Computational strategies to overcome antibiotic resistance: Exploring novel tetramic acids as Undecaprenyl Pyrophosphate Synthase (UPPS) inhibitors



Niamh Jiménez
University College
University of Oxford

A thesis submitted in partial fulfilment of
the requirements for the degree of
MSc(Res) Organic Chemistry

Trinity 2021

Acknowledgements

I would like to thank my supervisor Prof. Mark Moloney, who showed me great support throughout my time in Oxford and who kindly facilitated my change in direction from synthetic organic chemistry to computational studies. I would also like to express my gratitude to my collaborator Dr Fernanda Duarte, who introduced me to the field of computational chemistry and welcomed me into her research group. In light of the difficulties I experienced at Oxford, I would like to extend my gratitude towards the Oxford Disability Advisory Service and my personal Disability Advisor Julie Marsh. Thank you, Julie, for helping me to secure physical adjustments which allowed me to work in spite of my chronic illness. I am also incredibly grateful to my Disability Mentor, Julia Hill, whose support and friendship were an incredible source of comfort to me throughout my studies. Finally, I thank my beautiful family and friends, who were continually patient, loving and kind.

Table of Contents

Abstract **1**

List of Abbreviations **2**

Chapter 1. Introduction **5**

1.1. Antimicrobial resistance (AMR) 5

1.2. Small molecule antibiotic approaches to combatting AMR 7

1.2.1. Antibiotic Monotherapy 8

1.2.2. Antibiotic Combinations 13

1.3. Undecaprenyl Pyrophosphate Synthase (UPPS) as an essential antimicrobial target 16

1.3.1. Structure and function of UPPS 16

1.3.2. Inhibitors of UPPS 22

1.4. Tetramic acid (TA) antibiotics 32

1.4.1. TAs as UPPS Inhibitors 32

1.4.2. TAs as metal binders 36

1.5. Aims and motivation 39

Chapter 2: Methods **41**

2.1. Quantum Mechanics 41

2.2. Density Functional Theory (DFT) 43

2.3. Molecular Dynamics 46

2.3.1. Newton's equations of motion 46

2.3.2. Integrators 48

2.3.3. Thermodynamic ensembles and thermostats/barostats 50

2.3.4. Periodic Boundary Conditions (PBCs) 54

2.3.5. Treatment of long-range interactions: Coulombic and Lennard-Jones potentials 54

2.3.6. Principal Component Analysis (PCA) 58

2.4. Enhanced sampling methods 59

2.4.1. Umbrella Sampling (US) 60

2.4.2. Replica Exchange Molecular Dynamics (REMD) 61

2.5. Free Energy calculations 66

2.5.1. Scoring Functions (Astex Statistical Potential) 68

2.5.2. Linear Interaction Energy (LIE) 69

2.5.3. Alchemical free energy calculations 73

2.6. Binding pocket analysis 80

Chapter 3. Apo-UPPS as the antimicrobial target **83**

3.1 Previous literature work 83

3.2 Methods 85

3.2.1 Systems Preparation 85

3.2.2 MD Simulations 85

3.3 Data analysis 86

3.3.1 Determining the protonation state of active site residue H43 86
3.3.2 Elucidating the effect of differing H43 protonation state on key active site residues 87
3.3.3 Exploring the conformational flexibility of the hydrophobic tunnel ... 87
3.3.4 Investigating the effect of differing H43 protonation state on the conformational dynamics 88
3.4 Discussion 89
3.4.1 Determining the protonation state of active site residue H43 89
3.4.2 Elucidating the effect of differing H43 protonation state on key active site residues 92
3.4.3 Exploring the conformational flexibility of the hydrophobic tunnel ... 95
3.4.4 Investigating the effect of differing H43 protonation state on the conformational dynamics 97
Chapter 4. TAs as UPPS Inhibitors 101
4.1. Binding hypotheses 101
4.2. Methods 104
4.2.1. QM calculations 104
4.2.3. System preparation 104
4.2.4. MD simulations 108
4.2.5. Estimating ΔGbind using LIE 108
4.2.6. Data analysis 109
4.2.7. Estimating ΔGbind using US 109
4.2.8. Estimating ΔGbind using FEP-MBAR 110
4.2.9. Hydrophobic tunnel volume measurements 115
4.2.10. Temperature Replica Exchange (TRE) MD of UPPS apoenzyme . 115
4.3. Results and Discussion 116
4.3.1. Relative stability of TAs' keto-enol tautomers 116
4.3.2. MD analysis of binding sites 121
4.3.3. Characterisation of site 2 UPPS-TA complexes 137
Conclusions 146
References **..... 149**
Supporting Information **..... 163**

Abstract

The rise of multiple antibiotic resistance is a critical obstacle to the design and development of novel antimicrobial agents. The historical paradigm of monotherapies (i.e., drugs acting on single biological targets) has been replaced in recent years by multi-targeting therapies (e.g., antibiotic combinations and hybrids) which act on two or more targets. The efficacy of this multi-targeting approach is increased if the targets belong to divergent metabolic pathways, thereby minimising the likelihood of simultaneous dual-target resistance. Thus, the development of novel antibiotic scaffolds capable of inhibiting multiple, diverse enzymes is critical to addressing antibiotic resistance. Computational techniques, molecular dynamics (MD) simulations in particular, play a fundamental role in structurally elucidating protein-drug binding interactions and in predicting their binding affinities, tasks which are essential to the antibiotic drug design process. In this thesis, the mode of inhibition of selected 3-carboxamide and 3-acyl tetramic acids (TAs), which exhibit dual inhibition of bacterial undecaprenyl pyrophosphate synthase (UPPS) and RNA polymerase (RNAP), is investigated within UPPS using computational methods. UPPS is an essential bacterial enzyme involved in peptidoglycan synthesis of the bacterial cell wall and is hence a prime antibiotic drug target. In addition, the TA inhibitors are not only novel but contain a single chemical scaffold which is multi-targeting, both of which represent desirable attributes of the latest antibiotic drugs. In the absence of an X-ray crystal structure of the UPPS-TA complex, this work elucidates the most probable binding site of the TAs in UPPS, employing rigid docking and MD-based free energy calculations. The area of antibiotic drug design may benefit greatly from such computational predictions since the structural basis for inhibition of UPPS by the TAs remains poorly understood.

List of Abbreviations

AD	Applicability Domain
AMR	Antimicrobial Resistance
ASP	Astex Statistical Potential
BAR	Bennett Acceptance Ratio
BLI	β-Lactamase Inhibitor
CADD	Computer Aided Drug Design
COM	Centre-of-Mass
CPA	Cyclopiazonic Acid
CV	Collective Variable
DFT	Density Functional Theory
EAF	Exchange Attempt Frequency
EXP	Exponential Averaging
FEP	Free Energy Perturbation
FPP	Farnesyl Pyrophosphate
FPPS	Farnesyl Diphosphate Synthase
FsPP	Farnesyl Thiopyrophosphate
GGA	Generalised Gradient Approximation
GOLD	Genetic Optimisation for Ligand Docking
GPPS	Geranyl Diphosphate Synthase
GTO	Gaussian-Type Orbitals
HIV	Human Immunodeficiency Virus
HMBC	Heteronuclear Multiple Bond Correlation
HREMD	Hamiltonian Replica Exchange Molecular Dynamics
IPP	Isopentenyl Pyrophosphate

LCAO	Linear Combination of Atomic Orbitals
LDA	Local Density Approximation
LIE	Linear Interaction Energy
LJ-PME	Lennard Jones-Particle Mesh Ewald
LRA	Linear Response Approximation
MBAR	Multiple Bennett Acceptance Ratio
MD	Molecular Dynamics
MDR	Multi-Drug Resistant
MGGA	Meta-Generalised Gradient Approximation
MIC	Minimum Inhibitory Concentration
MMGBSA	Molecular Mechanics w. Generalised Born & Surface Area
MMPBSA	Molecular Mechanics w. Poisson Boltzmann & Surf. Area
MO	Molecular Orbital
MRSA	Methicillin-Resistant *Staphylococcus aureus*
MTB	*Mycobacterium tuberculosis*
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear Overhauser Effect spectroscopy
PBCs	Periodic Boundary Conditions
PBP	Penicillin-Binding Protein
PC	Principal Component
PCA	Principal Component Analysis
PES	Potential Energy Surface
PME	Particle Mesh Ewald
PMF	Potential of Mean Force
POVME	POcket Volume Measurer

PPPM	Particle-Particle-Particle Mesh
PVA	Pocket Volume Analysis
QM	Quantum Mechanics
RB	rotatable bond count
RC	Reaction Coordinate
REMD	Replica Exchange Molecular Dynamics
RESP	Restrained Electrostatic Potential
REST	Replica Exchange with Solute Tempering
RMSD	Root-Mean-Square Deviation
RMSF	Root-Mean-Square Fluctuation
RNAP	RNA Polymerase
SAR	Structure-Activity Relationship
SERCA	Sarcoplasmic and Endoplasmic Reticuli
SMD	Steered Molecular Dynamics
SMD	Solvation Model based on Density
STO	Slater-Type Orbitals
TA	Tetramic Acid
TI	Thermodynamic Integration
TREMD	Temperature Replica Exchange Molecular Dynamics
UPP	Undecaprenyl Pyrophosphate
UPPS	Undecaprenyl Pyrophosphate Synthase
US	Umbrella Sampling
WHAM	Weighted Histogram Analysis Method

Chapter 1. Introduction

1.1. Antimicrobial resistance

The spread of antimicrobial resistance (AMR), which has overtaken the emergence of novel resistance-breaking antibiotics, represents a significant threat to global health.[1–3] AMR may be defined as the adaptive capacity of a microorganism to resist drugs which previously eradicated it.[4] Core challenges for sustainable antibiotic development are the difficulty in developing resistance-breaking drugs with novel modes of action, the rapid resistance mechanisms exploitable by bacteria, and the lack of financial incentive for drug development.[5] Diminishing investment by the pharmaceutical industry arises from the high failure rate of antibiotic development (i.e., estimated 95%), coupled with a lack of profitability when compared to other therapeutic interventions.[5] New antibiotics must be reserved as a last resort treatment for multidrug-resistant (MDR) infections, which reduces their unit sales.[5] The dwindling antibiotic pipeline has profound public health implications.

The causes of endogenous resistance are manifold, including decreased bacterial cell permeability, increased antibiotic extrusion or efflux, target changes, and the inhibition of antibiotics by bacterial degradative enzymes, such as β-lactamases.[4,6–8] One of the most prominent examples of target-based AMR is the growth of resistance to methicillin in *S. aureus*, caused by the acquisition of the foreign *mecA* gene.[4] The latter encodes a modified version of a penicillin-binding protein (PBP), involved in peptidoglycan synthesis (i.e., bacterial cell wall biosynthesis), which exhibits low affinity for all β-lactams.[4] This renders PBP-

targeting β-lactam antibiotics ineffective against methicillin-resistant *Staphylococcus aureus* (MRSA).

Single-step mutation of bacterial targets, which are the product of single genes, leads to high-level resistance of monotherapies. Hence, single-target antibiotics, once glorified for their high target specificity, have more recently been found to be ineffective as sole agents, instead requiring administration as antibiotic combinations.[9,10] The latter include antibacterials such as rifampin, novobiocin, and trimethoprim, which target RNA transcription (i.e., RNA Polymerase (RNAP)), DNA replication (i.e., DNA gyrase B subunit), and folate synthesis (i.e., dihydrofolate reductase, FolA) respectively.[6] Thus, standard systemic monotherapy has more recently been found to require the use of resistance-breaking agents which target proteins encoded by multiple genes, such as β-lactam antibiotics (penicillin-binding proteins (PBPs); cell wall synthesis), tetracyclines (16S rRNA; protein synthesis), and fluoroquinolones (DNA gyrase, topoisomerase IV; DNA replication).[6,11,12] Although multitarget monotherapies are less susceptible to the development of endogenous resistance, exogenous resistance may still develop via horizontal gene transfer. As a strategy to counter growing resistance within the multitarget monotherapeutic framework, new pharmacophores may be appended to pre-existing drug scaffolds, thereby increasing the probability of engagement with multiple binding targets.[6] Moreover, it has been postulated that the capacity for multiple intramolecular and intermolecular drug-target interactions lies at the heart of combatting AMR.[6] This postulate is supported by the polypharmacological nature of naturally-occurring antimicrobial peptides, which tend to target complex bacterial structures, such as

the cell wall and/or cell membrane.[2] Thus, a paradigm shift has emerged from single-target monotherapies towards multitarget antibiotic hybrids and combinations.

1.2. Small molecule antibiotic approaches to combatting AMR

Within the established multitarget paradigm, three branches of investigation have been identified: a) single-pharmacophore drugs, b) dual-pharmacophore hybrids, and c) antibiotic combinations (i.e., typically comprising pre-existing, single-target inhibitors) (**Figure 1**).[6]

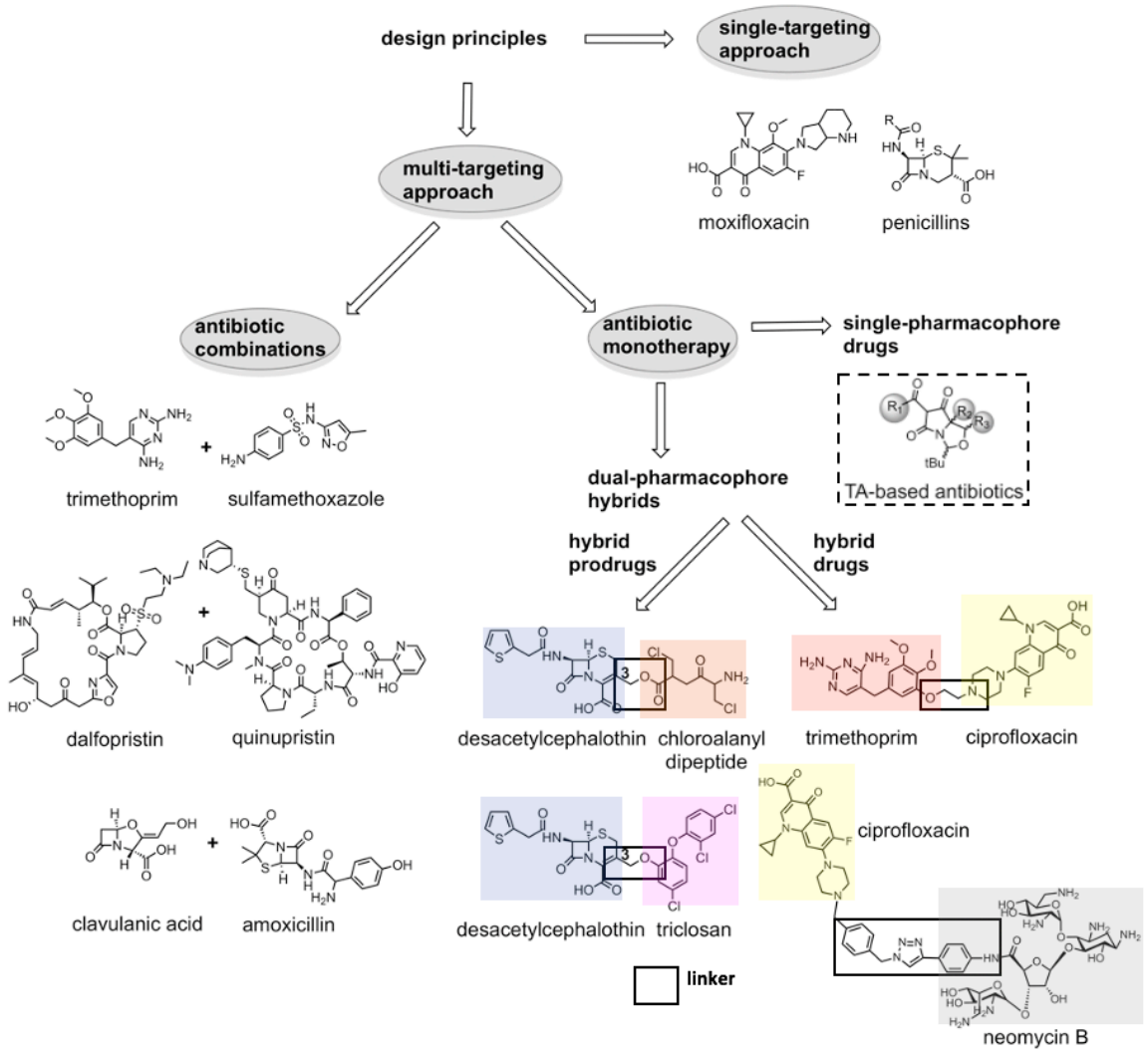


Figure 1: Different approaches to antibiotic drug design.

1.2.1. Antibiotic Monotherapy

Single-pharmacophore drugs

Multitarget single-pharmacophore drugs may be subdivided into those acting on related targets of the same metabolic pathway and those affecting targets of distinct metabolic pathways. The most prominent examples of the former are the β-lactam antibiotics (e.g., penicillins, **Figure 1**), which tend to target more than one PBP.[2,13] Other well-known antibiotics of this class are the quinolones (e.g., moxifloxacin, **Figure 1**), which target both topoisomerase II and IV, involved in DNA supercoiling and the decatenation of interlinked chromosomes respectively.[2,14] In general, antibacterials targeting multiple nodes of the same metabolic pathway exhibit less susceptibility towards resistance than those targeting a single structure or process. For example, a class of aminobenzimidazole-based compounds, developed by Vertex as dual inhibitors of the highly conserved ATP-binding subunits of DNA gyrase (GyrB) and topoisomerase IV (ParE), were found to be less susceptible towards resistance when compared to novobiocin, whose sole target was DNA gyrase (GyrB).[6,15]

Examples of monotherapies targeting the unrelated components of different pathways are less numerous. However, this polypharmacological antibiotic class is of particular utility as there is a lower probability of acquiring target-based resistance, which requires the adaptation of two different stress responses to counter the action of the antibiotic.[2] Polymyxins, such as polymyxin B and polymyxin E (colistin), target the inner and the outer membranes of multiresistant gram-negative bacteria.[2,16] Similarly, the anti-leprosy drug clofazimine (Lamprene) and the atypical tetracycline chelocardin each possess dual modes of action involving distinct target structures.[2] The former targets the guanine bases of

bacterial DNA as well as upregulating phospholipase A1 activity, while the latter inhibits the bacterial ribosome and depolarises the cytoplasmic membrane.[17,18]

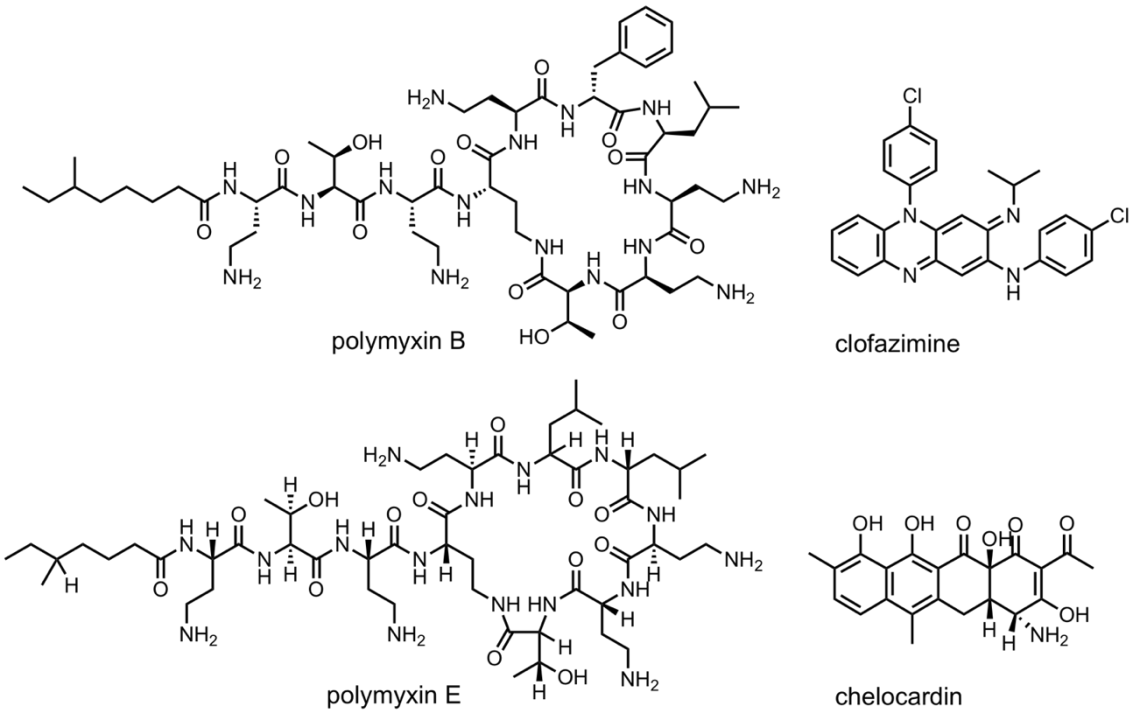


Figure 2: Chemical structures of selected multitarget single-pharmacophore drugs.

Dual-pharmacophore hybrids

Hybridisation involves the unification of two (usually pre-existing) pharmacophores of established biologically active agents, with retention of the activity of both constituents (i.e., "dual action" or *bimodality)*. The first hybrid was developed by Roche in the form of a class of β-lactam-fluoroquinolone molecules (**RO 24-8138**, **RO 24-4383**, **Figure 3**).[19] A molecular linker is typically used to join the parent drugs, with the mode of covalent attachment (i.e., cleavable or non-cleavable) defining the approach (i.e., *hybrid prodrug* versus *hybrid drug* respectively, **Figure 1**).[20] The physicochemical properties of the linker, as well as the length, are vital for antibacterial potency through expansion of the chemical space and through preservation of the separate pharmacophoric regions.[20] The hybrid integrates the resistance suppression of antibiotic combinations into a monotherapy, possessing a single pharmacokinetic profile. According to the

design postulate, hybrids eliminate resistance of the bacterial strain to one or even both of the hybrid constituents, as well as broadening the spectrum of activity.[6] Although each drug fragment is biologically active, the resulting hybrid may have a novel mode of action, pharmacological action or enhanced efficacy relative to its constituents.[20] Importantly, hybridisation has been proved to be a cost-effective strategy for redeeming the efficacy of legacy antibiotics against drug-resistant strains.[20]

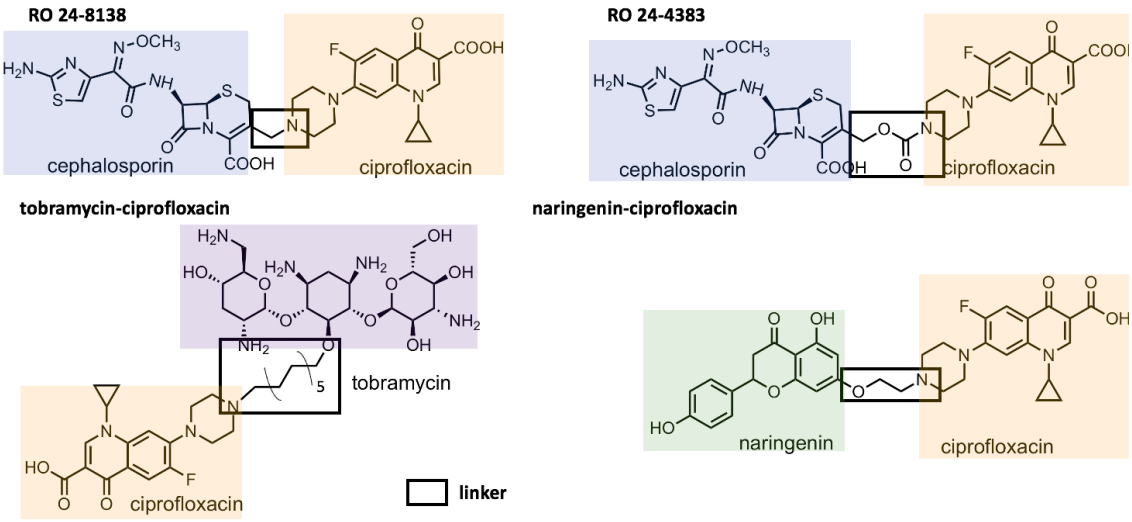


Figure 3: Chemical structures of selected hybrid molecules.

Moreover, antibiotic hybrids may be classified as either *hybrid prodrugs* or *hybrid drugs* (**Figure 1***).* Hybrid prodrugs are synthetic dual-pharmacophore constructs, in which the distinct pharmacophores are connected by a cleavable linker characterised by bacterium-specific lability.[20] The bacterium-specificity of this biotransformation is critical in distinguishing hybrid prodrugs from antibiotic-antibiotic or antibiotic-adjuvant combinations. Most reported hybrid prodrugs consist of a β-lactam antibiotic (i.e., cephalosporin) covalently bonded to a non-β-lactam antibiotic via the adjacent carbon 3.[20] Examples included hybrids built of desacetylcephalothin (cephalosporin) linked to a chloroalanyl dipeptide[21,22] or a chlorine-containing phenolic triclosan (**Figure 1**).[20,23] Although there are no

hybrid prodrugs in clinical use, the latter were successful in advancing to preclinical and clinical trials.[20,23–25] Hence, the challenge in designing a bacterium-specific linker, as well as the different pharmacokinetics of the functional units after cleavage, renders prodrugs less popular than their hybrid drug counterparts.[20]

On the other hand, hybrid drugs are distinguishable from hybrid prodrugs by the stability of their linker, which is resistant to enzymatic attacks of the host and the bacterium. Unlike the prodrug, no biotransformation of the unimolecular entity occurs upon penetration of the bacterial membrane, hence a single pharmacokinetic profile is retained.[20] It is generally acknowledged that a hybrid drug whose constituents each inhibit a different target, will not be able to inhibit both targets simultaneously.[20] Most hybrid drugs incorporate a fluoroquinolone, owing to its well defined structure-activity relationship (SAR) and broad-spectrum activity.[20] For example, a series of hybrids based on the linking of benzyl pyrimidines to fluoroquinolones (e.g., ciprofloxacin), developed by MerLion Pharmaceuticals Pte. Ltd, exhibited antibacterial activity against both gram-positive (e.g., *Staphylococcus aureus*, *Staphylococcus epidermidis*) and gram-negative (e.g., *E. coli*) bacteria.[20] Notably, trimethoprim-ciprofloxacin (BP-4Q-002) displayed potent inhibition of not only *S. aureus* (MIC 0.5 μg/ml) and *E. coli* (MIC 1 μg/ml) but was also active against ciprofloxacin-resistant *S. aureus* strain NRS19 (**Figure 1**).[20] This demonstrated the promising capacity of hybrid drugs to eliminate bacterial strains resistant to one of the parent drugs.

Other fluoroquinolone-based hybrids described in the literature include those linked with phenolic flavonoids and with aminoglycosides.[26,27] The linking of naturally occurring phenolic flavonoids with fluoroquinolones produced hybrids such as naringenin-ciprofloxacin, which displayed inhibition of MDR *E. coli*, MRSA, and amphotericin B-resistant *Candida albicans* (**Figure 3**).[26] The high antibacterial activity of these hybrids was attributed to efflux pump inhibition by the flavonoid component, which ensured a high intracellular accumulation of the drug through prevention of its expulsion.[26] Similarly, aminoglycoside-fluoroquinolones, such as neomycin B-ciprofloxacin hybrids, were active against MRSA and drug-resistant strains of *E. coli*, and had superior antibacterial profiles relative to either the parent neomycin B or the equimolar mixture of the hybrid components (**Figure 1**).[27] This demonstrated the capacity of the hybrid to benefit from properties beyond the individual components, in particular the evasion of efflux-mediated resistance pathways.[28] However, the aminoglycoside-fluoroquinolones, like many others, suffered from poor membrane permeability, likely owing to their high molecular weights (>600 g/mol).[20] Poor bacterial cell uptake compromises the utility of many such hybrid drugs *in vivo*, in addition to low penetration of the inner and outer membranes of gram-negative bacteria due to suboptimal physicochemical profiles.[20]

Furthermore, there may be potential for establishing a new paradigm in which the combination and hybrid approaches are coupled. This method was exemplified by the repurposing of the non-antibiotic anticancer drug mitomycin C against MDR gram-negative bacteria through its combination with a tobramycin-ciprofloxacin antibiotic hybrid TOB-CIP (**Figure 3**).[29] It was discovered that synergising

mitomycin C with TOB-CIP reduced its minimum inhibitory concentration (MIC) against the MDR isolates of *P. aeruginosa*, *E. coli*, *Klebsiella pneumoniae*, and *Enterobacter cloacae*.[29] Notably, neither hybrid component synergised with the anticancer drug individually.[29]

1.2.2. Antibiotic Combinations

Antibiotic resistance is intrinsic to bacterial ecologies, leading to the biosynthesis of antibiotic combinations by many strains in response to selective pressure.[30] Few naturally occurring antibiotics are single agents but are synthesised instead as combinations.[30,31] These combinations are encoded by adjacent biosynthetic gene clusters, likely transmitted horizontally as single mobile elements, thereby conferring on a strain the capacity for effective competition with other species.[30,32,33] Examples of such naturally occurring antibiotic combinations are those produced by *Streptomyces clavuligerus* and *Micromonospora carbonacea*. The former produces a combination of the β-lactam cephamycin and the BLI clavulanate (**Figure 4**), while the latter synthesises a combination of structurally divergent natural products (NPs) targeting the ribosome.[30] Thus, the inherently combinatorial nature of naturally occurring antibiotics may serve to inform antibiotic drug design in a high resistance climate.

Moreover, *synergism* arises when a drug combination's pharmacological action exceeds the anticipated additive effects of the individual components.[2] Synergistic antibiotic combinations may be classified as *congruous* (all constituents active), *syncretic* (one constituent active) and *coalistic* (constituents active together but not alone).[9] There are examples of both antibiotic-antibiotic (i.e., *congruent)* and antibiotic-adjuvant (i.e., *syncretic*) approaches. Fixed-dose orally-administered

congruent combinations in clinical use include Septra (trimethoprim:sulphamethoxazole in a 1:5 w/w ratio; **Figure 1**) and Synercid (dalfopristin:quinupristin in a 3:7 w/w ratio; **Figure 1**), as well as topical combinations such as bacitracin and polymyxin B, which offer broad spectrum coverage of gram-positive and gram-negative bacteria.[6,9] In addition, syncretic combinations are employed, comprising an antibiotic component and an antibiotic adjuvant, also known as a *resistance breaker*.[9,34] The adjuvant does not exhibit antibiotic activity but increases the efficacy of the antibiotic component through modification of either the bacterial physiology or the host biology. One prominent example is the β-lactam/β-lactamase inhibitor combination, in which the efficacy of the β-lactam antibiotic is enhanced by a non-antibiotic β-lactamase inhibitor (BLI), which counteracts the resistance mechanism (i.e., hydrolysing β-lactamase enzymes) of the bacteria.[9] The first syncretic combination which was made commercially available in fixed doses was Augmentin, a co-formulation of the BLI clavulanic acid with amoxicillin (**Figure 1**).[9]

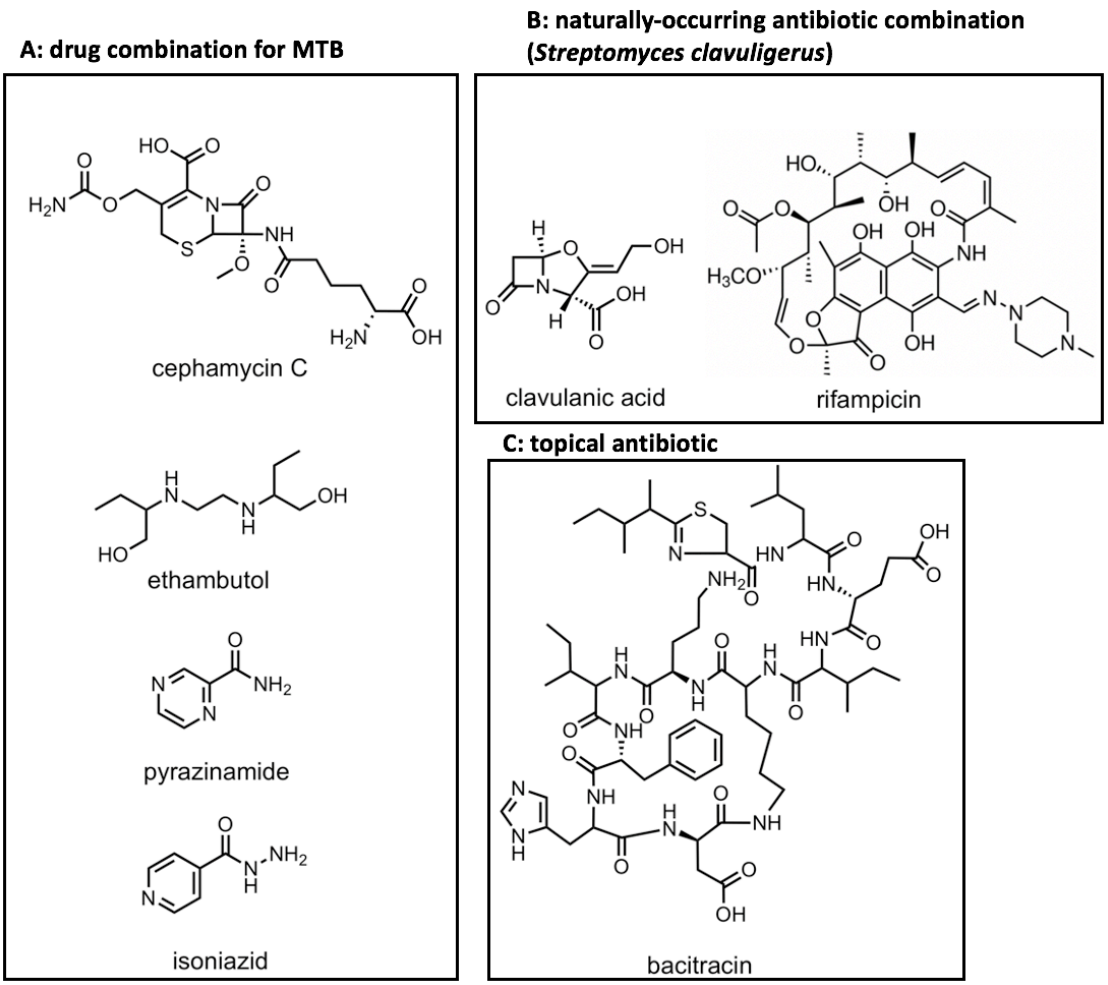


Figure 4: Chemical structures of various antibiotic combination components.

Furthermore, combination therapies may act on targets in different metabolic pathways, distinct nodes in the same metabolic pathway, or on the same target in diverse ways. Combinations of pre-existing single-target inhibitors are used in the treatment of Human Immunodeficiency Virus (HIV) and *Mycobacterium tuberculosis* (MTB), where the antibacterials target different metabolic pathways. MTB may be effectively treated using a combination of isoniazid (inhibitor of the enoylreductase subunit of fatty acid synthase), rifampicin (inhibitor of RNAP), ethambutol (inhibitor of arabinosyl transferases in cell wall biosynthesis), and pyrazinamide (unknown inhibition mechanism) (**Figure 4**).[30] If any of the drugs are administered as a single agent, the target is susceptible to single-step mutation.[6,30] However, the combination results in inhibition of at least three diverse pathways

simultaneously, thereby lowering the potential for resistance.[30] The less diversified approach of targeting different nodes in the same pathway is exemplified by Augmentin (i.e., clavulanic acid-amoxicillin) and Septra (i.e., sulfamethoxazole-trimethoprim; **Figure 1**). The latter combines inhibitors of dihydropteroate synthase and dihydrofolate reductase respectively, both of which are enzymes involved in the biosynthesis of the cofactor folic acid.[30] Finally, the least diversified approach, involving a single target, is exemplified by Synercid (**Figure 1**). The constituents of this congruent combination bind to adjacent sites of the 50S subunit of the bacterial ribosome, proximate to the peptidyl transferase center.[30,35]

1.3. Undecaprenyl Pyrophosphate Synthase (UPPS) as an essential antimicrobial target

1.3.1. Structure and function of UPPS

E and *Z*-Prenyltransferases catalyse the reaction of isopentenyl pyrophosphate (IPP) with an allylic pyrophosphate to generate *trans* and *cis* double bonds respectively. Although catalytically similar, the two classes of enzyme have been found not to share sequence homology.[36] UPPS is a *cis*-prenyltransferase enzyme which catalyses the consecutive condensation reactions of farnesyl pyrophosphate (FPP) with eight IPP molecules to yield undecaprenyl pyrophosphate (UPP**; Figure 7**). UPP is a carrier of lipid II across cell membranes for peptidoglycan synthesis, the principal component of bacterial cell walls. As an enzyme essential for bacterial survival, UPPS has long been considered a promising antimicrobial drug target.[37]

The active form of UPPS is a dimer. Ko *et al.* have described the structure as butterfly-like if viewed along the dyad axis, in which the catalytic region represents the "butterfly wings" and the pairing domain embodies the "butterfly's body" (**Figure 5: A**).[36] Each monomer is comprised of six β strands and seven α helices, arranged in a parallel α/β topology. Four β strands (βA-βB-βD-βC), as well as the α2 and α3 helices, encompass a long hydrophobic tunnel.[36] Based on site-directed mutagenesis studies and crystallographic data, the FPP and IPP binding sites have been found to reside at the top of the tunnel, where many conserved hydrophilic amino acids are located (i.e., N28, R30, H43, F70, S71, R194, E198).[38–41] X-ray crystal structures of apo and holo UPPS have revealed three predominant conformations of the enzyme: a closed state (FPP-bound; PDB ID: **1X06**, **4H8E**) possessing a kinked α3 helix, an open state (product-bound; PDB ID: **1UEH**, **2E98**) possessing a straight α3 helix, and an apo state (PDB ID: **3QAS**). The latter represents an intermediate conformation adopted in the absence of substrate or product (**Figure 5: B**).[41]

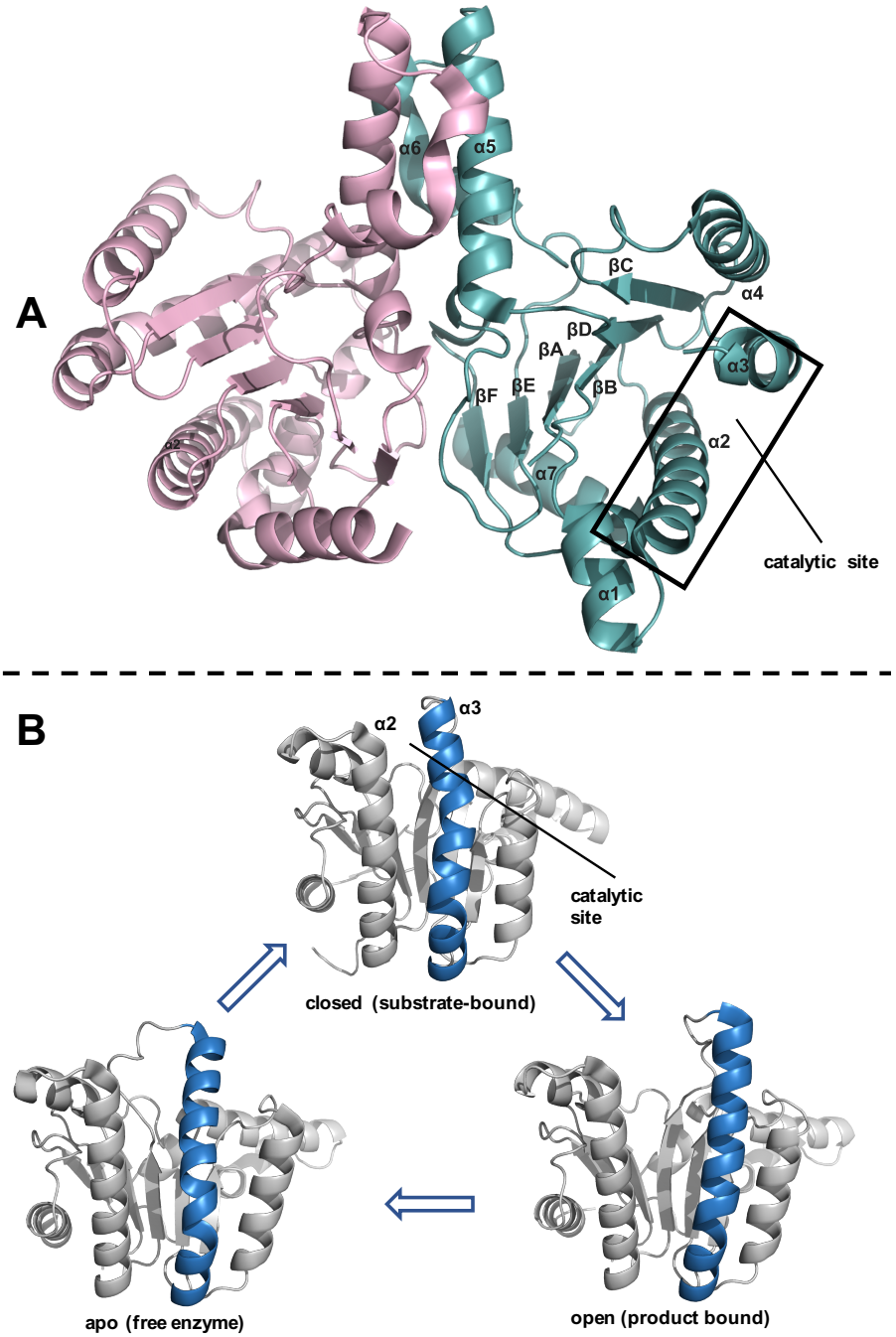


Figure 5: **A)** The structure of UPPS in ribbon representation (PDB ID: **4H8E**; Adapted from Ko (2001)).[36] **B)** The closed (PDB ID: **4H8E**), open (PDB ID: **1UEH**) and apo (PDB ID: **3QAS**) states of UPPS, each showing a distinct disposition of the α3 helix (red ribbon; Adapted from Guo (2005)).[41]

In *trans*-prenyltransferases, FPP binds to conserved aspartate rich DDXXD motifs via a magnesium ion (Mg^{2+}). Conversely, in *cis*-prenyltransferases, FPP binding does not require the metal ion, and the DDXXD motifs are absent.[42] While excessive concentrations of Mg^{2+} (e.g., 50 mM) have been found to impair UPPS activity, a concentration of 1 mM of Mg^{2+} is required for maximal performance.[41] Moreover, the X-ray crystal structure of UPPS-Mg^{2+}-IPP-FsPP (PDB ID: **1X06**)

reveals Mg^{2+} coordinated to the carboxylate of D26, the pyrophosphate moiety of FPP thio-analogue, farnesyl thiopyrophosphate (FsPP), and three water molecules (**Figure 6**). However, in all other X-ray crystal structures of UPPS complexed to FPP (PDB ID: **1V7U**, **5KH4**, **4H8E, 4Q9M**), Mg^{2+} is absent. In these cocrystal structures, the pyrophosphate group of FPP forms hydrogen bonds with multiple hydrophilic residues at the top of the hydrophobic tunnel, including the backbone nitrogen of G29 and R30, as well as the side chains of N28, R30, H43, R77 and R39 (**Figure 6**). In addition, important hydrophobic interactions are made between the farnesyl tail and residues of the α3 helix, including L85, L88 and F89 (**Figure 6**).[42] IPP has been postulated to occupy a site adjacent to FPP, forming hydrogen bonds with R194 and R200.[42]

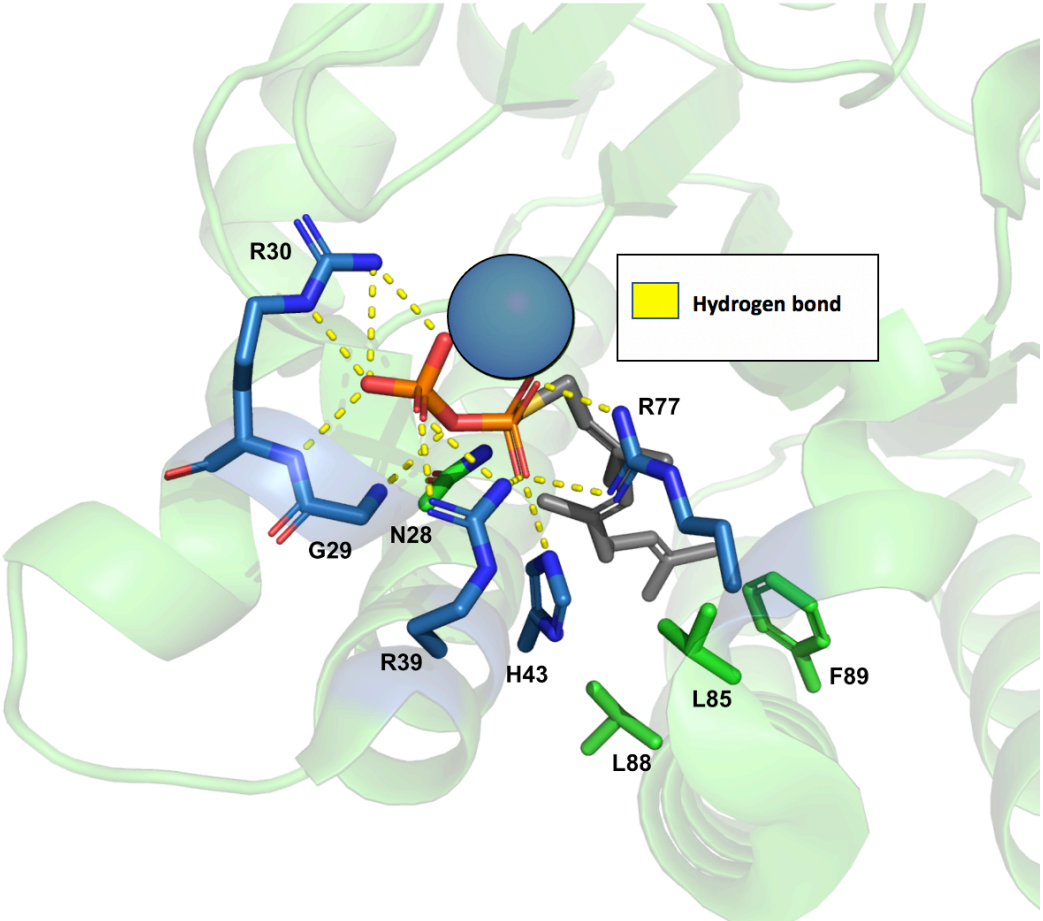


Figure 6: Farnesyl thiopyrophosphate (FsPP) in complex with *E. coli* UPPS and Mg^{2+} (PDB ID: **1X06**).

