, *Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration*. Entropy, 2014. **16**(1): p. 163-199.
42. Kumar, S., et al., *The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules .I. The Method*. Journal of Computational Chemistry, 1992. **13**(8): p. 1011-1021.
43. Newman, M.E.J., *The structure and function of complex networks*. Siam Review, 2003. **45**(2): p. 167-256.
44. Eargle, J. and Z. Luthey-Schulten, *NetworkView: 3D display and analysis of protein.RNA interaction networks*. Bioinformatics, 2012. **28**(22): p. 3000-1.
45. Danon, L., et al., *Comparing community structure identification*. Journal of Statistical Mechanics-Theory and Experiment, 2005.
46. Girvan, M. and M.E. Newman, *Community structure in social and biological networks*. Proc Natl Acad Sci U S A, 2002. **99**(12): p. 7821-6.
47. Newman, M.E., *Modularity and community structure in networks*. Proc Natl Acad Sci U S A, 2006. **103**(23): p. 8577-82.
48. Massova, I. and P.A. Kollman, *Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding*. Perspectives in Drug Discovery and Design, 2000. **18**: p. 113-135.

49. Miller, B.R., 3rd, et al., *MMPBSA.py: An Efficient Program for End-State Free Energy Calculations*. J Chem Theory Comput, 2012. **8**(9): p. 3314-21.
50. Nguyen, C.N., T.K. Young, and M.K. Gilson, *Grid inhomogeneous solvation theory: hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril*. J Chem Phys, 2012. **137**(4): p. 044101.
51. Petermann, E. and T. Helleday, *DNA replication-associated lesions: importance in early tumorigenesis and cancer therapy*. Biochem Soc Trans, 2007. **35**(Pt 5): p. 1352-4.
52. Helleday, T., *DNA repair as treatment target*. European Journal of Cancer, 2011. **47**: p. S333-S335.
53. Helleday, T., et al., *DNA repair pathways as targets for cancer therapy*. Nature Reviews Cancer, 2008. **8**(3): p. 193-204.
54. Adhikari, S., et al., *Targeting base excision repair for chemosensitization*. Anticancer Agents Med Chem, 2008. **8**(4): p. 351-7.
55. Gates, K.S., T. Nooner, and S. Dutta, *Biologically relevant chemical reactions of N7-alkylguanine residues in DNA*. Chem Res Toxicol, 2004. **17**(7): p. 839-56.
56. Sobol, R.W., et al., *Base excision repair intermediates induce p53-independent cytotoxic and genotoxic responses*. Journal of Biological Chemistry, 2003. **278**(41): p. 39951-39959.
57. Boiteux, S. and M. Guillet, *Abasic sites in DNA: repair and biological consequences in Saccharomyces cerevisiae*. DNA Repair, 2004. **3**(1): p. 1-12.
58. Hitomi, K., S. Iwai, and J.A. Tainer, *The intricate structural chemistry of base excision repair machinery: Implications for DNA damage recognition, removal, and repair*. DNA Repair, 2007. **6**(4): p. 410-428.

59. Memisoglu, A. and L. Samson, *Base excision repair in yeast and mammals*. Mutation Research, 2000. **451**(1-2): p. 39-51.
60. Seeberg, E., L. Eide, and M. Bjoras, *The base excision repair pathway*. Trends in Biochemical Sciences, 1995. **20**(10): p. 391-7.
61. Stivers, J.T., *Extrahelical damaged base recognition by DNA glycosylase enzymes*. Chemistry, 2008. **14**(3): p. 786-93.
62. Wyatt, M.D., et al., *3-methyladenine DNA glycosylases: structure, function, and biological importance*. Bioessays, 1999. **21**(8): p. 668-76.
63. Lau, A.Y., et al., *Crystal structure of a human alkylbase-DNA repair enzyme complexed to DNA: mechanisms for nucleotide flipping and base excision*. Cell, 1998. **95**(2): p. 249-58.
64. Pegg, A.E., *Repair of O(6)-alkylguanine by alkyltransferases*. Mutat Res. , 2000 **462**(2-3): p. 83-100.
65. Tubbs, J.L., et al., *Flipping of alkylated DNA damage bridges base and nucleotide excision repair*. Nature, 2009. **459**(7248): p. 808-13.
66. Alseth, I., et al., *A new protein superfamily includes two novel 3-methyladenine DNA glycosylases from Bacillus cereus, AlkC and AlkD*. Mol Microbiol, 2006. **59**(5): p. 1602-9.
67. Rubinson, E.H., et al., *An unprecedented nucleic acid capture mechanism for excision of DNA damage*. Nature, 2010. **468**(7322): p. 406-11.
68. Rubinson, E.H., et al., *A new protein architecture for processing alkylation damaged DNA: the crystal structure of DNA glycosylase AlkD*. J Mol Biol, 2008. **381**(1): p. 13-23.