Using the stopped-flow fluorescence technique, Chen *et al.* have probed the binding process of IPP and FsPP to *E. coli* UPPS.[43] Unlike the binding of FPP, IPP binding has been found to require Mg^{2+}.[41,43] It was discovered that FPP binding quenched the fluorescence of W91, which served as an indicator of the conformational change from the open (substrate-free) to the closed (substrate-bound) UPPS form.[43] Additionally, fluorescence quenching data has demonstrated that the order of substrate binding to *E. coli* UPPS is FPP followed by IPP.[43] The three-phase fluorescence change associated with mixing of IPP with the pre-incubated UPPS-FPP complex has been attributed to binding of IPP to the closed enzyme (first phase), followed by subsequent conformational changes of the enzyme during catalysis (second phase), and culminating in the final interconversion to the open state (third phase) with product release.[43]

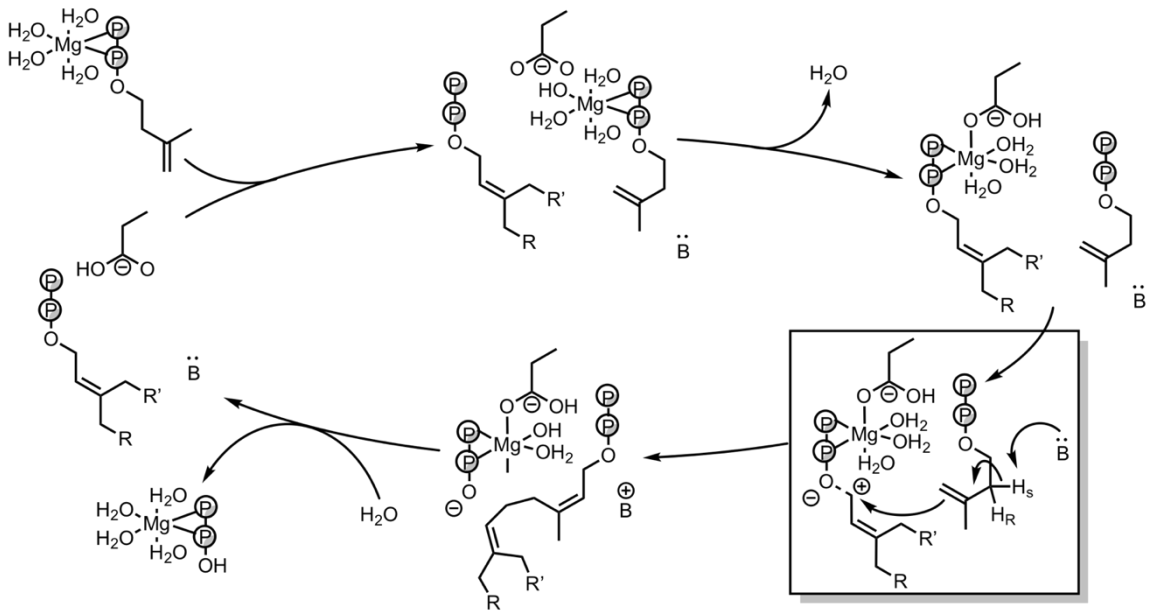


Figure 7: The proposed catalytic mechanism of UPP biosynthesis; key catalytic step highlighted.[41] *Adapted from Guo (2005).*

The proposed catalytic mechanism is shown in **Figure 7**. It commences with the binding of FPP at the top of the hydrophobic tunnel, anchored by hydrogen bonds to G29, R30, N28, and R39.[41] This is followed by binding of pre-complexed IPP-

Mg^{2+} adjacent to the FPP site. The metal ion is then transferred from IPP to FPP via D26, thereby activating the allylic pyrophosphate for nucleophilic attack by IPP. The vital role of D26 in the catalysis has been supported by the 1000-fold reduction in the k_{cat} elicited by the D26A mutation.[41] Additionally, the importance of H43, which has been postulated to serve as a proton donor for the pyrophosphate of FPP, was highlighted by a similar decrease in k_{cat} as a result of the H43A mutation.[41]

Furthermore, Chang *et al.* have established that the interconversion of the closed (catalytically *active*) and open (catalytically *inactive*) conformational states is mediated by a flexible loop (residues 72-83 in *E. coli* UPPS).[44] It has been suggested that the flexible loop facilitates the condensation of FPP with IPP by correctly orienting the two substrates relative to each other.[36] According to site-directed mutagenesis, the flexible loop has a critical role in IPP binding, since the N74A, E81A and S71A mutations revealed substantial decreases in IPP affinity.[45] Additionally, the length of the flexible loop is vital for the catalytic process, as demonstrated by Cheng *et al.* through the insertion and deletion of non-conserved amino acids, such as S72, S83, and V82. Specifically, fluorescence binding investigations have indicated that Insertion of alanine residues prevents the flexible loop from adopting the closed state and binding FPP.[44] Deletion of V82 and S83 or of S72 also prevents interconversion from the open to the closed state, and FPP binding has been found to be less efficient under such conditions.[44] The UPPS mutants generated by loop insertion or deletion, which adopted open conformations, have exhibited 10^4-fold lower UPPS activity versus the wild-type.[44]

Thus, it has been concluded that the closed conformation, possessing the kinked α3 helix, is the catalytically active form of the enzyme.

1.3.2. Inhibitors of UPPS

Many inhibitors of the prenyltransferases geranyl diphosphate synthase (GPPS) and farnesyl diphosphate synthase (FPPS) have also been found to be active against UPPS. X-ray crystal structures of *E. coli* UPPS in complex with allylic-diphosphate-mimicking bisphosphonates (dual inhibitors of UPPS and FPPS) have been reported by Gao *et al* (PDB ID: **2E98**, **2E99**, **2E9A**, **2E9C**, **2E9D**; **Figure 8**).[46] These inhibitors have been found to bind predominantly to four sites in each monomer (designated as *sites 1-4* in **Figure 8**), and exhibit IC_{50} values ranging from 0.59 μM to 1.1 μM (**Figure 8**).[47] Site 1, which corresponds to the FPP-binding (catalytic) site, is occupied in all bisphosphonate complexes. Four out of five of the crystallographic complexes feature interactions with key residues of site 1 (D26, N28, R39, H43) and site 4 (R51, R102 and H103) (**Figure 8**).[47]

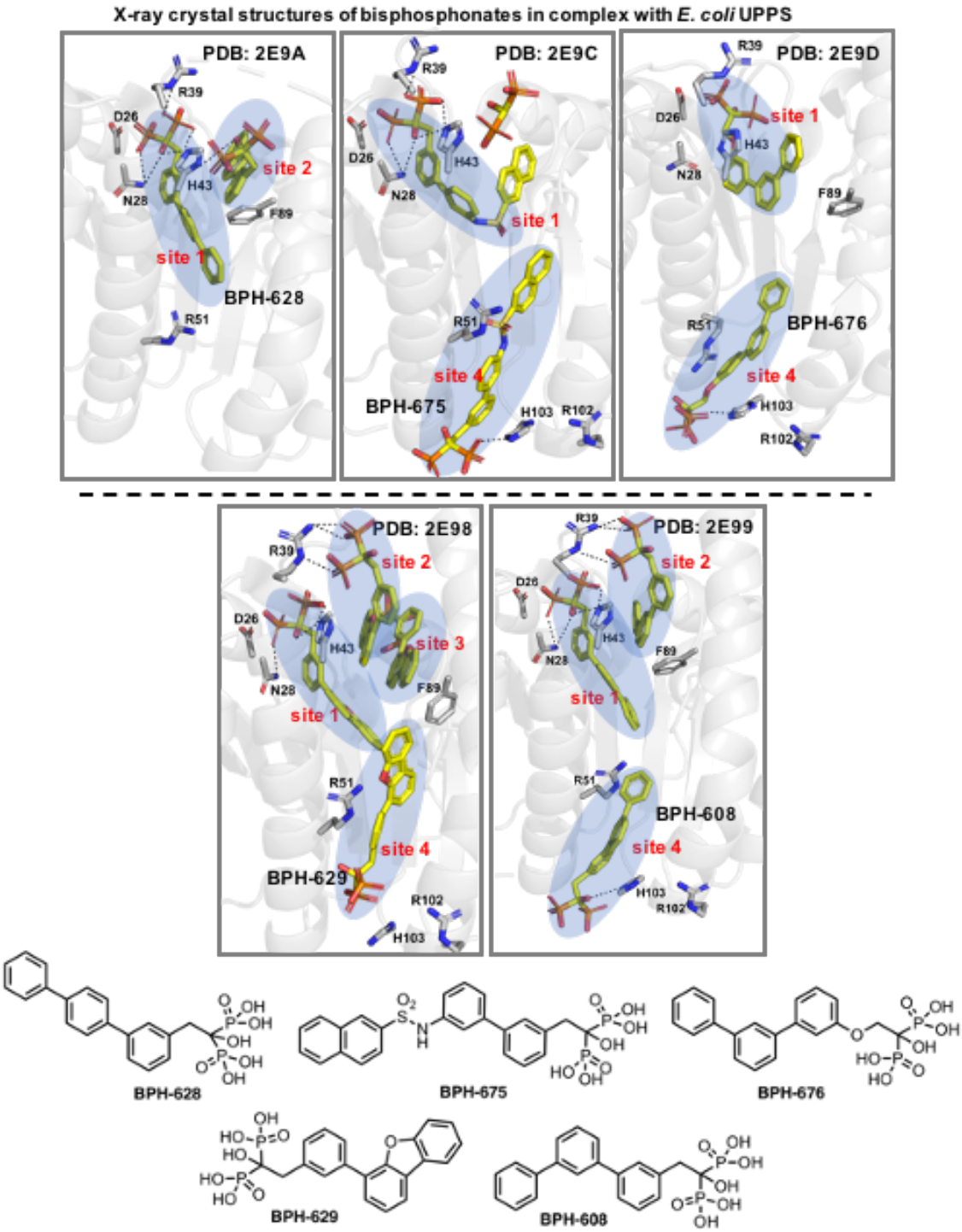


Figure 8: *E. coli* UPPS in complex with bisphosphonates BPH628 (PDB ID: **2E9A**), BPH-675 (PDB ID: **2E9C**), BPH676 (PDB ID: **2E9D**), BPH-629 (PDB ID: **2E98**), and BPH608 (PDB ID: **2E99**), including the chemical structures of all inhibitors.

Moreover, virtual screening by Durrant *et al.* has uncovered non-bisphosphonate (NB) dual inhibitors of FPPS and UPPS (**Figure 9**). Although the binding modes of the structurally divergent compounds **NB-5** and **NB-6** have not been elucidated by X-ray crystallography, the compounds exhibit potent inhibition of *E. coli* UPPS (i.e., IC_{50} values of 5.7 µM and 3.2 µM respectively).[48]

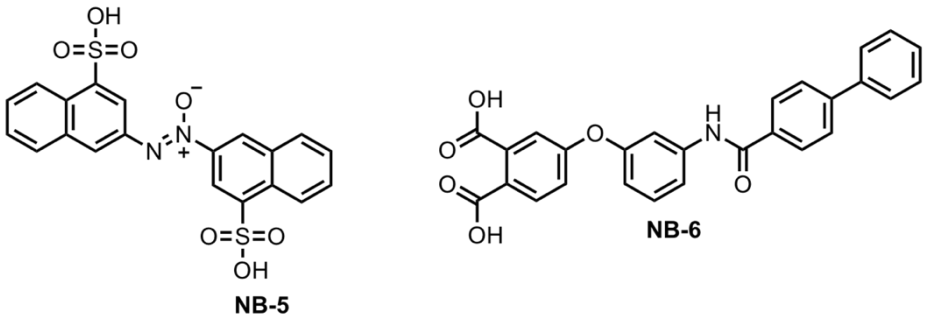


Figure 9: Chemical structures of compounds NB-**5** and NB-**6**.

Moving beyond GPPS and FPPS inhibitors, Zhang *et al.* have demonstrated that certain HIV-1 integrase (IN)-inspired inhibitors with the keto/diketoacid moiety exhibit activity against prenyltransferases UPPS and dehydrosqualene synthase. Dehydrosqualene synthase catalyses the condensation of two FPP molecules to form presqualene diphosphate and thence dehydrosqualene, which is ultimately transformed into the virulence factor staphyloxanthin. This discovery has been driven by the observation that both IN and prenyltransferases share an active site Mg^{2+}/Asp motif which interacts with the native substrate. Zhang and colleagues have found that the most potent UPPS inhibitor exhibits an IC_{50} of 0.24 µM against *E. coli* UPPS.[49]

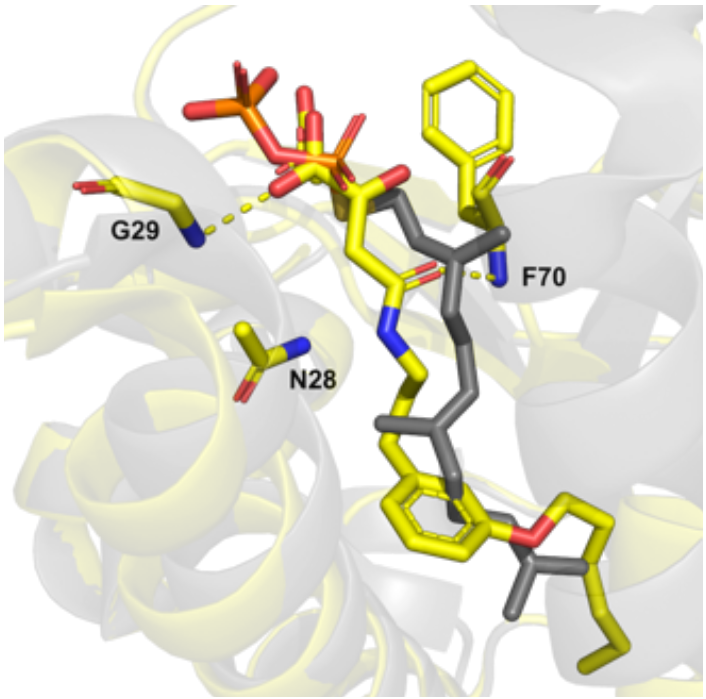


Figure 10: Superposition of the X-ray crystal structures of *E. coli* UPPS in complex with a diketoacid inhibitor (PDB ID: **3TH8**, yellow ribbon) and *E. coli* UPPS in complex with FsPP (PDB ID: **1X06**, grey ribbon).

The X-ray crystal structure of UPPS complexed with one such diketoacid has revealed that it occupies the FPP-binding site (i.e*.,* site 1) proximate to the key residues D26 and N28 (PDB ID: **3THB**; **Figure 10**).[49] Similar to the five bisphosphonate crystal structures presented by *Guo et al*., Mg^{2+} is absent from the complex. This is to be expected as it has been shown that the metal is not necessary for FPP binding to UPPS.

Further examples of UPPS active diketoacids (compounds **DKA-14, DKA-15**; **Figure 11: A**) have been identified by Zhu *et al*. via *in silico* high-throughput screening (PDB ID: **4H3C**, **4H3A**).[50] However, the X-ray crystal structures of UPPS complexed with the diketoacids **DKA-14** and **DKA-15** reveal binding to site 4, in contrast to the IN-inspired UPPS inhibitor, which has been found to bind to site 1.[49] In addition to the diketoacids, Zhu and colleagues have identified other structurally diverse classes of UPPS inhibitors, including benzoic acids, phosphonic acids, bisamidines and bisamines (**Figure 11**). Notably, the focus has not been limited to FPP-mimicking anionic inhibitors but also includes cationic species. For example, it has been found that the biphenyl bisamidine **C-16** is a potent inhibitor of UPPS (IC_{50} of 0.1 μM).[50]

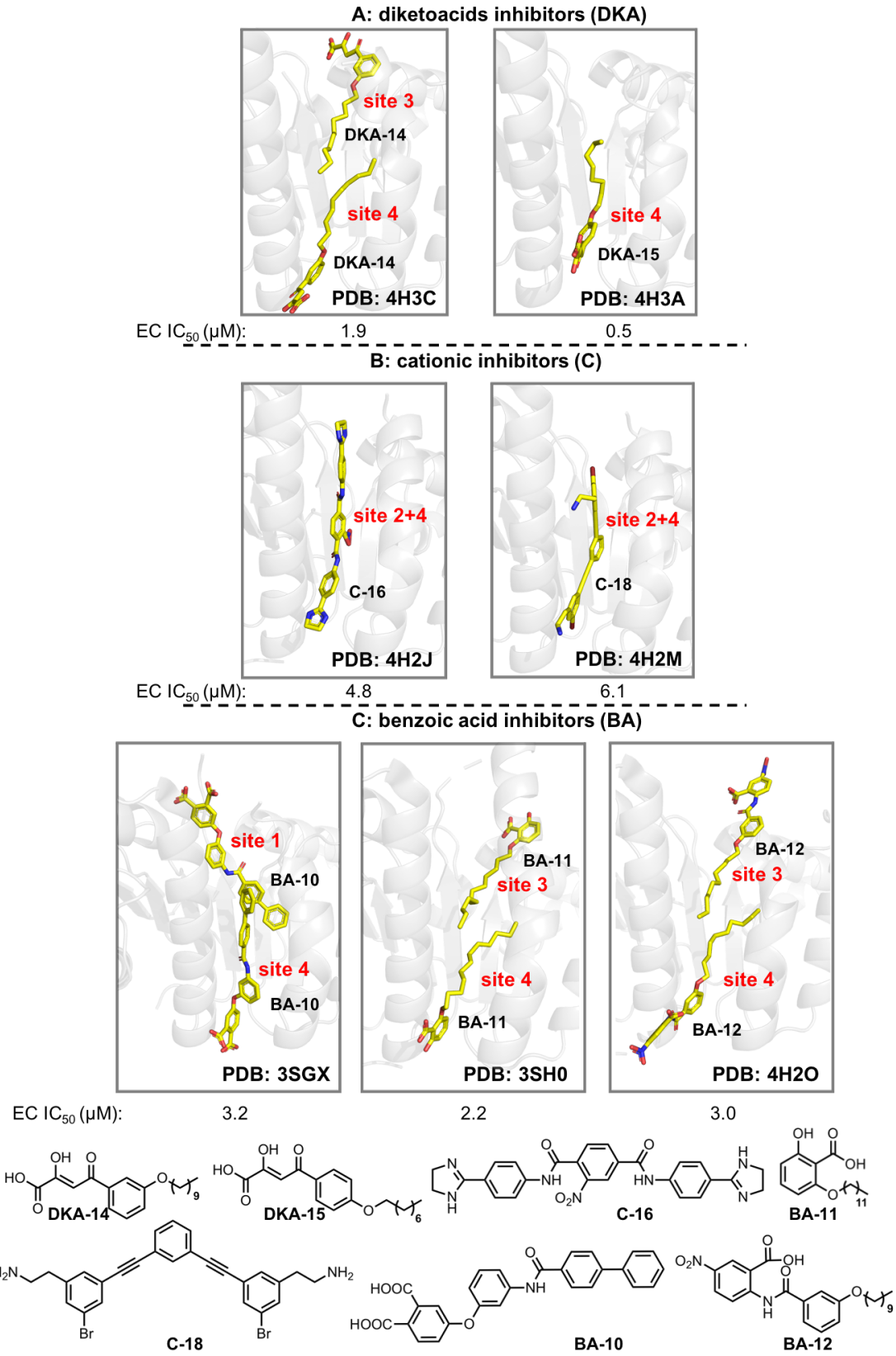


Figure 11: Crystal structures and IC_{50} data for UPPS active diketoacids (**A**), cationic inhibitors (**B**), and benzoic acid inhibitors (**C**), as well as the chemical structures of the inhibitors.

According to crystallographic data, bisamidine **C-16** and bisamine **C-18** occupy sites 2 and 4, with the polar moieties located at the surface of the enzyme and the

hydrophobic "linkers" buried inside the hydrophobic tunnel (**Figure 11: B**).[50] Moreover, the X-ray crystal structures of this diverse range of inhibitors in complex with UPPS has highlighted a possible correlation between the occupation of site 4 in UPPS and higher inhibitor potency, specifically amongst the benzoic acids and the diketoacids (**Figure 11: A, C**).[50]

Moreover, Wang and colleagues have demonstrated that certain lipophilic benzoic acids dually inhibit UPPS and the membrane protein undecaprenyl diphosphate phosphatase, also involved in peptidoglycan biosynthesis. When compared to phosphonic acids, benzoic acids have been found to exhibit much higher potency, and a good correlation has been established between cell growth inhibition and UPPS inhibition.[51] Additionally, the lipophilic benzoic acids synergise with known cell-wall biosynthesis inhibitors, including ampicillin, bacitracin, fosmidomycin, carbenicillin, and vancomycin.[51]

Similar to the benzoic acids are the anthranilic acids (AA), which have been found to inhibit *E. coli* UPPS by Gobec *et al.* They have implemented a structure-based virtual screening whereby the ZINC compound database has been filtered using a bisphosphonate-based pharmacophore, the selected compounds have been subjected to ensemble docking in MD generated conformational clusters, and the top-scoring compounds have been biologically evaluated.[52] Thus, **AA-1**, **AA-2** and **AA-3** (**Figure 12**) have been identified as active UPPS inhibitors, possessing IC_{50} values of 45 µM, 25 µM, and 24 µM respectively against *E. coli* UPPS (**Figure 12**).[52] Based on the ensemble docking, it has been postulated that the compounds occupy open conformations of UPPS, similar to the bisphosphonates. Additionally, Gobec and colleagues have proposed that the AAs are competitive UPPS

inhibitors, possessing the AA moiety as a pyrophosphate mimetic conjugated to distinct lipophilic scaffolds.

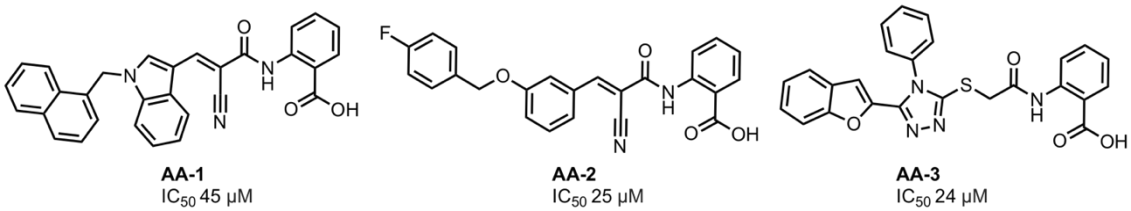


Figure 12: Chemical structures of anthranilic acids **AA-1, AA-2 and AA-3**.

Furthermore, many diversified UPPS-inhibiting scaffolds have been identified through the use of *in silico* screening. Three such novel UPPS inhibitor classes have been discovered by Sinko *et al*.: the rhodanines (4-oxo-2-thioxo-1,3-thiazolidines) **1**, dihydroxyphenyls **2**, and pyrimidinetriones **3** (**Figure 13**). The most potent of the three is rhodanine **1** (*S. aureus* UPPS IC_{50} of 2.6 µM), structurally similar to the aldolase reductase inhibitor epalrestat, which is used in the treatment of diabetic neuropathy.[37] Rhodanine **1** has also been shown to synergise with methicillin against MRSA, revealing its potential for use in antibiotic combination therapies.[37] Just as compound **1** mimicked the existing drug epalrestat, other drugs have been repurposed against UPPS. For example, Farha *et al*. have found that clomiphene, a widely used fertility drug, exhibits moderate *in vitro* inhibition of *S. aureus* (IC_{50} of 7.5 µM), and *E. coli* (IC_{50} of 15 µM) UPPS.

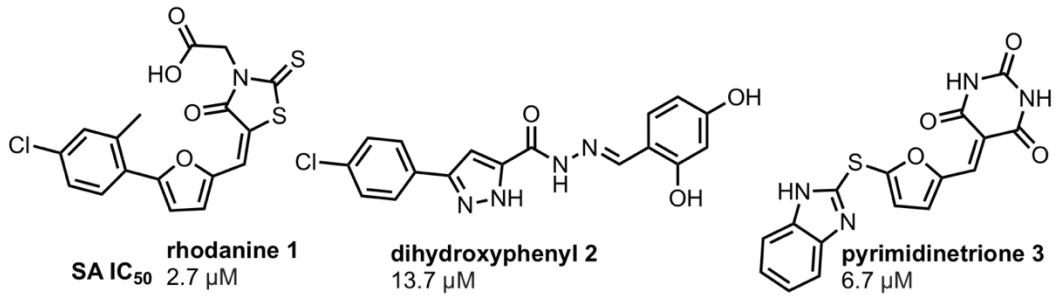


Figure 13: Chemical structure of novel cores rhodanine 1, dihydroxyphenyl 2 and pyrimidinetrione 3.

The X-ray cocrystal structure of *E. coli* UPPS-clomiphene reveals the inhibitor binding to a channel which connects site 1 with site 4, characterised primarily by

hydrophobic interactions of the chloride and ring moieties (PDB ID: **5CQJ**; **Figure 14: A**).[53] A single hydrogen bond is formed between the phenoxy oxygen of clomiphene and the side chain carboxylate of E96.[53]

Unlike the heretofore reported inhibitors, which bind to UPPS in the absence of FPP, Danley *et al.* have presented the first X-ray crystal structure of *S. pneumoniae* UPPS in complex with FPP and the allosteric inhibitor UK-106051 (PDB ID: **4Q9M**). The isoxazole-based inhibitor, which exhibits an IC_{50} of 0.50 μM against *S. pneumoniae* UPPS, binds to a site adjacent to FPP in the middle of the hydrophobic tunnel (**Figure 14: B**).[54] The inhibitor makes primarily hydrophobic contacts but a single hydrogen bond is formed between the oxygen of the isoxazole ring and Q53 (**Figure 14**). Additionally, a π-stacking interaction was observed between the isopropylphenyl moiety of the inhibitor and the terminal alkene of FPP. In this X-ray cocrystal structure, UPPS adopts a closed conformation equivalent to that of *E. coli* UPPS in complex with FsPP (PDB ID: **1X06**). Surface Plasmon Resonance binding experiments have supported the allosteric inhibition mechanism by demonstrating a 25-fold increase in binding affinity for UK-106051 in the presence of FPP.[54]

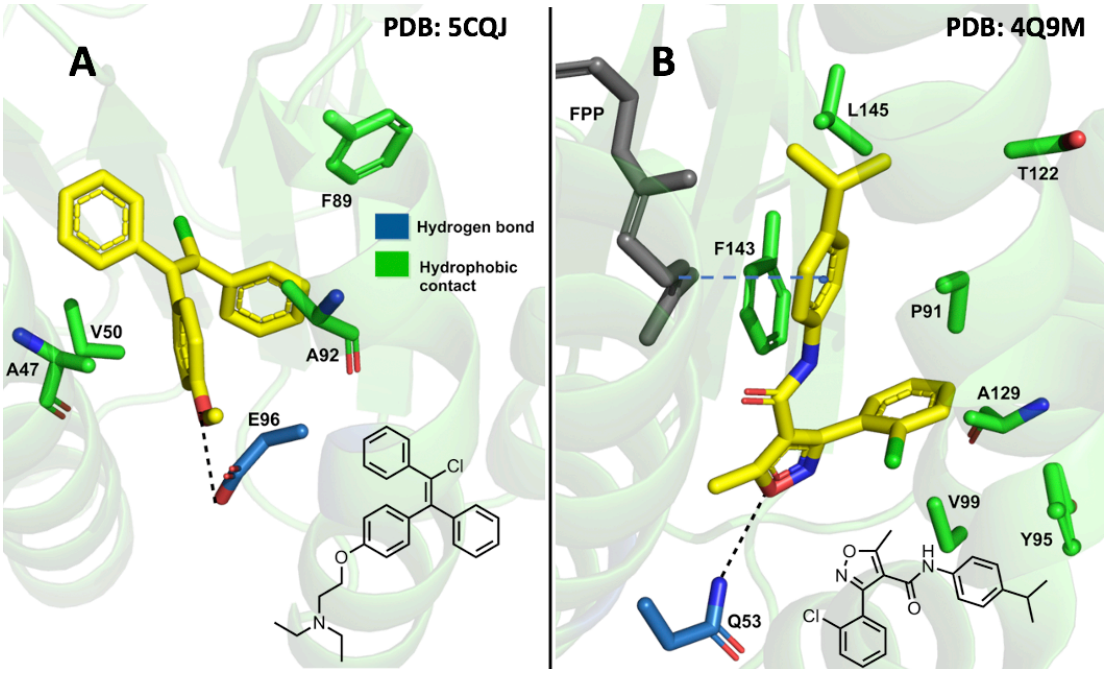


Figure 14: *E. coli* UPPS in complex with clomiphene (**A**). *S. pneumoniae* UPPS in complex with FPP (grey sticks) and the allosteric inhibitor UK-106051 (yellow sticks), showing the single hydrogen bond to Q53 and the π-stacking interaction with the terminal alkene of FPP (**B**).

Similar to the allosteric inhibitor UK-106051, a unique binding mode has been elucidated by a pyrazole-based inhibitor discovered by Concha *et al*. This novel structure exhibits potent inhibition of *S. aureus* UPPS (IC_{50} 0.19 µM).[55] The X-ray crystal structure of *S. pneumoniae* UPPS in complex with the inhibitor reveals binding to a hydrophobic pocket at the base of the FPP-binding site (PDB ID: **5KH5**). Analogous to inhibitor UK-106051, the binding of the pyrazole-based compound is driven mainly by π-stacking interactions between the benzyl isopropyl ether moiety and F143, as well as between the benzothiophene and the residues F94 and M49 (**Figure 15: B**).[55]

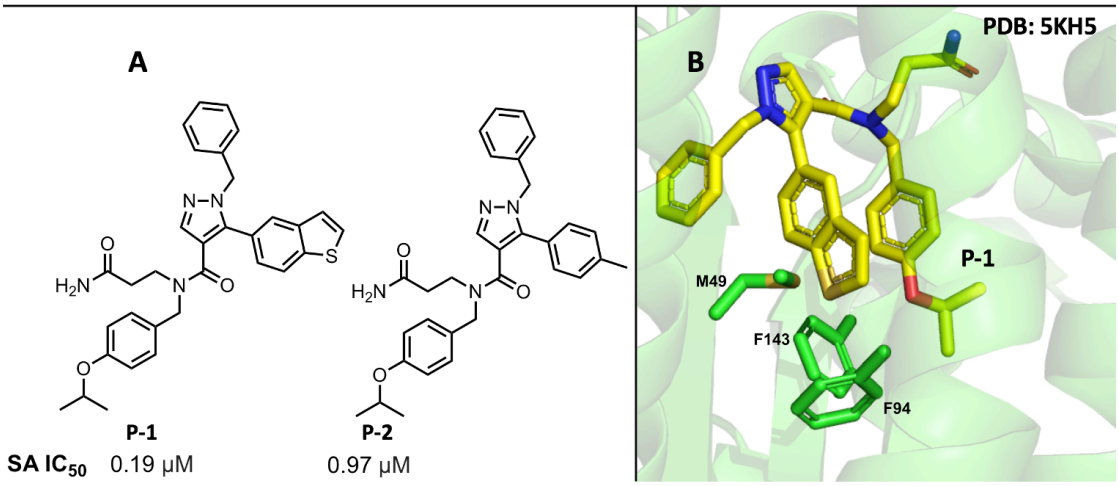


Figure 15: The chemical structures of compounds **P-1** and **P-2** (**A**). *S. pneumoniae* UPPS in complex with the pyrazole-based inhibitor (PDB ID: **5KH5**), showing π-stacking interactions with residues F94, F143 and M49 (**B**).

These studies demonstrate that structurally diverse inhibitors recognise different sites within the hydrophobic tunnel of UPPS, which has been broadly subdivided into sites 1-4, based on the binding modes of the bisphosphonates. Multiple binding in some cases suggests that the binding has the potential to be relatively non-specific. The FPP-mimicking bisphosphonates represent the largest and most studied class of UPPS inhibitor, exhibiting multiple binding (i.e., to the catalytic site 1 and one or more of the remote sites 2-4) of an expanded pocket conformation. The AAs are postulated to bind to site 1 in a similar mode within the expanded pocket state. Additionally, the IN-inspired diketoacids have been found to bind to site 1. On the other hand, a wide range of non-bisphosphonate inhibitors have been found to bind to the non-catalytic sites 2-4, including certain classes of keto/diketoacids, benzoic acids, phosphonic acids, and cationic species such as bisamidines and bisamines. Allosteric inhibitors, binding promiscuously to pockets within the hydrophobic tunnel, include an isoxazole-based inhibitor (PDB ID: **4Q9M**), the fertility drug clomiphene (PDB ID: **5CQJ**), and a pyrazole-based inhibitor (PDB ID: **5KH5**).

1.4. Tetramic acid (TA) antibiotics

1.4.1. TAs as UPPS Inhibitors

A diversified library of tetramic acid (TA), tetronic acid, and dihydropyridine-2-one-3-carboxamides, have been synthesised by Novartis based on a pharmacophore for UPPS inhibition modelled on the binding mode of FPP.[56] Docking of a TA-based high-throughput screen (HTS) hit, **1a**, in *E. coli* UPPS (PDB ID: **1V7U**) has highlighted the pharmacophoric importance of two hydrogen bond acceptors and a hydrophobic group (**Figure 16**).[42,57] The TAs have exhibited the highest antibacterial activity of the three classes investigated. Members of the TA library are characterised by MIC effectiveness in gram-positive species, as well as potent inhibition of *S. pneumoniae* UPPS. The proposed SAR has established that hydrophobic moieties in the R_1 position, such as isobutyl and 2-imidazolyl-methyl groups, reduce the half maximal inhibitory concentration (IC_{50}) relative to the unsubstituted case (**Figure 16**).[56] In addition, large aromatic motifs are preferred to aliphatic groups in the R_2 position, such as arylphenyl or heterocycloalkyl-phenyl groups.[56] Moreover, despite displaying submicromolar activity against *S. pneumoniae* UPPS, the MIC effectiveness of the TAs is substantially reduced in the presence of blood serum.[56]

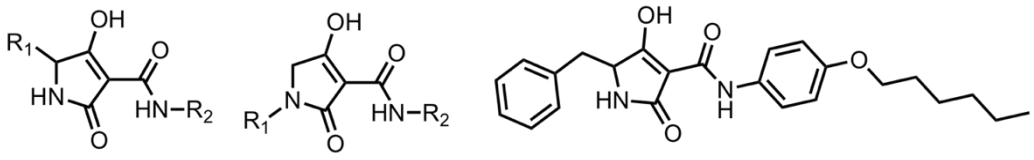


Figure 16: The chemical structure of TA-based design scaffolds.

Beyond the synthetic TAs reported by Peukert *et al*., a broad range of antibacterial NPs possess the TA motif, including streptolydigin, reutericyclin and signermycin B (**Figure 17**).[58–61] A NP-inspired design approach by Moloney *et al.* has led to the discovery of TA dual inhibitors of bacterial UPPS and RNAP, involved in gene

transcription and peptidoglycan synthesis respectively (**Figure 17**).[62] The TA dual inhibitors harbour potential as multitarget single-pharmacophore antibiotics, characterised by a single novel chemical scaffold and enzyme targets belonging to different metabolic pathways. Development of such novel chemical scaffolds able to inhibit multiple biological targets is critical to address antibiotic resistance.[6,12,63–66] Whilst there is precedent for multitarget UPPS inhibitors, such as the UPPS and FPPS-inhibiting bisphosphonates, there is no precedent for multitarget UPPS inhibitors whose additional enzymatic target(s) belongs to a distinct, unrelated metabolic pathway. Prior to the discovery of dual UPPS and RNAP inhibiting TAs, the multitarget UPPS inhibitors have targeted related prenyltransferases, involved in catalysing prenyl chain elongation in terpene biosynthesis.

According to *in vitro* binding data reported by Moloney *et al.*, monocyclic and bicyclic 3-carboxamide TAs and bicyclic 3-acyl TAs are active against *S. pnuemoniae* UPPS, of which the bicyclic 3-carboxamide TAs have been proved to be the most potent. The strength of UPPS inhibition is dependent on the identity of the pendant carboxamide/acyl R_1 substituent. By contrast, only bicyclic variants of both TA classes have displayed RNAP inhibition, with 3-acyl TAs identified as the most potent. The 3-enamine TAs have been found to be inactive against both UPPS and RNAP, suggesting that the oxygen of the 3-acyl and 3-carboxamide derivatives is critical for UPPS and/or RNAP affinity. While the *tert*-butyl group at the 8-position has been found to be key for both RNAP and UPPS inhibition, substituents at the 5 (R_2) and 6 (R_3) positions, as well as the nature of the heteroatom at the 7-position, exhibit minimal effects on the activity. In general, the identity of R_1 in

both 3-carboxamide and 3-acyl TAs is important, with bulky aromatic or lipophilic aliphatic substituents favoured by both enzymes.