69. Varnai, P. and R. Lavery, *Base flipping in DNA: Pathways and energetics studied with molecular dynamic simulations*. Journal of the American Chemical Society, 2002. **124**(25): p. 7272-7273.
70. Huang, N., N.K. Banavali, and A.D. MacKerell, *Protein-facilitated base flipping in DNA by cytosine-5-methyltransferase*. Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(1): p. 68-73.
71. Song, K., et al., *An Improved Reaction Coordinate for Nucleic Acid Base Flipping Studies*. Journal of Chemical Theory and Computation, 2009. **5**(11): p. 3105-3113.
72. Giudice, E., P. Varnai, and R. Lavery, *Base pair opening within B-DNA: free energy pathways for GC and AT pairs from umbrella sampling simulations*. Nucleic Acids Res, 2003. **31**(5): p. 1434-43.
73. Priyakumar, U.D. and A.D. MacKerell, Jr., *Computational approaches for investigating base flipping in oligonucleotides*. Chemical Reviews, 2006. **106**(2): p. 489-505.
74. Bergonzo, C., et al., *Energetic preference of 8-oxoG eversion pathways in a DNA glycosylase*. Journal of the American Chemical Society, 2011. **133**(37): p. 14504-6.
75. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD*. Journal of Computational Chemistry, 2005. **26**(16): p. 1781-802.
76. Lavery, R., et al., *Conformational analysis of nucleic acids revisited: Curves+*. Nucleic Acids Res, 2009. **37**(17): p. 5917-29.
77. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics*. Journal of Molecular Graphics & Modelling, 1996. **14**(1): p. 33-38.

78. Rubinson, E.H., P.P. Christov, and B.F. Eichman, *Depurination of N7-Methylguanine by DNA Glycosylase AlkD Is Dependent on the DNA Backbone*. *Biochemistry*, 2013. **52**(42): p. 7363-5.
79. Stivers, J.T. and Y.L. Jiang, *A mechanistic perspective on the chemistry of DNA repair glycosylases*. *Chem Rev*, 2003. **103**(7): p. 2729-59.
80. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water*. *J. Chem. Phys.*, 1983. **79**: p. 926-935.
81. Case, D., et al., *AMBER11*. 2011.
82. Hornak, V., et al., *Comparison of multiple amber force fields and development of improved protein backbone parameters*. *Proteins-Structure Function and Bioinformatics*, 2006. **65**(3): p. 712-725.
83. Pérez, A., et al., *Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of \pm^3 Conformers*. *Biophysical Journal*, 2007. **92**(11): p. 3817-3829.
84. Wang, J., et al., *Development and testing of a general amber force field*. *Journal of Computational Chemistry*, 2004. **25**(9): p. 1157-74.
85. Wang, J., et al., *Automatic atom type and bond type perception in molecular mechanical calculations*. *J Mol Graph Model*, 2006. **25**(2): p. 247-60.
86. Stephens, P.J., et al., *Ab-Initio Calculation of Vibrational Absorption and Circular-Dichroism Spectra Using Density-Functional Force-Fields*. *Journal of Physical Chemistry*, 1994. **98**(45): p. 11623-11627.
87. Gaussian 03, R.C., Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani,

- G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A.; Gaussian, Inc., Wallingford CT, 2004.
88. Kale, L., et al., *NAMD2: Greater scalability for parallel molecular dynamics*. Journal of Computational Physics, 1999. **151**(1): p. 283-312.
89. Essmann, U., et al., *A smooth particle mesh Ewald method*. J. Chem. Phys., 1995. **103**: p. 8577-8593.
90. Grossfield, A., "WHAM: the weighted histogram analysis method", version 2.0.4, <http://membrane.urmc.rochester.edu/content/wham>.
91. Conrad, B. and S.E. Antonarakis, *Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease*. Annu Rev Genom Hum G, 2007. **8**(1): p. 17-35.
92. Conant, G.C. and K.H. Wolfe, *Turning a hobby into a job: How duplicated genes find new functions*. Nat Rev Genet, 2008. **9**(12): p. 938-950.
93. Copley, S., *Enzymes with extra talents: moonlighting functions and catalytic promiscuity*. Curr Opin Chem Biol, 2003. **7**(2): p. 265-272.

94. Tawfik, O.K. and S. Dan, *Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective*. *Annu Rev Biochem*, 2010. **79**(1): p. 471-505.
95. Zhang, J., et al., *Evolution of Minimal Specificity and Promiscuity in Steroid Hormone Receptors*. *PLoS Genet*, 2012. **8**(11): p. e1003072.
96. Morohashi, K., T. Baba, and M. Tanaka, *Steroid hormones and the development of reproductive organs*. *Sex Dev*, 2013. **7**(1-3): p. 61-79.
97. Blaustein, J.D., *Steroid hormone receptors: long- and short-term integrators of the internal milieu and the external environment*. *Horm Metab Res*, 2012. **44**(8): p. 563-8.
98. Webster, N., *The hormone-binding domains of the estrogen and glucocorticoid receptors contain an inducible transcription activation function*. *Cell*, 1988. **54**(2): p. 199-207.
99. Hollenberg, S.M., et al., *Colocalization of DNA-binding and transcriptional activation functions in the human glucocorticoid receptor*. *Cell*, 1987. **49**(1): p. 39-46.
100. Yang-Yen, H.-F., et al., *Transcriptional interference between c-Jun and the glucocorticoid receptor: Mutual inhibition of DNA binding due to direct protein-protein interaction*. *Cell*, 1990. **62**(6): p. 1205-1215.
101. Schüle, R., et al., *Functional antagonism between oncoprotein c-Jun and the glucocorticoid receptor*. *Cell*, 1990. **62**(6): p. 1217-1226.
102. Surjit, M., et al., *Widespread negative response elements mediate direct repression by agonist-liganded glucocorticoid receptor*. *Cell*, 2011. **145**(2): p. 224-241.
103. Luisi, B.F., et al., *Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA*. *Nature*, 1991. **352**(6335): p. 497-505.
104. Hudson, W.H., C. Youn, and E.A. Ortlund, *The structural basis of direct glucocorticoid-mediated transrepression*. *Nat Struct Mol Biol*, 2012. **20**(1): p. 53-58.