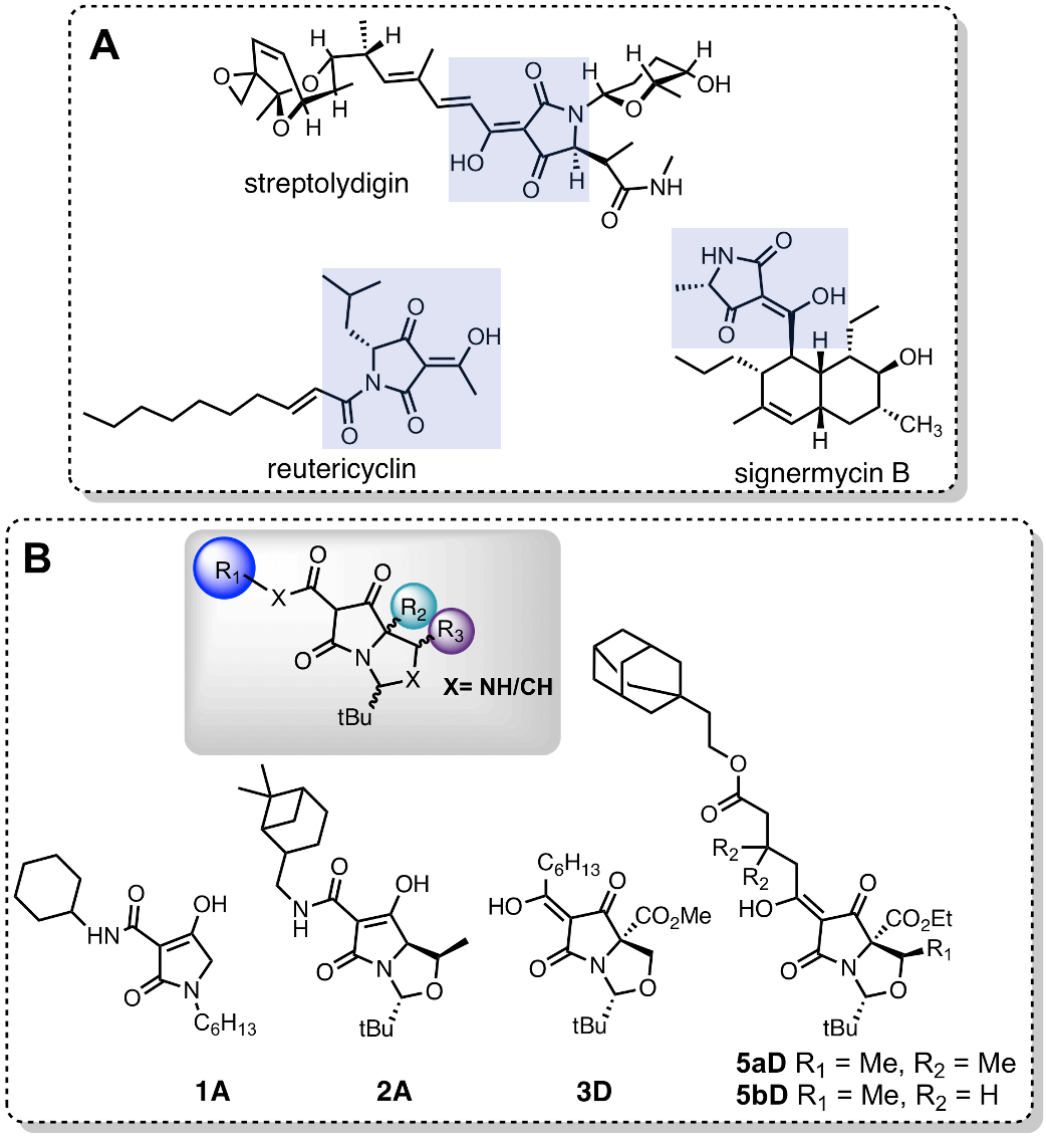


Figure 17: Structures of key TA-containing natural products (**A**) and synthetic TAs of interest from the library of Moloney *et al* (**B**).

Lee and colleagues challenged the competitive model of UPPS inhibition by the TAs, which has been postulated by Novartis and Moloney *et al*., demonstrating through biophysical experiments that TA-based UPPS inhibitors may be allosteric in nature.[67] The location of TA binding to UPPS has been investigated by Lee *et al*. using a TA-based photoprobe (**Figure 18**). Using photo-cross-linking and tandem mass spectrometry, the amino acid sequence (AGMEALQTV) of the peptide

modified by the mass of the TA photoprobe in *S. pnuemoniae UPPS* has been identified (**Figure 18**).[67] Thus, it has been suggested that TAs bind to an allosteric site adjacent to the FPP binding site, since the precise site of photoprobe attachment is Met49 on the α2 helix.[67] In addition, fluorescence spectroscopy investigations employing an FPP fluoroprobe have revealed the order of binding of the natural substrates and the TA inhibitor to *S. pneumoniae* UPPS.[67] TA binding takes place only in the absence of the other substrates, followed by the binding of either FPP or IPP. This observation was in contrast to the well-defined sequential order of substrate binding to *E. coli* UPPS, in which FPP binds first.[67] The resultant tertiary complex (i.e., UPPS-TA-IPP or UPPS-TA-FPP) cannot bind to any further substrates and is therefore unproductive, giving rise to UPPS inhibition.

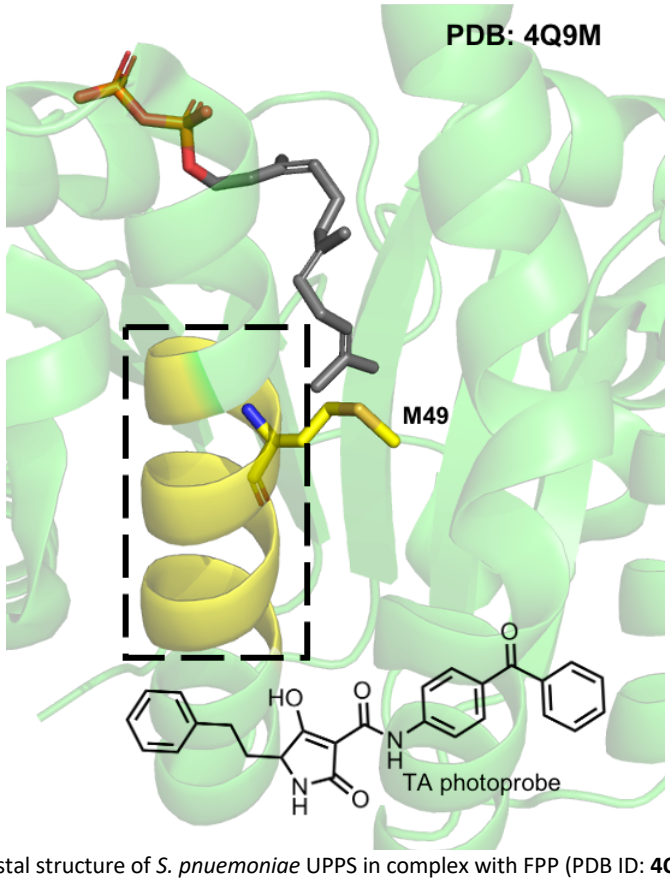
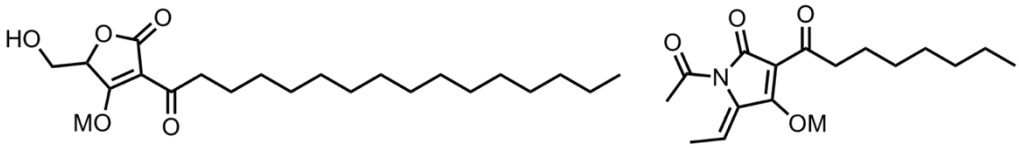


Figure 18: X-ray crystal structure of *S. pnuemoniae* UPPS in complex with FPP (PDB ID: **4Q9M**), indicating the photoprobe-modified sequence or site of attachment (yellow ribbon).

1.4.2. TAs as metal binders

3-Acyl TAs have been reported to occur naturally as covalent metal chelates.[68] Specifically, chelation of divalent metal ions by the ketonic oxygen of position 4 (*E*-isomer) or position 2 (*Z*-isomer) and the oxygen of the exocyclic acyl moiety at the 3-position is prevalent.[69] Isolation of TA-based NPs typically requires acidification to release the free acids, suggesting that the TAs are more stable in the form of metal adducts.[70] One of the most prominent examples is the phytotoxin tenuazonic acid, which has been isolated from the fungus *Phoma Sorghina* as mixtures of calcium and magnesium salts, as well as in the form of transition metal complexes with Cu(II), Ni(II) and Fe(III).[71–73] Further examples include the antibiotic magnesidin and the toxic indole TA cyclopiazonic acid (CPA). Kohl *et al.* have isolated magnesidin from *Pseudomonas magnesiorubra nov. sp*. as 1:1 mixtures of magnesium chelates.[71] Crude extracts of CPA isolated from cultures of *Penicillium cyclopium* exist as divalent metal chelates prior to acidification.[74]

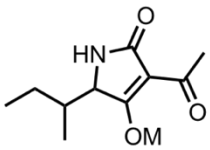


RK-682

$M = H^+, Na^+, 1/2\ Ca^{2+}, 1/2\ Mg^{2+}$

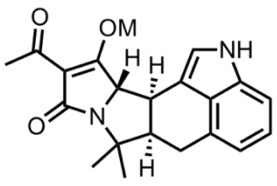
magnesidin A

$M = 1/2\ Mg^{2+}$



tenuazonic acid

$M = K^+, Na^+, 1/2\ Ca^{2+}, 1/2\ Mg^{2+}, 1/2\ Mn^{2+}, 1/3\ Fe^{3+}, 1/2\ Cu^{2+}, 1/2\ Ni^{2+}, 1/2\ Zn^{2+}$



cyclopiazonic acid

$M = H^+, 1/2\ Mg^{2+}, 1/2\ Mn^{2+}$

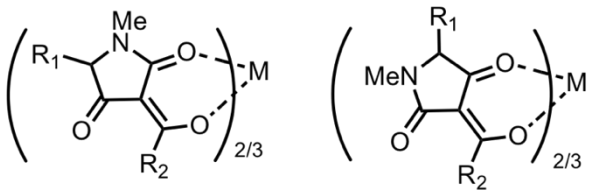
Figure 19: Chemical structures of RK-682[72], magnesidin A[75], tenuazonic acid[68,69], and cyclopiazonic acid.[74,76]

Moreover, the metallophoric nature of the TAs is vital in the consideration of their *in vivo* mode of action, toxicity and assay.[74] TAs are highly acidic owing to the presence of the tricarbonylmethine motif (estimated pKa 2.5-4.0), which gives rise to enolization into four tautomers (**Figure 36**). Since deprotonation under physiological conditions (pH ≈ 7.4) is expected, the question of whether TAs exert their biological effects as free acids or as metal chelates becomes relevant.[68] The enhanced bioactivity of certain TA metal chelates relative to their non-chelated counterparts has been attributed to changes in their physicochemical profiles. Specifically, an increase in lipophilicity may lead to improved cell permeation through membrane lipid bilayers.[68]

The increase in biological activity of the TA metal chelates has also been rationalised by their capacity to mimic phosphate groups in the binding sites of kinases and phosphatases. One such example is the 3-palmitoyl-5-hydroxymethyl TA RK-682, which has been isolated as the sodium and calcium chelates of *Actinomycetes* and *Streptomycetes*, and which inhibits HIV-1 protease and other tyrosine phosphatases.[72] Additionally, CPA has been found to target the Ca^{2+}-ATPase of the sarcoplasmic and endoplasmic reticuli (SERCA), showing competitive inhibition with ATP and Ca^{2+}. Inhibition by the TA was achieved by binding to and stabilising the inactive conformation of SERCA, which exhibits decreased Ca^{2+} affinity. Crystallographic data has confirmed that the TA inhibited SERCA as the metal chelate (i.e., Mg^{2+}, Mn^{2+}).[68,77]

Furthermore, the melophlins, a class of antimicrobial 3-acyl-N-methyl TAs isolated from the marine sponge *Melophlus sarasinorum*, provides another example of the synergistic enhancement of the bioactivity of TAs by metal fragments. Biersack *et*

al. have demonstrated that certain pharmacologically significant melophlins (e.g., melophlin A and C; **Figure 20**) display enhanced antiproliferative and antimicrobial action in the form of metal chelates. More specifically, growth inhibition of the melophlin C-resistant bacterium *M. Luteus* by gallium, lanthanum, and ruthenium complexes of melophlin C has been detected. Additionally, the magnesium and zinc complexes of melophlin A have been shown to exhibit antitumour activity against select human cancer cell lines.[78]



***Z*-isomer** ***E*-isomer**

melophlin A: R_1 = H, R_2 = *n*-$C_{15}H_{31}$
melophlin C: R_1 = Me, R_2 = $(CH_2)_3CHMe$
$(CH_2)_6Me$

M = Ga^{3+}, La^{3+}, Ru^{3+}, Mg^{2+}, Zn^{2+}

Figure 20: Chemical structures of melophlins A and C.[78]

1.5. Aims and motivation

Given the current AMR crisis, there is a growing need to identify novel, multitarget antibiotics. [6,20,79] Among them, TAs have been recognised as promising multitarget antibiotics.[37,62,80,81] They act by inhibiting UPPS, an essential bacterial enzyme involved in peptidoglycan synthesis. However, the structural basis for their inhibition of UPPS remains poorly understood, thereby hampering efforts to design and optimise novel TA antibiotics. The present work seeks to shed light on the binding mode of the TAs in UPPS using computational techniques, and hence to employ this knowledge in establishing a relationship between the specific TA binding and the enzyme's conformational dynamics. To achieve this goal, the following aspects will be examined: the conformational dynamics of the apoenzyme, the active tautomer of the TA inhibitors, the TA binding site in UPPS, and the conformational dynamics of the TA-UPPS complexes.

This work will commence with investigation of the apoenzyme using MD; pocket volume analysis (PVA) will be used to characterise the hydrophobic tunnel and principal component analysis (PCA) will be applied to identifying its main collective motions (**Chapter 3**). For the sake of completeness, the conformational space accessed by the apoenzyme will be examined under the condition of different H43 protonation states, since the latter is a conserved active site residue of catalytic importance.

Following study of the apoenzyme, the conformational dynamics of UPPS induced by TA binding will be explored. Firstly, the keto-enol tautomerism of the TAs will be examined, and the lowest energy tautomers of the 3-acyl, 3-carboxamide and 3-enamine TAs will be identified. To fulfil the aim of identifying the active tautomers involved in UPPS binding, the relative stabilities of a subset of TAs in water will be calculated by DFT (**Chapter 4.3.1**). This will provide insights into the form taken by selected TAs when interacting with their target.

Having identified the active tautomer of each of a selection of TA inhibitors, their potential binding site in the enzyme will be investigated. In the absence of a UPPS-TA co-crystal structure or a confirmed inhibition mode, the study will draw from literature precedents, centring on the investigation of three purported TA binding

sites in UPPS, employing docking, MD simulations, and free energy calculations (**Chapter 4.3.2**).

Subsequently, the conformational dynamics of the proposed UPPS-TA complexes will be investigated relative to the characteristic dynamics of the apoenzyme. Hence, this will illuminate the effect of complexation on the conformational space occupied by the enzyme. Employing PCA and PVA techniques, the effect of complexation on the dynamics of a key flexible loop, involved in the interconversion of the open and closed states, will be explored (**Chapter 4.3.3**).

Chapter 2: Methods

In this chapter, the fundamental concepts and theory underpinning the computational techniques employed in the work presented herein are reviewed. This will provide the theoretical knowledge required to contextualise the results of the thesis.

2.1. Quantum Mechanics

In Quantum Mechanics (QM), a wavefunction (ψ) is a mathematical construct which models a system comprised of nuclei and electrons. Properties of the electronic system (i.e., *observables*) may be extracted as eigenvalues through the application of operators to the wavefunction. The most prominent operator is the Hamiltonian operator ($\widehat{H}$), which returns the energy (E) of the system as its eigenvalue. The latter is a function of both the electronic (r) and nuclear (R) coordinate vectors respectively:[82]

$$\hat{H}_{\psi}(r, R) = E_{\psi}(r, R) \tag{1}$$

The Born Oppenheimer approximation, typically invoked in solving the Schrödinger equation, decouples the nuclear and electronic dynamics such that the nuclear motion can be neglected on the timescale of the electronic motion (**2**):[82]

$$\hat{H}_{\psi}(r, R) = E_{\psi}(r)E_{\psi}(R) \tag{2}$$

Thus, the resulting electronic Hamiltonian is a function of the 3N coordinates of the electrons only and consists of three terms (**3**): the electronic kinetic energy, the electron-nuclear interaction (also denoted the *external potential*), and the

electron-electron repulsion. While the kinetic energy and the electron-electron repulsions are universally constant, V_{ext} is system-dependent. [82]

$$\hat{H} = -\frac{\hbar^2}{2}\sum_{i}^{N}\frac{1}{m_i}\Delta_i^2 + \sum_{i}^{N}V_{\text{ext}}(r_i) + \sum_{i=1}^{N}\sum_{j>1}U(r_i, r_j) \qquad (3)$$

For single nuclei atomic systems, the wavefunction adopts the form of the hydrogenic orbitals (i.e., 1s, 2s, 2p etc.). Molecular systems are described by molecular orbital (MO) wavefunctions (*basis sets*), which are constructed from the linear combination of atomic orbitals (LCAO).[82] The MO (ψ) is constructed from the linear combination of N wavefunctions (∅), each bearing a coefficient, c:

$$\psi = \sum_{i=1}^{N} c_i \emptyset_i \qquad (4)$$

Similarly, a multi-electron system can be constructed from the linear combination of N single-electron Hamiltonians ($\widehat{h_i}$), each one satisfying its own Schrödinger equation:

$$\hat{H} = \sum_{i=1}^{N} \hat{h}_i \qquad (5)$$

Moreover, the Hamiltonian of the Hartree Product wavefunction, constructed from a series of single electron wavefunctions, returns the total energy of the system as the sum of the individual eigenvalues (**6**). However, this treatment neglects the electron correlation.[82]

$$\hat{H}_{\psi_{HP}} = (\sum_{i=1}^{N} \epsilon_i)\psi_{HP} \qquad (6)$$

2.2. Density Functional Theory (DFT)

Density Functional Theory (DFT) provides an approximate solution to the many-body Schrödinger equation, which cannot be solved analytically. This approach reduces a 3N-dimensional problem to a 3-dimensional problem through approximation of the ground state energy as a functional of the electronic density, which depends only upon the (x, y, z) Cartesian coordinates. This one-to-one correspondence between the electronic density and the wavefunction was proposed by Thomas and Fermi.[83,84] Subsequently, Hohenberg and Kohn established the practical basis of DFT through the recognition that the external potential V_{ext}, and hence the Hamiltonian of the system, is uniquely determined by the ground state density (n_0).[85] Further, for any V_{ext}, a functional of the energy in terms of the electronic density may be defined ($E = E[n(r)]$), such that the electron density which minimises this functional is the exact ground state density (**7**).

$$E = min_n \int V_{\text{ext}}(\overrightarrow{\mathrm{r}}) n(\overrightarrow{\mathrm{r}}) d^3(\overrightarrow{\mathrm{r}}) + F[n(\overrightarrow{\mathrm{r}})] \qquad (7)$$

$F[n(\vec{r})]$ represents the universal constant of the energy functional, which is independent of N, mass and the nuclear charge, including components such as the kinetic energy and the classic Coulomb interactions (**7**).

While Hohenberg-Kohn DFT is in principle exact, the practical implementation of DFT requires transformation of the interacting-particle Hamiltonian with a more readily soluble non-interacting ancillary. The external potential associated with the non-interacting particle Hamiltonian is denoted by the effective potential or the Kohn Sham potential. The Kohn Sham potential (ν_{ext}) of the non-interacting system and V_{ext} of the interacting system are determined by the same electronic

density. Further, the energy of the system which has electron density $\rho(r)$ may be decomposed into the kinetic energy of the non-interacting system ($T_{ni}\,[\rho(r)]$), the nuclear-electron interaction ($V_{ne}\,[\rho(r)]$), the classical electron-electron repulsion ($V_{ee}\,[\rho(r)]$), and two correction terms ($\Delta T\,[\rho(r)]$ + $\Delta V_{ee}\,[\rho(r)]$). The latter corrections compensate for the *interacting* electron kinetic energy and the non-classical QM components of the repulsion energy (i.e., *exchange-correlation*) respectively.

$$E[\rho(r)] = T_{ni}[\rho(r)] + V_{ne}[\rho(r)] + V_{ee}[\rho(r)) + \Delta T[\rho(r)] + \Delta V_{ee}[\rho(r)] \qquad (8)$$

Moreover, the Kohn Sham approach replaces the complex interacting-particle Hamiltonian with a series of fictitious, non-interacting single-particle orbitals (i.e., *Kohn Sham orbitals*), where v_H and $\boldsymbol{v_{xc}}$ represent the classical Coulomb energy and the exchange-correlation energy respectively (**9**).[86]

$$(-\frac{1}{2}\Delta^2 + v_{\mathrm{ext}}(r) + v_H(r) + v_{\mathrm{xc}}(r))\psi_i(r) = \varepsilon_i\psi_i(r) \qquad (9)$$

Furthermore, the utility of DFT depends on the approximation employed for the exchange-correlation functional, $v_{xc}\,[\rho]$. A hierarchy of DFT methods have been developed based on different treatments of $v_{xc}\,[\rho]$, including Local Density Approximation (LDA), Generalised Gradient Approximation (GGA), meta-GGA (MGGA), and hybrid functionals (**Figure 21**).[87]

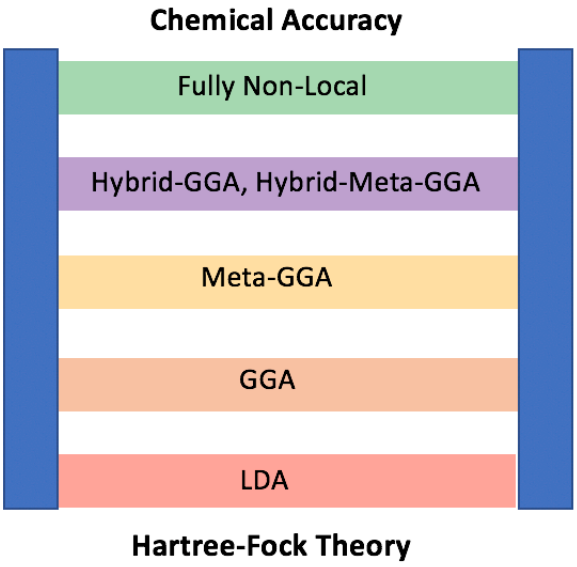


Figure 21: A hierarchical representation (i.e., "Jacob's ladder") of the five generations of DFT functionals.[87] Movement up the ladder corresponds to an increase in the accuracy of the approximations of the density functionals. LDA is the simplest density functional, based on the assumption that v_{xc} at any point in space is a function of the electron density at that point. GGA methods represent the second class, in which v_{xc} is a function of both the electron density and the gradient. Meta-GGA are a third class, which incorporate additional terms such as the kinetic energy density. Hybrid density functionals represent a class above, comprised of a combination of the exchange correlation of GGA and the Hartree-Fock exchange. Within this class, the meta-versions also include the kinetic energy density.

Choice of basis set is an important determinant of DFT efficacy. While larger basis sets may provide a more accurate description of the electronic structure, these require greater computational effort. In early DFT, Slater type orbitals (STOs) represented the hydrogenic *s, p, d* orbitals. However, Gaussian type orbitals (GTOs), also known as Gaussian *primitive*s, later replaced STOs owing to an increase in computational efficiency. As GTOs decay faster than STOs (i.e., with r^2 as opposed to r), contractions were formed from linear combinations of GTOs. Contracted GTOs approximated the shape of the STO function, while still achieving higher computational speed. Pople basis sets are commonly used, where the number of primitives constituting the GTO contractions in the inner shell (n) and in the double zeta (xy) or triple zeta (xyz) valence shell is encoded in their designations of n-xyG or n-xyzG. These are split valence basis sets, in which a single contracted basis function represents the core orbitals, while the valence orbitals

are split into two or more basis functions. In the current study, the 6-311G** basis set was used for all single point energy calculations. This is an example of a split valence, triple zeta basis set, in which 6 primitives constituted the inner shell and (3+1+1) primitives constituted the valence shell.

Finally, Pople basis sets may be augmented by polarisation functions (*/**) and diffuse (+/++) functions, which influence the shape of the electron density surrounding an atom. For example, the basis set designation 6-311G** indicates the addition of d-type polarisation functions to all heavy atoms, with p-functions on the hydrogens.

2.3. Molecular Dynamics

2.3.1. Newton's equations of motion

The motion of a particle in an N-particle system may be quantitatively described by Newton's second law of motion as a differential equation (**10**).

$$m_i \frac{d^2 r_i}{dt^2} = F_i(r_1, ..., r_N, \frac{d}{dt} r_i). \qquad (10)$$

Since the particle i experiences a force due to all other particles in the system, F_i depends on the positions of the latter (i.e., $r_1, \dots, r_N$), as well as the velocity of particle i ($\frac{d}{dt}$r$_i$). In addition, each particle may experience an external force due to the environment or an external agent. The solution to the second order differential equation, which governs the particle's motion, is uniquely identified by specifying the velocity and the position at a certain instant. Thus, the classical dynamics of the particle evolves deterministically from two initial conditions: the initial position, $r(0)$, and the initial velocity, $v(0)$. Fulfilment of these initial conditions requires description of the particle as a point in 6N-dimensional *phase*

space, i.e., a Cartesian space comprising three momenta (p) and three position (r) variables: $x = (r_1, \ldots, r_N, p_1, \ldots, p_N)$.

Moreover, throughout the particle's progression in phase space, the total energy, comprised of the kinetic and potential energies, is conserved (**11**).

$$\frac{d}{dt}\frac{1}{2}mv^2 + V(r) = mva + \frac{dV}{dt} = mva + \frac{dV}{dr}\frac{dr}{dt} = ma - F = 0 \qquad (11)$$

When the number of interacting particles exceeds two, the motion of the system cannot be provided analytically, and MD is employed as an efficient method for solving the equations of motion numerically for representative macroscopic systems (i.e., 10^5-10^6 atoms). MD permits integration of the equations of motion in time using computational algorithms or *integrators*. These algorithms evolve the system in phase space along a *trajectory*.

Hence, a MD simulation or trajectory involves four basic steps: initialisation, integration, equilibration, and averaging. **Initialisation** refers to selection of the initial conditions i.e., the positions and the velocities. Taking a biomolecular system as the example, the Cartesian coordinates are most commonly provided by X-ray crystal structures, and the velocities may be randomly assigned from the Maxwell-Boltzmann distribution at a temperature of choice (**12**).

$$P(v) = \frac{m}{2\mu kT}^{\frac{3}{2}} 4\mu v^2 e^{\frac{-mv^2}{2kT}} \qquad (12)$$

Following initialisation, **integration** of the equations of motion to provide an accurate trajectory of the biomolecule in phase space is performed using computational algorithms. Subsequently, **equilibration** is vital in order for the

system to lose memory of its initial conditions. Allowing the system to equilibrate ensures that the calculated expectation values of physical observables (e.g., thermodynamic properties) do not depend on the initial conditions. Finally, **averaging** describes the calculation of these physical observables throughout a trajectory. Under the hypothesis of *ergodicity*, it is assumed that the temporal average along a trajectory is equal to the ensemble-average over the phase-space (**13**). Hence, the expectation value of a thermodynamic property A can be accurately reproduced by integrating over a trajectory, provided the trajectory visits all regions of phase space (**13**).

$$\langle A \rangle = \frac{1}{M} \sum_{i=1}^{M} A(p^N, r^N) \qquad (13)$$

2.3.2. Integrators

The computational algorithms which integrate the equations of motion in MD must be time-reversible and robust. The simplest MD integrator is the *Verlet* algorithm, which provides a numerical solution to the equations of motion using finite difference methods. The central principle of this method is the calculation of future positions and velocities (at time $t + \Delta t$) from a knowledge of the current positions and velocities (at time t). It requires the discretization of time into small steps, each separated by a fixed interval of Δt, during which the force is assumed to be constant. Using the Verlet algorithm, the velocity (v) and acceleration (a) are represented as the first and second derivatives of the position (r), in Taylor expansion series (**14**-**16**).

$$r(t + \Delta t) = r(t) + v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 + \frac{1}{6}b(t)\Delta t^3 + O(\Delta^4)... \qquad (14)$$

$$r(t+\Delta t) = r(t) - v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 - \frac{1}{6}b(t)\Delta t^3 + O(\Delta^4)... \quad (15)$$

The sum of these two quantities furnishes the position at time t:

$$r(t+\Delta t) = 2r(t) - r(t-\Delta t) + a(t)\Delta t^2 + O(\Delta t^4)... \quad (16)$$

Importantly, the velocity is not present explicitly in the Verlet algorithm but may be calculated from the expression **17**.

$$v(t) = \frac{r(t+\Delta t) - r(t-\Delta t)}{2\Delta t} \quad (17)$$

The Verlet integrator calculates the coordinates with an accuracy of $O(\Delta t^4)$. Hence, the accuracy may be tuned by variation of the timestep, Δt. In general, timesteps on the order of femtoseconds are used (i.e., 10^{-15}). In the current study, a timestep of 2 femtoseconds was employed in all MD simulations.

Other MD integrators such as the *velocity Verlet* and the *leap-frog algorithm* are improved versions of the simple Verlet algorithm with the advantage of explicit inclusion of the velocity. The leap-frog integrator was employed in the MD simulations presented here. In the latter, the velocities are calculated at time $t + \frac{1}{2}\Delta t$ in order to furnish the positions at $t + \Delta t$ (**18**-**19**).

$$v(t+\frac{\Delta t}{2}) = v(t-\frac{\Delta t}{2}) + a(t)\Delta t \quad (18)$$

$$r(t+\Delta t) = r(t) + v(t+\frac{\Delta t}{2})\Delta t \quad (19)$$

Hence, the velocity at time t may be approximated using the expression **20**.

$$v(t) = \frac{1}{2}\lfloor v(t+\frac{\Delta t}{2} + v(t-\frac{\Delta t}{2})\rfloor \quad (20)$$

The limitation of this approach is that the velocities and the positions are not known at the same time, which renders calculation of the total energy at any single time point difficult.

2.3.3. Thermodynamic ensembles and thermostats/barostats

In simulations, certain intensive thermodynamic properties, such as temperature (**T**) and pressure (**P**), may be held constant. Straightforward integration of the equations of motion furnishes the microcanonical ensemble, in which the number of particles (N), the volume (V), and the energy (E) are held constant. However, other important thermodynamic ensembles which are more representative of true experimental conditions include the canonical (NVT) and isobaric-isothermal (NPT) ensembles. Depending on the choice of ensemble, it is necessary to augment the MD integrator with *thermostats* and/or *barostats*, which hold constant T and P respectively.

Thermostats rely typically on velocity manipulation or *scaling*. Since the temperature is related to the kinetic energy (KE), it is possible to control T by controlling the velocities or momenta (**21**-**23**).

$$\mathrm{KE} = \frac{1}{2}\sum_i m_i v_i^2 = \frac{1}{2}\sum_i \frac{p_i^2}{m_i} \tag{21}$$

$$\mathrm{KE} = \frac{3N}{2}kT \tag{22}$$

$$T = \frac{1}{3Nk}\sum_i \frac{p_i^2}{m_i} \tag{23}$$

Approaches to thermostatting include constrained (*velocity rescaling*), stochastic (L*angevin*), and extended system (e.g., Nosé-Hoover). In velocity rescaling (*v*-

rescale), the velocities at time T are scaled with the parameter λ in order to approximate the desired temperature, T_d (**24**).

$$v_i^{\text{new}} = v_i^{\text{old}}\lambda \tag{24}$$

$$\lambda = (\frac{T_d}{T_{MD(T)}})^{\frac{1}{2}} \tag{25}$$

The velocities are scaled every timestep or at a periodic interval which may be specified. The V-rescale thermostat was used in the MD simulations presented here, unless otherwise stated. Unlike the related *Berendsen* thermostat, it produces the correct kinetic energy distribution, and hence rigorously generates the canonical ensemble. This permits calculation of the correct thermodynamic averages.

The Berendsen thermostat, the simplest algorithm typically employed for equilibration, describes weak coupling between the system and an external heat bath (**26**).

$$v_i^{\text{new}} = v_i^{\text{old}}\lambda \tag{24}$$

$$\lambda = (1 + \frac{\Delta t}{\tau}(\frac{T_d}{T_{MD(T)}} - 1))^{\frac{1}{2}} \tag{26}$$

The strength of this coupling is controlled by the parameter τ, designated the *rise time*. Strong coupling is characterised by small τ values, while weak coupling is indicated by large values of τ. For simulations with a timestep of 1-2 fs, τ typically ranges from 0.1 to 0.4 ps. The advantage of this thermostat and its applicability to equilibration simulations is the exponential decay of the temperature deviation to the desired temperature. However, this thermostat is not appropriate for data

collection, as it fails to generate the correct canonical ensemble due to suppression of fluctuations in the kinetic energy.

In addition to V-rescale, other thermostats which generate the correct thermodynamic ensemble include the *Anderson* thermostat and extended methods such as the Nosé-Hoover. When using the Anderson thermostat, atoms are randomly selected and assigned new velocities from the Maxwell-Boltzmann distribution at the desired temperature (**12**). Using this approach, the collision frequency, f, describes the number of times atoms are selected for velocity reassignment. This equates to the strength of the coupling to the external heat bath. On the other hand, the Nosé-Hoover thermostat introduces the heat bath as an additional degree of freedom to the Hamiltonian.

Finally, the temperature may be controlled using Langevin dynamics, an example of a stochastic method. Using this approach, the equations of motion are modified by the addition of a friction coefficient (γ_i) or *viscosity*, and random forces (f_i) (**27**-**29**).

$$\frac{dq_i}{dt} = \frac{dT}{dp_i} \tag{27}$$

$$\frac{dp_i}{dt} = -\frac{dV}{dq_i} \tag{28}$$

$$\frac{dp_i}{dt} = -\frac{dV}{dq_i} - \gamma_i p_i + f_i \tag{29}$$

The temperature is controlled by applying random forces to the system, generated from the Gaussian distribution with variance $ð_i^2$ (**30**). This is performed every N steps according to a specified collision frequency.

$$\sigma_i^2 = \frac{2m_i\gamma_i k_B T}{\Delta t} \tag{30}$$

Contrary to thermostats, barostats control the pressure by scaling the volume of the simulation box. The virial equation of state relates the pressure and volume of the system (**31**).

$$\frac{PV_m}{RT} = 1 + \frac{B}{V_m} + \frac{C}{V_m^2} + \ldots \tag{31}$$

Using the Berendsen barostat, the box vectors are modified by the factor μ in order to match the desired pressure, P_d (**32**-**33**).

$$L_i^{\text{new}} = L_i^{\text{old}}\mu \tag{32}$$

$$\mu = (1 - \frac{\Delta t}{\tau}\gamma(P_d - P_{MD(T)}))^{\frac{1}{3}} \tag{33}$$

According to this approach, γ represents the compressibility of the system, while τ represents the strength of the coupling to the hypothetical pressure bath (**33**). Thus, larger values of τ indicate weaker coupling. Similar to the Berendsen thermostat, the Berendsen barostat was employed in this work solely for equilibration, since the exact NPT ensemble cannot be generated. Thus, this barostat does not permit accurate calculation of thermodynamic properties from trajectories.

Furthermore, the *Parrinello-Rahman* barostat was used in the present work for production simulations as it reproduced the correct NPT ensemble. Similar to the Nosé-Hoover thermostat, the Hamiltonian was modified by the addition of extra degrees of freedom. The extra degree of freedom represented the volume of the

box, whose sides could be independent of each other and whose vectors could be nonorthogonal. This permitted anisotropic pressure coupling to be performed.

2.3.4. Periodic Boundary Conditions (PBCs)

Real macroscopic systems (i.e., 10^{23} atoms) cannot be subjected to MD, as in practice only a finite number of degrees of freedom may be simulated. However, through the use of PBCs, a finite number of molecules can exactly reproduce the behaviour of an infinite system. This is achieved if a molecule in the primary simulation cell interacts with other independent molecules (i.e., the closest periodic images) up to its maximum range of dynamical correlation. This maximum range is half the length of the primary simulation box and satisfies what is called the *minimum image convention*. Thus, PBCs ensure that molecules at the surface of the simulation box are modelled as thought they were in the bulk through the formation of a periodic array without boundaries.

While the simplest periodic system to programme is the cubic box, it is not the most computationally efficient. Other box shapes reduce the computational burden of surplus solvent, such as the *truncated octahedron* and the *rhombic dodecahedron*, which occupy 0.770 and 0.707 of the cubic volume respectively. In the current work, MD simulations were carried out in the *rhombic dodecahedron*, unless otherwise stated.

2.3.5. Treatment of long-range interactions: Coulombic and Lennard-Jones potentials

In classical MD, the number of interactions, which may be decomposed into van der Waals and electrostatic components, scales with the square of the number of atoms (N^2). The interactions between the dipole moments are modelled by the

Lennard-Jones potential, which decays rapidly with r^{-6} distance dependence. In MD, a *cutoff* is introduced such that solely the interactions within the specified distance are computed. A *neighbour list* is updated at specified time intervals in order to identify the interacting particles which are within this cutoff distance. Introduction of the cutoff does not result in substantial error for the van der Waals interactions, since the Lennard Jones energy is close to zero at the cutoff distance. Introduction of a *switching function* renders the Lennard Jones energy equal to zero at the cutoff. Although the error is small, this approximation can affect the accuracy of free energy calculations, hence various correction schemes exist. In GROMACS, cutoff artifacts are reduced by the analytical correction for energy and pressure. However, this correction is imperfect since it assumes an isotropic solvent distribution outside of the cutoff, which does not accurately model large biomolecules. Hence, other correction schemes have been developed, such as the numerical correction for non-isotropic systems by Shirts *et al.*, denoted as "EXP-LR". Additionally, Wennberg *et al*. introduced the *Lennard Jones particle mesh Ewald* (LJ-PME) method, which may be implemented within GROMACS and removes the requirement for an added correction.[88]

On the other hand, interactions between point charges are modelled by the Coulomb potential energy, which decays with r^{-1} distance dependence (**34**). For electrostatic interactions, the Coulomb energy is still significant at the cutoff distance, and hence a large error, as well as discontinuities in the potential energy surface (PES), are introduced. Additionally, the sum of all pairwise electrostatic interactions converges extremely slowly (**35**). In expression **35**, i and j are the interacting particles and n represents all replicate cells.

$$E = \frac{1}{4\pi\varepsilon_0} \sum_{(i,j)} \frac{q_i q_j}{|r_{ij}|} \tag{34}$$

$$E = \frac{1}{4\pi\varepsilon_0} \frac{1}{2} \sum_{n} \sum_{i=1}^{N} \sum_{j=1}^{N}{}' \frac{q_i q_j}{|r_{ij} + nL|} \tag{35}$$

Hence, to combat this limitation, Ewald introduced the concept of transforming a slow converging sum into two quickly converging components. This concept may be described mathematically using expression **36** and forms the basis of the Ewald method.

$$\frac{1}{r} = \frac{f(r)}{r} + \frac{1 - f(r)}{r} \tag{36}$$

This method evaluates the long-range electrostatic potential, subject to PBCs, due to a periodic lattice of point charges. The technique may be modelled by a one-dimensional ionic lattice, with a charge distribution given in **Figure 22**. The given lattice may be decomposed into two rapidly converging components (**Figure 22**: B and C) computed in real and reciprocal space respectively. The original point charges (**Figure 22**: A) are screened by Gaussian charge clouds of opposite sign, thereby producing an auxiliary lattice of short-range interactions (**Figure 22**: B). To compensate for the introduction of these screening charges (**Figure 22**: B), Gaussian charge clouds of the same sign as the original (slow converging) point charge distribution are introduced (**Figure 22**: C). The charge distribution of the latter (**Figure 22**: C) is computed by taking the Fourier transform (FT) of the rapidly converging charge series (**Figure 22**: B) in reciprocal space. Thus, the original slow converging charge distribution is replaced by the fast-converging sum of the point

charges and screening Gaussian charges (**Figure 22**: B) plus the compensating Gaussian charges (**Figure 22**: C).

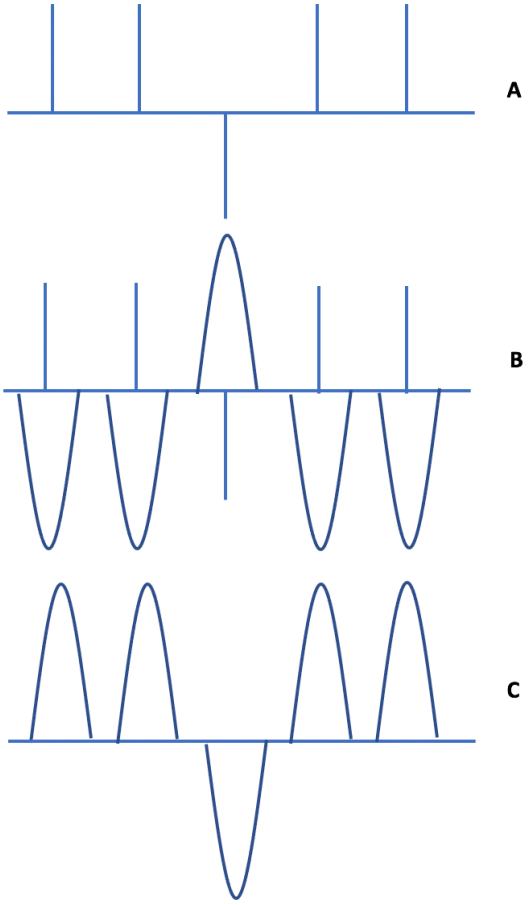


Figure 22: Graphical illustration of the Ewald summation method.[89] Long-range Coulombic interactions are slow-converging and difficult to compute (**A**). Thus, Gaussian charge distributions of opposite charge are introduced to screen the original point charges (**B**). This yields an interaction between point charges which is of short-range character and easy to compute in real space. The first set of Gaussian charge distributions is counteracted by a second set (**C**), which is computed by taking the FT of the first set (**B**) in reciprocal space. The slow-converging charge distribution (**A**) may be computed from the fast-converging sum of distributions **B** and **C**.

Furthermore, the Ewald method reduces the scaling of the interactions with respect to the number of atoms from N^2 to $N\,ln\,N$. Highly accurate and widely used methods based on the Ewald technique include the *particle mesh Ewald* (PME) and *particle-particle-particle mesh* (PPPM). In the present study, PME was employed in all MD simulations.

2.3.6. Principal Component Analysis (PCA)

PCA reduces the dimensionality of information through reduction of the number of variables. Thus, new uncorrelated variables (i.e., *principal components*) may be constructed from linear combinations of the initial variables, such that the first principal component (PC) captures the most data. In MD, PCA is a valuable tool for decomposing the dynamics of a molecule into the dominant modes of motion. Initial variables may represent all α-carbon motions throughout the MD trajectory, while the PCs may correspond to correlated vibrational modes or collective motions of groups of atoms.[90]

PCA extracts the dominant modes of motion from MD trajectories.[91–93] Prior to PCA analysis, the molecular system must undergo translation to its average geometrical center and least squares fit superimposition onto a reference structure, in order to remove translational and rotational motion.[90,9292] The input data for the PCA are the Cartesian coordinates of the system throughout the simulation, which are used to generate a $3N \times 3N$ covariance matrix and to reconstruct the configurational space. Eigenvectors, as vectorial representations of the modes of motion, are computed through diagonalisation of the matrix.[90,94,95] The latter correspond to the PCs of the motion, equivalent to the directions containing the highest variance (i.e., largest motion). Eigenvalues are the coefficients of the eigenvectors which quantify the variance contained in each PC. Thus, the eigenvalues represent the energetic contribution to the total dynamics of a particular PC.[90] Ranking of the eigenvalues in descending order provides the PCs in order of their significance, (i.e., from first PC to the last). Thus, the eigenvectors associated with lower eigenvalues may be discarded, leading to the generation of a *feature vector* or matrix of reduced dimensionality. One useful

application of PCA is projection of the trajectory onto a new feature space defined by the axes of two selected PCs, thereby highlighting the time-dependent motions of all other components in the particular vibrational mode represented by the selected PCs.[90]

2.4. Enhanced sampling methods

Important conformations of biomolecules may be separated by high energy barriers which prevent high rates of barrier crossing or transition, thereby leading to *kinetic trapping*.[96] In order to overcome these barriers and sample the conformational space sufficiently, enhanced sampling methods have been developed.

These methods tend to fall into two broad categories. The first involves the addition of a bias potential to the Hamiltonian of the system in order to accelerate MD sampling. This category, which includes methods such as *umbrella sampling* (US) and *metadynamics*, requires the definition of *collective variables* (CVs), along which the MD simulation is guided.[97] However, identification of CVs is rarely trivial and CV-free approaches, such as the use of generalised ensembles, constitute the second category of enhanced sampling methods. The latter perform random walks in the temperature or potential energy space, thereby giving rise to a canonical probability distribution with broader sampling.[97] Of the generalised-ensemble methods, *parallel tempering* or *replica exchange molecular dynamics* (REMD) is one of the most widely used.

2.4.1. Umbrella Sampling (US)

US is a popular technique for computing the ΔG_{bind} of protein-ligand systems.[98] Establishing an association-dissociation pathway is necessary prior to performing US calculations, and methods such as *steered molecular dynamics* (SMD), metadynamics, or *adaptive biasing force* (ABF) are employed.[98] The initial step in calculating the free energy according to the US method, is defining the CV. In calculating the binding free energy of protein-ligand complexes, the CV corresponds to the centre-of-mass (COM) distance between the protein and the noncovalent ligand. The function, ξ, also referred to as the *reaction coordinate* (RC), describes the CV in terms of the position coordinates (i.e., r^3N).[99] The free energy along the RC, $A(\xi)$, also known as the *potential of mean force* (PMF), is related to the probability of finding the system in the vicinity of a given ξ (**37**).[99]

$$A(\xi) = -k_B T \ln P(\xi) \tag{37}$$

In US, $A(\xi)$ may be obtained through MD simulations in which biasing potentials force sampling in the vicinity of the chosen CV.[99,100] Hence, the Hamiltonian of the system is comprised of the Hamiltonian of the unbiased system, $H_0(\Gamma)$, and the biasing potential, $W_i(\xi)$, each of which are functions of the microscopic state of the system, Γ (**38**).[99]

$$H_i(\Gamma) = H_0(\Gamma) + W_i(\xi) \tag{38}$$

To facilitate efficient sampling, the CV is divided into individual MD windows, each representing specific protein-ligand COM distances. Biasing potentials are applied to restrict sampling to regions in the vicinity of these COM distances (i.e., regions approximating the center of each window, ξ_i^0). Typically, the biasing potentials take the form or harmonic functions, and as such, they may be defined as:[99]

$$W_i(\xi) = \frac{k_i}{2}(\xi - \xi_i^0)^2, i = 1, ..., R \quad (39)$$

The harmonic force constant, the total number of US windows (R), and the window spacing (i.e., distances between windows) are all important determinants of efficient sampling.[99] Additionally, accuracy may be improved by increasing convergence of the US simulations. Thus, enhanced sampling approaches may be incorporated, such as the use of *replica exchange umbrella sampling* (REUS) or integration of the *Markov model*.[101–104]

Moreover, using biased simulations, it is possible to estimate the unbiased probability of finding the system in the given ξ_i state (**40**).[99]

$$P_i^b(\xi_i) = P^u(\xi_i) \exp \frac{W_i(\xi) - F_i}{k_B T} \quad (40)$$

Thus, once $P^u(\xi_i)$ has been calculated for each US window, the unbiased free energy may be computed in each window using equation **40**. This requires that R simulations are run in order to obtain the total unbiased probability and hence the total unbiased free energy for the system. Combination of all $P^u(\xi_i)$ estimates, and hence the reconstruction of the unbiased PMF, is achieved using *weighted histogram analysis* (WHAM).[105–107] The resulting total binding free energy is affected by the degree of overlap between US windows, as well as the convergence of all MD simulations.

2.4.2. Replica Exchange Molecular Dynamics (REMD)

First developed by Okamoto and co-workers, REMD is a hybrid method which integrates MD and the Monte Carlo algorithm.[108,109] In the original version of

REMD, temperature REMD (TREMD), many independent simulations of the same system (i.e., *replicas*) are run in parallel, each at a different temperature (**Figure 23**). The simulations at higher temperature accelerate the conformational sampling of the trajectory at the physical temperature, from which calculations of the desired physical observable are performed.[97]

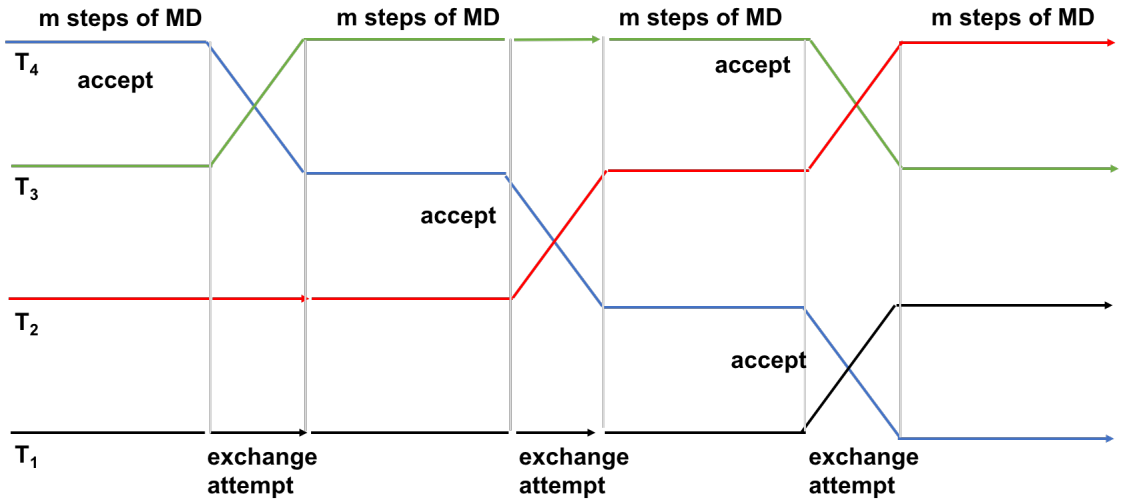


Figure 23: Illustration of the REMD cycle, involving exchange attempts of system coordinates between four different temperature replicas. Adapted from Earl and Deem (2005).[110]

Interactions between replicas are performed through coordinate swapping attempts at a specified time interval or *exchange attempt frequency* (EAF).[97] Swapping, which is achieved by the exchange of temperatures of two replicas, is equivalent to the exchange of conformations between two adjacent temperatures (**Figure 23**).[109] In order to accept the attempted swap, the potential energy distributions of the adjacent temperature replicas must overlap sufficiently.[109] Successful exchange attempts require rescaling of the velocities to match the new temperatures, thereby ensuring the correct kinetic energy corresponding to the target temperature (**41**).[96]

$$v_{\text{new}} = v_{\text{old}} \sqrt{\frac{T_{\text{new}}}{T_{\text{old}}}} \qquad (41)$$

During the attempted exchange of replica i (at temperature T_n) with replica j (at temperature T_m), the detailed balance condition is fulfilled, where $p(i \longrightarrow j)$ represents the transition probability and $P_n(i)$ defines the probability of state i at temperature T_n (**42**).[96]

$$P_n(i)P_m(j)p(i,n \longrightarrow j,m) = P_m(i)P_n(j)p(j,m \longrightarrow i,n) \qquad (42)$$

The exchange probability is governed by the Metropolis criterion, such that:

$$p = \min(1, \exp\{(\beta_m - \beta_n 19)(E(q^{[i]}) - E(q^{[j]})\}) \qquad (43)$$

where $q^{[i]}$ represents the position of replica i, β = $\frac{1}{k_B T}$, and $E(q^{[i]})$ denotes the potential energy of replica i.[96]

Moreover, the EAF is a critical determinant of TREMD efficiency and significant debate exists around its optimal value.[96,97,109] When the EAF approximates zero, the sampling is equivalent to that of the conventional canonical ensemble. It has been postulated that the value of the EAF should be equal to or greater than the potential energy autocorrelation time.[111,112] On the other hand, it is argued that the sampling efficiency is optimised by increasing values of the EAF.[96,113] According to Qi *et al.*, the generally accepted range of the EAF is 500-1500 integration steps, considering a timestep of approximately 2 fs.[109] An EAF value of 1000 was employed in all TREMD simulations of the present study.

Furthermore, the most important limitation of REMD is its inefficiency in simulating large biomolecules in explicit solvent.[96,97] Large systems require many hundreds of replicas owing to increasing numbers of degrees of freedom.

Specifically, the number of replicas required scales with $\boldsymbol{O(f^{1/2})}$, where f represents the number of degrees of freedom.[96] In addition to its inefficiency with respect to increasing system size, REMD is also inefficient as it samples all degrees of freedom of the system.[97] Conversely, CV-based enhanced sampling methods, such as US, restrict sampling to the RC of interest.

Hamiltonian REMD (HREMD) is a variation of the original temperature-based REMD devised by Fukunishi *et al.*[114] In HREMD, the variable condition to be exchanged among replicas is not the temperature but the Hamiltonian.[115] The Hamiltonian may be defined as the sum of the kinetic energy, $K(p)$, and the potential energy, $E(q)$, where p and q denote the momentum and coordinate vectors (**44**).[115]

$$H(q,p) = \mathrm{KE}(p) + E(q) \tag{44}$$

In HREMD, the replicas differ solely in their potential energy form of the Hamiltonian. Exchange between the replicas i and j, simulated with the Hamiltonians H_m and H_n respectively, may be described as:[115]

$$\Omega = \{H_m(q_i,p_i), H_n(q_j,p_j)\} \longrightarrow \Omega' = \{H_n(q_i,p_i), H_m(q_j,p_j)\} \tag{45}$$

Additionally, the exchange probability satisfies the detailed balance condition (**46**).[115]

$$w(\Omega \longrightarrow \Omega') \equiv_{\min} [1, \exp(-\Delta)] \tag{46}$$

$$\Delta \equiv \beta[E_n(q_i) - E(q_i) + E(q_j) - E(q_j)] \tag{47}$$

Thus, the advantage of this approach is that exchanges between replicas are independent of the unmodified part of the Hamiltonian, depending only on the modified component of the potential energy.[114–117]

In HREMD, the dynamic properties of the system may be influenced by modifications to the interparticle nonbonded interactions, as components of the potential energy function.[115] The use of modified forcefield parameters for these nonbonded interactions may in turn enhance the conformational sampling.[115] Many variations of HREMD, based on unique treatments of the nonbonded interactions, have been developed. While studies by Fukunishi *et al*. centered on simulation of biomolecules in implicit solvent or in a vacuum, a commonly used adaptation for aqueous systems is *replica exchange with solute tempering* (REST). In the REST method, developed by Liu *et al*., the potential energy scales with temperature such that the solute of interest 'heats up' while the solvent remains cold.[118] Thus, the acceptance probability for REST is independent of the number of explicit water molecules.[118]

Hence, the HREMD implementation addresses limitations of TREMD. Firstly, there is no requirement for velocity rescaling upon acceptance of replica exchanges during constant temperature HREMD.[115] In addition, the modified part of the Hamiltonian, the potential energy, is restricted to a subset of the total system which encompasses a smaller number of the total degrees of freedom. This is an advantage of scaling with the potential energy, an *extensive* property, as opposed to the temperature, an *intensive* property; selection of only the most relevant coordinates of the system for 'heating' results in a smaller number of replica simulations.[114,117]

2.5. Free Energy calculations

The free energy difference between two thermodynamic states is directly related to the ratio of the probabilities of these states.[119] Taking the bound and unbound states of a protein-ligand complex as an example, each state is described by a partition function Q, which describes a specific statistical ensemble corresponding to a particular free energy.[120,121] In the canonical (NVT) ensemble (where U = internal energy, S = entropy), the Helmholtz free energy may be expressed as:[121]

$$A = -k_B T \ln Q_{NVT} = U - TS \tag{48}$$

On the other hand, in the isothermal-isobaric (NPT) ensemble, the Gibbs free energy is defined as:

$$G = A + PV = U - TS + PV = H - TS \tag{49}$$

Under isothermal-isobaric conditions, the Gibbs free energy of a single ligand binding event may be expressed in terms of the dissociation constant (k_d) of the complex (**50**).

$$\Delta G_{\text{bind}} = -k_B T \ln \frac{K_d}{c^\circ} \tag{50}$$

In expression **50**, c° is the standard concentration, defined as 1 molar by convention.[122] The binding free energy of 1 mole of the ligand may be calculated by replacing k_B with the gas constant, $N_A k_B$**.** Moreover, the dissociation constant K_d is equivalent to the ratio of the probabilities for the system existing in the bound or unbound states.[122] This probability and hence the ΔG_{bind} may be determined by the ratio of the partition functions of the latter, each associated with a specific free energy.[122] However, the partition functions of the two states cannot be computed analytically but must be solved using MD simulations, through sampling of the configurational phase space.

A hierarchy of physics-based free energy calculation methods exist, most striving to achieve a balance between precision and computational efficiency. The fastest and least accurate of these methods are scoring functions, widely employed in docking software, which typically predict the ΔG_{bind} of a protein-ligand complex using a physical potential. Using the scoring approach, inaccuracy arises from the inability to represent the conformational flexibility of the biomolecule using a single X-ray crystal structure or representation of the complex. In addition, only the bound state is considered. Since the binding free energy is determined by a conformational ensemble, methods employing MD simulations increase the predictive precision. Thus, an alternative to scoring algorithms is the more rigorous MD-based approach or *pathway* methods, which may be subdivided into *alchemical*, *endpoint* and *geometrical* calculations.

Alchemical calculations, which rely on a thermodynamic cycle whose endpoints are interpolated by non-physical *alchemical* states, typically achieve the highest level of accuracy[123]. Using this approach, the ΔG_{bind} is given by the sum of the free energy changes obtained from simulations of all non-physical intermediates. Conversely, endpoint methods compute ΔG_{bind} from simulations of only the physical endpoint states, without intermediates.[124] These include the linear interaction energy (LIE) and the molecular mechanics with Poisson Boltzmann or generalised Born and surface area (MMPBSA/MMGBSA) methods.[125–130] Finally, geometrical calculations recover ΔG_{bind} along a predetermined physical pathway by applying a force which extracts the ligand from the protein cavity and hence calculates the work required for this association/dissociation process.[123,131,132] Widely used steered methods in this class include US and metadynamics.[133,134]

2.5.1. Scoring Functions (Astex Statistical Potential)

Docking is a molecular modelling method which predicts the binding orientation (i.e., *pose*) of ligands in protein binding sites. Most commonly employed docking programs include DOCK, FlexX, Autodock and Genetic Optimisation for Ligand Docking (GOLD).[135–140] GOLD was the docking program employed in the present study.[139,140] For each ligand, a search algorithm is employed to select the preferred binding poses, with are ranked in descending order of a scoring function. Scoring functions are simplified potentials which predict the approximate binding affinity of the protein-ligand complex. The high computational speed of scoring functions renders them particularly useful in virtual screening.[141] In general, the higher the score associated with a specific binding pose, the higher the predicted binding affinity. GOLD provides four scoring functions: ChemPLP (Piecewise Linear Potential), GoldScore, ChemScore and ASP (Astex Statistical Potential). ASP was the scoring function used in the docking studies presented here.

In general, scoring functions may be empirical (e.g., ChemPLP), derived from molecular mechanics force fields (e.g., GoldScore),[142–144] or trained by regression against experimental binding data[145,146] (e.g., ChemScore). ASP belongs to an additional category of scoring functions based on knowledge-based or *statistical potentials* and may be compared to others such as PMF and DrugScore.[141] The underlying basis of these atom-atom distance potentials is comparison of the observed number of protein-ligand atom contacts to the expected number of protein-ligand atom contacts in the absence of an interaction.[141] In ASP, the expected number of atom-atom contacts in the hypothetical state of no interaction is denoted the *reference state*, and is informed by the Protein Data

Bank (PDB) of experimental protein-ligand complexes.[141,147] Atom-atom contacts are reduced to radial distribution functions and hence the number of contacts between protein atom type i and ligand atom type j is expressed as a function of their separation, r.[141] Thus, the statistical potential between the atom types i and j takes the following general form:

$$\mathrm{StatScore}(i,j,r) = -\ln \frac{n_{\mathrm{obs}}(i,j,r)}{n_{\mathrm{exp}}^{\mathrm{StatScore}}(i,j,r)} \qquad (51)$$

In expression **51**, $n_{\mathrm{exp}}^{\mathrm{StatScore}}$ represents the reference state.[141]

Furthermore, the form of the statistical potential is modified through the addition of volume corrections, which account for the volume occupied by the ligand.[141] Calculations of the expected number of protein-ligand atom contacts must use only the available volume (i.e., not that already occupied by the ligand).

2.5.2. Linear Interaction Energy (LIE)

LIE is an end-point method which computes ΔG_{bind} from the difference of polar and non-polar interaction energies of the initial unbound and final bound states.[127,148,149148] LIE furnishes greater accuracy than scoring approaches but is less accurate than alchemical methods.[148] Hence, this trade-off between accuracy and computational efficiency, renders LIE particularly useful for high-throughput virtual screening. Additionally, the method can be applied to flexible proteins, in which multiple protein-ligand conformations contribute to the ΔG_{bind}.[150]

The LIE method was first developed by Aqvist *et al*., in which computation of ΔG_{bind} was centred upon the transfer process of the ligand from the solvent medium to that of the solvated biomolecule, where translation and rotation were restricted.[149,151,152] In this thermodynamic cycle, the transition of the fully

interacting, separate ligand and protein to the fully interacting ligand-protein complex is interpolated by non-interacting unphysical intermediates (**Figure 24**).[153] Thus, $\Delta G_{\text{free}}^{\text{polar}}$ and $\Delta G_{\text{bound}}^{\text{polar}}$ describe the free energy change associated with turning on the electrostatic intermolecular interactions of the solvated, unbound ligand, and that of the bound ligand respectively (**Figure 24**).[153] The contribution from entropic confinement is contained in the nonpolar term, $\Delta\Delta G^{\text{nonpolar}}$.[153] Based on this thermodynamic cycle, ΔG_{bind} is computed from the average differences in the polar and nonpolar energies of the physical endpoint states (i.e., bound and unbound) using expression **52**.

$$\Delta G_{\text{bind}} = (\Delta G_{\text{bound}}^{\text{polar}} - \Delta G_{\text{free}}^{\text{polar}}) + \Delta\Delta G_{\text{bind}}^{\text{nonpolar}} = \Delta\Delta G_{\text{bind}}^{\text{polar}} + \Delta\Delta G_{\text{bind}}^{\text{nonpolar}} \quad (52)$$

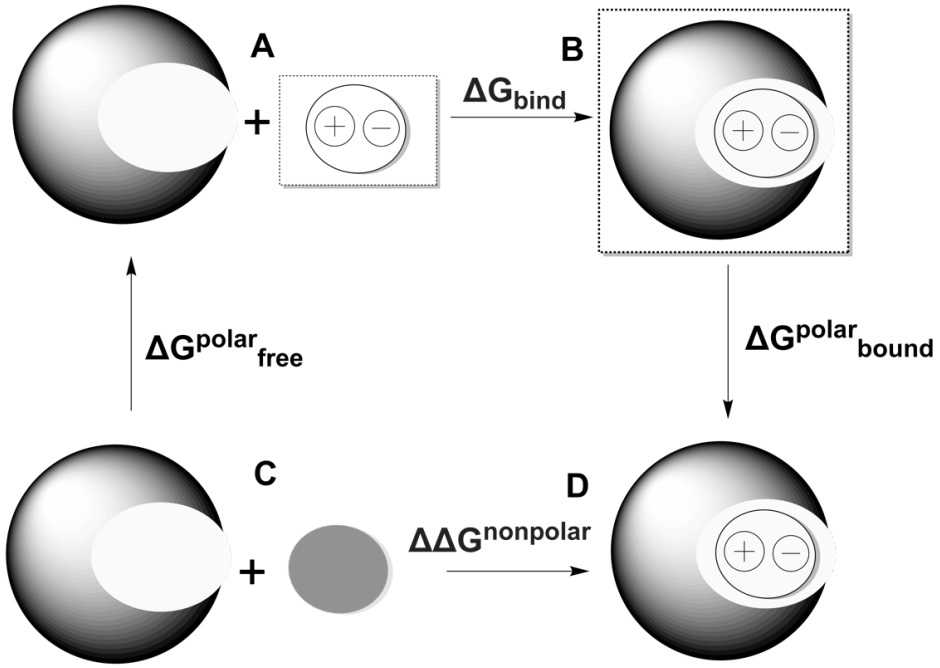


Figure 24: The thermodynamic cycle used in calculating ΔG_{bind} by the LIE method. **A** and **B** represent the physical unbound and bound states of the protein-ligand complex respectively. **C** and **D** represent the two non-physical intermediate states, i.e., the free (**C**) and bound (**D**) ligand without any intermolecular electrostatic interactions. Adapted from Åqvist (2012).[127]

More specifically, computation of ΔG_{bind} requires taking the difference of the sum of the average van der Waals and electrostatic interaction energies of the free and bound ligand, quantities which are calculated from MD simulations of the solvated ligand and the solvated ligand-protein complex. The formulae for calculation of

both the electrostatic and van der Waals contributions are derived from the *Linear Response Approximation* (LRA), with addition of the unbound ligand as an important reference state (**53**-**54**).[148,153]

$$\Delta G_{\text{bind}}^{\text{polar}} = \beta\{\langle U_{l-2}^{el}\rangle_b - \langle U_{l-2}^{el}\rangle_f\} = \beta\Delta\langle U_{l-s}^{el}\rangle \quad (53)$$

$$\Delta G_{\text{bind}}^{\text{nonpolar}} = \alpha\{U_{l-2}^{vdW}\rangle_b - \langle U_{l-2}^{vdW}\rangle_f\} = \alpha\Delta\langle U_{l-s}^{vdW}\rangle + \gamma \quad (54)$$

Thus, the LIE method combines the van der Waals and electrostatic contributions, which are derived from the MD ensemble-averaged interaction energies, and scales these by the α and β parameters to establish the overall LIE equation **55**.[153]

$$\Delta G_{bind} = \alpha\Delta\langle U_{l-s}^{vdW}\rangle + \beta\Delta\langle U_{l-s}^{el}\rangle + \gamma \quad (55)$$

The α parameter has been determined empirically to have the value 0.18, while the γ parameter is an offset parameter reflecting the hydrophobicity of the binding site.[148,153] The latter is required for absolute free energies of binding but is not necessary for computing relative values.[153] Binding free energy calculations have revealed deviation of the β parameter from its LRA value of 0.5, depending on the chemical nature of the ligand, which in turn gives rise to variations in the hydrogen bonding interactions with the solvent.[125,126,154] For example, *free energy perturbation* (FEP) calculations of chemically diverse ligands performed by Aqvist and Hansson culminated in the assignment of the values 0.5, 0.43, 0.37 and 0.33 to charged ligands, neutral ligands, and ligands exhibiting one or two hydroxyl groups respectively.[126] Further refinements to these rules were later introduced by Almlof *et al.*, who attributed specific values of β according to the number and types of functional groups present (e.g., alcohols, primary and secondary amines, amides, and carboxylic acids).[125]

Moreover, calibration of the α and β scaling factors for a specific protein-ligand system, using a training set of compounds with known experimental affinities, may enhance accuracy relative to the use of pre-assigned values.[148,155] In particular, this calibration has been employed in the iterative LIE scheme developed by Stjernschantz and Oostenbrink, in which the contributions from multiple MD simulations, starting from different protein-ligand configurations, were quantified in a Boltzmann-weighted scheme (**56**).[150]

$$\Delta G_{\text{bind}} = \alpha \sum_{i}^{N} W_i(\Delta \langle U_{l-s}^{vdW} \rangle) + \beta \sum_{i}^{N} W_i(\Delta \langle U_{l-s}^{el} \rangle) \qquad (56)$$

This strategy was useful for flexible proteins such as the cytochrome P450s, which bind ligands in a variety of orientations.[150,156] Thus, the relative contributions of multiple docked poses, each employed as starting points for MD-LIE analysis, were considered. This reduced the importance of selecting the most probable single pose. The statistical weighting scheme introduced by Stjernschantz and Oostenbrink was refined by Peric and Hassler, through incorporation of multiple protein conformations and many replicas of the protein-ligand configurations selected.[156,157]

Furthermore, advancements in the LIE method include use of an implicit solvent model and automation.[158] The latter is of particular relevance to high-throughput applications, in which statistically weighted models of LIE may be implemented automatically, including definition of the *applicability domain* (AD).[148] As the calibration is performed using a training set, the model's applicability for new predictions is defined by a specific chemical space and set of interactions. Hence,

confidence in the validity of new predictions must involve AD assessment, which has been integrated into automated models such as eTOX ALLIES.[159]

2.5.3. Alchemical free energy calculations

Alchemical free energy calculations describe a set of methods which compute the free energy difference between thermodynamic states of a system, as a function of changes to the chemical structure or identity of the system. Since the free energy is a state function, the free energy difference is independent of pathway.[160] Hence, in alchemical calculations, the free energy difference between end points (i.e., bound and unbound states) is interpolated by multiple *unphysical* intermediate states.[160–163] The most prominent methods for calculating the free energy difference between two thermodynamic states are FEP and *thermodynamic integration* (TI).

The FEP method is derived from the Zwanzig relationship (**57**), in which the Boltzmann-weighted potential energy difference between two states (e.g., 0 and 1) is calculated as an ensemble average over the MD configurations sampled by a single state (i.e., state 0).[164]

$$\Delta G_{0,1} = -k_B T \ln\langle e^{-\beta(U_{1(x)} - U_{0(x)}\rangle_0} \qquad (57)$$

The free energy difference is independent of the kinetic energy contribution to the Hamiltonian, as the temperature and masses remain unaltered. Calculation of this free energy difference between the potentials U_0 and U_1, over a coordinate and momentum space (x), is known as *exponential averaging* (EXP).[122] While the equation represents the forward transformation (i.e., $0 \longrightarrow 1$), the reverse transformation is theoretically equivalent. In practice, differences between ΔG of the forward and reverse transformations are referred to as *hysteresis* and indicate

poor convergence.[122] EXP is severely limited by its poor convergence as a function of sampling.[119] Additionally, its accuracy is restricted to cases where the final state phase space, V_j, is a subset of the phase space of the initial simulated state, V_i.[119]

In cases where the endpoints exhibit minimal phase space overlap, the overall transformation may be decomposed into a series of alchemical intermediates. Each pair of alchemical intermediates should exhibit good phase space overlap and may be subjected to a free energy calculation.[119,160] Each alchemical intermediate lies along a continuous pathway of states in a thermodynamic transformation. These discrete states are parametrised by the coupling parameter λ, and collectively form a series between the initial (λ = 0) and final states (λ = 1). Each λ state represents an independent MD simulation. More specifically, λ scales the coulombic charges, Lennard Jones parameters or restraint forces along separate transformation pathways, each associated with a specific alchemical intermediate.[122] An example of such a pathway is the turning on (i.e., *coupling*) of electrostatic interactions between a protein and a ligand, where λ = 0 and λ = 1 represent states in which the electrostatic interactions are turned off and turned on respectively. Hence, the sum of the individual free energy differences between each alchemical state i and its successor $i + 1$, each belonging to a λ-parametrised pathway, yields the total free energy difference between the physically meaningful initial (1) and final states (K) respectively (**58**).[119]

$$\Delta G_{1,K} = \sum_{i=1}^{k-1} \Delta G_{i,i+1} \tag{58}$$

The Bennett Acceptance Ratio (BAR) is a statistically optimal free energy method which relies on configurational sampling from the forward and reverse

transformations of each pair of adjacent λ states.[165–167] Unlike EXP, this method permits calculation of free energy differences between states with low phase space overlap through inclusion of multiple intermediate states. The *multistate Bennett Acceptance Ratio* (MBAR), proposed by Shirts *et al.*, considers the free energy differences of all intermediate states simultaneously, as opposed to simply two states in the BAR approach.[122,167] Showing lower variance than either the EXP or BAR methods, MBAR is generally classed as one of the most effective estimators for alchemical free energy calculations.[122] A Python implementation of MBAR, provided by Chodera *et al*., was used in the present work: https://github.com/choderalab/pymbar.[162]

TI is a method which lies between EXP and BAR. Similar to EXP, it computes the ensemble average from a single state, neglecting the configurations sampled in the neighbouring states. Similar to BAR/MBAR, multiple intermediate λ states are required to calculate the free energy difference. However, contrary to the previous methods, the free energy difference between endpoints is given by the derivative of the potential energy with respect to the coupling parameter λ, as opposed to the potential energy differences. Thus, the free energy difference may be calculated using formula **59**:[122]

$$\Delta G_{0,1} = \int_{\lambda=0}^{\lambda=1} \frac{\partial U(\lambda, x)}{\partial \lambda} \lambda \qquad (59)$$

Both integration and perturbation methods require a thermodynamic cycle of alchemical intermediates connecting the physically meaningful end states (**Figure 25**). The free energy cycle employed in the present study features eight thermodynamic states, comprised of the bound and unbound states (**Figure 25**, A

and H respectively), and six alchemical intermediates (**Figure 25,** B-G). Similar to that proposed by Mobley *et al*., the cycle commences with calculation of the free energy change associated with addition of orientational restraints to the fully interacting protein-ligand complex ($\boldsymbol{\Delta G^{PL}_{restr}}$).[168] Restraints are applied to the ligand to restrict its accessible sampling volume and hence to aid convergence of the simulations. Thus, the fully interacting, restrained complex is transformed to a non-interacting protein and ligand by scaling the electrostatic (ΔG^{PL}_{elec}) and van der Waals (ΔG^{PL}_{LJ}) interactions to zero *via* independent simulation of two nonphysical states (**Figure 25**, B-C). This process of removing the intermolecular interactions between the ligand and its environment is termed *decoupling*. State 4, in which the non-interacting ligand is restrained to the protein binding site, is equivalent to state 5, in which the latter is restrained in solution. The free energy change (ΔG^{L}_{restr}) associated with removal of the restraints is calculated analytically (**Figure 25**, E⟶F). Finally, the unrestrained, fully decoupled ligand (**Figure 25**, F) is transformed into the fully interacting ligand by turning on the van der Waals (ΔG^{L}_{LJ}) and electrostatic interactions (ΔG^{L}_{elec}) *via* two alchemical intermediates (**Figure 25**, G-H), thereby closing the cycle. Hence, the total binding free energy (ΔG_{bind}) is recovered by the sum of all major steps in the cycle, including a correction for the 1M standard state (**60**).

$$\Delta G_{\mathrm{bind}} = \Delta G^{\mathrm{PL}}_{\mathrm{restr}} + \Delta G^{\mathrm{PL}}_{\mathrm{elec}} + \Delta G^{\mathrm{PL}}_{\mathrm{LJ}} + \Delta G^{L}_{\mathrm{restr}} + \Delta G^{L}_{\mathrm{LJ}} + \Delta G^{L}_{\mathrm{elec}} \qquad (60)$$

Furthermore, in the thermodynamic cycle highlighted, the use of restraints is critical to preventing the so-called “wandering ligand” problem.[122,163,168] This describes the propensity for the unrestrained, fully decoupled ligand to drift away from the protein binding site into the large volume of the simulation box. This

leads to convergence issues owing to the large configurational phase space accessible to the ligand, which must find a physically meaningful path back to the protein cavity.[122] However, the use of restraints requires a standard state correction, as the volume no longer represents the volume of the whole simulation box.[122,169]

The restraint scheme developed by Boresch is most widely employed in alchemical free energy calculations, as it restricts the ligand's orientation with respect to the protein and provides an analytical solution to ΔG^{L}_{restr}.[170] Harmonic restraints, consisting of one distance, two angle and three dihedrals, restrict the movement of three chosen atoms of the protein with respect to three chosen atoms of the ligand. The analytical solution not only reduces the required number of simulations, but also includes the standard state correction (**61**).[122,168,170]

$$\Delta G^{L}_{\mathrm{restr}} - k_B T \ln \frac{V}{V^{\circ}} = k_B T \ln[\frac{8\pi^2 V^{\circ}}{r^2_{aA} \sin\theta_a \sin\theta_A} \frac{(k_{r_{aA}} k_{\theta_a} k_{\theta_A} k_{\phi_{ba}} k_{\phi_{aA}} k_{\phi_{AB}})^{\frac{1}{2}}}{(2\pi k_B T)^2}] \quad (61)$$

Conversely, $\boldsymbol{\Delta G^{PL}_{restr}}$ is calculated numerically *via* MD simulations, in which the force constants for the six harmonic restraints are scaled with the parameter λ_{restr} between their specified value and zero.[122]

Improving convergence in alchemical free energy calculations may be achieved not only with restraints but with enhanced sampling methods, such as HREMD. Woods *et al*. have combined TREMD with TI calculations, as well as HREMD with FEP (and TI) calculations effectively to improve sampling.[171,172] In these schemes, each λ state represents a replica in HREMD, involving periodic exchanges between λ states. Meng and colleagues demonstrated that HREMD is intrinsically a FEP algorithm.[160] This was evident through comparison of the Zwanzig relationship

(**57**) with the mathematical expression for the transition probability in HREMD (**62**).[160] The latter governs exchange between replicas 1 and 2 with the corresponding potential energies U_1 and U_2.

$$w(q_1 \rightarrow q_2) =_{\min} [1, e^{-|\frac{(U_2(q_1)+U_1(q_1)-U_1(q_2)-U_2(q_2))}{k_B T}|} \quad (62)$$

Thus, computing the transition probability in HREMD yields an FEP calculation, through calculation of the ensemble averages of the potential energy differences from the configurations of both replica simulations 1 and 2.[160] In effect, FEP is equivalent to a HREMD calculation without exchange attempts between λ replicas.[160] In the work presented, the HREMD/FEP method was employed to aid convergence and enhance sampling in all alchemical free energy calculations.

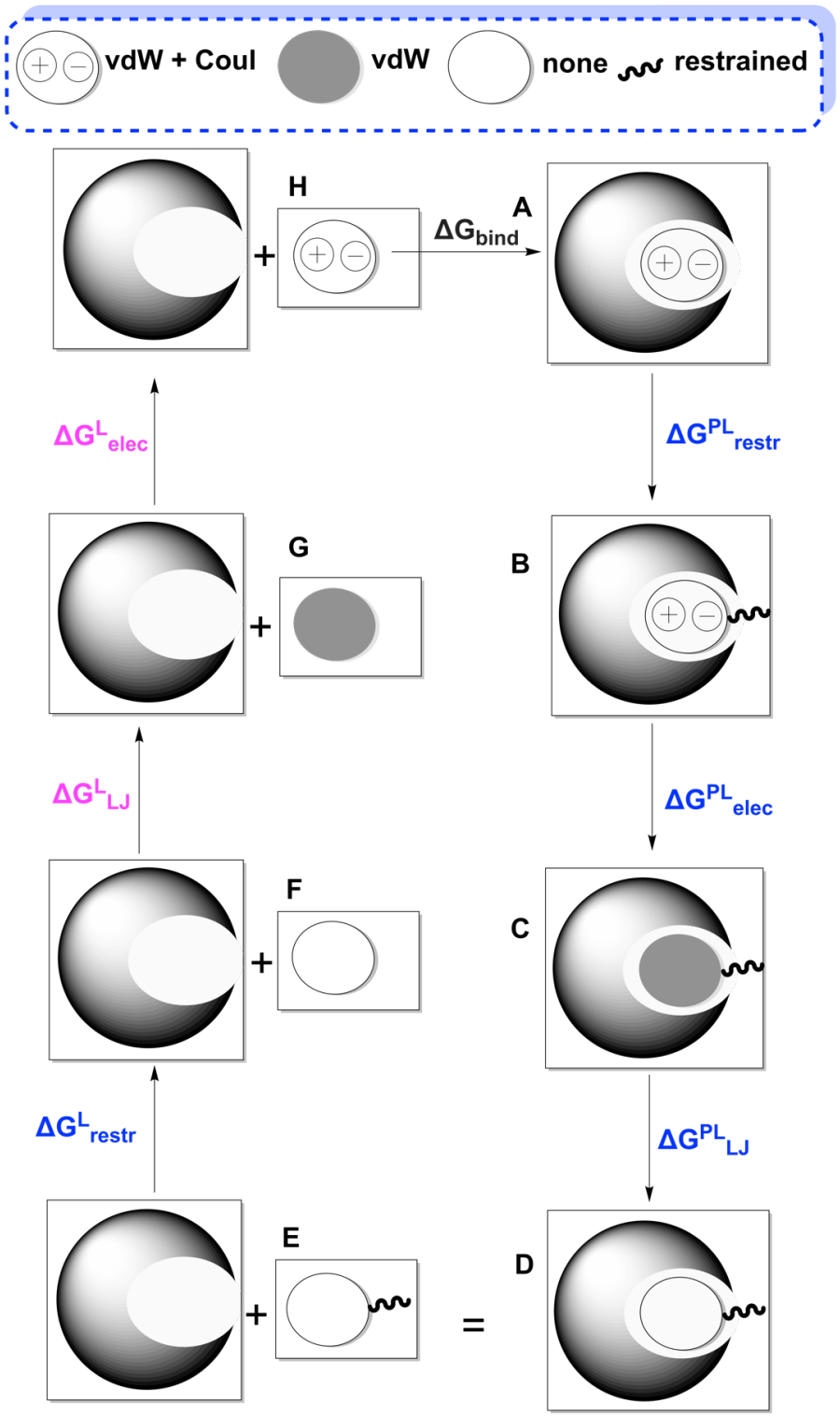


Figure 25: The thermodynamic cycle for alchemical binding free energy calculations, divided into complex (blue) simulations and ligand (violet) simulations. Starting from the fully coupled UPPS-TA complex (**1**), the Boresch-type distance and orientational restraints are applied (**2**, ΔG^{PL}_{restr}), followed by the consecutive decoupling of coulombic (**3,** ΔG^{PL}_{elec}) and van der Waals (**4,** ΔG^{PL}_{LJ}) interactions. The free energy change associated with removal of the Boresch-type restraints from the fully decoupled complex is calculated analytically (**6,** ΔG^{L}_{restr}). The restrained and fully decoupled UPPS-TA system (**4**) is equivalent to the fully decoupled and restrained TA in water (**5**). Subsequently, in the separate set of TA-only simulations (violet), the TA is gradually re-coupled to its van der Waals interactions in water (**7,** ΔG^{L}_{LJ}), followed by re-coupling of the coulombic interactions in water (**8,** ΔG^{L}_{elec}), thereby closing the cycle. Adapted from Mobley *et al.* (2006).[168]

2.6. Binding pocket analysis

Pocket volume is a widely used descriptor of the overall shape and size of enzyme active sites. It can be employed to characterise differences in the conformational dynamics and size of the active site between a wild-type and a mutant enzyme, as well as between the liganded and unliganded enzyme states.[173–175] Additionally, it can be used to monitor active site changes during the course of a catalytic reaction.[175] Examination of active site pocket dynamics constitutes a critical part of many ligand-protein binding studies, including more recent investigations led by Lopez *et al*., Gonzalez *et al*., and Creutznacher et *al*.[173,176–178]

Chen *et al*. highlighted differences in the evolution of the active site volume between the wild-type and the Y68A mutant of catechol O-methyltransferase during catalysis (i.e., transfer of a methyl group from S-adenosylmethionine to a catechol).[175] Similarly, Mariutti *et al*. demonstrated how a change in cavity shape between the wild-type and P115Q mutant of the arginine repressor caused a change in its specificity from Y to R.[173] This study presents one of the limitations of employing pocket volume as a descriptor of active site dynamics: PVA is restricted to providing descriptions of the global shape of the active site but fails to distinguish local conformational changes.