105. Kadmiel, M. and J.A. Cidlowski, *Glucocorticoid receptor signaling in health and disease*. Trends Pharmacol Sci, 2013. **34**(9): p. 518-30.
106. Chantong, B., et al., *Mineralocorticoid and glucocorticoid receptors differentially regulate NF-kappaB activity and pro-inflammatory cytokine production in murine BV-2 microglial cells*. J Neuroinflammation, 2012. **9**: p. 260.
107. Berger, S., et al., *Mineralocorticoid receptor knockout mice: pathophysiology of Na⁺ metabolism*. Proc Natl Acad Sci U S A, 1998. **95**(16): p. 9424-9.
108. Arora, V.K., et al., *Glucocorticoid receptor confers resistance to antiandrogens by bypassing androgen receptor blockade*. Cell, 2013. **155**(6): p. 1309-22.
109. Hudson, W.H. and E.A. Ortlund, *The structure, function and evolution of proteins that bind DNA and RNA*. Nat Rev Mol Cell Biol, 2014. **15**(11): p. 749-760.
110. Hudson, W.H., C. Youn, and E.A. Ortlund, *Crystal structure of the mineralocorticoid receptor DNA binding domain in complex with DNA*. PLoS ONE, 2014. **9**(9): p. e107000.
111. Meijsing, S.H., et al., *DNA Binding Site Sequence Directs Glucocorticoid Receptor Structure and Activity*. Science, 2009. **324**(5925): p. 407-410.
112. Baumann, H., et al., *Refined solution structure of the glucocorticoid receptor DNA-binding domain*. Biochemistry, 1993. **32**(49): p. 13463-71.
113. Carroll, S.M., E.A. Ortlund, and J.W. Thornton, *Mechanisms for the Evolution of a Derived Function in the Ancestral Glucocorticoid Receptor*. PLoS Genet, 2011. **7**(6): p. e1002117.
114. McKeown, Alesia N., et al., *Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module*. Cell, 2014. **159**(1): p. 58-68.
115. Beato, M., *Gene regulation by steroid hormones*. Cell, 1989. **56**(3): p. 335-44.

116. Eick, G.N., et al., *Evolution of minimal specificity and promiscuity in steroid hormone receptors*. PLoS Genet, 2012. **8**(11): p. e1003072.
117. Watson, L.C., et al., *The glucocorticoid receptor dimer interface allosterically transmits sequence-specific DNA signals*. Nat Struct Mol Biol, 2013. **20**(7): p. 876-83.
118. Flicek, P., et al., *Ensembl 2014*. Nucleic Acids Res, 2013. **42**(D1): p. D749-D755.
119. Bridgham, J.T., E.A. Ortlund, and J.W. Thornton, *An epistatic ratchet constrains the direction of glucocorticoid receptor evolution*. Nature, 2009. **461**(7263): p. 515-9.
120. Heck, S., et al., *I κ B α -independent downregulation of NF- κ B activity by glucocorticoid receptor*. EMBO J, 1997. **16**(15): p. 4698-4707.
121. Sharma, S. and A. Lichtenstein, *Dexamethasone-induced apoptotic mechanisms in myeloma cells investigated by analysis of mutant glucocorticoid receptors*. Blood, 2008. **112**(4): p. 1338-45.
122. Tao, Y., C. Williams-Skipp, and R.I. Scheinman, *Mapping of glucocorticoid receptor DNA binding domain surfaces contributing to transrepression of NF-kappa B and induction of apoptosis*. J Biol Chem, 2001. **276**(4): p. 2329-32.
123. Liberman, A.C., et al., *Compound A, a dissociated glucocorticoid receptor modulator, inhibits T-bet (Th1) and induces GATA-3 (Th2) activity in immune cells*. PLoS ONE, 2012. **7**(4): p. e35155.
124. Aharoni, A., et al., *The 'evolvability' of promiscuous protein functions*. Nat Genet, 2004.
125. Breen, M.S., et al., *Epistasis as the primary factor in molecular evolution*. Nature, 2012. **490**(7421): p. 535-538.
126. Bloom, J.D. and F.H. Arnold, *In the light of directed evolution: Pathways of adaptive protein evolution*. Proc Natl Acad Sci U S A, 2009. **106**(Supplement_1): p. 9995-10000.