Moreover, Potshangbam *et al.* demonstrated that variations in pocket size volume between mutants of dihydropteroate synthase (DHPS) could be correlated with the docking and free energy of binding (i.e., MMGBSA) results of potential DHPS inhibitors.[174] This was of particular interest to the present study as it examined the effect of active site shape and size on inhibitor binding efficacy.[174]

Binding site analysis is a growing adjunct to *computer aided drug design* (CADD), concerned primarily with the detection, dynamical characterisation, and druggability prediction of binding sites (i.e., *pockets*).[179,180] The inherent flexibility of pockets may be investigated through comparison of the different conformations provided by X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and MD simulations.[181] The latter supplies many thousands of snapshots of the pocket throughout thermal motion, and the collective may be filtered for pharmacologically relevant binding conformations. Hence, a smaller subset of MD clusters, representative of the full conformational diversity of the pocket, may be selected for ensemble docking of ligands.[182–185] Transient pockets may be identified, and different geometries of the same pocket may be uniquely recognised by disparate structural classes of inhibitor.[182,186] Owing to the utility of this approach in CADD, a multitude of pocket-studying algorithms have been developed, including fpocket[187,188], TRAPP[189,190], PocketAnalyzer[191], trj-cavity[192], Epoch[193] and Volsite[194].

Moreover, there is an important distinction between pocket detection and pocket analysis, each of which exploit different types of algorithm.[181] Pocket detection is concerned with the comparison of rotation-invariant representations of a query pocket and a set of known druggable pockets.[181] Conversely, pocket analysis does not use rotation-invariant representations, and is specifically adapted to analyse many thousands of aligned frames for finer structural differences (e.g., in geometry or volume).[181] Additionally, different pocket-studying programs use different mathematical representations for the binding pocket, the most common of which are the *grid*/*voxel-based* and *alpha sphere-based* representations.[181]

POcket Volume MEasurer (POVME) is an example of a widely used, grid-based pocket analysis tool.[181,182,195]

The POVME algorithm performs four basic tasks on a pre-aligned MD trajectory.[182] Alignment of the trajectory is necessary as the program assumes that there is no rotational and translational motion. First, a user-specified *inclusion region*, encompassing the binding site, is specified from a pre-aligned trajectory. Additionally, subregions of the binding site may be excluded if they fall outside a user-defined region denoted the *convex hull*.[182] The convex hull of the receptor atoms bordering the inclusion region is specified by a gift-wrapping algorithm.[182,196] Secondly, the inclusion region is filled with equispaced points (i.e., 1.0 Å apart by default), creating a volume-grid file. Thirdly, deletion of any volume grid point which is sufficiently close to an atom (i.e., the van der Waal's radius plus a default tolerance of 1.09 Å) is performed.[182] Collections of isolated or discontiguous points are also removed. Finally, the volume of the binding site is computed based on the remaining grid points, each of which is equispaced and hence represents a discrete volume.

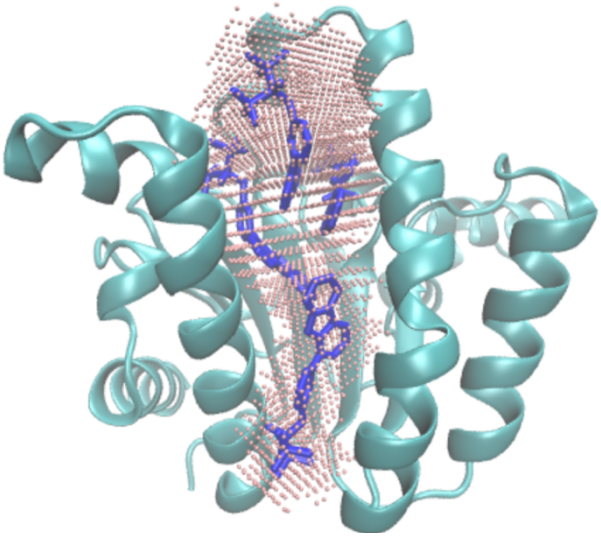


Figure 26: The inclusion region encompassing sites 1-4 (as defined by four copies of the bisphosphonate) of the hydrophobic tunnel of *E. coli* UPPS (PDB ID: **2E98**).

Chapter 3. Apo-UPPS as the antimicrobial target

3.1 Previous literature work

H43, an active site residue in *E. coli* UPPS, is believed to be a proton donor to the departing pyrophosphate of FPP in the catalytic mechanism (**Figure 7**).[44,186] Computational investigations of *E. coli* UPPS led by McCammon *et al.* established that differing H43 protonation states yielded differences in global flexibility and the conformations of important active site residues (**Figure 27**).[186] The HIP43 system was found to exhibit greater root-mean-square fluctuation (RMSF). In addition, the flexible loop residue R77 was found to sample the FPP and IPP binding sites more extensively in the HID43 simulations.[186] R77, which was postulated to transfer the pyrophosphate between substrates FPP and IPP, was only found deep in the active site when H43 was in the HID state.[45,186] According to McCammon, R77 was therefore likely to exhibit increased catalytic efficacy in the HID state due to greater sampling of the active site.

Moreover, there is literature precedent for applying the pocket volume descriptor to the study of UPPS. Using MD, McCammon has demonstrated that the probability of the UPPS apoenzyme visiting closed conformational states is high, while that of visiting the open *expanded state* bound by the bisphosphonates (PDB ID: **2E98**) is almost negligible.[186] In MD simulations starting from the open state of *E. coli* UPPS (PDB ID: **2E98**), McCammon observed significant shrinkage of the active site pocket upon ligand removal (i.e. 1032 $Å^3$ compared to an average of 332 $Å^3$ throughout the MD simulations). Subsequent clustering of the apoenzyme trajectory indicated that the majority of structures possessed reduced volume pockets, except for the single instance of an expanded pocket conformation (939

Å^3).[186] Based on the finding that the crystallographic bisphosphonate-bound conformation (PDB ID: **2E98**) was rarely populated in the apoenzyme's trajectory, McCammon proposed that induced fit changes were at least partially responsible for adoption of the bisphosphonate-bound UPPS conformation. Taking highly populated MD clusters, McCammon and colleagues identified a closed state conformation which, upon docking of TAs, reproduced the binding mode of FPP. Similarly, the expanded pocket conformation sampled once during the MD simulations was found to reproduce the binding mode of the bisphosphonates.[186] Thus, it was evident that the active site was highly mobile, allowing structurally divergent inhibitors to specifically bind to different conformational states of apo-UPPS, which were described by different active site pocket volumes. In the case of the bisphosphonates, the open state was sparsely populated in the MD trajectories of the apoenzyme and hence McCammon proposed that a population shift of the binding pocket size was in operation upon inhibitor binding.

Furthermore, PCA conducted by McCammon demonstrated that HID43-UPPS and HIP43-UPPS sampled different regions of conformational space with very little overlap.[186] However, McCammon's results were based on single HID43 and HIP43 MD simulations (each 85 ns of MD) without enhanced sampling, as opposed to multiple concatenated replicas. Here we examine the data generated from the simulated HID43 and HIP43 apo-systems (PDB ID: **3QAS**), each over 500 ns of MD.

3.2 Methods

3.2.1 Systems Preparation

A homodimer of subunit A of *E. coli* apo UPPS (PDB ID: **3QAS,** res. 1.70 Å) was generated in Chimera owing to missing residues 73-89 in subunit B.[197] Residues H43, N59, Q163, Q168, and Q179 were flipped according to Molprobity predictions.[198] Caps were added at the N (ACE) and C (NMe) termini in PyMOL due to missing terminal residues. All crystallographic waters were retained.

3.2.2 MD Simulations

All MD simulations were performed using the GROMACS package (v. 2019.4)[199] with the OPLSA forcefield.[144] Each system was centered in a rhombic dodecahedral box (1.0 nm buffer region), solvated (TIP3P water)[200], neutralised with 2 Cl ions, and minimised using the steepest descent algorithm until the maximum force was 886 kJ mol^{-1} nm^{-1}. From the minimised structure, five independent simulations were initiated by random velocities from a Maxwell distribution at 300 K. In each case, restrained equilibration was performed at fixed volume and temperature (NVT ensemble). Temperature annealing was carried out for 1 ns (1 fs timestep) such that the temperature was increased from 0 K to 300 K in 50 K increments every 200 ps. Peptide and ligand heavy atoms were position restrained by a small harmonic force (i.e., 1000 kJ mol^{-1} nm^{-2}). Subsequently, the replicate structures were simulated in the NPT ensemble with gradual release of the restraints (1000 kJ mol^{-1} nm^{-2}, 500 kJ mol^{-1} nm^{-2}, 250 kJ mol^{-1} nm^{-2}, 0 kJ mol^{-1} nm^{-2}) over 1 ns. For both steps of the equilibration, the protein and non-protein atoms were coupled to separate temperature baths using the Berendsen weak coupling method.[201] Following this, production MD was conducted in the NPT ensemble in the absence of restraints for 100 ns. For data collection, the V-rescale thermostat and the

Parrinello-Rahman barostat were employed to maintain the temperature at 300 K and to regulate isotropically the pressure at 1.0 bar respectively. All bonds involving hydrogens were constrained using the LINCS algorithm.[202]

3.3 Data analysis

3.3.1 Determining the protonation state of active site residue H43

Using PROPKA[203], the ionisation state of H43 in *E. coli* apo-UPPS (PDB ID: **3QAS**) was determined at 7.4 (**Table S1**). The stability of the system for the different protonation states of H43 (**Figure 27**) was established by monitoring the RMSD of H43 and its surrounding residues (i.e., within 3.5 Å of H43) during 500 ns of cumulative MD simulation per protonation setup. Conformational clustering was performed using the GROMOS clustering algorithm.[204] Following fitting of the trajectories to the alpha carbons of all UPPS subunit A residues (excepting the C and N-termini), clustering was performed with root-mean-square (RMS) differences of all subunit A residues. The conformations of H43 in the three highest cluster centroids (i.e., accounting for approximately 90% of all representative structures collectively) were inspected. The difference in the root-mean-square deviation (RMSD) between the HID and HIP trajectories was rationalised by the presence of distinct hydrogen bonding patterns to the imidazole side chain of the histidine for each protonation state. Furthermore, hydrogen bonds were identified, and distances were measured using the readHBmap python script and gmx distance module within GROMACS.

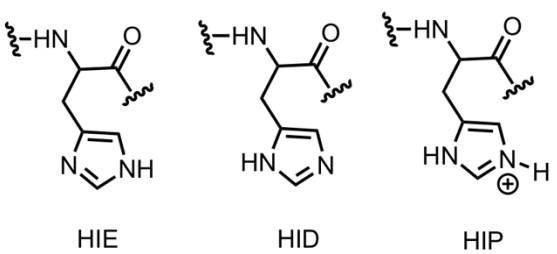


Figure 27: The three possible protonation states of histidine at pH 7.

3.3.2 Elucidating the effect of differing H43 protonation state on key active site residues

The mean RMSD of the flexible loop (i.e., residues 72-82 in *E. coli* UPPS) in the apo-HID43 and apo-HIP43 systems was calculated from each of the 500 ns concatenated MD trajectories. Conformational clustering allowed for visual inspection of the conformations of the key active site residues, R77 and R39.

3.3.3 Exploring the conformational flexibility of the hydrophobic tunnel

PVA was carried out on each of the concatenated (500 ns) apo-HID43 and apo-HIP43 simulations using POVME 3.0.[181] Frames were extracted from each of the five trajectories every 20 ps and aligned in Visual Molecular Dynamics (VMD) software[205] with the X-ray crystal structure of *E. coli* open UPPS (PDB ID: **2E98**, res. 1.90 Å). The volume of the hydrophobic tunnel (encompassing sites 1-4) was specified by the total volume occupied by the bisphosphonate molecules in each of the four sites (**Figure 26**).[181] This measured volume, known as the *inclusion region*, consisted of the grid points nearest each bisphosphonate atom, which were extended out in all directions by 3 Å. The inclusion region was flooded with equidistant points each spaced 1.0 Å apart. Thus, each point represented a region of space of 1 $Å^3$ in volume. Any grid points which were within 1.09 Å of the van der Waals surface of the receptor atoms were removed. The same positioning of grid points was used for all pre-aligned structures. In addition to calculating the mean hydrophobic tunnel volumes of the apo-HID43 and apo-HIP trajectories, the tunnel volumes of the crystal structures representing open (PDB ID: **2E98**), closed (PDB ID: **4H8E**, res. 1.30 Å) and apo (PDB ID: **3QAS**) UPPS were computed.

3.3.4 Investigating the effect of differing H43 protonation state on the conformational dynamics

PCA was carried out within GROMACS for both the MD trajectories of apo-UPPS and the UPPS-TA complexes. PCA extracted the dominant modes of motion of the protein or protein-ligand complex from the MD trajectories. First, the overall translational and rotational motions in the MD trajectory were eliminated by least-squares-fit superimposition onto the C-alpha skeleton of the minimised crystallographic input (PDB ID: **3QAS**), which served as the reference structure. Following this, a covariance matrix of the atomic fluctuations was generated and diagonalized to produce a set of eigenvectors and eigenvalues describing the direction and strength (i.e., energetic contribution) of each component of the collective motion. The PCs were the highest-amplitude collective motions. The HID and HIP sets of trajectories were concatenated separately, and each concatenated trajectory was projected onto the PCs 1 and 2, which accounted for the main modes of collective motion sampled by the MD trajectories. In the resulting two-dimensional projection, each point represented the protein conformation associated with an individual simulation frame. Hence, the distribution of points indicated the configurational subspace sampled along the PCs extracted from the MD trajectories. Further, the trajectories were filtered along selected eigenvectors to extract the extreme conformations indicating the type of collective motion. These extreme projections were used to generate porcupine plots, which showed the direction and strength of the PCs for the backbone atoms.

3.4 Discussion

3.4.1 Determining the protonation state of active site residue H43

In this section, the most probable protonation state of H43 is proposed based on the residue's mean RMSD in each of the HID43 and HIP43 simulations of *E. coli* apo-UPPS (PDB ID: **3QAS**).

The protonation state of key active site residues has an effect on the conformational dynamics of an enzyme's active site and the interactions profile of the inhibitor-enzyme complex. Of particular interest in the present study is the active site histidine (residue H43 in *E. coli* UPPS).

Since the pKa of the imidazole side chain of histidine approximates 7, it is equally probable that one or both of the nitrogen atoms, ND1 and NE2, is protonated at neutral pH. Hence, three protonation states of histidine are possible at pH 7, and originate from protonation of ND1, NE2, or both nitrogen atoms; these are HID, HIE and HIP (**Figure 27**).[206]

Although the electron density map of the crystallographic structure of the apoenzyme was well-defined (PDB ID: **3QAS; Figure S1**), indicating that the crystal was of good quality, the protonation state of the active site histidine was unknown. According to PROPKA predictions, the H43 residue existed in the HID form in the apoenzyme (PDB ID: **3QAS**, pKa = 5.5) at pH 7.4 (i.e., physiological conditions).[203,207] However, previous studies have associated incorrect protonation states with larger RMSD of the histidine residue and/or larger RMSD of surrounding residues relative to the correct protonation assignment.[206] The MD simulations of the HID and HIP systems established that H43, as well as the surrounding residues, exhibited the largest RMSD in the HID state (**Figure 28: D**).

Thus, based on the RMSD principle, the MD results suggested that H43 of apo-UPPS was most likely to exist in the protonated (i.e., HIP) form under physiological conditions, an observation which was inconsistent with the preliminary PROPKA pKa prediction.

Conformational clustering of the HID and HIP simulations of apo-UPPS revealed different conformations of H43 depending on its protonation state. It was observed that the imidazole side chain of H43 was oriented upwards and parallel to the α2 helix in the HIP simulations (**Figure 28: B**) but was positioned orthogonal to the α2 helix in the HID simulations (**Figure 28: A**). The former conformation of H43 was closest to that of the histidine residue in the starting crystal structure (PDB ID: **3QAS; Figure 28: B**, yellow sticks). This observation was reinforced by the mean RMSD of H43, which was lowest in the HIP simulations (**Figure 28: D**).

Furthermore, a single hydrogen bond was observed between the backbone nitrogen of H43 and the active site residue R39 in the highest populated MD cluster of the HID trajectory (**Figure 28: A**). No hydrogen bonds were formed between the imidazole moiety and the α2 helix, whose conformation differed significantly from the starting crystallographic conformation (PDB ID: **3QAS; Figure 28: A**).

By contrast, hydrogen bonds were formed by the amino and carbonyl groups of H43 with the carbonyl and amino groups of A40 and A47 respectively in the HIP trajectory (**Figure 28: B**). Hydrogen bonds to A40 involved not only the backbone nitrogen of H43 but also the protonated imidazole nitrogen ND1. This latter interaction held the histidine in an orientation which approximated its crystallographic starting conformation (PDB ID: **3QAS; Figure 28**, yellow sticks).

In conclusion, it was found that the H43 residue was most likely to adopt the HIP protonation state in the apoenzyme. This result was rationalized by the lower RMSD of the backbone histidine residue in the HIP state, due to partial anchoring of H43 parallel to the α2 helix by a hydrogen bond to A40.

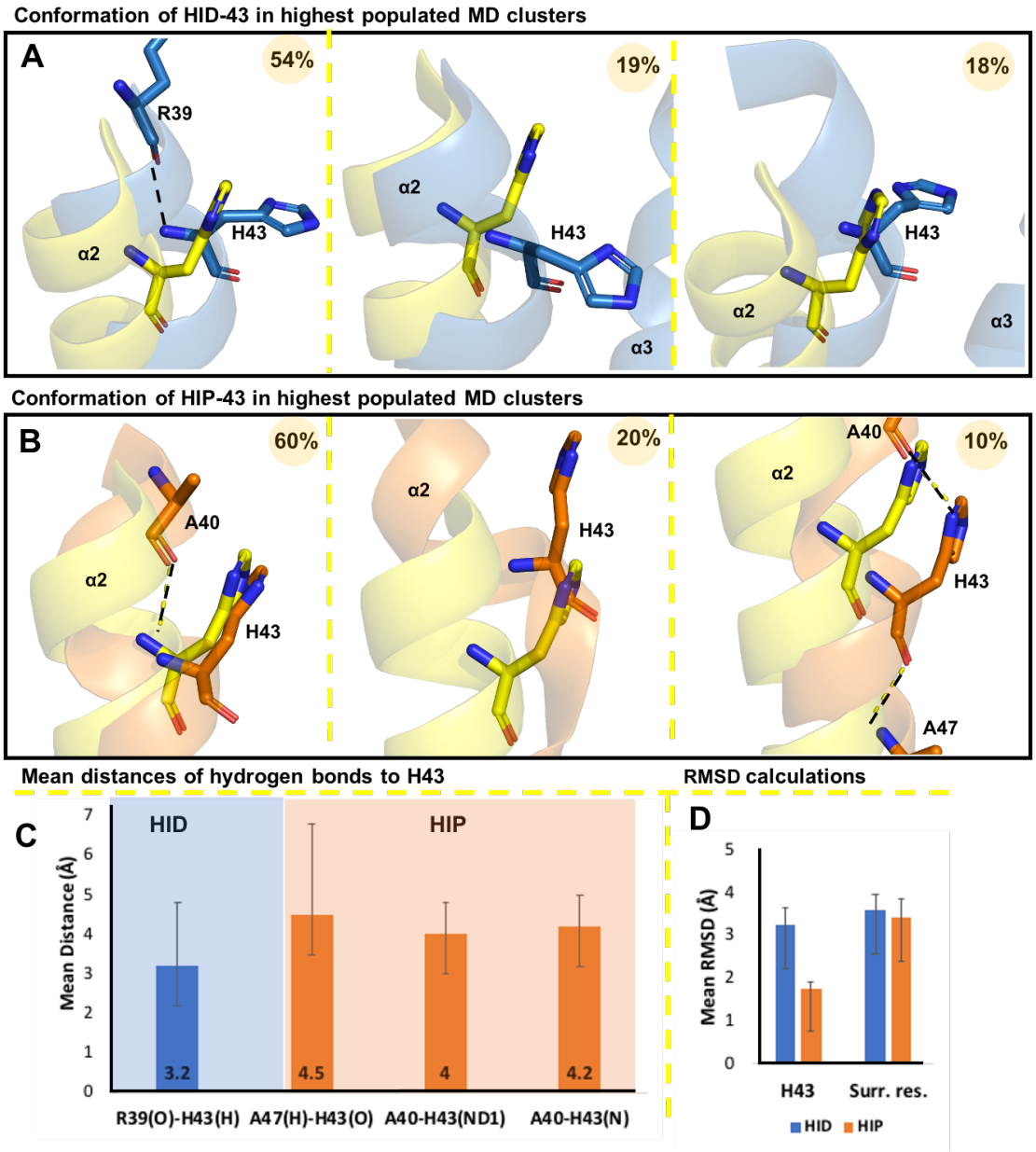


Figure 28: The first, second and third MD clusters of HID-43 (blue ribbon, **A**) and HIP-43 (orange ribbon, **B**) variants of apo-UPPS, showing the conformation of H43 relative to the starting crystal structure (PDB ID: **3QAS**, yellow sticks). Mean distances (including standard deviations) of hydrogen bonds to H43 are given, calculated from each of the 500 ns HID and HIP trajectories (**C**). Mean RMSD and standard deviations of H43 and the surrounding residues (i.e., within 3.5 Å of H43) are presented (**D**).

3.4.2 Elucidating the effect of differing H43 protonation state on key active site residues

In this section, differences in the conformational dynamics of the flexible loop and conserved active site residues between the apo-HID43 and apo-HIP43 systems are presented. Specifically, the effects of the H43 protonation state on the conformations of the residues R77 and R39 are outlined.

It was discovered that the RMSF of the HIP43 and HID43 systems were almost equivalent (**Figure S2**). Similarly, the mean RMSD of the flexible loop (i.e., residues 72-82 in *E. coli* UPPS) in each protonation setup was found to be equal (**Figure 29: B**). However, closer examination of the three highest populated MD clusters for each protonation setup revealed differences in the flexible loop's conformation. It was observed that the latter occupied the active site to a greater extent in the HIP state (**Figure 29: A**). In the third cluster, this movement was mediated by residues W75, N76 and R77, which were pulled inwards towards the active site and the α2 helix, when H43 was in the HIP state (**Figure 29: C**). Hence, according to these observations, the protonation state of H43 affected the conformations of specific flexible loop residues, but the global flexibility (i.e., RMSD) was equivalent.

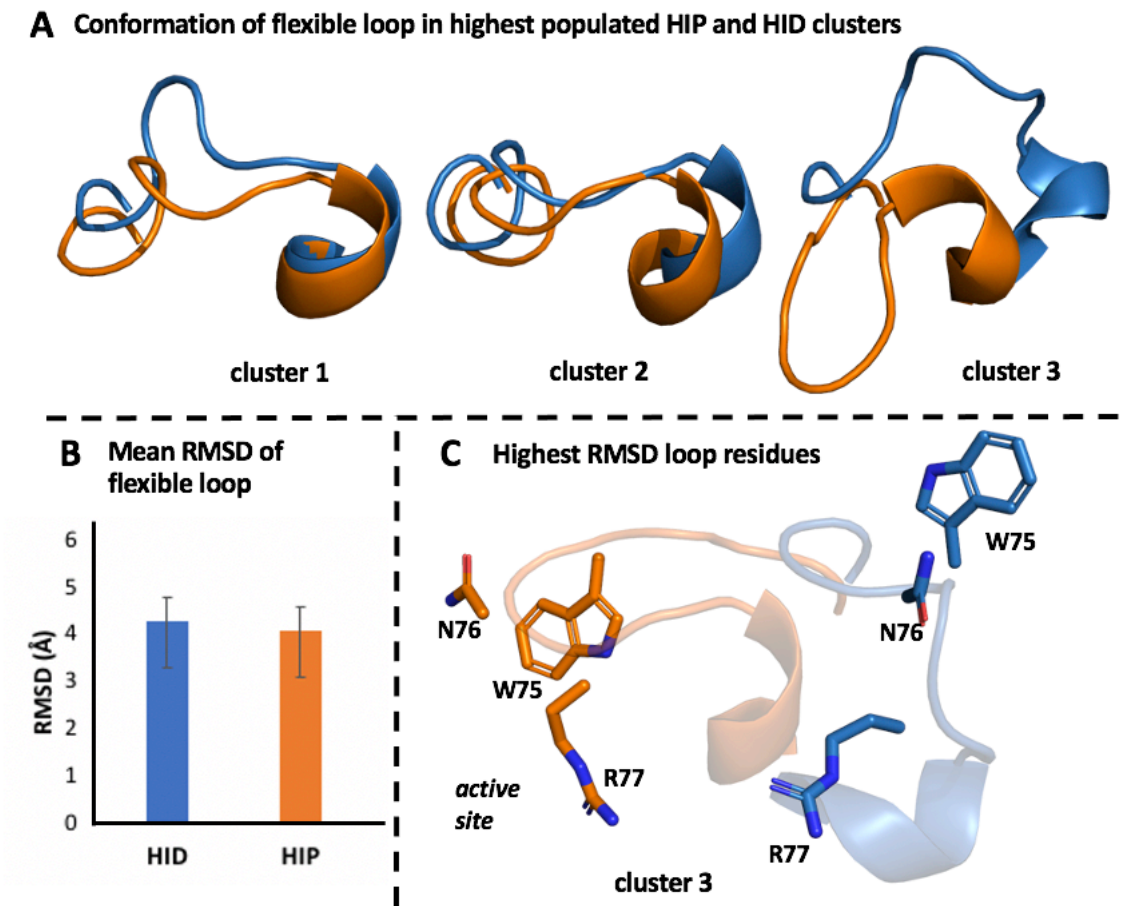


Figure 29: Superposition of the flexible loop (i.e., residues 72-82) from MD clusters 1-3 extracted from the 500 ns MD simulations for each of the HID (blue ribbon) and HIP (orange ribbon) protonation states (**A**). The mean RMSD, including the standard deviations, of the flexible loop across each 500 ns trajectory (**B**). Superposition of the third HIP and HID clusters, showing the conformations of highest RMSD residues, W75, N76, and R77 (**C**).

The effect of differing H43 protonation states on the catalytically significant flexible loop residue R77 was investigated.[186] According to conformational clustering, the conformations of R77 in the first and second clusters were similar (i.e., each directed towards the active site), but comparison of the third clusters revealed that R77 of the HIP43 simulation pointed deeper into the active site (**Figure 30: A**). Contrary to McCammon's observations, results of the present study suggest that R77 accessed the active site in both protonation state setups but showed a propensity for deeper occupation of the active site in the HIP state. Moreover, the residue exhibited an equal mean RMSD in both the HID and the HIP states (**Figure S3**), indicating that the H43 protonation state did not affect its overall flexibility.

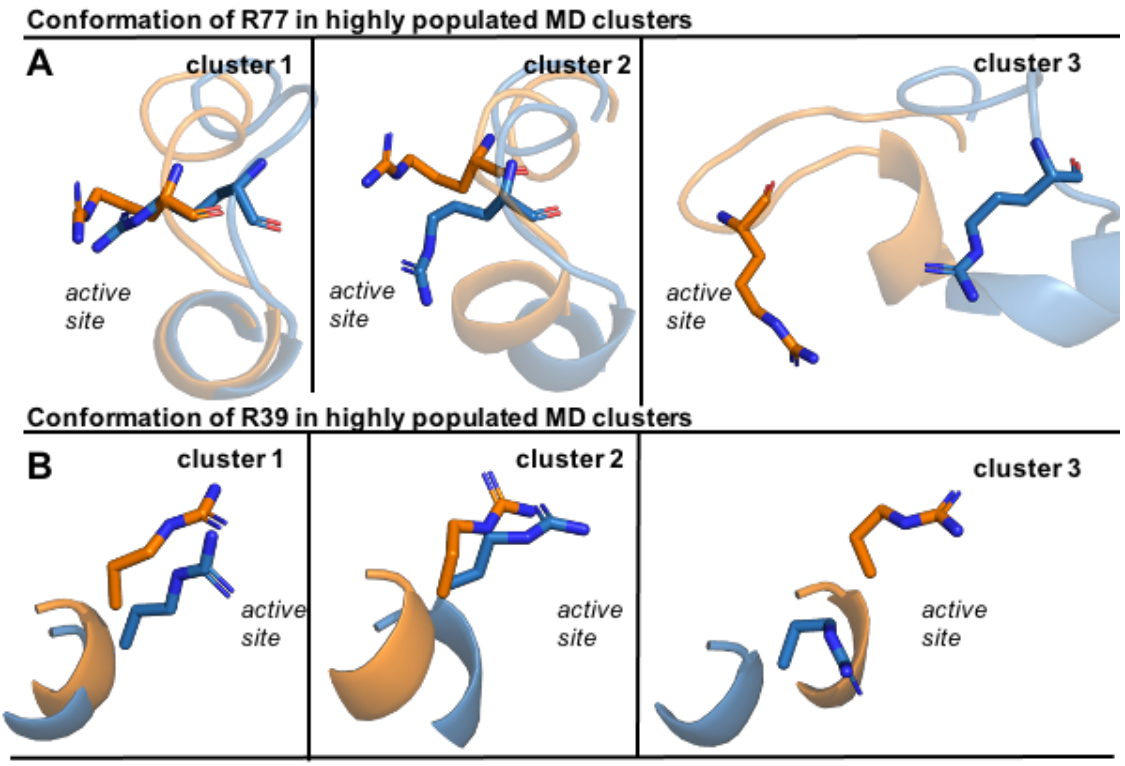


Figure 30: Superposition of the highest populated MD clusters of the HID (blue ribbon) and HIP (orange ribbon) trajectories, showing the conformations of R77 (**A**). Superposition of the highest populated MD clusters of the HID (blue ribbon) and HIP (orange ribbon) trajectories, showing the conformations of R39 (**B**).

Similarly, the effect of differing H43 protonation states on the conformation of key active site residue R39 was investigated. MD clustering revealed that the conformations of the residue R39, which has been postulated to facilitate FPP binding and charge neutralization, were equivalent in the first and second clusters (**Figure 30: B**).[41] However, comparison of the third highest populated clusters revealed that R39 exhibited deeper occupation of the active site in the apo-HIP43 trajectory (**Figure 30: B**), while showing a propensity to point away from the active site in the apo-HID43 trajectory. Similar to R77, R39 was characterised by an equal mean RMSD in both the HID43 and HIP43 simulations (**Figure S4**), which indicated that differing H43 protonation states affected the residue's conformation but not its relative magnitude of displacement. Such conformational differences have the potential to affect the interactions profile of any inhibitor-UPPS complex. Thus, it

was suggested that the HIP43 state, in which R39 had a slightly stronger propensity to point into the active site, may give rise to more favourable interactions.

Furthermore, the global flexibility (i.e., RMSD) of the HIP43 and HID43 apo-systems was unaffected by differing H43 protonation states (**Figure S3**). However, the conformational dynamics of the flexible loop and the active site residues R77 and R39 were affected. This in turn was likely to affect the interaction of the apoenzyme with a potential competitive inhibitor.

3.4.3 Exploring the conformational flexibility of the hydrophobic tunnel

In this section, the results of PVA reveal the conformational dynamics which characterise the hydrophobic tunnel of unliganded UPPS.

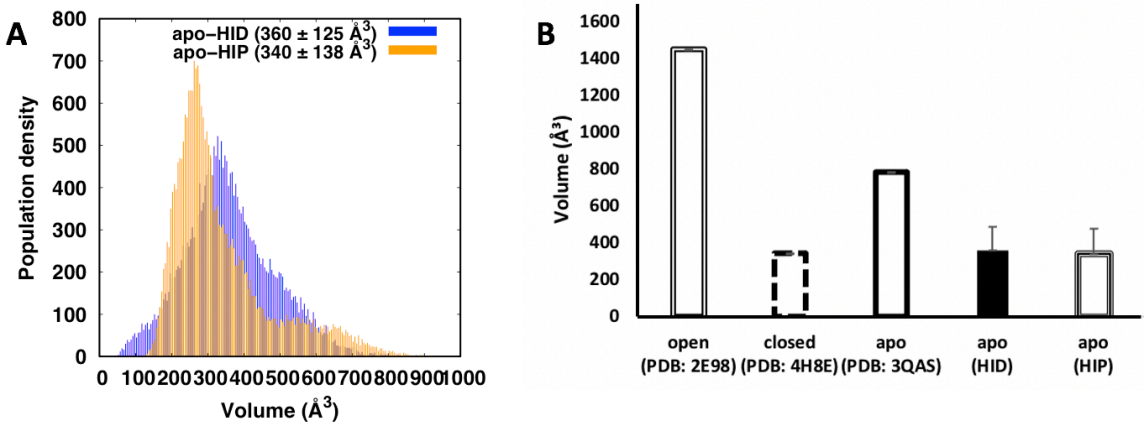


Figure 31: Histogram representation of the conformational states accessed by apo-HID and apo-HIP MD simulations of *E. coli* apo-UPPS (PDB ID: **3QAS**) (**A**). The mean volume and standard deviations of the hydrophobic tunnel are presented for the apo-HID and apo-HIP simulations, in addition to the fixed volumes of the X-ray crystal structures of UPPS in the open (PDB ID: **2E98**), closed (PDB ID: **4H8E**) and apo (PDB ID: **3QAS**) forms (**B**).

In the present study, pocket volume was utilised as an indicator of changes in the global shape of the hydrophobic tunnel (encompassing the binding cavity) upon complex formation to a TA-based inhibitor. Characterising the intrinsic tunnel flexibility of the apoenzyme was necessary before complexation-induced changes could be established. Thus, it was found that the mean volumes of the hydrophobic tunnel in the apo-HID43 and apo-HIP43 simulations were 360 Å^3 ±

125 Å^3 and 340 Å^3 $\pm$ 138 Å^3 respectively. Hence, according to this result, differing H43 protonation states exerted an almost negligible influence over the global pocket shape and flexibility. A noteworthy observation was that the mean tunnel volumes approximated the tunnel volume of closed *S. aureus* UPPS (PDB ID: **4H8E**) (**Figure 31: B**). This suggested that kinking of the α3 helix inwards, resulting in closure of the hydrophobic tunnel (and the active site) was intrinsic to the conformational dynamics of the apoenzyme. Moreover, the results of PVA indicated similarities in the global flexibility of the binding cavity but did not eliminate the possibility of local conformational differences (i.e., in the conformations of key active site residues), such as were observed in the varying conformations of R77 and R39.

In addition, it was found that the pocket size of the largest volume conformation (965 Å^3) of all HID43 and HIP43 structures was significantly smaller than that of the bisphosphonate-bound *E. coli* open-UPPS (PDB ID: **2E98,** 1448 Å^3) (**Figure 31: B**). This was significant as it suggested that the bisphosphonate-populated open form of UPPS involved an induced conformational change.

In conclusion, the PVA results confirmed McCammon's observation that the apoenzyme had an intrinsic propensity to adopt closed state conformations.[186] In the absence of a ligand, it was demonstrated that closure of the hydrophobic tunnel (encompassing the active site) was intrinsic to the enzyme's conformational dynamics. An important implication of this observation was recognition of the *expanded pocket* observed in the bisphosphonate-bound crystal structure (PDB ID: **2E98**) as a probable induced-fit conformational change of the active site.

3.4.4 Investigating the effect of differing H43 protonation state on the conformational dynamics

In this section, differences in the global conformational dynamics of the HID43 and HIP43 systems are explored using PCA. The results are presented through two-dimensional projections of the trajectories onto the first (PC1) and second (PC2) eigenvectors, as well as porcupine plots revealing the correlated motion associated with PC1 and PC2.

PCA analysis revealed significant overlap between the HID43 and HIP43 simulations (**Figure 32: A**). This inconsistency could be rationalised by a greater number of simulations per protonation state in the present study. In addition, the HID43 trajectory was found to sample large regions of conformational space which were not accessed by the HIP43 simulation, and vice versa. This indicated that the differing H43 protonation state gave rise to differences in the apoenzyme's main modes of correlated motion. Thus, this sampling of conformationally distinct subspaces validated McCammon's observation that differing H43 protonation states affected the enzyme's global dynamics.[186]

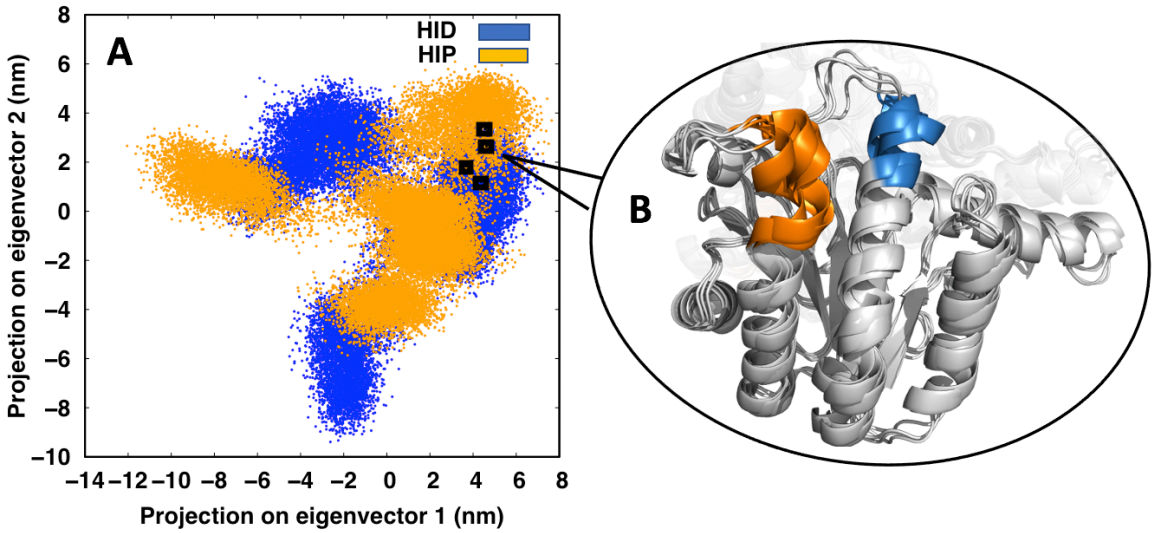


Figure 32: PCA calculated from the concatenated HID (blue) and HIP (orange) trajectories of apo-UPPS (**A**). Selected closed structures from the HID simulations with volumes of 319 Å^3 are shown (**B**).

Additionally, the two-dimensional projection plot indicated that closed states of UPPS, characterised by equal volumes of the hydrophobic tunnel, exhibited different conformational dynamics. Four conformations of HID-UPPS with the same volume (i.e., 319 $Å^3$) were found to sample different regions of the conformational space by PCA (**Figure 32: B**). This highlighted the limitations of the pocket volume approach. The latter measured the overall shape of the hydrophobic tunnel but failed to differentiate between structures of equal volume, which possessed subtle conformational differences in key structural elements such as the α-helices 2 and 3 (**Figure 32: B**).

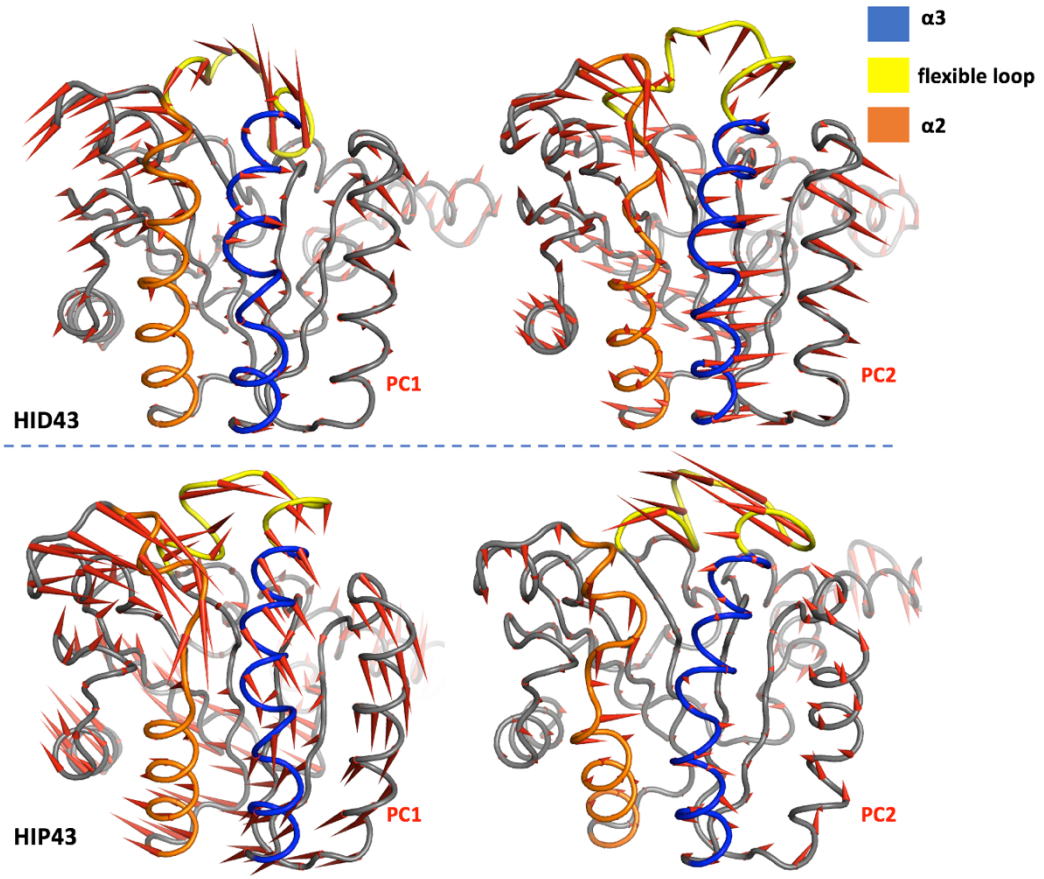


Figure 33: Porcupine plots of subunit A of apo-UPPS in the HID43 and HIP43 states, generated from the extreme projections of PC1 and PC2. The α helices 2 and 3, as well as the flexible loop (residues 72-82), are coloured orange, blue and yellow respectively.