127. Ortlund, E.A., et al., *Crystal structure of an ancient protein: evolution by conformational epistasis*. Science, 2007. **317**(5844): p. 1544-8.
128. Bloom, J.D., L.I. Gong, and D. Baltimore, *Permissive secondary mutations enable the evolution of influenza oseltamivir resistance*. Science, 2010. **328**(5983): p. 1272-5.
129. Gong, L.I., M.A. Suchard, and J.D. Bloom, *Stability-mediated epistasis constrains the evolution of an influenza protein*. Elife, 2013. **2**: p. e00631.
130. Natarajan, C., et al., *Epistasis among adaptive mutations in deer mouse hemoglobin*. Science, 2013. **340**(6138): p. 1324-7.
131. Harms, M.J. and J.W. Thornton, *Historical contingency and its biophysical basis in glucocorticoid receptor evolution*. Nature, 2014. **512**(7513): p. 203-7.
132. Anderson, D.W., A.N. McKeown, and J.W. Thornton, *Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites*. Elife, 2015. **4**: p. e07864.
133. Tokuriki, N. and D.S. Tawfik, *Protein Dynamism and Evolvability*. Science, 2009. **324**(5924): p. 203-207.
134. Harms, M.J., et al., *Biophysical mechanisms for large-effect mutations in the evolution of steroid hormone receptors*. Proc Natl Acad Sci U S A, 2013. **110**(28): p. 11475-11480.
135. Wilson, C., et al., *Kinase dynamics. Using ancient protein kinases to unravel a modern cancer drug's mechanism*. Science, 2015. **347**(6224): p. 882-6.
136. Romero, P.A. and F.H. Arnold, *Exploring protein fitness landscapes by directed evolution*. Nat Rev Mol Cell Biol, 2009. **10**(12): p. 866-876.
137. Adams, P.D., et al., *PHENIX: a comprehensive Python-based system for macromolecular structure solution*. Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 2): p. 213-21.

138. McCoy, A.J., et al., *Phaser crystallographic software*. J Appl Crystallogr, 2007. **40**(4): p. 658-674.
139. Emsley, P., et al., *Features and development of Coot*. Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 4): p. 486-501.
140. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis*. J Comput Chem, 2004. **25**(13): p. 1605-12.
141. Joosten, R.P., et al., *The PDB_REDO server for macromolecular structure model optimization*. IUCrJ, 2014. **1**(Pt 4): p. 213-20.
142. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water*. The Journal of Chemical Physics, 1983. **79**(2): p. 926-935.
143. Perez, A., et al., *Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers*. Biophys J, 2007. **92**(11): p. 3817-29.
144. Wang, J.M., P. Cieplak, and P.A. Kollman, *How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?* J Comput Chem, 2000. **21**(12): p. 1049-1074.
145. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD*. J Comput Chem, 2005. **26**(16): p. 1781-802.
146. Tuckerman, M., B.J. Berne, and G.J. Martyna, *Reversible Multiple Time Scale Molecular-Dynamics*. J Chem Phys, 1992. **97**(3): p. 1990-2001.
147. Forester, T.R. and W. Smith, *SHAKE, rattle, and roll: Efficient constraint algorithms for linked rigid bodies (vol 19, pg 102, 1998)*. J Comput Chem, 2000. **21**(2): p. 157-157.
148. Essmann, U., et al., *A Smooth Particle Mesh Ewald Method*. J Chem Phys, 1995. **103**(19): p. 8577-8593.

149. Sethi, A., et al., *Dynamical networks in tRNA:protein complexes*. Proc Natl Acad Sci U S A, 2009. **106**(16): p. 6620-5.
150. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics*. J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.
151. Newman, M.E.J., *A measure of betweenness centrality based on random walks*. Social Networks, 2005. **27**(1): p. 39-54.
152. Floyd, R.W., *Algorithm 97: Shortest path*. Commun ACM, 1962. **5**(6): p. 345.
153. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput*. Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
154. Guindon, S. and O. Gascuel, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood*. Syst Biol, 2003. **52**(5): p. 696-704.
155. Anisimova, M. and O. Gascuel, *Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative*. Syst Biol, 2006. **55**(4): p. 539-52.
156. Yang, Z., S. Kumar, and M. Nei, *A new method of inference of ancestral nucleotide and amino acid sequences*. Genetics, 1995. **141**(4): p. 1641-50.
157. Yang, Z., *PAML: a program package for phylogenetic analysis by maximum likelihood*. Comput Appl Biosci, 1997. **13**(5): p. 555-6.
158. Hanson-Smith, V., B. Kolaczkowski, and J.W. Thornton, *Robustness of ancestral sequence reconstruction to phylogenetic uncertainty*. Mol Biol Evol, 2010. **27**(9): p. 1988-99.
159. Musille, P.M., J.A. Kohn, and E.A. Ortlund, *Phospholipid – Driven gene regulation*. FEBS Letters, 2013.