Moreover, porcupine plots generated from the extreme projections of PC1 and PC2 in the HID43 and HIP43 trajectories revealed that the most dynamic structural elements were the helices α1-4 and the flexible loop (**Figure 33**). Within each plot,

the direction of the arrows indicated the direction of motion and the length of the arrows indicated the magnitude of the motion (**Figure 33**). Thus, the predominant collective motion identified from these projections was the opening and closing of the hydrophobic tunnel via movement of the α2 helix, the α3 helix and the flexible loop, a movement which McCammon designated the *breathing motion* (**Figure 33**).[186] Consistent with nuanced differences in the conformational dynamics, the specific correlated motions associated with PC1 and PC2 of the concatenated HID43 and HIP43 trajectories were found to be subtly different.[186] The HIP43 simulations were characterised by more pronounced movements of the flexible loop into and out of the active site, as well as accentuated motion of the α2 helix inwards towards the active site (**Figure 33**).

In conclusion, PCA revealed that differing H43 protonation states affected the conformational dynamics of the enzyme as a whole. In addition, it was concluded that UPPS structures of equal pocket volume were characterised by subtle conformational differences of the α-helices (**Figure 32**).

Based on all the above results, in which the techniques of conformational clustering, RMSD measurement, PCA, and PVA were employed, four key characteristics of the apoenzyme were revealed. Firstly, it was shown that the active site histidine (H43) was most likely to adopt the HIP protonation state. In addition, it was discovered that the conformational dynamics of the flexible loop and the catalytically active residues R77 and R39 were influenced by differing H43 protonation states. This was likely to affect the interactions profile of any potential (competitive) inhibitor-UPPS complex. The third observation was that the apoenzyme had an intrinsic propensity to adopt closed state conformations,

characterised by closure of the hydrophobic tunnel. Lastly, differing H43 protonation states affected the conformational dynamics of the most mobile structural elements, the α helices. Apoenzyme structures of equivalent volume were found to possess subtle differences in the conformations of $\alpha 2$ and $\alpha 3$ helices.

Chapter 4. TAs as UPPS Inhibitors

4.1. Binding hypotheses

In the absence of a UPPS-TA X-ray co-crystal structure, three potential binding sites of the TAs in UPPS have been investigated and are described in this section (**Figure 34**). Sites 1 and 2 are models of competitive inhibition based on the binding mode of diketoacid-based UPPS inhibitors[49] and McCammon's previous computational work[186] respectively, while site 3 is a model of allosteric inhibition based on the physical experiments of Lee *et al*.[67]

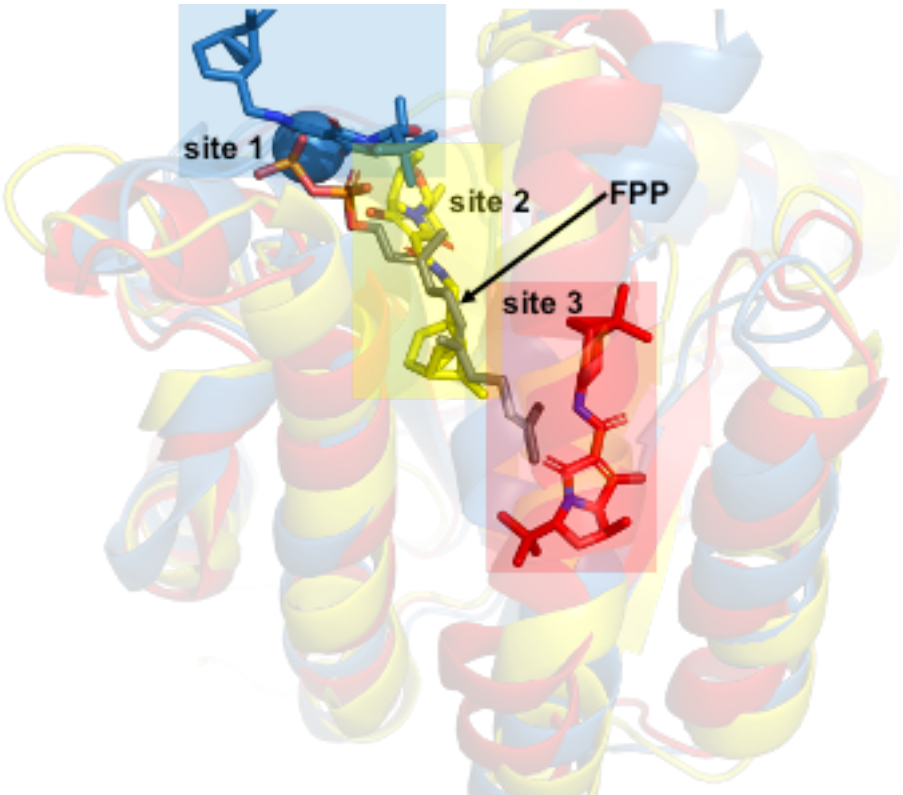


Figure 34: Proposed binding sites 1-3 in the UPPS enzyme, superimposed on the crystallographic binding mode of the native ligand, FPP (PDB ID: **4H8E**).

Site 1

Our first binding model, for which there is no literature precedent, postulates that the TAs bind to UPPS as their magnesium chelates, since they are known to be strong metal chelators.[49]

In this model of TA-UPPS complexation, the position of the active site Mg^{2+} cofactor is reproduced from the X-ray crystal structure of a Mg-bound UPPS-FsPP complex (PDB ID: **1X06**; **Figure 34**: site 1).[41,68,73,75,77,208] In this crystal structure, the Mg^{2+} is coordinated to the pyrophosphate head group of FsPP. Thus, it was conceived that the carbonyl oxygen of the exocyclic 3-acyl group and the ring 4-carbonyl could coordinate with the metal in an analogous manner. Further rationale was drawn from the binding mode of the HIV-1 IN-inspired diketo inhibitors of UPPS, in which the diketoacid headgroup occupied the place of the pyrophosphate of FsPP in the active site.[49] Moreover, metal chelation under physiological conditions was deemed likely, considering the known acidity of the TAs (pKa 2-5), whose equilibrating tautomeric enol and deprotonated enolate forms have been reported to be strong divalent metal chelators.[68]

Site 2

Our second binding model, inspired by McCammon's computational study[186], proposes that the TAs are competitive inhibitors of UPPS, i.e., they occupy the FPP-binding site (**Figure 34**: site 2) as free acids.[57,186] This hypothesis has not been proved experimentally, but is a prevalent assumption based on the original design pharmacophore, which mimics the native ligand FPP.[56] In addition, multiple conserved active site residues in the catalytic site may account for the broad spectrum inhibitory effect of the TAs. At least fifteen residues of the catalytic site have been found to be conserved in UPPS of a range of bacterial species (**Figure S5**).

Site 3

Based on the work of Lee *et al*, we hypothesised that the free acid TAs bind to an allosteric site adjacent to the active site (**Figure 34**: site 3).[67] This work comprised a single photo-cross-linking experiment, in which a TA probe was found to bind to Met49, adjacent to the FPP-binding site in *S. pneumoniae* UPPS. No further biophysical evidence for the allosteric mode of inhibition of TAs in UPPS has yet been published. Moreover, it has been found that critical residues of the proposed allosteric site in *S. pneumoniae* UPPS are not conserved in UPPS of other key bacterial species (**Figure S5**). According to this observation, one might predict sole inhibition of *S. pnuemoniae* UPPS by the TAs (assuming a strong dependence on amino acid sequence), but in fact the TAs exhibit *in vitro* activity against a range of gram-positive bacteria.

The remainder of this chapter will explore the keto-enol tautomerism of the TAs and their most probable mode of inhibition in UPPS, using DFT measurements and MD-based free energy calculations respectively. Additionally, the conformational dynamics of the proposed UPPS-TA complexes will be characterised, using PVA and PCA.

4.2. Methods

4.2.1. QM calculations

QM calculations were performed using the Gaussian 16 software.[209] Molecular geometries for each of the structurally simplified compounds **1**-**6** (**Figure 37**; **Figure S6**) were optimised at the PBE0-D3BJ[210]/6-311++G(d,p) level of theory[210–212] , with water described using the and the implicit SMD model[213]. This methodology was chosen as it yielded results which were the most consistent with the NMR interpretation.

4.2.2. Ligand parametrisation

TA ligands were parametrised using the OPLSA forcefield. Geometries were optimised at the SMD-PBE0/6-311++G(d,p) level of theory within Gaussian 16. Using these geometries, partial charges of the ligand were obtained using the Restrained Electrostatic Potential (RESP) fitting procedure. The restrained ESP fit was performed in AmberTools19 of Antechamber. GROMACS topologies were generated from Amber ones using AnteChamber PYthon Parser interfacE (ACPYPE; v.0.1.1).[214]

4.2.3. System preparation

site 1 complexes

Owing to missing residues 73-89 in chain B of *E. coli* apo-UPPS (PDB ID: **3QAS**, res.: 1.7 Å), a homodimer of chain A was generated in Chimera. Residues H43, N59, Q163, Q168, and Q179 were flipped, employing Molprobity predictions. Caps were added at the N (ACE) and C (NMe) termini in PyMOL owing to missing residues. Crystallographic waters were preserved.

Following a similar approach to McCammon *et al.*, each of the geometry-optimised, deprotonated TAs **1A**, **2A** and **3A** were docked into the catalytic site of

the four most highly populated cluster centroids of the *E. coli* apo-HID UPPS trajectory (PDB ID: **3QAS,** res. 1.70 Å) using the ASP scoring function in GOLD.[140,186] Prior to the docking, a magnesium ion was included in the active site of the selected cluster centroids, based on the location of the metal in the UPPS-IPP-FsPP-Mg^{2+} complex (PDB ID: **1X06, res. 1.90** Å).

Visual inspection of the highest ranked poses of TAs **1A**, **2A** and **3A** in each of the four clusters revealed that cluster 3 placed the TAs closest to FsPP in the catalytic site (**Figure 35**). Hence, this cluster conformation was selected on the basis that it best reproduced the binding mode of the native substrate.

Furthermore, the resulting complexes were manually aligned with Mg^{2+} chelated FsPP (PDB ID: **1X06**), such that the charged oxygen at position-4 and the oxygen of the pendant carbonyl at position-3 of the tetramate were overlapping with the pyrophosphate headgroup of FsPP (**Figure S7**). This alignment placed the pendant 3-acyl or 3-carboxamide group in each of two possible orientations **(Figure S7**). Thus, six inhibitor-UPPS structures were generated for subsequent MD-LIE calculations in which UPPS subunit A contained a single inhibitor chelated to Mg^{2+}.

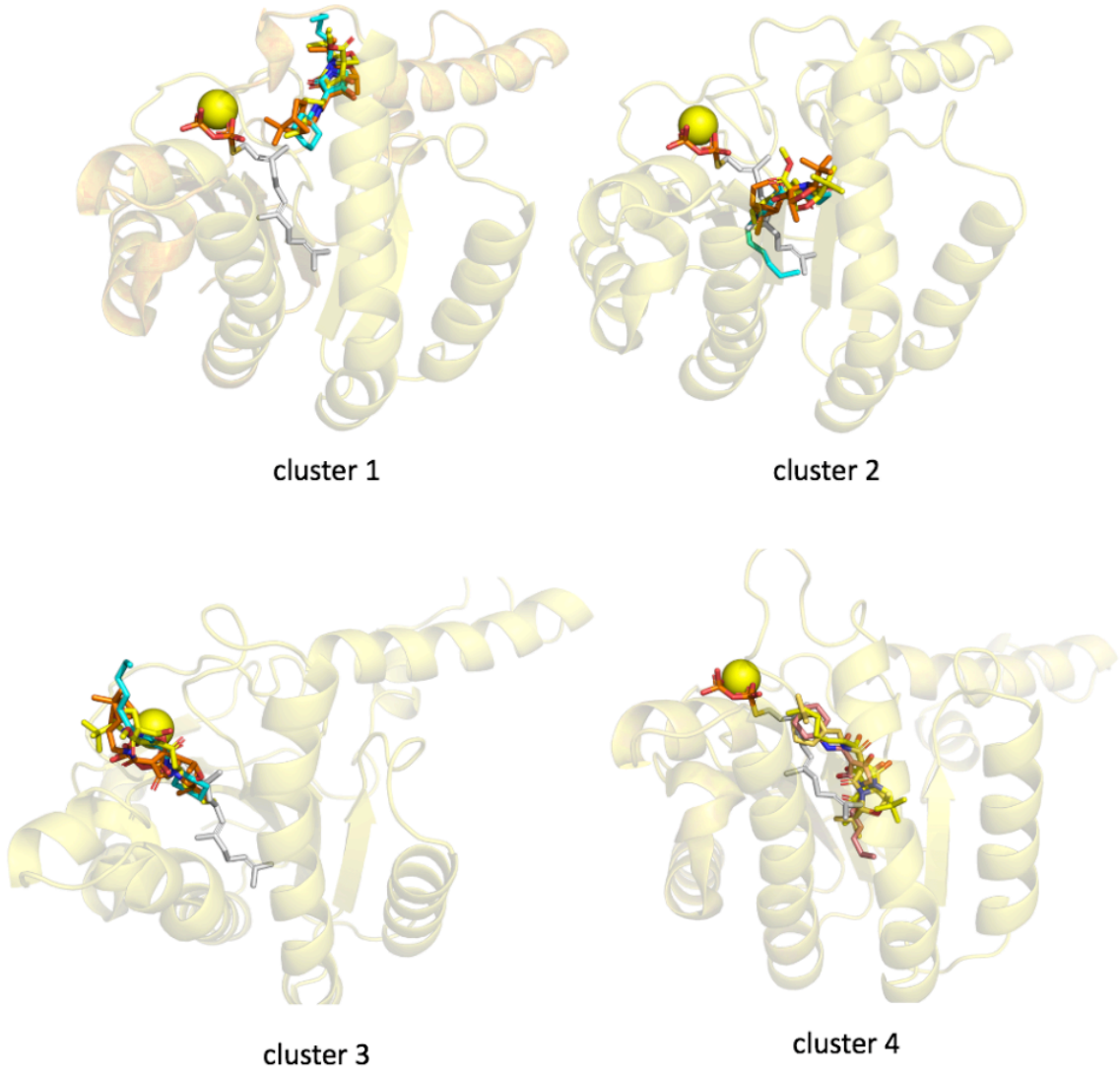


Figure 35: Highest ranked docking poses of the TAs **1-3** in clusters 1-4, extracted from MD simulations of apo-HID-UPPS (PDB ID: **3QAS**).

site 2 complexes

A homodimer of subunit A of *S. aureus* HID-UPPS in the closed form (PDB ID: **4H8E**) was generated in PyMOL using the crystallographic symmetry operations contained in the original PDB file. Residues N35, N90, N98, H187 and Q244 were flipped according to Molprobity predictions.[198] Missing atoms in D121 and N175 were modelled in Chimera. Caps were added at the N (ACE) and C (NMe) termini in PyMOL owing to missing terminal residues. Crystallographic waters were preserved.

The geometry optimised free acid TAs (**1A**, **2A**, **3D**, **5aD** and **5bD**) were docked into the catalytic site occupied by FPP employing the ASP scoring function in GOLD. The resulting poses were clustered within 2.5 Å RMSD. The best-ranked poses of each cluster were employed as starting protein-ligand configurations for MD-LIE calculations, yielding 9 poses for **1A** (**Figure S9**), 9 poses for **2A** (**Figure S10**), 9 poses for **3D** (**Figure S11**), 1 pose for **5aD** and 2 poses for **5bD.**[199] Poses for the TAs **1A**, **2A** and **3D** were simulated in triplicate, while poses for the compounds **5aD** and **5bD** were simulated in sextriplicate, employing different starting velocities in GROMACS. All MD simulations were carried out using GROMACS version 2019.4, unless otherwise stated.

site 3 complexes

Closed *S. pneumoniae* UPPS in complex with FPP and a methylisoxazole-based allosteric inhibitor **UK-106051** (PDB ID: **4Q9M,** res. 2.06 Å) was selected as the protein template.[54] The N and C termini were capped with ACE and NMe fragments respectively due to missing terminal residues. Residues N30, Q39, Q53, N136 and Q236 of subunit A, and residues N30, H45, Q53, N97, H103 and N136 of subunit B, were flipped in accordance with Molprobity predictions. Crystallographic waters were retained.

Docking of the geometry-optimised free acids **1A**, **2A** and **3D** was performed into the allosteric region encompassing the amino acid sequence (residues 46-56: **AGMEALQTVTK**) identified by Lee *et al.* using ASP in GOLD (**Figure S8**).[140] The highest ranked single pose for each TA was subjected to TREMD (over a narrow temperature range of 310-316 K) within GROMACS, followed by conformational

clustering.[199] The highest populated cluster of each TA-UPPS complex was used as the input to subsequent MD-LIE calculations (**Figure S12**).

4.2.4. MD simulations

All complexes were simulated according to the procedure described in **Section 3.2.2**. Each MD simulation was performed in duplicate: one simulation where the active site histidine was singly protonated (HID) and one where the latter was doubly protonated (HIP). In addition, the TAs **1A**, **2A**, **3A**, **3D**, **5aD** and **5bD** were simulated independently in explicit solvent for 50 ns.

4.2.5. Estimating ΔG_{bind} using LIE

The free energies of binding (ΔG_{bind}) were determined using the LIE method, as developed by Åqvist.[127,149] The difference of the electrostatic interaction energies (ΔV^{Ele}) and the van der Waals counterpart (ΔV^{VdW}) were extracted from the final thirty percent of the trajectories of the TA-UPPS complexes and the solvated TAs (**Tables S2-S9**). Each ΔG_{bind} was calculated as the difference between the sum of the interaction energies of the complex and that of the solvated TA.

For site 2 complexes, where multiple poses were simulated in triplicate or sextuplicate, the relative contributions (W_i) of each pose (and its replicas) to the total calculated ΔG_{bind} of the complex were quantified via Boltzmann-weighting (**63**).

$$W_i = \frac{e^{\frac{-\Delta G_i}{k_B T}}}{\sum_i e^{\frac{-\Delta G_i}{k_B T}}} \tag{63}$$

$$\Delta G_{\text{bind}} = \alpha \sum_i w_i \Delta V_i^{\text{VdW}} + \beta \sum_i w_i \Delta V_i^{\text{Ele}} \tag{64}$$

4.2.6. Data analysis

For the highest-weighted trajectories of each UPPS-TA complex, conformational clustering was performed, employing a C^{α} RMSD cutoff to determine structurally similar clusters. The trajectories were fit to the alpha carbons of a selection of residues within 10 Å of the active site, with RMSD calculation of these selected residues carried out by the GROMOS clustering algorithm.[204] In each case, the three highest populated cluster accounted for more than 80% of all conformations sampled throughout the trajectory.

Furthermore, hydrogen bonds were identified, and distances measured using the readHBmap python script and gmx distance module within GROMACS.

4.2.7. Estimating ΔG_{bind} using US

System preparation

The final snapshots of the UPPS-**2A** complexes of sites 1-3 (**Section 4.2.4**) were used as the starting configurations for SMD simulations. The structures were inserted into rectangular boxes of dimensions sufficient to satisfy the minimum image convention and to provide space for pulling of the inhibitor along the z axis within GROMACS version 2018.3 (**Figure S13**).

SMD simulation

Neutralising counterions were added and steepest descent minimization was performed. Subsequently, the system was equilibrated in the NPT ensemble for 100 ps using weak Berendsen pressure and temperature coupling and position restraints on all heavy atoms (1000 kJ mol^{-1} nm^{-2}). The position restraints were removed and an SMD simulation was conducted in which the TA was pulled away from UPPS along the z axis over 500 ps using a spring constant of 1000 kJ mol^{-1} nm^{-}

^{2} and a pull rate of 0.01 nm ps^{-1}. Alternative spring constants (i.e., 600 kJ mol^{-1} nm^{-2} and 2000 kJ mol^{-1} nm^{-2}) and pull rates (i.e., 0.001 nm ps^{-1} and 0.005 nm ps^{-1}) were trialled but the latter were selected to expedite data collection without inducing artifacts. The Nosé-Hoover thermostat and Parrinello-Rahman barostat were employed in the SMD simulation to maintain the correct NPT ensemble.

US simulation

Snapshots were extracted every 1 ps from the 500 ps SMD trajectory to generate starting configurations for US windows. For each complex, an asymmetric distribution of sampling windows was used, such that there was a higher concentration of windows centered at shorter COM distances (**Figure S14**). This enabled sufficient overlap between neighbouring US windows, a necessary indicator of convergence. Moreover, this resulted in 59, 47 and 33 US windows (each of 1 ns MD) for the complexes of sites 1-3 respectively.

Data analysis

Analysis of the results was performed using WHAM.[105,106,119] This evaluated a PMF along the RC, ξ, using the gmx wham protocol. The binding free energy was determined as the difference between the maximum and minimum values of the PMF curve. The error of computations was evaluated over 100 rounds of bootstrapping analysis.

4.2.8. Estimating ΔG_{bind} using FEP-MBAR

System preparation

FEP-MBAR calculations for the UPPS-**2A c**omplexes in sites 1-3 were performed employing the final snapshots of previous MD simulations (**Section 4.2.4**). Summing up the free energy changes clockwise around the non-physical

thermodynamic cycle, beginning at the bound-state (**1**) and ending at the unbound-state (**8**), yielded the negative of the standard free energy of binding, as described by Mobley *et al*. (**Figure 25**).[168]

The TA's relative position and orientation with respect to the protein was restrained using one distance, two angles and three dihedral harmonic potentials with force constant of 10 kcal mol^{-1} $Å^{-2}$ (deg^{-2}), as described by Boresch *et al.*[170] The addition of restraints to the fully coupled TA-UPPS complex was evaluated numerically using MBAR, while the release of restraints from the fully decoupled complex was calculated analytically with inclusion of a standard-state correction.[215]

The coulombic interactions were decoupled using a linear alchemical pathway with $\Delta\lambda = 0.10$ (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0), followed by the decoupling of the van der Waals interactions using $\Delta\lambda = 0.05$ (0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.00) and a soft-core potential, both with restraints implemented. The same procedure was applied to the independent solvated TA without restraints. For the addition of the TA restraints, seventeen non-uniformly distributed λ values were employed (0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.50, 0.60, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.00). Thus, a total of forty-eight windows for the complex simulations and thirty-one windows for the ligand simulations were used. Obtaining the relative binding free energies for each site complex required seventy-nine atomistic MD simulations (each of 2 ns), resulting in a total simulation time of 158 ns.

MD simulation

For each window, 50 000 step steepest descent minimization was performed, followed by 250 ps of equilibration in the canonical ensemble with harmonic position restraints applied to the peptide and solute heavy atoms with a force constant of 1000 kJ mol^{-1} nm^{-2}. Temperature was coupled using Langevin dynamics with 300 K as the reference temperature. A 250 ps unrestrained simulation in the isothermal-isobaric ensemble was then performed, in which the temperature was coupled using Langevin dynamics and the pressure was regulated using the Parrinello-Rahman barostat. Subsequently, data was collected over 1-6 ns of unrestrained HREMD, which was carried out in the isothermal-isobaric ensemble, using the same temperature and pressure coupling scheme. Electrostatics were treated using the PME algorithm with a PME cutoff (*rcoulomb*) of 12 Å, a *pme-order* of 6, a relative tolerance (*ewald_rtol*) of 10^{-6}, and a fourier spacing of 1.2 Å. A switch function (*rvdw-switch*) of 10 Å was used for vdW interactions and the P-LINCS constraint algorithm was applied solely to H-bonds. The GROMACS long range dispersion correction for energy and pressure was applied. The length of the HREMD simulation was determined by the extent of phase space overlap achieved between λ replicas, which could be quantified using the overlap matrix generated through the MBAR estimator of the alchemical_analysis.py tool (available from the python package pymbar). 100 000 swaps (-*nex*) between any state pair were attempted every 1000 (-*replex*) time steps, following the Gibbs sampling scheme proposed by Chodera and Shirts.[216]

Data analysis

The results were analysed using MBAR implemented in the analysis_alchemical.py tool, which incorporated automated detection of the equilibration period in each window, as well as subsampling of the data to include only the uncorrelated data-points through calculation of the autocorrelation-time.[217]

A semi-empirical charge-transformation correction, developed by Rocklin *et al* was applied to the site 1 UPPS-2A.[218,219] Use of PBCs and PME in FEP calculations involving charged ligands (i.e., which have not been neutralised), results in changes in the net charge of the system during coulombic (de)coupling of the ligand. This in turn produces inaccuracies in the estimated ΔG_{bind} owing to strong dependencies on the size of the simulation box.[219] Thus, addition of the charge-transformation correction to the uncorrected $\Delta G_{\mathrm{bind,raw}}$ ensures that the resultant $\Delta G_{\mathrm{bind,corr}}$ is independent of the box size.[218,219]

Implementation of the correction required the FEP simulations to be carried out in a cubic box in the canonical ensemble. PBCs were applied but no neutralising counterions were added. Separate PQR files of the complex and the free TA were generated from extracted frames of the charged trajectories, using the *-mead* option passed to *gmx editconf* in GROMACS. Explicit waters were removed and the resultant PQR files were each treated separately as input to APBS 1.5., for numerical solvation of the Poisson-Boltzmann equation for three non-periodic systems.[220]

Three residual integrated potentials (RIPs) were calculated from the ABPS output using the python scripts provided by Olsson *et al*., denoted I_p, I_L and $I_{\mathrm{L,hom}}$, and

used in the analytical calculation of the five corrections encompassing the total charge-transformation correction.[218] This included a correction for periodicity-induced net-charge interactions (ΔG_{NET}), a correction for periodicity-induced net-charge undersolvation (ΔG_{USV}), a correction for RIP effects (ΔG_{RIP}), an empirical correction (ΔG_{EMP}), and a correction for discrete solvent effects (ΔG_{DSC}).[218,219] In the first calculation, the protein atoms were charged while the ligand atoms remained uncharged (I_p), and for the final two calculations the ligand was charged while the protein was uncharged. The latter two calculations were differentiated by the value of the dielectric constant of the solvent, which was equal to that of the bulk solvent (ε_s = 97 for TIP3P water, I_L) and the internal dielectric constant (ε_s = 1, $I_{\mathrm{L,hom}}$) respectively. The $\Delta G_{\mathrm{bind,corr}}$ was calculated using the following expressions:

$$\Delta G_{\mathrm{NET}} + \Delta G_{\mathrm{USV}} = \frac{-\xi_{\mathrm{LS}}((Q_P + Q_L)^2 - Q_P^2)}{8\pi\varepsilon_0(\varepsilon_s L)} \tag{65}$$

$$\Delta G_{\mathrm{RIP}} = \frac{(I_P - I_L)(Q_P + Q_L) - I_P Q_P}{L^3} \tag{66}$$

$$\Delta G_{\mathrm{EMP}} = -\frac{16\pi^2}{360\pi\varepsilon_0}(1 - \frac{1}{\varepsilon_s})[(Q_P + Q_L)^2 - Q_P^2]\frac{R_L^5}{L^6} \tag{67}$$

$$\Delta G_{\mathrm{DSC}} = \frac{\gamma_s Q_L N_s}{6\varepsilon_0 L^3} \tag{68}$$

$$R_L = \sqrt{\frac{I_L - I_{\mathrm{L,hom}}}{\frac{4\pi}{24\pi\varepsilon_0}(1 - \frac{1}{\varepsilon_s})|Q_L|}} \tag{69}$$

Thus, $\Delta G_{\mathrm{bind,corr}}$ was given by the sum of these five components and the uncorrected $\Delta G_{\mathrm{bind,raw}}$ from the original unmodified FEP calculations. The final measure of binding free energy, represented by the free energy difference

between the UPPS-TA complex and independent TA ($\Delta\Delta G_{\mathrm{bind,corr}}$) was computed using the following formula:

$$\Delta\Delta G_{\mathrm{bind,corr}} = \Delta\Delta G_{\mathrm{bind,raw}} + \Delta\Delta G_{\mathrm{NET}} + \Delta\Delta G_{\mathrm{USV}} + \Delta\Delta G_{\mathrm{RIP}} + \Delta\Delta G_{\mathrm{EMP}} + \Delta\Delta G_{\mathrm{DSC}} \quad (70)$$

4.2.9. Hydrophobic tunnel volume measurements

The three 100 ns MD simulations of UPPS-**2A** complexes of site 1 (PDB ID: **3QAS**), site 2 (PDB ID: **4H8E**) and site 3 (PDB ID: **4Q9M**) were subjected to PVA using POVME 3.0[181], according to the same procedure as outlined in **Section 3.3.3.**

4.2.10. Temperature Replica Exchange (TRE) MD of UPPS apoenzyme

TREMD simulations of *E. coli* apo-UPPS (PDB ID: **3QAS**) were performed in. GROMACS[199] with the OPLSA forcefield.[144] Each system was centered in a rhombic dodecahedral box (1.0 nm buffer region), solvated (TIP3P water)[200], neutralised with 2 Cl ions, and minimised using the steepest descent algorithm until the maximum force was 886 kJ mol^{-1} nm^{-1}. Eighteen replicas spanning the 310-324 K temperature range were initiated from the minimised structure. For each replica, restrained equilibration was performed at fixed pressure and temperature (NPT ensemble) for 200 ps (1 fs timestep). Peptide and ligand heavy atoms were position restrained by a small harmonic force (i.e., 1000 kJ mol^{-1} nm^{-2}), and the protein and non-protein atoms were coupled to separate temperature baths using the Berendsen weak coupling method.[201] Subsequently, each replica was subjected to 5 ns REMD in the NPT ensemble for a total combined simulation time of 90 ns. For data collection, the V-rescale thermostat and the Parrinello-Rahman barostat were employed to maintain the temperature at the specified T and to regulate isotropically the pressure at 1.0 bar respectively. All bonds involving hydrogens were constrained using the LINCS algorithm.[202] Exchanges were

attempted every 10 ps and the average exchange ratio for each replica was found to lie between 0.25 and 0.41.

4.3. Results and Discussion

4.3.1. Relative stability of TAs' keto-enol tautomers

In this section, the keto-enol tautomerism of the TAs is discussed, drawing upon ^{13}C NMR assignment and the established literature.[62,70,221–224] Using DFT calculations, approximate trends in the relative tautomeric stabilities of selected TAs in solution are proposed, informing subsequent docking and MD studies.

TAs potentially exist as an equilibrating mixture of four tautomers (**Figure 36: A**).[62,69,70,221,223–226] Internal tautomers (A and B, C and D) have been identified as single averaged hybrids of the external tautomers (AB and CD) by NMR spectroscopy (**Figure 36: A**). The external tautomers, which are differentiable, each possess distinct chemical shifts in ^{1}H and ^{13}C NMR spectra. Additionally, the *exo*-enol tautomers (B and D) have been distinguishable from the *endo*-enol tautomers (A and C) by Heteronuclear Multiple Bond Correlation (HMBC) analysis.[62]

Moreover, ^{13}C NMR assignment of 3-acyltetramates has been facilitated by the chemical shifts of the C(2) and C(4) carbonyls. The hydrogen bonded carbonyls (i.e., C(4) in tautomer B and C(2) in tautomer D) have resonated downfield of the free carbonyls (i.e., C(4) in tautomers C and D and C(2) in tautomers A and B) (**Figure 36: A**). The free carbonyls (i.e., C(4) in tautomers C and D) have in turn resonated at lower field relative to their enolic carbonyl counterparts (i.e., C(4) in tautomer A). In addition, the hydrogen-bonded carbonyls of the exocyclic 3-acyl

group in tautomers A and C have been found to be more deshielded than their corresponding enolic carbonyls in tautomers B and D (**Figure 36: A**).[70]

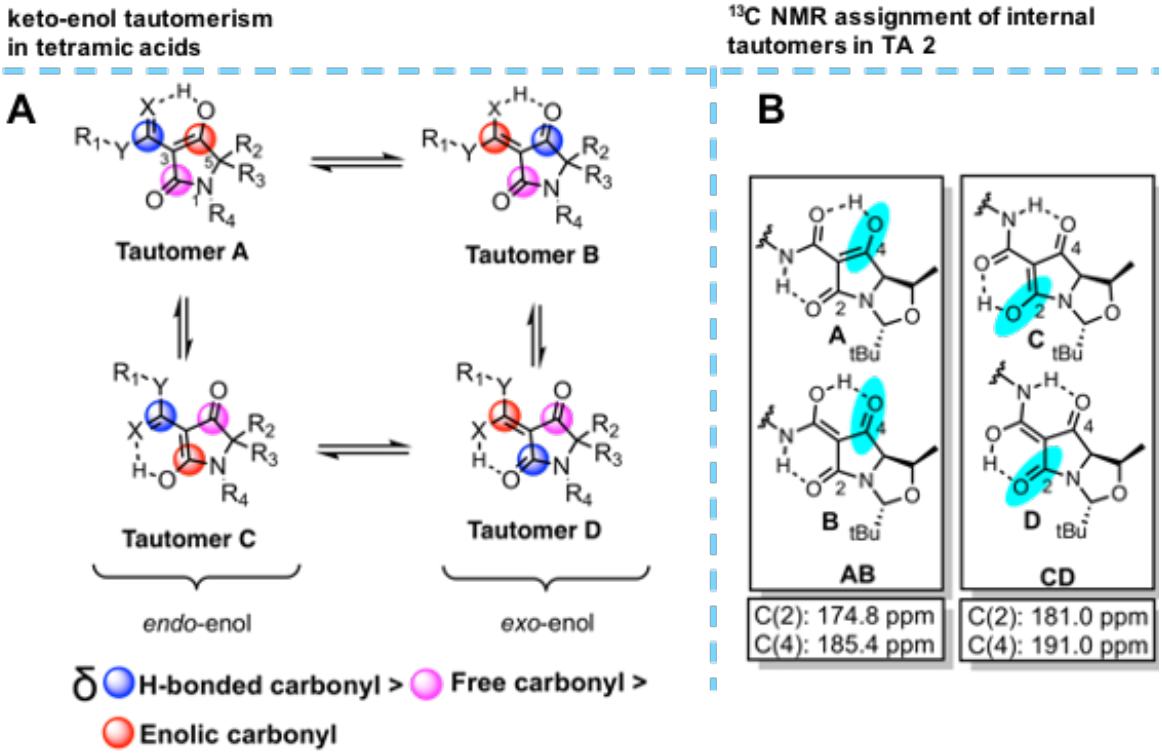


Figure 36: Keto-enol tautomerism in TAs (**A**). Identification of the internal tautomers within the external tautomer pairs **AB** and **CD** of TA **2** using ^{13}C NMR spectroscopy (**B**).

Besides evaluating the ratio of the external tautomeric pairs (i.e., AB and CD), it has also been possible to deduce the ratio of internal tautomers using the ^{13}C NMR shifts. In the AB external pair, the chemical shift of C(4) has been found to be lower for the hydrogen-bonded enolic carbonyl (tautomer A) than for the hydrogen-bonded keto-carbonyl (tautomer B).[62] Similarly, in the CD pair, the C(2) keto-carbonyl (tautomer D) has been found to resonate downfield of its enolic carbonyl counterpart (tautomer C).[62]

Taking compound **2** as an example, ^{13}C NMR chemical shifts of C(2) and C(4) were useful diagnostic indicators of the identity of the internal tautomers (**Figure 36: B**). This identification was facilitated by comparison of the ^{13}C NMR chemical shifts of carbons C(4) and C(2) in the averaged hybrid signals with the chemical shifts

belonging to known internal tautomers. Thus, for the external signal AB, C(4) resonated at 185.4 ppm, which was diagnostic of the hydrogen-bonded enolic carbonyl in tautomer A (i.e., the signal was upfield of the hydrogen-bonded keto counterpart in tautomer B). Using the same approach, C(2) of the external signal CD resonated at 181.0 ppm, a low field chemical shift indicative of the keto carbonyl of tautomer D (i.e., the shift was equivalent to the corresponding carbon in tautomer D of compound **3**).[62]

NMR spectroscopic analyses in deuterated chloroform ($CDCl_3$) have established that the *exo*-enol tautomer is preferred by the 3-acyl and 3-enamine TAs, while the *endo*-enol tautomer is predominant for the 3-carboxamide TAs.[62,70,227] Spectroscopic analysis of a set of monocyclic 3-acyl TAs by Jeong *et al*. has revealed that N-H and N-alkyl forms favour the *exo*-enol tautomer D (>80% D: <20% B).[70] Additionally, HMBC correlations for the bicyclic 3-acyltetramate **3** have revealed a tautomeric ratio of 80% D: 20% B **(Figure 37)**.[62] Similar spectroscopic investigation of a series of monocyclic 3-carboxamide TAs by Jeong *et al.* have indicated that tautomerism is not strongly affected by N(1)-substitution, but favours the *endo*-enol tautomer A in all cases (>80%), with only minor contributions from tautomer D.[227] Further HMBC and Nuclear Overhauser Effect Spectroscopy (NOESY) correlation data of cysteine-derived 3-carboxamide TAs presented by Panduwawala *et al.* has revealed that analogues of this class exist predominantly as tautomer A.[224] These collective observations have been consistent with the NMR determined internal tautomeric ratios of the compounds **1**, **2** and **4**, which exist predominantly as tautomer A (80% A: 20% D) (**Figure 37**).[62]

While the spectroscopic evidence has pointed to definite trends in TA tautomeric stability, the present study sought to further validate these observations using DFT analysis. Literature precedents for such DFT investigation exist for the 3-acyl and 3-carboxamide TAs. Semiempirical calculations of the ground state energies of monocyclic N-H and N-Me substituted 3-acyl TAs at the AM1 approximation by Gromak *et al*. have predicted the order of stability for these systems as: D>B>C>A.[222] According to this approach, the reported difference in energy between D and B ranged from 0.5-2 kcal/mol, while that between C or A and D fell within the range of 4-7.8 kcal/mol.[222] Furthermore, Jeong *et al*. has computed that the lowest energy tautomers for monocyclic 3-acyl TAs are tautomers D and B, with relative differences of 0.8-1.5 kcal/mol, employing DFT methods (B3LYP/6-31G* in Spartan 02).[70] Jeong applied the same level of theory to calculation of the ground state energies of tautomers A-D for monocyclic, N-Me substituted 3-carboxamide TAs. According to this investigation, the order of relative stability has been reported as A>D>B>C, in which differences between tautomer A (0.0 kcal/mol) and the tautomers D, B and C were 0.5 kcal/mol, 1 kcal/mol and 2.3 kcal/mol respectively.[227]

Furthermore, results of the present study suggested that DFT was a poor estimator of the relative stabilities of the 3-carboxamide TA tautomers, predicting energies which were inconsistent with the NMR interpretation (**Figure 37**). According to NMR, the 3-carboxamide TAs preferred the *endo*-enol form, specifically tautomer A. Conversely, DFT predicted tautomers A, B and D to be of equal energy and tautomer C to be relatively disfavoured. However, results for the 3-acyl and 3-enamine TAs were in good alignment with the NMR predictions and the literature

precedents. 3-Acyl-TAs **3** and **5**, as well as 3-enamine TA **6**, exhibited a preference for the *exo*-enol tautomers B and D, while *endo*-enol tautomers A and C were relatively disfavoured.

Based on the trends deduced from both NMR and DFT, representative compounds **1** and **2** (3-carboxamide TAs) were taken as tautomer A and compounds **3**, **5a** and **5b** (3-acyl TAs) were taken as tautomer D in the subsequent docking-MD study. Critically, it was acknowledged that this was an approximation, as transfer from the solvent medium to the enzyme environment is likely to affect the tautomeric equilibrium.

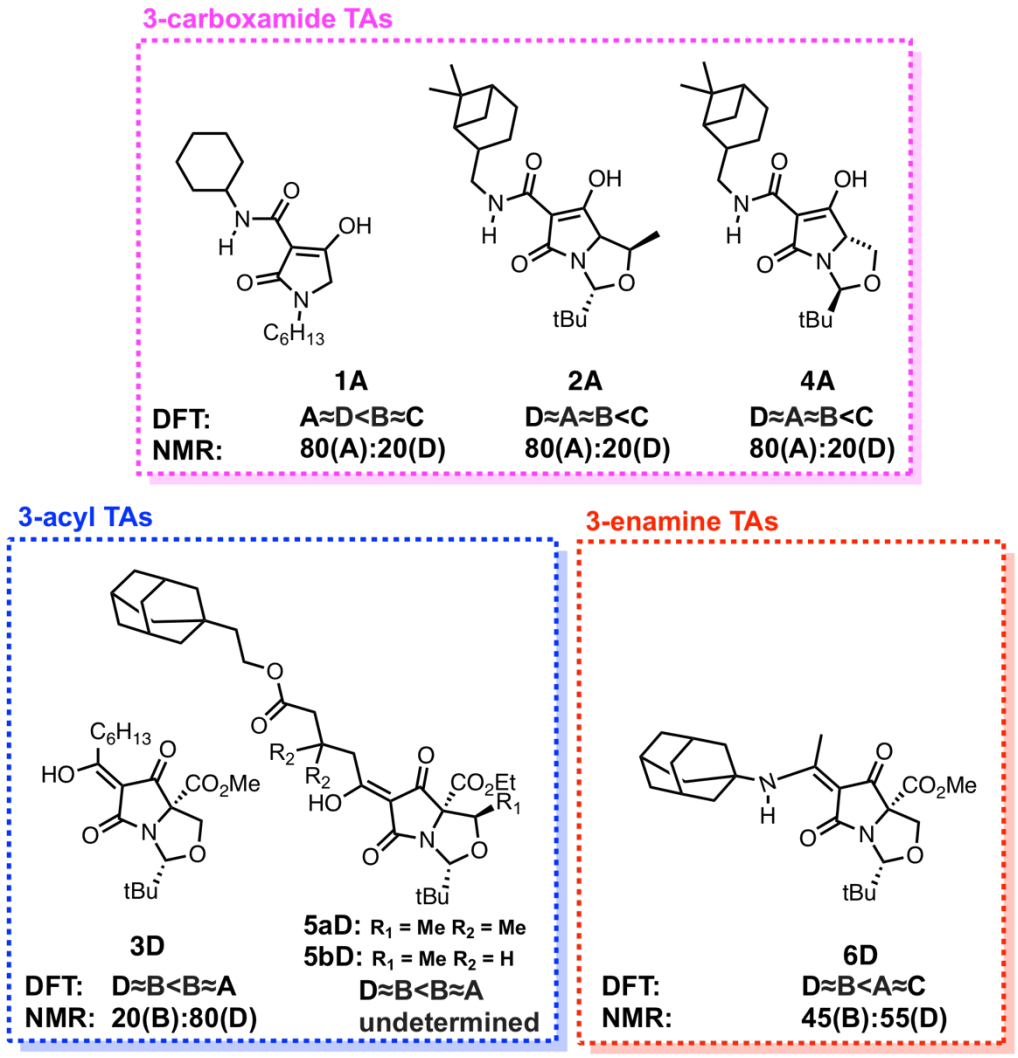


Figure 37: Comparison of the tautomeric ratios determined by NMR and the relative stabilities computed by DFT for selected TA compounds **1-6.**

MD analysis of binding sites

Following the study of TA inhibitor stability, the preferred TA binding site was identified using MD-based techniques. Initially, the LIE method was applied to FPP-bound UPPS to ensure that its predictions about this specific system were reliable. Subsequently, an MD-LIE study involving UPPS and selected TAs was conducted in two parts. First, the relative affinities of the TA **2A** for the proposed binding sites 1-3 were quantified (**Figure 34**). This was followed by affinity calculations for an expanded TA subset comprising compounds **1-3**.

Elucidating the TA binding site using LIE

LIE method validation using UPPS-FPP system

To validate the MD-LIE approach for this system, the method was applied to two crystallographic UPPS-FPP binding modes (PDB ID: **1V7U**), one in which FPP occupied the catalytic site (site 1) and one in which FPP occupied the remote site 4. A reliable method should predict higher affinity for site 1 relative to site 4. Consistent with this expectation, results of the MD-LIE calculations revealed that site 1 was preferred to site 4 by 27 kcal/mol (**Figure 38**). Despite the large number of hydrogen bonds and hydrophobic interactions observed (**Figure 38**), the ΔG_{LIE} was found to be positive in both cases. However, since the aim here was to determine the ranking of the proposed binding sites 1-3, the relative rather than absolute free energies of binding were considered. The same approach was then applied to compound **2A** in each of the proposed binding sites 1-3 (**Figure 34**).

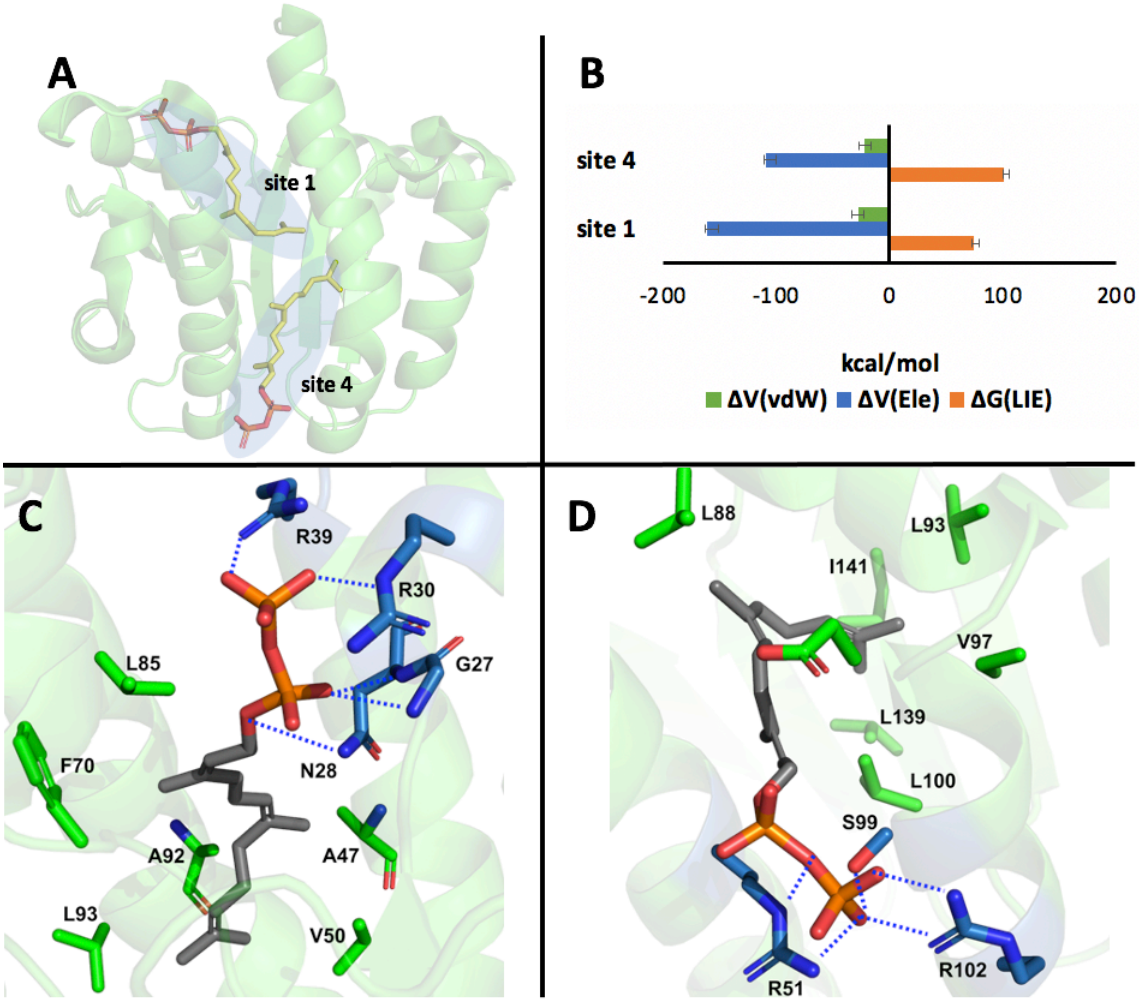


Figure 38: Site 1 (catalytic site) and site 4 in *E. coli* closed UPPS (PDB ID: **1V7U**) (**A**). LIE results for sites 1 and 4 (**B**). The main hydrogen bonds (blue sticks) and hydrophobic interactions (green stick) between FPP and sites 1 (**C**) and 4 (**D**).

*Preferred protonation states of active site histidine in UPPS-**2A** complexes*

Since it could not be assumed that the apoenzyme's preference for the HIP state (**Section 3.4.1**) remained upon binding of the inhibitor **2A**, the effect of different active site histidine protonation states on the ΔG_{LIE} and histidine RMSD were explored for each site. It was hypothesised that the protonation state yielding the most favourable ΔG_{LIE} (and the lower value of histidine RSMD) would be preferred, although more rigorous FEP-based methods will be required to confirm this approximation.

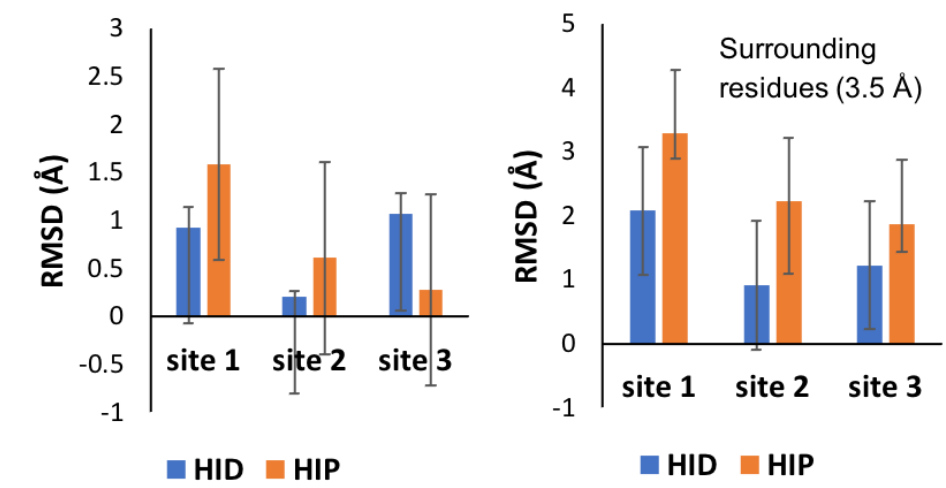


Figure 39: Mean RMSD of H43 (H50 in *S. aureus* – site 2; H45 in *S. pneumoniae* – site 3) and the surrounding residues calculated from the 100 ns MD simulations of the UPPS-**2A** complexes for sites 1-3.

Contrary to this hypothesis, results of the protonation state study indicated that there was no clear correlation between the RMSD of the histidine and the ΔG_{LIE} of the UPPS-**2A** complex. It was observed that there was a trend for the mean RMSD of the histidine in the UPPS-**2A** trajectories to be lowest in the HID form for sites 1 and 2, but highest in the HID state for site 3 (**Figure 39**).

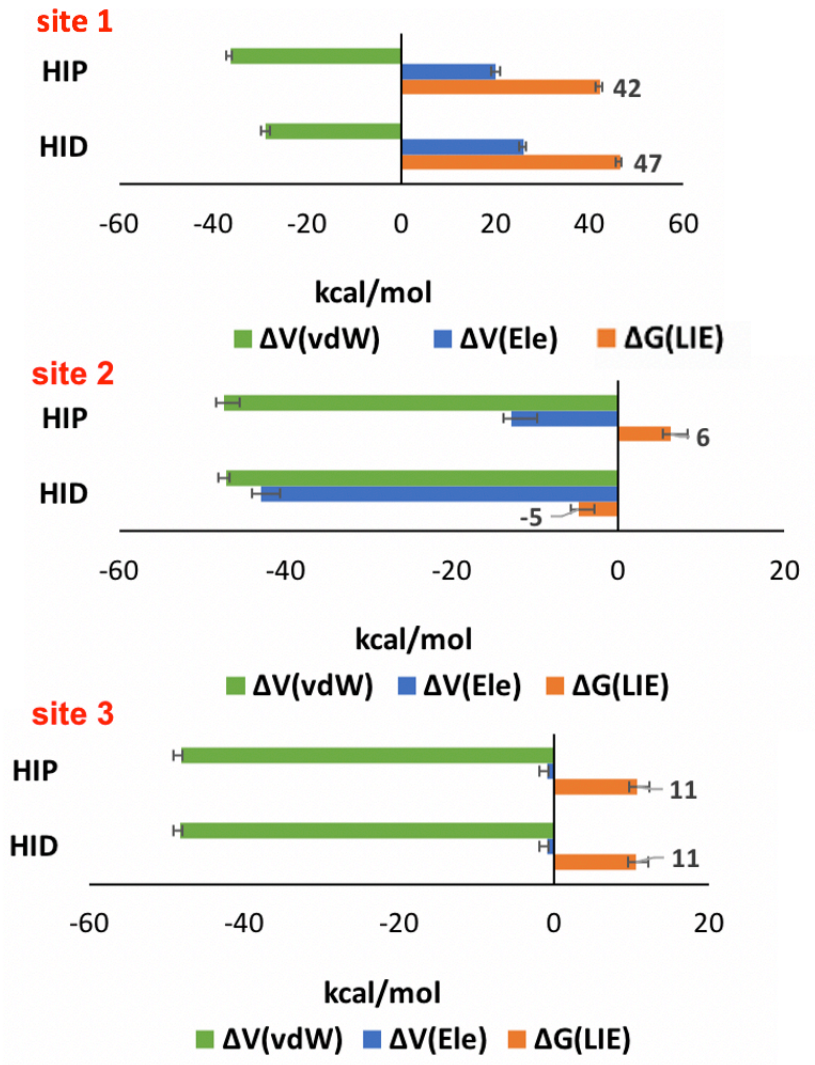


Figure 40: MD-LIE results for the HID and HIP trajectories of UPPS-**2A** in sites 1-3.

Moreover, differing histidine protonation states yielded negligible ($\Delta\Delta G_{LIE}$ = 4 kcal/mol) and no ($\Delta\Delta G_{LIE}$ = 0 kcal/mol) differences in the affinity of compound **2A** for the proposed binding sites 1 and 3 respectively (**Figure 40**). However, a larger affinity difference was observed for binding site 2 ($\Delta\Delta G_{LIE}$ = 11 kcal/mol; **Figure 40**), in which the HID state appeared to be preferred (**Figure 40**). This affinity difference was attributed to stronger hydrogen bonds between compound **2A** and the residues D33, N35, and R84 in the HID simulation (**Figure 41**). This was reflected in the higher electrostatic interaction energy associated with the HID simulation relative to its HIP counterpart ($\Delta\Delta V^{Ele}$ = -30 kcal/mol; **Figure 40**). In particular, it was observed that in the HID simulation R84 was situated deep in the

active site, available for hydrogen bond formation more frequently (**Figure 41: C**). This was consistent with the relative mean distances of electrostatic interaction between compound **2A** and the arginine residue (**Figure 41: D**); in the HID simulations, a hydrogen bond was observed (3 Å), whereas in the HIP simulation, only a weak electrostatic interaction (7 Å) was formed. Thus, based on these results, only the HID simulations were considered in all subsequent MD-LIE binding site comparisons.

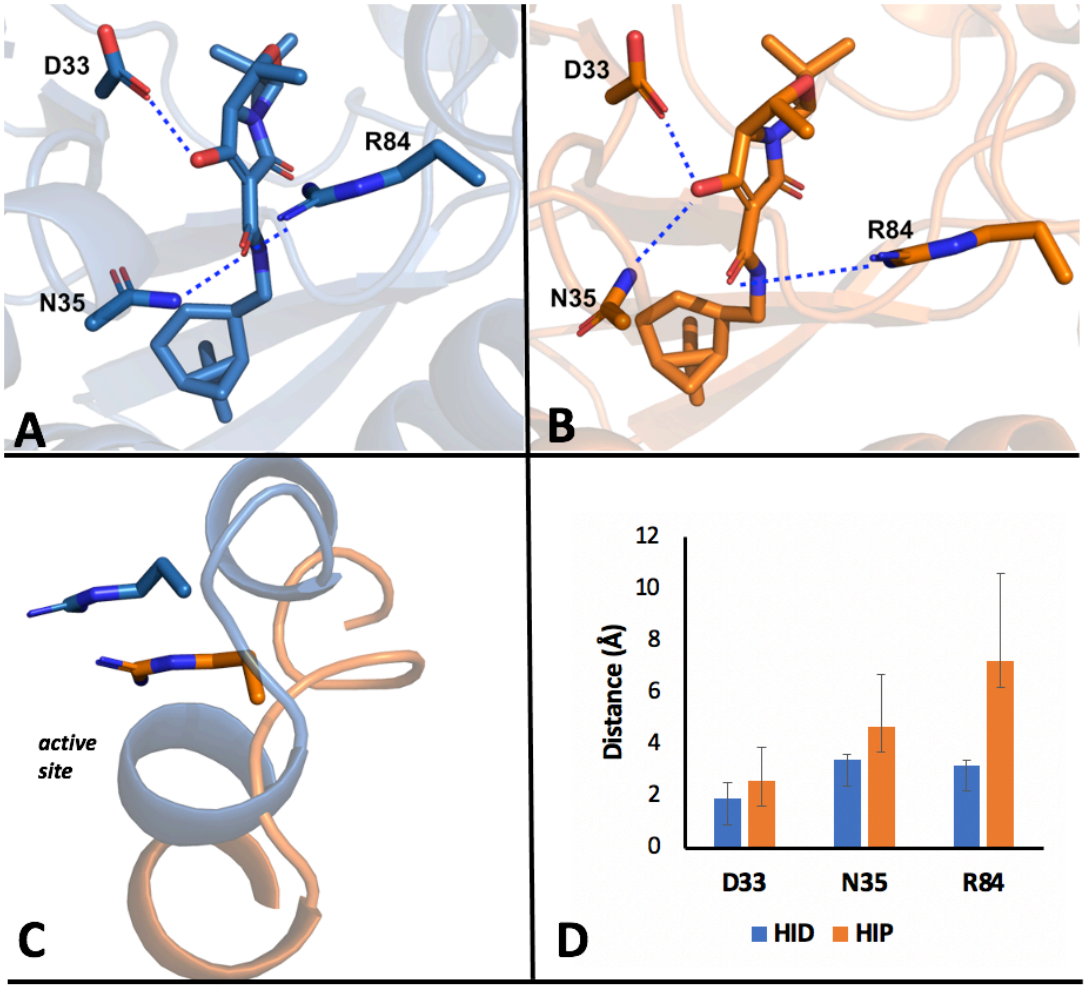


Figure 41: Main hydrogen bonds formed in the HID (**A**) and HIP (**B**) trajectories of the site 2 UPPS-**2A** complex. Differences in the conformations of the flexible loop (residues 79-89 in *S. aureus* UPPS, PDB ID: **4H8E**) and key residue R84 between the highest populated clusters of the HID and HIP trajectories (**C**). Average distances of electrostatic interactions, including standard deviations, for the HID and HIP trajectories of the site 2 UPPS-**2A** complex.

*MD-LIE study of UPPS-**2A** complexes*

In the MD-LIE binding site comparison utilising compound **2A**, no correlation was found between the docking score originally obtained for the complexes and the

ΔG_{LIE} values of the simulated complexes (**Figure S15**).[140] This suggested that while docking was useful for generating starting UPPS-TA conformations, it provided a poor measure of the TA binding affinity.

Results of this investigation ranked the binding sites such that site 2 (ΔG_{LIE} -4.7 ± 1.8 kcal/mol) > site 3 (ΔG_{LIE} 10.7 ± 1.6 kcal/mol) > site 1 (ΔG_{LIE} 46.6 ± 0.3 kcal/mol; **Figure 42**). The total absence of hydrogen bonds in sites 1 and 3 rationalised the TA's preference for site 2, in which modelling predicted the formation of hydrogen bonds with the residues N81, R84, N35 and D33 (**Figure 42**). Although site 1 overlapped with site 2, it was positioned at the top of the hydrophobic tunnel, remote from many of the hydrophobic residues located deeper inside the tunnel. Hence, the TAs were seen to form fewer interactions (**Figure S16**); this was consistent with its being the least favoured site according to LIE analysis.

Moreover, the TA formed fewer hydrogen bonds and hydrophobic interactions in site 2 than the native substrate, FPP. Modelling predicted that FPP formed twelve hydrogen bonds, primarily between the pyrophosphate head group and the hydrophilic residues (i.e., R84, R46, R37, H50, N35) at the top of the hydrophobic tunnel, but also with G34 and G36 (**Figure 38: C**). Additionally, the isoprenoid tail of FPP was observed to make van der Waals interactions with hydrophobic residues such as phenylalanine and isoleucine (i.e., F145, F99, I92, I57).

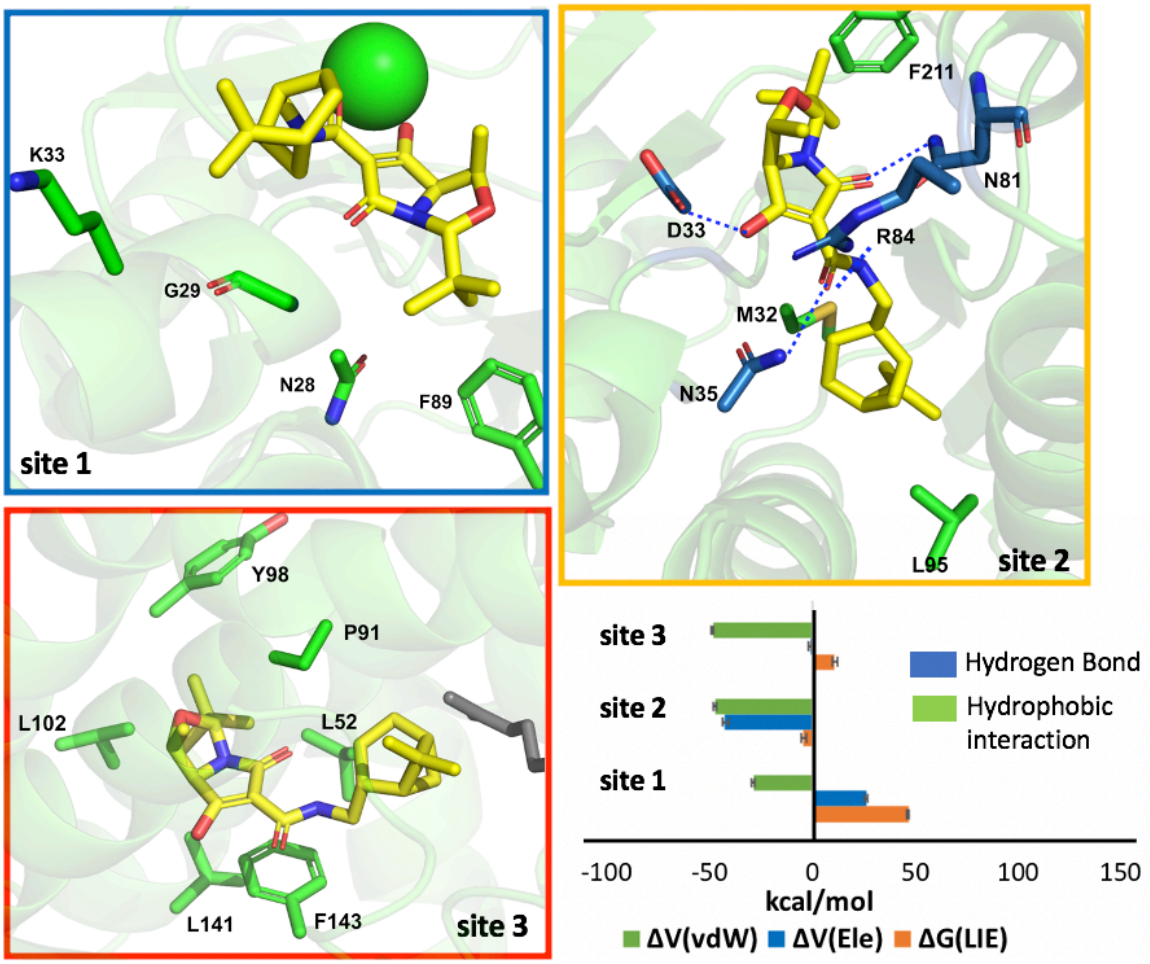


Figure 42: Interactions of the TA **2A** with UPPS in sites 1-3, in addition to the LIE results are presented.

*Hydrophobic tunnel volume of UPPS-**2A** complexes*

In addition to the MD-LIE study of the TA **2A** in sites 1-3, PVA was carried out on each of the complexes. The mean volumes of the hydrophobic tunnel were found to be 391 $Å^3 \pm 65$ $Å^3$, 264 $Å^3 \pm 27$ $Å^3$, and 436 $Å^3 \pm 60$ $Å^3$ for the UPPS-**2A** complex in sites 1-3 respectively (**Figure 43**). This result demonstrated that all three UPPS-2A complexes sustained closed conformational states whose volumes fell within the range of volumes accessed by the apoenzyme (**Figure 31**). This confirmed McCammon's hypothesis that TAs occupied a closed state of UPPS.[186] In addition, it suggested that the enzymatic structural change required upon TA binding was less than that required for binding of the bisphosphonates, which occupied an

expanded state whose volume exceeded any conformation sampled by the apoenzyme.

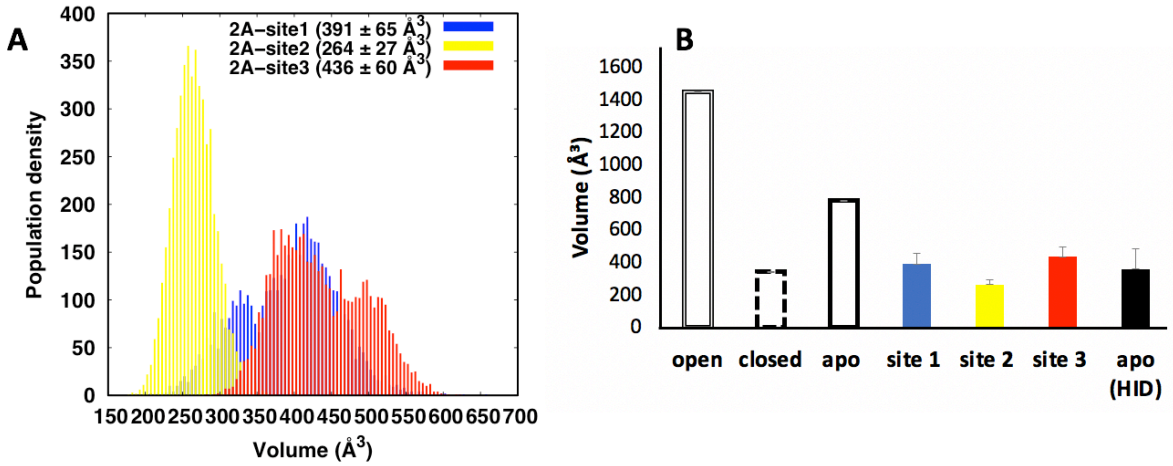


Figure 43: Histogram representation of the hydrophobic tunnel volumes sampled throughout the 100 ns MD simulations of the UPPS-**2A** complex for sites 1-3 (**A**). The mean volume and standard deviations of the hydrophobic tunnel for these complexes, in addition to the fixed volumes of the X-ray crystal structures of UPPS in the open (PDB ID: **2E98**), closed (PDB ID: **4H8E**) and apo (PDB ID: **3QAS**) forms (**B**).

Furthermore, combining results from the binding site comparison and analysis of the hydrophobic tunnel volume, the TA **2A** was found to exhibit high affinity for site 2, which has the smallest volume and least flexible hydrophobic tunnel. The affinity ranking of the binding sites, according to this single-compound study was found to be site 2 (ΔG_{LIE} -4.7 ± 1.8 kcal/mol) > site 3 (ΔG_{LIE} 10.7 ± 1.6 kcal/mol) > site 1 (ΔG_{LIE} 46.6 ± 0.3 kcal/mol). In addition, all UPPS-**2A** complexes were found to occupy the closed states of UPPS.

*Expansion of MD-LIE study to compounds **1** and **3***

The second part of the MD-LIE study included compounds **1** and **3** in the analysis, in order to determine whether the same trend in relative affinities was reproduced by different TA inhibitors. In this part of the study, correlations were noted between pIC_{50} data and ΔG_{LIE} as an approximative measure of the degree of correspondence between calculation and experiment, however it was

acknowledged that larger datasets would be necessary to confirm such relationships.

In site 1, all values of ΔG_{LIE} were large and positive (**1A** 28.5 ± 1.5 kcal/mol, **2A** 46.6 ± 0.3 kcal/mol, **3A** 38.3 ± 0.3 kcal/mol; **Figure 44**). The LIE determined binding free energies were inconsistent with the *in vitro* assay results; no correlation (r = -0.09) was found between the pIC_{50} and the ΔG_{LIE} values (**Figure S17**). Modelling predicted that compound **1A** formed hydrogen bonds with G27, R30 and R39, while the TA **3A** interacted with R39 and N28 (**Figure 44**). However, TA **2A** did not form any hydrogen bonds in site 1. The only noteworthy hydrophobic interaction observed for the TAs was that with F89 (**Figure 44**). Thus, based on the low numbers of interactions and the large positive ΔG_{LIE} values, the chelation model (site 1) was rejected.

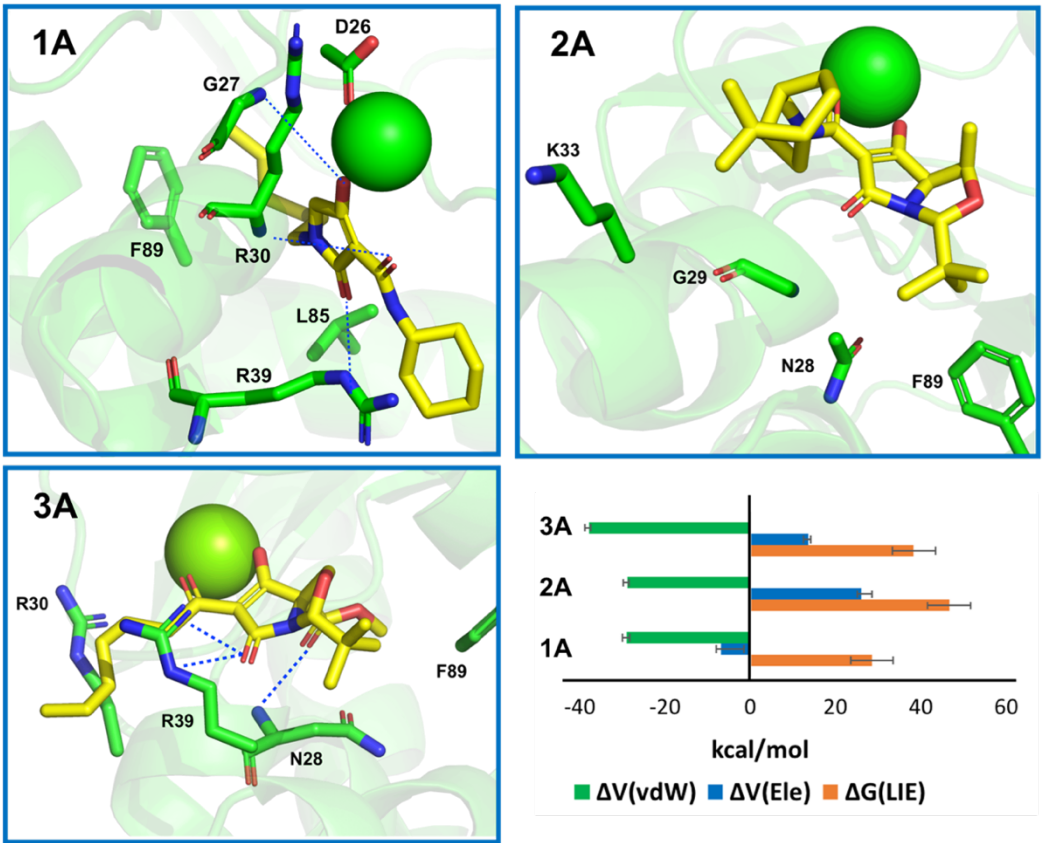


Figure 44: The highest populated MD clusters of UPPS-**1A**, UPPS-**2A** and UPPS-**3A** complexes for site 1, showing the main hydrogen bonds and hydrophobic interactions. The LIE results for the three complexes are provided.

In site 2, the TAs **1A** and **2A** were found to exhibit higher UPPS affinity than the TA **3D**, and this was rationalised by the presence of higher numbers of hydrogen bonds (**Figure 45**). Modelling predicted that the TA **1A** formed three hydrogen bonds (N35, Y75, R84) and hydrophobic interactions were observed between the cyclohexyl moiety and the residues D33, I31, and E80 (**Figure 45**). Similarly, the TA **2A** formed four hydrogen bonds (D33, N35, R84, and N81), as well as hydrophobic contacts with the residues L95, M32 and F211 (**Figure 45**). By contrast, modelling predicted that only a single hydrogen bond was formed between compound **3D** and N81. Binding of the latter was dominated by hydrophobic interactions, which were observed between the *tert*-butyl moiety and the residues D33 and Y75, as well as between the n-pentyl group and the residues I57, L95, and P96 (**Figure 45**).

Moreover, these observations were consistent with the *in vitro* assay results, in which TA **3D** was found to be inactive, while the compounds **1A** and **2A** showed high inhibitory activity (ΔG_{LIE} **1A** -0.6 ± 1.5 kcal/mol, **2A** -4.7 ± 1.8 kcal/mol, **3D** 4.9 ± 0.3 kcal/mol: **Figure 45**). Although the confirmation of any agreement between calculation and experiment requires a larger dataset, the strong negative correlation (i.e., r = **-0.89; Figure S17**) between the pIC_{50} data and the calculated binding free energies was encouraging (**Table S10**).

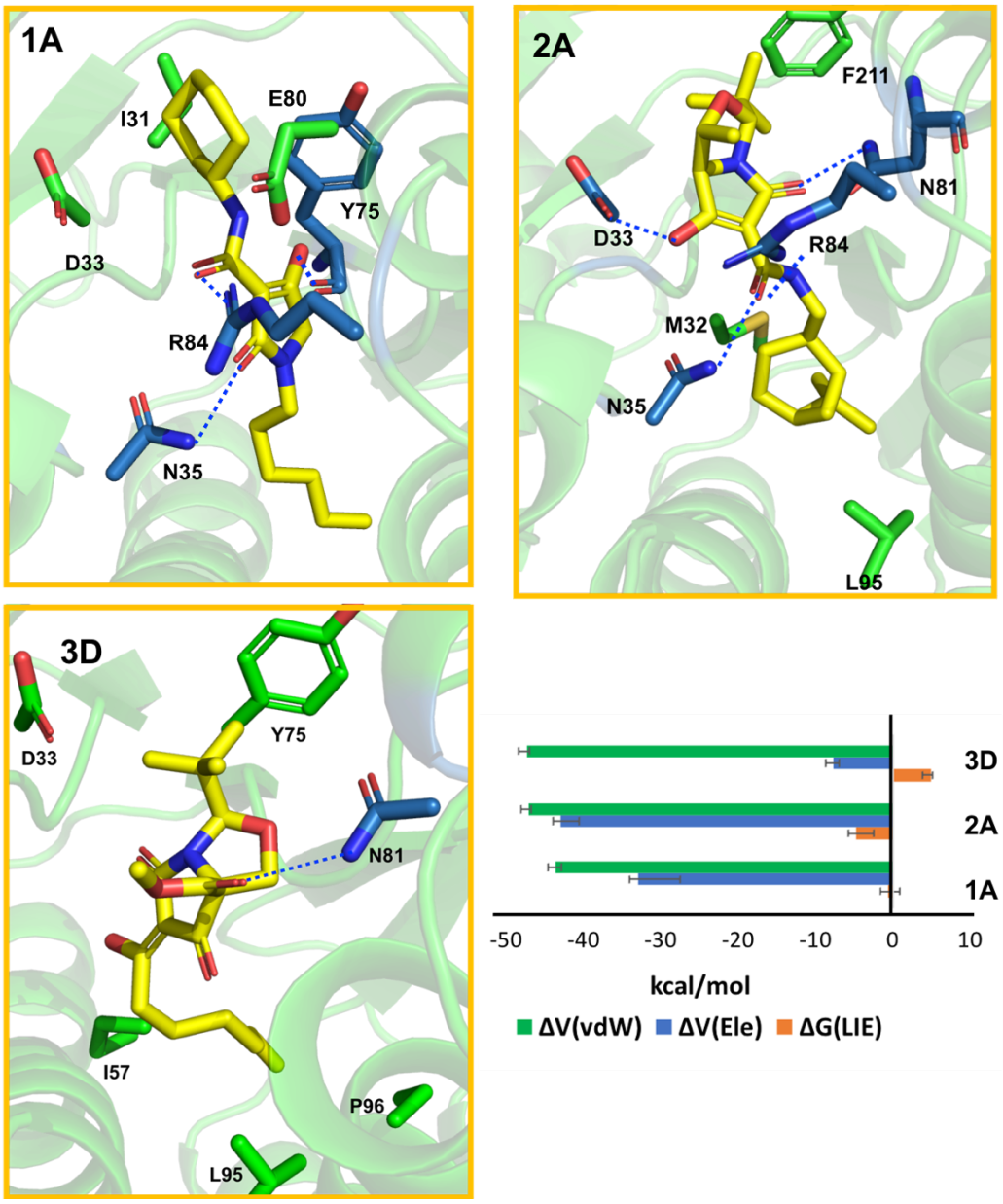


Figure 45: The highest populated MD clusters of UPPS-**1A**, UPPS-**2A** and UPPS-**3D** complexes for site 2. The LIE results for the three complexes are provided.

In site 3, binding of the TAs was found to be dominated by van der Waals interactions, consistent with the location of the proposed allosteric pocket deep within the hydrophobic tunnel. Site 3 offered little opportunity for the polar TA framework to interact with hydrophilic residues, when compared to site 2 (**Figure 45**). However, despite the absence of hydrogen bonds, the abundance of hydrophobic contacts yielded more favourable values of ΔG_{LIE} relative to site 1 (**1A** 10.0 ± 1.1 kcal/mol, **2A** 10.7 ± 1.6 kcal/mol, **3D** 6.5 ± 0.2 kcal/mol; **Figure 46**). Between five and seven hydrophobic contacts were made with each of the TAs,

centring around the pendant hydrophobic moiety and the tert-butyl group (**Figure 46**). No hydrogen bonds were formed, except by the compound **1A** to N146 (**Figure 46**). The correlation between the pIC_{50} data and the ΔG_{LIE} values was positive (r **= 0.98; Figure S17**), indicating that the calculated affinity decreased as the *in vitro* affinity increased. Although the dataset would need to be increased for any such correlations to be significant, the qualitative evidence (i.e., observed interactions) and the relative ΔG_{LIE} values suggested that site 2 was preferred by the TAs.

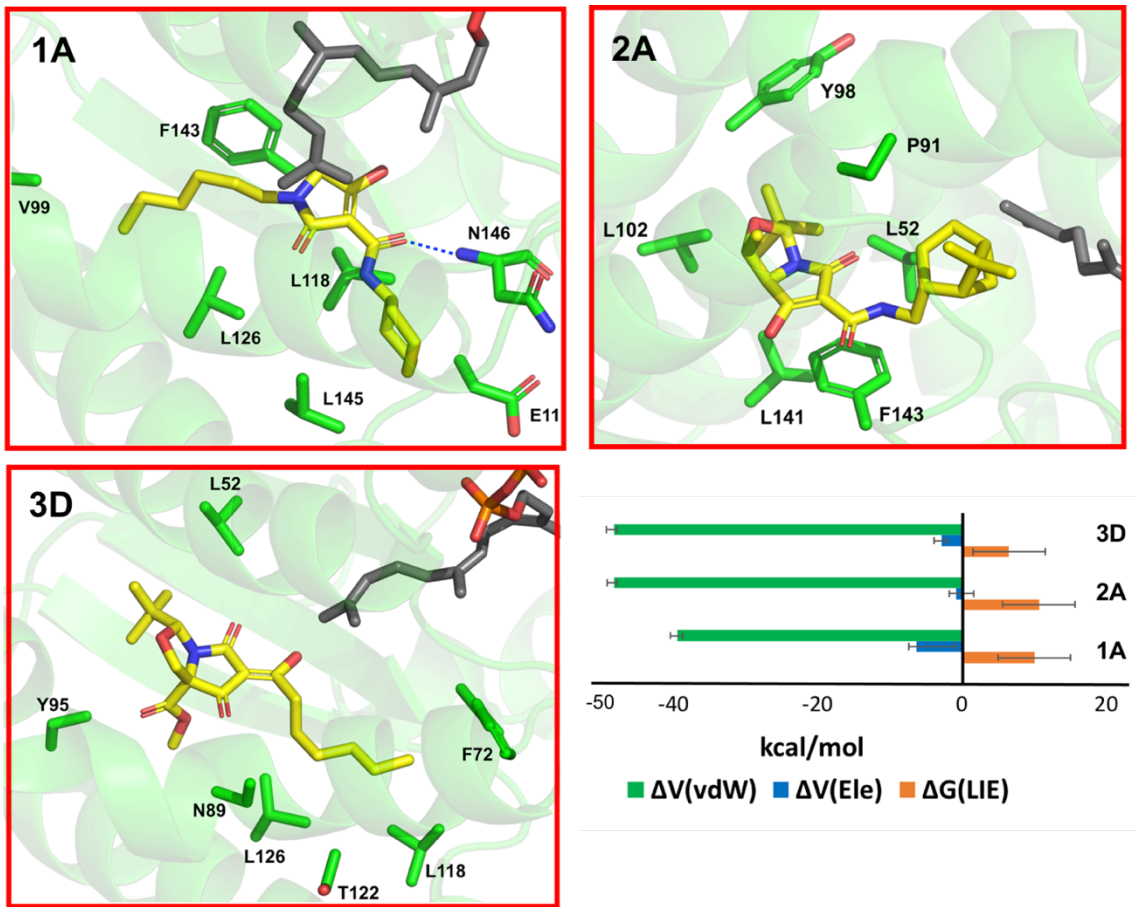


Figure 46: The highest populated MD clusters of UPPS-**1A**, UPPS-**2A** and UPPS-**3D** complexes for site 3. The LIE results for the three complexes are provided.

In conclusion, the binding site ranking provided by the 3-compound MD-LIE investigation was in alignment with the ranking of the initial single-compound study, i.e., site 2 > site 3 > site 1. Site 2, which had the most favourable interactions profile, appeared to be the preferred binding site.

Validating the TA binding site using US

The relative affinity of the TA **2A** for proposed binding sites **1-3** was probed using US. This method provided possible pathways of ligand-protein dissociation through SMD trajectories. The primary limitation of this approach was the flexibility of UPPS, since US is generally applied to rigid binding sites at the enzyme's surface.