160. Blind, R.D., et al., *The signaling phospholipid PIP3 creates a new interaction surface on the nuclear receptor SF-1*. Proc Natl Acad Sci U S A, 2014. **111**(42): p. 15054-9.
161. Liu, S., et al., *A diurnal serum lipid integrates hepatic lipogenesis and peripheral fatty acid use*. Nature, 2013. **502**(7472): p. 550-4.
162. Walker, A.K., et al., *A conserved SREBP-1/phosphatidylcholine feedback circuit regulates lipogenesis in metazoans*. Cell, 2011. **147**(4): p. 840-52.
163. Fernandez-Marcos, P.J., J. Auwerx, and K. Schoonjans, *Emerging actions of the nuclear receptor LRH-1 in the gut*. Biochim Biophys Acta, 2010.
164. Moore, D. *Targeting nuclear receptors to treat type 2 diabetes*. in *14th International Congress of Endocrinology*. 2010. Kyoto, Japan.
165. Oosterveer, M.H. and K. Schoonjans, *Hepatic glucose sensing and integrative pathways in the liver*. Cell Mol Life Sci, 2013.
166. Zhang, C., et al., *Liver receptor homolog-1 is essential for pregnancy*. Nature Medicine, 2013. **19**(8): p. 1061-1066.
167. Gerrits, H., et al., *Reversible infertility in a liver receptor homologue-1 (LRH-1)-knockdown mouse model*. Reprod Fertil Dev, 2014. **26**(2): p. 293-306.
168. Kelly, V.R., et al., *Dax1 Up-Regulates Oct4 Expression in Mouse Embryonic Stem Cells via LRH-1 and SRA*. Molecular endocrinology (Baltimore, Md.), 2010. **24**: p. 1-11.
169. Venteclef, N., et al., *Metabolic nuclear receptor signaling and the inflammatory acute phase response*. Trends in endocrinology and metabolism: TEM, 2011: p. 1-11.
170. Stein, S. and K. Schoonjans, *Molecular basis for the regulation of the nuclear receptor LRH-1*. Curr Opin Cell Biol, 2014. **33C**: p. 26-34.

171. Lee, J.M., et al., *A nuclear-receptor-dependent phosphatidylcholine pathway with antidiabetic effects*. *Nature*, 2011.
172. Bolado-Carrancio, A., et al., *Activation of nuclear receptor NR5A2 increases Glut4 expression and glucose metabolism in muscle cells*. *Biochem Biophys Res Commun*, 2014. **446**(2): p. 614-9.
173. Mamrosh, J.L., et al., *Nuclear receptor LRH-1/NR5A2 is required and targetable for liver endoplasmic reticulum stress resolution*. *eLife*, 2014. **3**(0): p. e01694-e01694.
174. Whitby, R.J., et al., *Identification of small molecule agonists of the orphan nuclear receptors liver receptor homolog-1 and steroidogenic factor-1*. *Journal of medicinal chemistry*, 2006. **49**: p. 6652-5.
175. Whitby, R.J., et al., *Small Molecule Agonists of the Orphan Nuclear Receptors Steroidogenic Factor-1 (SF-1, NR5A1) and Liver Receptor Homologue-1 (LRH-1, NR5A2)*. *Journal of medicinal chemistry*, 2011. **1**.
176. Busby, S., et al., *Discovery of Inverse Agonists for the Liver Receptor Homologue-1 (LRH1; NR5A2)*, in *Probe Reports from the NIH Molecular Libraries Program*. 2010: Bethesda (MD).
177. Benod, C., et al., *Structure-based discovery of antagonists of nuclear receptor LRH-1*. *J Biol Chem*, 2013. **288**(27): p. 19830-44.
178. Musille, P.M., et al., *Antidiabetic phospholipid-nuclear receptor complex reveals the mechanism for phospholipid-driven gene regulation*. *Nat Struct Mol Biol*, 2012. **19**(5): p. 532-7, S1-2.
179. Goodwin, B., et al., *A regulatory cascade of the nuclear receptors FXR, SHP-1, and LRH-1 represses bile acid biosynthesis*. *Molecular cell*, 2000. **6**: p. 517-26.

180. Zhi, X., et al., *Structural insights into gene repression by the orphan nuclear receptor SHP*. Proc Natl Acad Sci U S A, 2014. **111**(2): p. 839-44.
181. Goodwin, B., et al., *Differential regulation of rat and human CYP7A1 by the nuclear oxysterol receptor liver X receptor-alpha*. Mol Endocrinol, 2003. **17**(3): p. 386-94.
182. Lu, T.T., et al., *Molecular basis for feedback regulation of bile acid synthesis by nuclear receptors*. Molecular cell, 2000. **6**: p. 507-15.
183. Ortlund, E.A., et al., *Modulation of human nuclear receptor LRH-1 activity by phospholipids and SHP*. Nat Struct Mol Biol, 2005. **12**(4): p. 357-63.
184. Otwinowski, Z. and W. Minor, [20] *Processing of X-ray diffraction data collected in oscillation mode*. 1997. **276**: p. 307-326.
185. Murshudov, G.N., et al., *REFMAC5 for the refinement of macromolecular crystal structures*. Acta Crystallogr D Biol Crystallogr, 2011. **67**(Pt 4): p. 355-67.
186. Echols, N., et al., *Graphical tools for macromolecular crystallography in PHENIX*. J Appl Crystallogr, 2012. **45**(Pt 3): p. 581-586.
187. Nicholls, R.A., et al., *Conformation-independent structural comparison of macromolecules with ProSMART*. Acta Crystallogr D Biol Crystallogr, 2014. **70**(Pt 9): p. 2487-99.
188. B. R. Brooks, R.E.B., B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations*. Journal of Computational Chemistry, 1983. **4**: p. 187-217.
189. Glykos, N.M., *Software news and updates. Carma: a molecular dynamics analysis program*. Journal of Computational Chemistry, 2006. **27**(14): p. 1765-8.

190. Sablin, E.P., et al., *Structural basis for ligand-independent activation of the orphan nuclear receptor LRH-1*. *Molecular cell*, 2003. **11**: p. 1575-85.
191. Wang, W., et al., *The crystal structures of human steroidogenic factor-1 and liver receptor homologue-1*. *Proceedings of the National Academy of Sciences of the United States of America*, 2005. **102**: p. 7505-10.
192. Krylova, I.N., et al., *Structural analyses reveal phosphatidyl inositols as ligands for the NR5 orphan receptors SF-1 and LRH-1*. *Cell*, 2005. **120**: p. 343-55.
193. Blind, R.D., M. Suzawa, and H.A. Ingraham, *Direct Modification and Activation of a Nuclear Receptor-PIP2 Complex by the Inositol Lipid Kinase IPMK*. *Science Signaling*, 2012. **5**(229).
194. Schaaf, G., et al., *Functional anatomy of phospholipid binding and regulation of phosphoinositide homeostasis by proteins of the sec14 superfamily*. *Mol Cell*, 2008. **29**(2): p. 191-206.
195. Kanno, K., et al., *Structure and function of phosphatidylcholine transfer protein (PC-TP)/StarD2*. *Biochim Biophys Acta*, 2007. **1771**(6): p. 654-62.
196. Schouten, A., et al., *Structure of apo-phosphatidylinositol transfer protein alpha provides insight into membrane association*. *EMBO J*, 2002. **21**(9): p. 2117-21.
197. Nagy, L. and J.W. Schwabe, *Mechanism of the nuclear receptor molecular switch*. *Trends Biochem Sci*, 2004. **29**(6): p. 317-24.
198. Li, Y., M.H. Lambert, and H.E. Xu, *Activation of nuclear receptors: a perspective from structural genomics*. *Structure (London, England : 1993)*, 2003. **11**: p. 741-6.
199. Sablin, E.P., et al., *Structural basis for ligand-independent activation of the orphan nuclear receptor LRH-1*. *Mol Cell*, 2003. **11**(6): p. 1575-85.

200. Levy, R.M., et al., *Quasi-harmonic method for studying very low frequency modes in proteins*. Biopolymers, 1984. **23**(6): p. 1099-112.
201. Skjaerven, L., A. Martinez, and N. Reuter, *Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit*. Proteins, 2011. **79**(1): p. 232-43.
202. Atkins, W.M., *Biological messiness vs. biological genius: Mechanistic aspects and roles of protein promiscuity*. J Steroid Biochem Mol Biol, 2014.
203. Burendahl, S., E. Treuter, and L. Nilsson, *Molecular dynamics simulations of human LRH-1: the impact of ligand binding in a constitutively active nuclear receptor*. Biochemistry, 2008. **47**(18): p. 5205-15.
204. Musille, P.M., et al., *Divergent sequence tunes ligand sensitivity in phospholipid-regulated hormone receptors*. J Biol Chem, 2013. **288**(28): p. 20702-12.
205. Lee, Y.K., et al., *Phosphorylation of the hinge domain of the nuclear hormone receptor LRH-1 stimulates transactivation*. J Biol Chem, 2006. **281**(12): p. 7850-5.
206. Chanda, D., Y.B. Xie, and H.S. Choi, *Transcriptional corepressor SHP recruits SIRT1 histone deacetylase to inhibit LRH-1 transactivation*. Nucleic Acids Res, 2010. **38**(14): p. 4607-19.
207. Chalkiadaki, A. and I. Talianidis, *SUMO-dependent compartmentalization in promyelocytic leukemia protein nuclear bodies prevents the access of LRH-1 to chromatin*. Mol Cell Biol, 2005. **25**(12): p. 5095-105.
208. Hughes, T.S., et al., *Ligand and receptor dynamics contribute to the mechanism of graded PPARgamma agonism*. Structure, 2012. **20**(1): p. 139-50.

209. Martinez, L., I. Polikarpov, and M.S. Skaf, *Only subtle protein conformational adaptations are required for ligand binding to thyroid hormone receptors: simulations using a novel multipoint steered molecular dynamics approach*. J Phys Chem B, 2008. **112**(34): p. 10741-51.
210. Batista, M.R. and L. Martinez, *Dynamics of nuclear receptor Helix-12 switch of transcription activation by modeling time-resolved fluorescence anisotropy decays*. Biophys J, 2013. **105**(7): p. 1670-80.
211. Mackinnon, J.A., et al., *Allosteric mechanisms of nuclear receptors: insights from computational simulations*. Mol Cell Endocrinol, 2014. **393**(1-2): p. 75-82.
212. Bledsoe, R.K., et al., *Crystal structure of the glucocorticoid receptor ligand binding domain reveals a novel mode of receptor dimerization and coactivator recognition*. Cell, 2002. **110**(1): p. 93-105.
213. Gee, A.C. and J.A. Katzenellenbogen, *Probing conformational changes in the estrogen receptor: evidence for a partially unfolded intermediate facilitating ligand binding and release*. Mol Endocrinol, 2001. **15**(3): p. 421-8.
214. Kojetin, D.J. and T.P. Burris, *Small molecule modulation of nuclear receptor conformational dynamics: implications for function and drug discovery*. Mol Pharmacol, 2013. **83**(1): p. 1-8.
215. Nile, A.H., V.A. Bankaitis, and A. Grabon, *Mammalian diseases of phosphatidylinositol transfer proteins and their homologs*. Clin Lipidol, 2010. **5**(6): p. 867-897.
216. Kang, H.W., et al., *Regulatory role for phosphatidylcholine transfer protein/StarD2 in the metabolic response to peroxisome proliferator activated receptor alpha (PPARalpha)*. Biochim Biophys Acta, 2010. **1801**(4): p. 496-502.

217. Pommier, Y., et al., *Tyrosyl-DNA-phosphodiesterases (TDP1 and TDP2)*. DNA Repair (Amst), 2014. **19**: p. 114-29.
218. Liu, C., J.J. Pouliot, and H.A. Nash, *Repair of topoisomerase I covalent complexes in the absence of the tyrosyl-DNA phosphodiesterase Tdp1*. Proc Natl Acad Sci U S A, 2002. **99**(23): p. 14970-5.
219. Zeng, Z., et al., *TDP2/TTRAP is the major 5'-tyrosyl DNA phosphodiesterase activity in vertebrate cells and is critical for cellular resistance to topoisomerase II-induced DNA damage*. J Biol Chem, 2011. **286**(1): p. 403-9.
220. Caldecott, K.W., *Tyrosyl DNA phosphodiesterase 2, an enzyme fit for purpose*. Nat Struct Mol Biol, 2012. **19**(12): p. 1212-3.
221. Pommier, Y., et al., *DNA topoisomerases and their poisoning by anticancer and antibacterial drugs*. Chem Biol, 2010. **17**(5): p. 421-33.
222. Gomez-Herreros, F., et al., *TDP2-dependent non-homologous end-joining protects against topoisomerase II-induced DNA breaks and genome instability in cells and in vivo*. PLoS Genet, 2013. **9**(3): p. e1003226.
223. Schellenberg, M.J., et al., *Mechanism of repair of 5'-topoisomerase II-DNA adducts by mammalian tyrosyl-DNA phosphodiesterase 2*. Nat Struct Mol Biol, 2012. **19**(12): p. 1363-71.
224. Shi, K., et al., *Structural basis for recognition of 5'-phosphotyrosine adducts by Tdp2*. Nat Struct Mol Biol, 2012. **19**(12): p. 1372-7.
225. Adhikari, S., et al., *Development of a novel assay for human tyrosyl DNA phosphodiesterase 2*. Anal Biochem, 2011. **416**(1): p. 112-6.

226. Raymond, A.C., B.L. Staker, and A.B. Burgin, Jr., *Substrate specificity of tyrosyl-DNA phosphodiesterase I (Tdp1)*. J Biol Chem, 2005. **280**(23): p. 22029-35.
227. Murai, J., et al., *Tyrosyl-DNA phosphodiesterase I (TDPI) repairs DNA damage induced by topoisomerases I and II and base alkylation in vertebrate cells*. J Biol Chem, 2012. **287**(16): p. 12848-57.
228. Bahmed, K., K.C. Nitiss, and J.L. Nitiss, *Yeast Tdp1 regulates the fidelity of nonhomologous end joining*. Proc Natl Acad Sci U S A, 2010. **107**(9): p. 4057-62.
229. Raoof, A., et al., *Toxoflavins and deazaflavins as the first reported selective small molecule inhibitors of tyrosyl-DNA phosphodiesterase II*. J Med Chem, 2013. **56**(16): p. 6352-70.
230. Ihlenfeldt, W.D., et al., *Enhanced CACTVS browser of the Open NCI Database*. J Chem Inf Comput Sci, 2002. **42**(1): p. 46-57.
231. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints*. J Mol Biol, 1993. **234**(3): p. 779-815.
232. Morris, G.M., et al., *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility*. J Comput Chem, 2009. **30**(16): p. 2785-91.
233. Roe, D.R. and T.E. Cheatham, *PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data*. Journal of Chemical Theory and Computation, 2013. **9**(7): p. 3084-3095.
234. Hawkins, P.C., et al., *Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database*. J Chem Inf Model, 2010. **50**(4): p. 572-84.

235. Hawkins, P.C., A.G. Skillman, and A. Nicholls, *Comparison of shape-matching and docking as virtual screening tools*. J Med Chem, 2007. **50**(1): p. 74-82.
236. Tresadern, G., D. Bemporad, and T. Howe, *A comparison of ligand based virtual screening methods and application to corticotropin releasing factor 1 receptor*. J Mol Graph Model, 2009. **27**(8): p. 860-70.
237. Marchand, C., et al., *Deazaflavin Inhibitors of Tyrosyl-DNA Phosphodiesterase 2 (TDP2) Specific for the Human Enzyme and Active against Cellular TDP2*. ACS Chem Biol, 2016.
238. Miller, B.R., et al., *MMPBSA.py: An Efficient Program for End-State Free Energy Calculations*. Journal of Chemical Theory and Computation, 2012. **8**(9): p. 3314-3321.
239. Gohlke, H., C. Kiel, and D.A. Case, *Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes*. J Mol Biol, 2003. **330**(4): p. 891-913.
240. Hornak, V., et al., *Comparison of multiple Amber force fields and development of improved protein backbone parameters*. Proteins, 2006. **65**(3): p. 712-25.
241. Forester, T.R. and W. Smith, *SHAKE, rattle, and roll: Efficient constraint algorithms for linked rigid bodies (vol 19, pg 102, 1998)*. Journal of Computational Chemistry, 2000. **21**(2): p. 157-157.
242. Bayly, C.I., et al., *A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges - the Resp Model*. Journal of Physical Chemistry, 1993. **97**(40): p. 10269-10280.
243. Frisch, M.J., et al., *Gaussian 09*. 2009, Gaussian, Inc.: Wallingford, CT, USA.

244. Marchand, C., et al., *Biochemical assays for the discovery of TDPI inhibitors*. Mol Cancer Ther, 2014. **13**(8): p. 2116-26.
245. Wang, Z. and Q. Zhang, *Genome-wide identification and evolutionary analysis of the animal specific ETS transcription factor family*. Evol Bioinform Online, 2009. **5**: p. 119-31.
246. Kopp, J.L., et al., *Unique and selective effects of five Ets family members, Elf3, Ets1, Ets2, PEA3, and PU.1, on the promoter of the type II transforming growth factor-beta receptor gene*. J Biol Chem, 2004. **279**(19): p. 19407-20.
247. Wei, G.H., et al., *Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo*. EMBO J, 2010. **29**(13): p. 2147-60.
248. Hollenhorst, P.C., D.A. Jones, and B.J. Graves, *Expression profiles frame the promoter specificity dilemma of the ETS family of transcription factors*. Nucleic Acids Res, 2004. **32**(18): p. 5693-702.
249. Hollenhorst, P.C., et al., *Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ETS gene family*. Genes Dev, 2007. **21**(15): p. 1882-94.
250. Hollenhorst, P.C., et al., *DNA specificity determinants associate with distinct transcription factor functions*. PLoS Genet, 2009. **5**(12): p. e1000778.
251. He, G., et al., *Heterogeneous dynamics in DNA site discrimination by the structurally homologous DNA-binding domains of ETS-family transcription factors*. Nucleic Acids Res, 2015. **43**(8): p. 4322-31.

252. Wang, S., et al., *Mechanistic heterogeneity in site recognition by the structurally homologous DNA-binding domains of the ETS family transcription factors Ets-1 and PU.1*. J Biol Chem, 2014. **289**(31): p. 21605-16.
253. Poon, G.M., *Sequence discrimination by DNA-binding domain of ETS family transcription factor PU.1 is linked to specific hydration of protein-DNA interface*. J Biol Chem, 2012. **287**(22): p. 18297-307.
254. Stephens, D.C. and G.M. Poon, *Differential sensitivity to methylated DNA by ETS-family transcription factors is intrinsically encoded in their DNA-binding domains*. Nucleic Acids Res, 2016.
255. Garvie, C.W., J. Hagman, and C. Wolberger, *Structural studies of Ets-1/Pax5 complex formation on DNA*. Mol Cell, 2001. **8**(6): p. 1267-76.
256. Kodandapani, R., et al., *A new pattern for helix-turn-helix recognition revealed by the PU.1 ETS-domain-DNA complex*. Nature, 1996. **380**(6573): p. 456-60.
257. Poon, G.M. and R.B. Macgregor, Jr., *Base coupling in sequence-specific site recognition by the ETS domain of murine PU.1*. J Mol Biol, 2003. **328**(4): p. 805-19.
258. Case, T.M.a.D.A., *Modeling unusual nucleic acid structures*. Molecular Modeling of Nucleic Acids, 1998: p. 379-393.
259. D.A. Case, V.B., J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu and P.A. Kollman, *AMBER14*. University of California, San Francisco, 2014.

260. Morozov, A.V., et al., *Protein-DNA binding specificity predictions with structural models*. Nucleic Acids Res, 2005. **33**(18): p. 5781-98.
261. Hudson, W.H., et al., *Distal substitutions drive divergent DNA specificity among paralogous transcription factors through subdivision of conformational space*. Proc Natl Acad Sci U S A, 2016. **113**(2): p. 326-31.