From the SMD simulations of UPPS complexed with **2A** in sites 1-3, it was found that α1, α3 and α4 displayed the highest RMSD between the first and last frames (**Figure 47**). For the site 1 complex, α1 displayed the highest RMSD (**Figure 47**: 0.3 Å), while for the site 2 and site 3 complexes, α3 (**Figure 47**: 2.1 Å) and α4 (**Figure 47**: 3.2 Å) showed the highest RMSD respectively. These backbone RMSD values were compared to the RMSD of the corresponding α helices in the four highest populated conformational clusters of a TREMD trajectory of the apoenzyme (PDB ID: **3QAS**; **Section 4.2.10**). The latter provided indication of the extent of mobility of the α helices in the absence of a ligand. It was found that the magnitude of RMSD of the α1 helix in the site 1 complex and the α3 helix in the site 2 complex fell within the RMSD range of expected motion characteristic of the apoenzyme (i.e., relative to the starting crystallographic conformation; PDB ID: **3QAS**). Conversely, the backbone RMSD of the α4 helix in the site 3 complex exceeded this range by almost 3 Å (**Figure 47**). From these results, it appeared that the structural reorganisation required for binding to sites 1 and 2 fell within the intrinsic conformational dynamics of the apoenzyme. However, the displacement of the α4 helix required for binding of compound **2A** to site 3 appeared to represent a significant induced conformational change. Hence, this suggested that

site 3 was less accessible than sites 1 and 2 for inhibitor binding as it required more substantial conformational change.

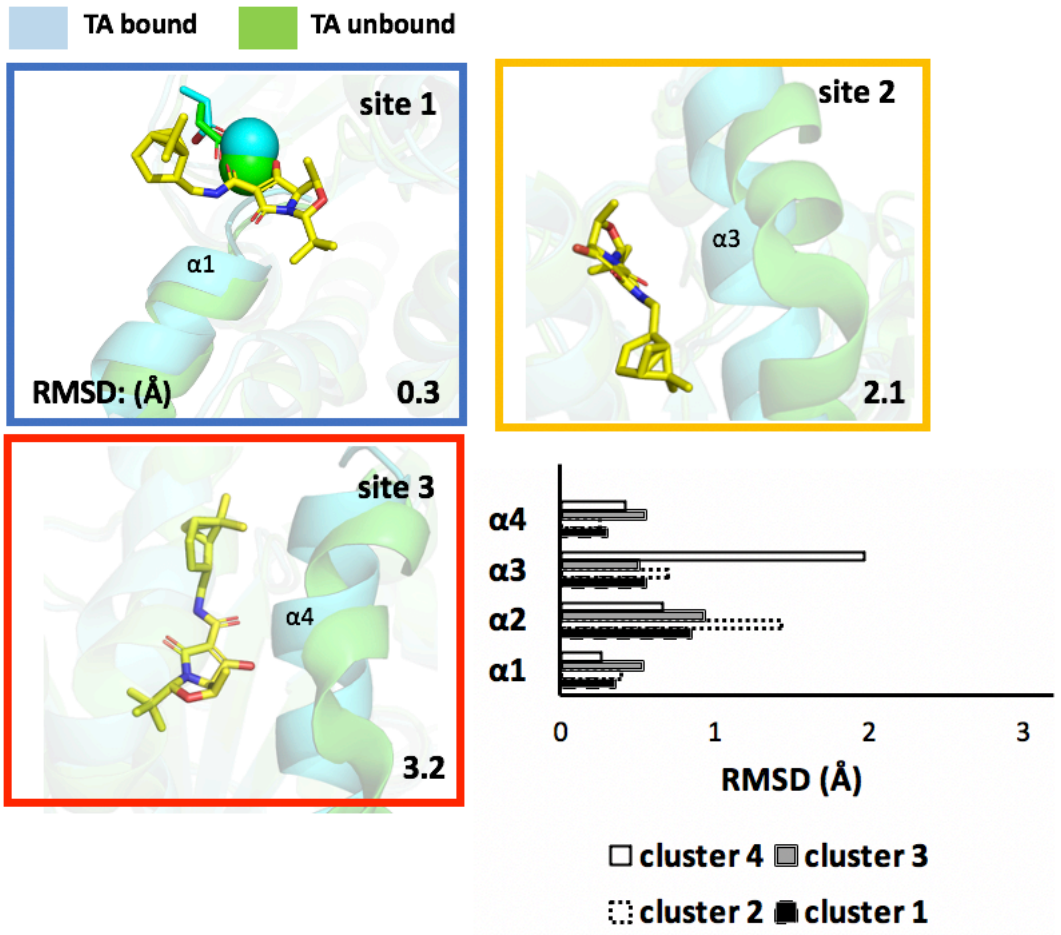


Figure 47: Superposition of the most mobile α helix from the first (cyan ribbon) and last (green ribbon) frames of the SMD trajectory for each site, including its corresponding backbone RMSD. The backbone RMSD of the α helices 1-4 of the highest populated MD clusters of apo-UPPS (relative to the starting crystal structure, PDB ID: **3QAS**), taken from the TREMD simulation.

Moreover, the binding site ranking provided by the US approach was in agreement with that supplied by the MD-LIE method: site 2 (ΔG_{US} -53.3 ± 1.9 kcal/mol) > site 3 (ΔG_{US} -43.3 ± 1.2 kcal/mol) > site 1 (ΔG_{US} -37.4 ± 1.0 kcal/mol) (**Figure 48**). This established further confidence of the affinity model in which site 2 was the preferred TA binding site. However, caution was necessary in the interpretation of the results, which were only meaningful in relative rather than absolute terms.

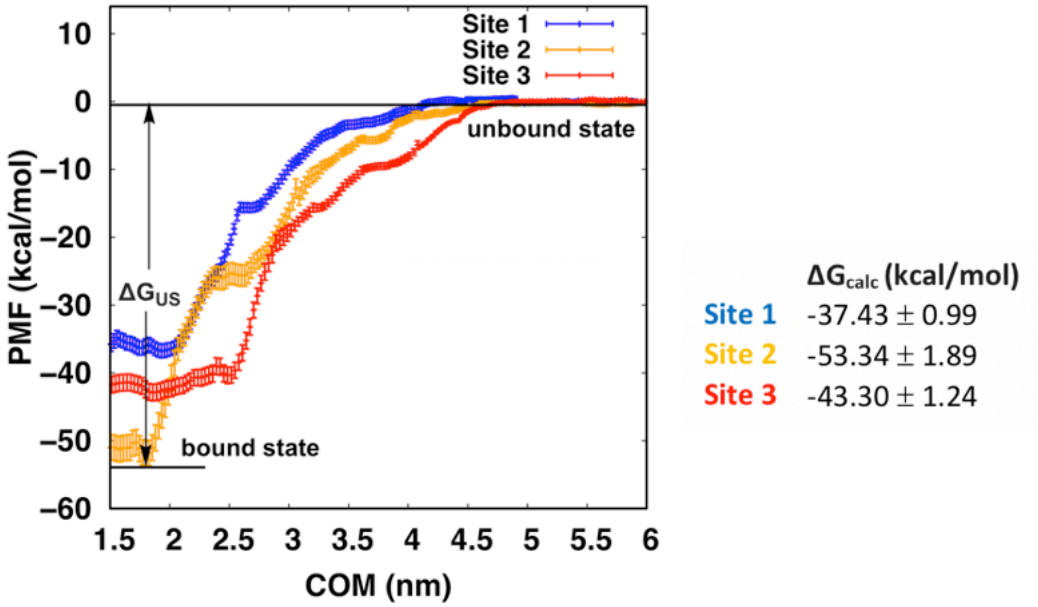


Figure 48: US results for the UPPS-**2A** complexes for sites 1-3. The ΔG_{bind} was computed as the difference between the smallest and lowest values of the PMF curves. The statistical error was evaluated using 100 rounds of bootstrapping analysis.[106,107]

Validating the TA binding site using FEP-MBAR

Using an approach similar to Biggin *et al*, the affinity of the TA **2A** in sites 1-3 was evaluated via alchemical free energy calculations employing atomistic, solvent-explicit MD.[163] Inputs for the FEP calculations were the last frames of the highest-ranked (by the prior LIE analysis) UPPS-**2A** complexes for each site. The binding free energies were computed using a thermodynamic cycle in which the bound and unbound states were interpolated by nonphysical alchemical intermediates (refer to *Methods*).[163,218,228,229]

The results of the FEP-MBAR study indicated that site 2 was the preferred binding site of the TA compound **2A** (refer to *Supporting Info* for full breakdown of the energetic contributions*: Table S11*). While sites 1 and 3 were relatively close in energy (i.e., $\Delta G \cong 3$ kcal/mol), the differences in binding free energy between sites 1 and 2 (i.e., $\Delta G \cong 7$ kcal/mol) and sites 3 and 2 (i.e., $\Delta G \cong 12$ kcal/mol) were more significant.

Although all three methods of free energy calculation demonstrated that site 2 gave rise to the lowest binding free energy, the FEP-MBAR method predicted a different binding site ranking relative to the LIE and US methods. The LIE and US methods predicted that site 2 > site 3 > site 1, while FEP predicted that site 2 (ΔG_{FEP} -22.1 ± 0.3 kcal/mol) > site 1 (ΔG_{FEP} -13.5 ± 0.5 kcal/mol) > site 3 (ΔG_{FEP} -10.4 ± 0.3 kcal/mol). It was acknowledged that further work in FEP/MBAR would be necessary to confirm the precise binding site ranking. This might be achieved by gaining greater accuracy through multiple repetition of the UPPS-**2A** complex calculations, thereby measuring the uncertainty due to finite sampling in addition to the MBAR estimate of uncertainty. Additionally, the calculations might be repeated for other TAs to ensure that this result was not unique to the specific structural framework of the selected TA compound **2A**. In the case of the charged compound, the accuracy of the semi-analytical correction for electrostatic finite size effects could be improved by calculating the residual integrated potential (RIP) for a greater number of frames.[219] This in turn would increase the precision of the binding free energy obtained for site 1.

Combining results of the LIE, US and FEP studies, it was concluded that site 2 was the most probable TA binding site in UPPS. Since site 2 overlapped with the catalytic site (i.e., FPP-binding site), it was further postulated that the TAs are competitive inhibitors of UPPS.

4.3.2. Characterisation of site 2 UPPS-TA complexes

Correlation between pIC_{50} and calculated affinities

Following identification of site 2 as the preferred TA binding site, the compounds **5aD** and **5bD** were added to the LIE investigation, increasing the dataset from three to five TAs (**Figure 49**). A moderate negative correlation (r = -0.6) was observed between the pIC_{50} data and the ΔG_{LIE} values of compounds **1**, **2**, **3**, **5a** and **5b** (**Figure 49**). Although the HID protonation state generally gave rise to more favourable values of ΔG_{LIE} (**Figure S18**), the UPPS-**1A** complex was an exception. In the latter case, MD-LIE analysis and conformational clustering suggested that the preferred protonation state of the active site histidine was HIP, owing to a more favourable interactions profile (**Figure S29**, **S30**). Considering these differences increased the Pearson coefficient from -0.6 to -0.8 (**Figure 49: C**).

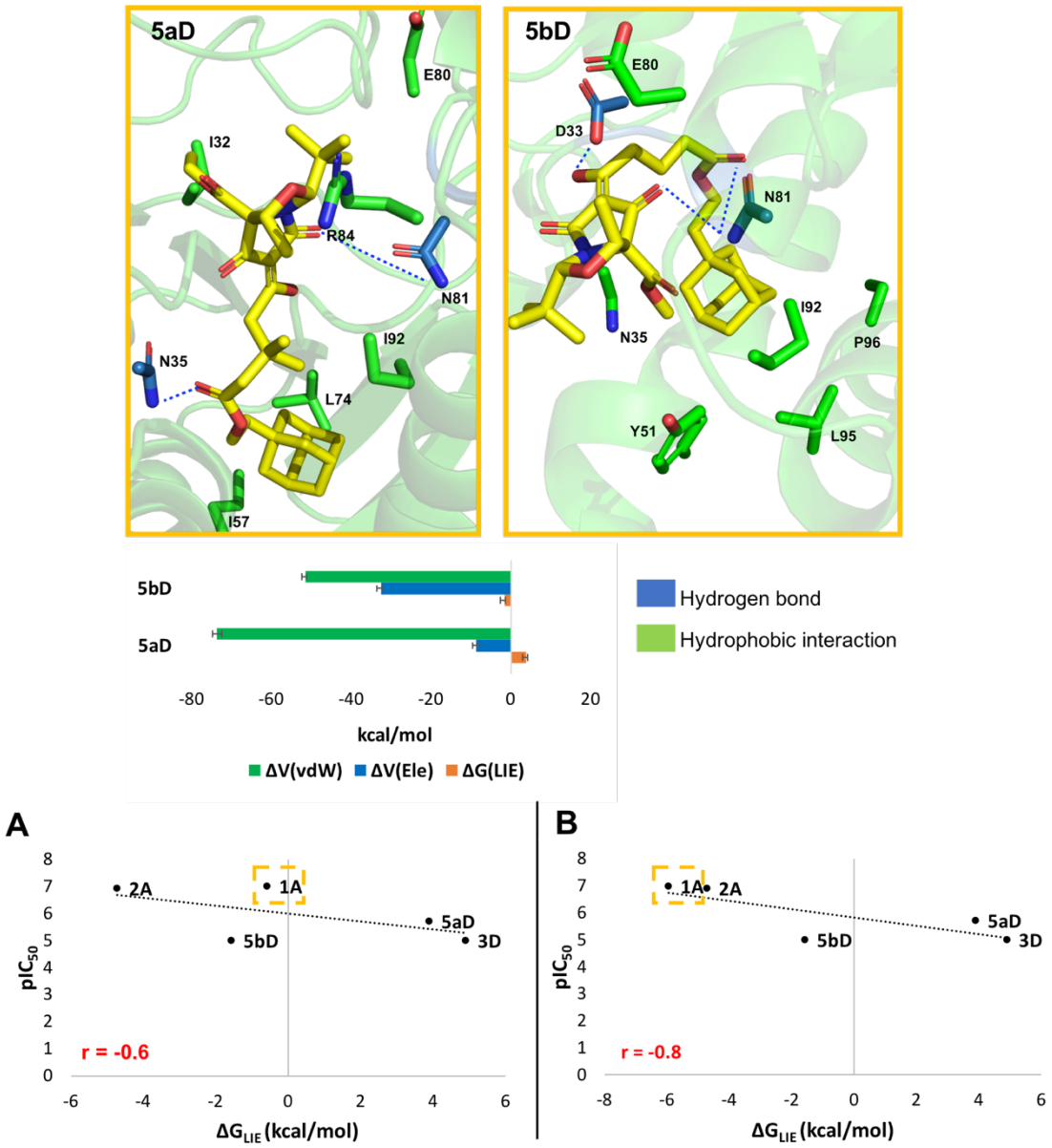


Figure 49: The main hydrogen bond and hydrophobic interactions in the highest populated conformational clusters of the UPPS-5aD and UPPS-5bD trajectories, including their LIE results. Correlation between pIC_{50} and ΔG_{LIE} for compounds 1, 2, 3, 5a and 5b, all in the HID state (**A**). Correlation between pIC_{50} and ΔG_{LIE} for compounds 1, 2, 3, 5a and 5b, all in the HID state excepting UPPS-1A, which was in the HIP state (**B**).

Moreover, it was possible using the LIE method to distinguish between the strong (**1A**, **2A**) and weak binders (**3D**, **5aD**), with the exception of compound **5bD**, which was a weak binder by *in vitro* IC_{50} but a strong binder according to LIE prediction. The precise affinity of **1A** was poorly estimated by the LIE method, which predicted that it was a weaker binder than **2A**, when in fact it exhibited approximately equal affinity according to *in vitro* IC_{50}. The TAs **1A** and **2A** exhibited alternate binding

modes, such that the pendant 3-carboxamide group, bearing the cyclohexyl and myrtanyl groups respectively, pointed out of the hydrophobic tunnel in the case of **1A** and into the tunnel in the case of **2A** (**Figure 45**). Similarly, divergent binding modes were observed for **5aD** and **5bD**, which differed only by the presence of two methyl groups β to the enol, but whose relative affinity for site 2 differed by 5.5 kcal/mol (**Figure 49: A**). This was rationalised by the location of **5aD** (which differed structurally from **5bD** solely by the presence of two additional methyl groups) deeper in the hydrophobic tunnel, giving rise to the formation of one less hydrogen bond than its counterpart **5bD**. (**Figure 50)**.

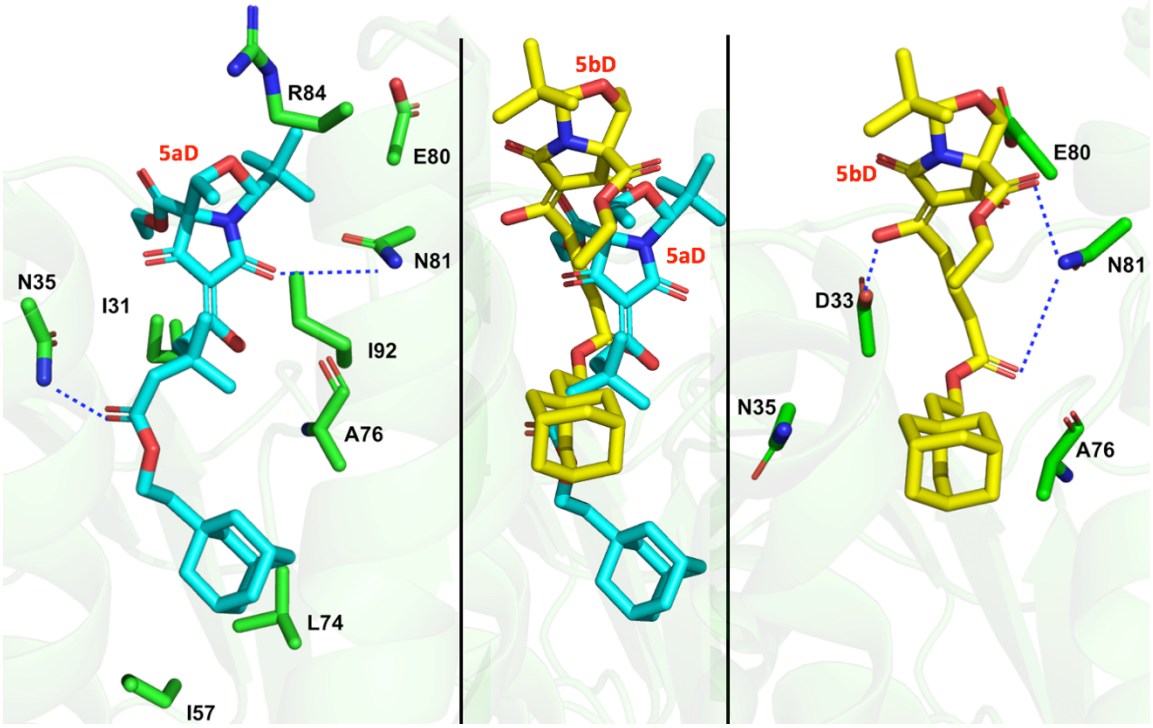


Figure 50: Contrasting the binding modes of **5aD** and **5bD** in site 2. Interacting residues are shown in sticks, with hydrogen bonds indicated (blue dashed line). Sticks shown without the blue dashed line represent hydrophobic interactions.

Conformational dynamics of site 2 UPPS-TA complexes

The global flexibility of the site 2 UPPS-TA complexes was evaluated using PCA. Following this, the volume of the hydrophobic tunnel was measured to characterise local conformational dynamics of the complexes. The effect of differing histidine protonation states was also investigated.

Global Flexibility

Porcupine plots revealed that PC1 of the UPPS-TA complexes showed higher flexibility in the HIP trajectory (**Figure S19**). This was also supported by the larger RMSF of chain A in the HIP simulations (**Figure S2**). Similar to the apoenzyme, the highest mobility was associated with the α helices bridging the hydrophobic tunnel (i.e., α1, α2, α3), with the exception of UPPS-**3D** which exhibited highest mobility in the α5 helix (**Figure S19**). In the UPPS-**1A** and UPPS-**2A** complexes, the greatest motion was associated with the α1 and α2 helices in the HIP state (**Figure S19**). UPPS-**3D** was the only complex to exhibit low mobility of the α helices 1-3 in both the HID and the HIP states of PC1 (**Figure S19**), which might account for the ΔG_{LIE} being equal for the two protonation states. In the UPPS-**5aD** complex, PCA and RMSF measurements revealed that the greatest motion was associated with α helices 2-4 in the HIP state, while for the complex UPPS-**5bD**, it was the α1 helix.

Thus, the identity of the TA and the differing active site histidine protonation states were found to affect the global conformational dynamics of the UPPS-TA complexes. The HIP complexes showed greatest global mobility. This was confirmed by the relative RMSF (**Figure S20**) and the differing PC1 motions revealed through PCA (**Figure S19, Figure S21**).

Local flexibility – hydrophobic tunnel

PVA results indicated that the TAs occupied closed conformational states of UPPS irrespective of the histidine protonation state, thereby reinforcing McCammon's observations that TAs prefer the UPPS closed state.[1] Hydrophobic tunnel volumes of the UPPS-TA complexes varied within the range of closed states 260-500 Å^3

(**Figure S22**), depending on the identity of the TA. Additionally, it was discovered that the HIP complexes were slightly larger than their HID counterparts (**Figure S22**). Thus, the identity of the TA and the histidine protonation state were found to yield subtle variations in the flexibility of the hydrophobic tunnel (i.e., as evidenced by differences in the closed state mean volumes).

Moreover, no statistically significant correlation was found between TA affinity and the pocket volume (**Figure 51: B**). The only noteworthy correlation was that observed between the measured pocket volume and the rotatable bond count (RB; **Figure 51: A**), which suggested that the pocket size increased as the number of rotatable bonds increased. This rationalised the observation that the TA **2A**, possessing the lowest RB, occupied the smallest pocket volume (264 $\pm$ 27 Å^3; **Figure S22**). It was concluded that the all UPPS-TA complexes adopted closed state conformations, with the HIP complexes showing slightly higher flexibility of the hydrophobic tunnel.

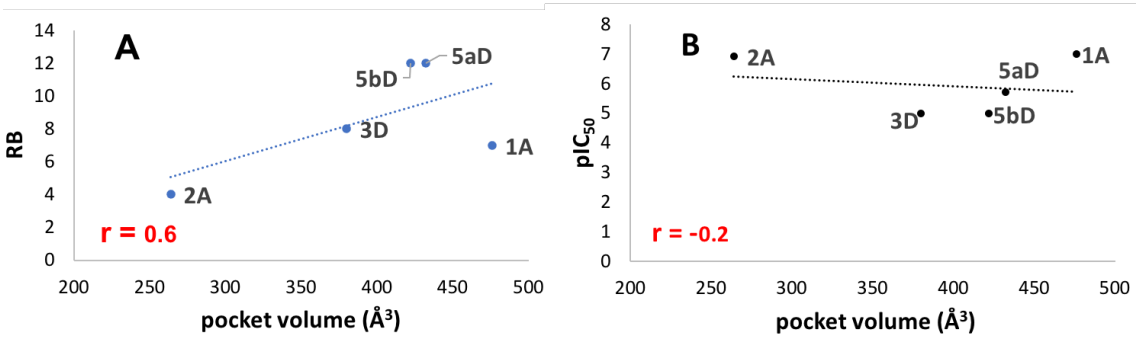


Figure 51: Moderate correlation between the mean pocket volume and the rotatable bond count (RB) of all UPPS-TA complexes (**A**). Negligible correlation between the pIC_{50} and the pocket volume of all UPPS-TA complexes (**B**).

Dynamics of the "flexible loop"

Evaluation of the conformational dynamics of the UPPS-TA complexes would not be complete without specific treatment of the flexible loop. The latter was critical for interconversion of the open and closed states via manipulation of the α3 helix

bridging the binding site.[2,62] The RMSD of the flexible loop (**Figure S23**), as well as the RMSF per residue (**Figure S2**), was found to be approximately equivalent for the HIP and HID trajectories of the apoenzyme (PDB ID: **3QAS**). This indicated that the histidine protonation state did not have a significant effect on the relative magnitude of the flexible loop's mobility in the absence of a ligand. However, it was shown that complexation affected the flexible loop dynamics. The closed conformation of UPPS (PDB ID: **4H8E**) exhibited lower RMSD of the flexible loop than the apoenzyme, while the closed UPPS-FPP complex exhibited the lowest RMSD of the flexible loop (**Figure S24**). Thus, it was concluded that the flexible loop was stabilised by the adoption of the closed conformation and even further stabilised by the presence of a ligand.

Upon examination of the UPPS-TA complexes, the mean RMSD of the flexible loop was approximately equivalent for the two histidine protonation states, except for the UPPS-**5bD** complex (**Figure S23)**. However, differences in the RMSF of the conserved flexible loop residues were recorded (**Figure S25**). Lower RMSF was observed for the HID simulations, except in the case of UPPS-**1A** and UPPS-**3D.** Although a larger TA subset would be required for confirmation, this pattern suggested that the protonation state associated with highest TA affinity was also associated with a lower RMSF of the flexible loop. Taking the UPPS-**1A** complex as an example, the HIP protonation state gave rise to the most favourable value of ΔG_{LIE} according to the MD-LIE analysis, hence the lower RMSF of the flexible loop in the HIP simulation was consistent with this deduction. (**Figure S25**).

Moreover, a strong positive correlation (r = **0.90**) was observed between the LIE-determined free energies of binding and the time-averaged RMSD of the flexible loop in all UPPS-TA complexes (**Figure 52**). Additionally, a strong negative correlation (r = **-0.83**) was observed between the pIC_{50} data and the RMSD of the flexible loop of the complexes (**Figure 52**). Thus, it was found that the flexible loop exhibited least fluctuation (i.e., the highest stabilization) in the highest-ranking UPPS-TA complexes, thereby indicating that increased affinity for UPPS gave rise to lower mobility of the flexible loop.

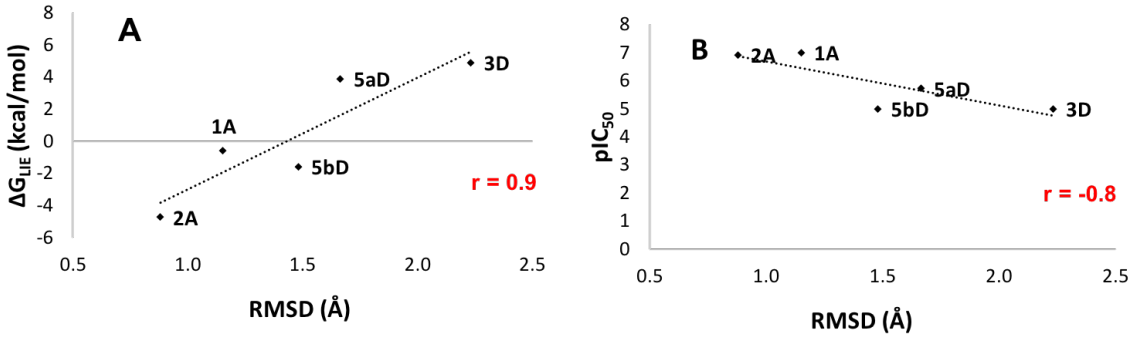


Figure 52: Correlation between ΔG_{LIE} and the RMSD of the flexible loop of all UPPS-TA complexes (**A**). Correlation between the pIC_{50} and the RMSD of the flexible loop of all UPPS-TA complexes (**B**).

Conformation of the α3 helix

Study of the flexible loop required consideration of the conformation of the α3 helix, which was also critical to interconversion of the open and closed states. An important difference in the conformation of the α3 helix in the apo and holo closed states was discovered. Specifically, the highest populated apo MD clusters of closed volume (i.e., 300-400 Å^3) were found to adopt a kinked conformation of the α3 helix (**Figure 53: C**) distinct from its conformations in the crystallographic FPP-bound UPPS (PDB ID: **4H8E**, **1X06**) and the UPPS-TA complexes. Ensemble docking of the TA **2A** into these closed state apo clusters did not reproduce the binding mode of FPP (**Figure 53: B**), in spite of the fact that the apo clusters possessed volumes in the range of the FPP-bound UPPS structure (**Figure 53**: 304

$\pm$ 38 Å^3). Contrastingly, docking of the compound **2A** into the FPP-bound closed structures (PDB ID: **4H8E**, **1X06**) accurately reproduced the FPP binding mode in the catalytic site (**Figure 53: A**). In contrast to McCammon's observations, this suggested that the active site volume was not the sole determinant of TA binding to UPPS.[1]

Moreover, it was found that complexation to FPP induced a conformational change in the α3 helix which was absent from the intrinsic conformational dynamics of the apoenzyme (**Figure 53: C**). This conformational change appeared to be necessary for binding of the TAs to site 2 in the binding mode of FPP. Thus, while it appears to be true that structurally diverse inhibitors recognise a specific set of receptor conformational states, it seems that pocket volume is not a sufficient descriptor of these distinct states. It is likely that inhibitors recognise not only a characteristic volume of the binding site but also induce a specific conformation of the α3 helix (**Figure 53: C**).

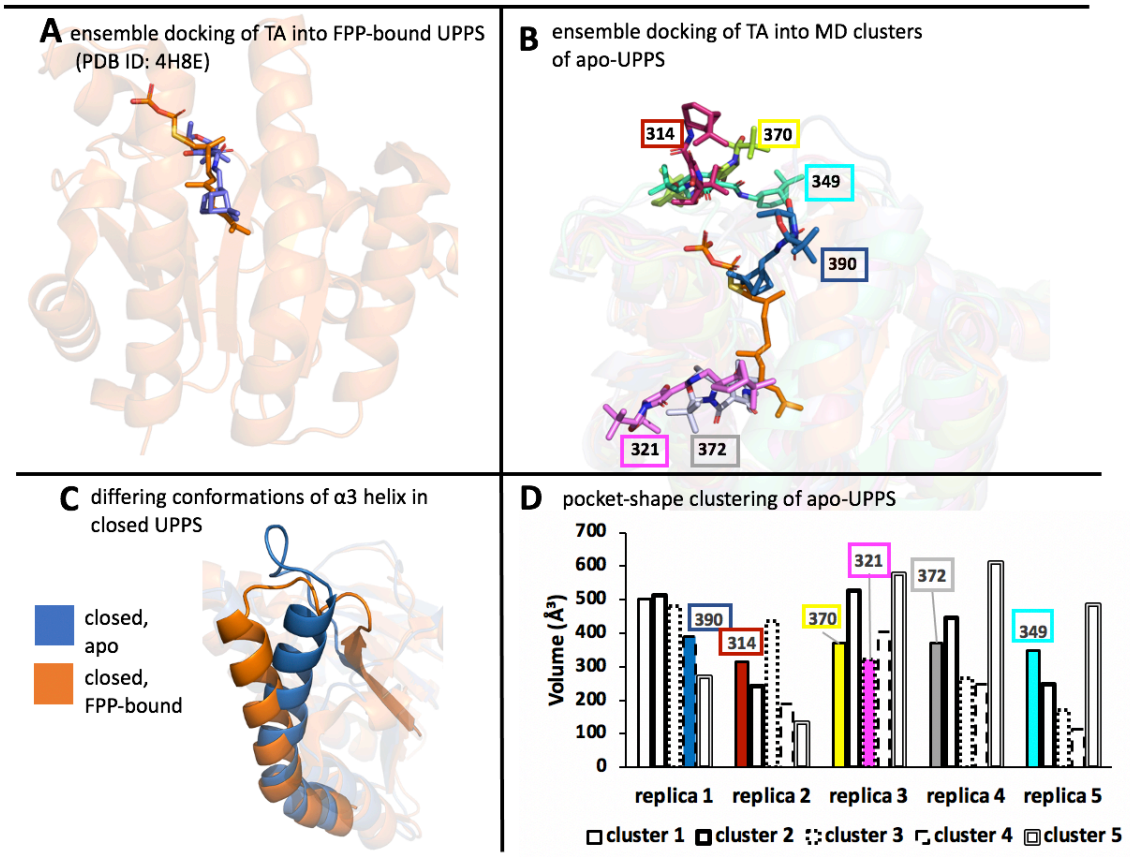


Figure 53: Closed holo state (PDB ID: **1X06**) (**A**). The superposition of closed apo cluster centroids in complex with the highest-ranked poses of **2A** from the ensemble docking study (**B**). The superposition of FPP-bound UPPS (PDB ID: **4H8E**, orange ribbon) and a representative closed state from the apo-HID simulations (cyan ribbon), highlighting the difference in the conformation of the α3 helix (**C**). The volumes of the five highest populated conformational states extracted from each of the five independent 100 ns MD simulations of apo (HID)-UPPS through clustering by POVME (**D**).[12] The clusters whose pocket volumes fell within the FPP-bound range, and which were used in the ensemble docking study, were highlighted (**D**).

Conclusions

In this work, MD-based free energy calculations were employed in the computational prediction of the affinity profile of TA-based ligands for the UPPS enzyme, without prior knowledge of the crystallographic structure of the UPPS-TA complexes. Such predictions remain a major challenge in drug design, but with systematic MD investigations, guided by experimental binding data, these problems can be effectively tackled. It is demonstrated how MD-based approaches, in the cases considered here, reveal important structural insights into UPPS-TA binding. What follows is a summary of the most noteworthy conclusions.

Effect of differing active site histidine protonation states on UPPS-TA binding

HIP is the most probable protonation state of the active site histidine in apo-UPPS, while in the UPPS-TA complexes (site 2), the state appears to depend on the identity of the TA ligand (**Figure S18**). Differing protonation states of the active site histidine influence the local dynamics of the flexible loop (**Figure 29**) and key active site residues (**Figure 30**) in the apoenzyme. Such conformational and dynamical differences between protonation states also affected the interaction profiles and the relative calculated affinities of the TAs for site 2 of UPPS (e.g., compound **2A**, **Figure 40**).

Identification of the most probable TA binding site within UPPS

Following a series of extensive docking and MD studies, it was found that the TAs (i.e., compounds **1-3**) showed the highest affinities for site 2 (the catalytic site) according to LIE, US and FEP free energy approaches. Thus, all three computational methods supported the competitive model of inhibition, in which the free (i.e., non-metal chelated) TA binds to the catalytic site. Furthermore, the calculated

affinities of the TAs for the catalytic site were found to be in good alignment with the experimental binding data (**Figure 49**).

Conformational dynamics of the UPPS-TA complexes

In the absence of the TAs, closure of the hydrophobic tunnel was shown to be central to the conformational dynamics, while in the presence of the TAs, this closed structure was preserved (i.e., throughout long MD simulations). This suggested that the TA inhibitors preferred closed conformational states of UPPS, consistent with the observations of McCammon.[1] However, the local flexibility of the hydrophobic tunnel was found to depend on the identity of the TA, since different ligands produced variations in the mean volume (**Figure S22**). The correlation between the rotatable bond count of the selected TAs and the volume of the hydrophobic tunnel suggested that an increase in the inhibitor flexibility elicited an increase in the flexibility of the binding site (**Figure 51**).

Moreover, it was found that UPPS conformations which were globally similar (i.e., approx. equal volumes of the hydrophobic tunnel) were characterised by local differences in the conformation of the α-3 helix (**Figure 32; 53: C**). For example, complexation of UPPS to FPP induced a conformational change in the α3 helix which was different to that seen in the equivalent closed apo state (**Figure 53: C**). This specific local change in conformation appears to be necessary for binding of the TAs to site 2 in the mode of the native substrate FPP. Additionally, it was discovered that an increase in the UPPS affinity of the TA was correlated with a decrease in the RMSD of the flexible loop, which is involved in the interconversion of the open and closed conformational states (**Figure 52**). This suggests that higher affinity TAs exert a stabilising effect on the closed conformational state.

Finally, although the results of the present study are encouraging, further investigation is required to evaluate the effectiveness of these predictions. Since LIE is computationally cheap, and a good correlation with the IC_{50} binding data was achieved for the cases reported here, it is envisaged that this combined docking and MD-LIE approach could be used to inform future TA-inhibitor design efforts.

Future Work

Future work might center on two different tasks. Firstly, FEP calculations (representing the highest level of rigor) would need to be performed on a larger set of structurally diverse (dual inhibiting) TAs, in order to confirm the correlation between experimental binding affinity (i.e., pIC_{50}) and calculated affinities (i.e., ΔG_{LIE}) for site 2. This would further validate our proposal that the TAs exhibit a competitive mode of inhibition of UPPS. Secondly, the binding of the TAs in RNAP could be investigated using an analogous computational approach (i.e., MD simulations and free energy calculations. These structural investigations could be guided by the X-ray crystal structures of RNAP holoenzyme in complex with the 3-acyl TA containing antibiotic streptolydigin (PDB: **1ZYR, 2A6H**). This second series of MD studies would allow for the modes of inhibition of the TAs to be fully elucidated in both targets. However, access to far greater computational resources will be necessary to simulate RNAP (holoenzyme residue count ≈ 7590) relative to UPPS (total residue count ≈ 512) owing to its large size.

References

1 S. R. Palumbi, H. A. Mooney, J. Lubchenco and J. M. Melillo, *Science*, 2001, **293**, 1786–90.

2 D. A. Gray and M. Wenzel, *ACS Infect. Dis.*, 2020, **6**, 1346–1365.

3 K. R. Hardie, *Emerg. Top. Life Sci.*, 2020, **4**, 129–136.

4 S. F. Nadeem, U. F. Gohar, S. F. Tahir, H. Mukhtar, S. Pornpukdeewattana, P. Nukthamna, A. M. Moula Ali, S. C. B. Bavisetty and S. Massa, *Crit. Rev. Microbiol.*, 2020, **46**, 1–22.

5 C. Årdal, M. Balasegaram, R. Laxminarayan, D. McAdams, K. Outterson, J. H. Rex and N. Sumpradit, *Nat. Rev. Microbiol.*, 2020, **18**, 267–274.

6 L. L. Silver, *Nat. Rev. Drug Discov.*, 2007, **6**, 41–55.

7 D. Du, X. Wang-Kan, A. Neuberger, H. W. van Veen, K. M. Pos, L. J. V. Piddock and B. F. Luisi, *Nat. Rev. Microbiol.*, 2018, **16**, 523–539.

8 K. Bush and P. A. Bradford, *Nat. Rev. Microbiol.*, 2019, **17**, 295–306.

9 M. Tyers and G. D. Wright, *Nat. Rev. Microbiol.*, 2019, **17**, 141–155.

10 A. R. M. Coates, Y. Hu, J. Holt and P. Yeh, *Expert Rev. Anti. Infect. Ther.*, 2020, **18**, 5–15.

11 H. Brötz-Oesterhelt and N. A. Brunner, *Curr. Opin. Pharmacol.*, 2008, **8**, 564–573.

12 M. Gajdács, *Molecules*, 2019, **24**, 892.

13 D.-J. Scheffers and M. G. Pinho, *Microbiol. Mol. Biol. Rev.*, 2005, **69**, 585–607.

14 K. Drlica, M. Malik, R. J. Kerns and X. Zhao, *Antimicrob. Agents Chemother.*, 2008, **52**, 385–392.

15 N. Mani, C. H. Gross, J. D. Parsons, B. Hanzelka, U. Müh, S. Mullin, Y. Liao, A. L. Grillot, D. Stamos, P. S. Charifson and T. H. Grossman, *Antimicrob. Agents Chemother.*, 2006, **50**, 1228–1237.

16 T. Velkov, K. D. Roberts, R. L. Nation, P. E. Thompson and J. Li, *Future Microbiol.*, 2013, **8**, 711–724.

17 J. L. Arbiser and S. L. Moschella, *J. Am. Acad. Dermatol.*, 1995, **32**, 241–247.

18 J. J. Stepanek, T. Lukežič, I. Teichert, H. Petković and J. E. Bandow, *Biochim. Biophys. Acta - Proteins Proteomics*, 2016, **1864**, 645–654.

19 N. H. Georgopapadakou and A. Bertasso, *Antimicrob. Agents Chemother.*, 1993, **37**, 559–565.

20 R. Domalaon, T. Idowu, G. G. Zhanel and F. Schweizer, *Clin Microbiol Rev*, 2018, **31**, https://doi.org/10.1128/CMR.00077-17.

21 S. Mobashery and M. Johnston, *Biochemistry*, 1987, **26**, 5878–5884.

22 S. Mobashery, M. Johnston and S. A. Lerner, *J. Am. Chem. Soc.*, 1986, **108**, 1685–1686.

23 Q. Li, J. Y. Lee, R. Castillo, M. S. Hixon, C. Pujol, V. R. Doppalapudi, H. M. Shepard, G. M. Wahl, T. J. Lobl and M. F. Chan, *Antimicrob. Agents Chemother.*, 2002, **46**, 1262–1268.

24 G. W. Stone, Q. Zhang, R. Castillo, V. R. Doppalapudi, A. R. Bueno, J. Y. Lee, Q. Li, M. Sergeeva, G. Khambatta and N. H. Georgopapadakou, *Antimicrob. Agents Chemother.*, 2004, **48**, 477–483.

25 R. N. Jones, A. L. Barry and C. Thornsberry, *Antimicrob. Agents Chemother.*, 1989, **33**, 944–950.

26 Z. P. Xiao, X. D. Wang, P. F. Wang, Y. Zhou, J. W. Zhang, L. Zhang, J. Zhou, S. S. Zhou, O. Hui, X. Y. Lin, M. Mustapa, A. Reyinbaike and H. L. Zhu, *Eur. J. Med. Chem.*, 2014, **80**, 92–100.

27 V. Pokrovskaya, V. Belakhov, M. Hainrichson, S. Yaron and T. Baasov, *J. Med. Chem.*, 2009, **52**, 2243–2254.

28 K. K. Wang, L. K. Stone, T. D. Lieberman, M. Shavit, T. Baasov and R. Kishony, *Mol. Biol. Evol.*, 2016, **33**, 492–500.

29 R. Domalaon, D. Ammeter, M. Brizuela, B. K. Gorityala, G. G. Zhanel and F. Schweizer, *Front. Microbiol.*, 2019, **10**, 1556.

30 M. A. Fischbach, *Curr. Opin. Microbiol.*, 2011, **14**, 519–523.

31 G. L. Challis and D. A. Hopwood, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 14555–14561.

32 J. M. Ward and J. E. Hodgson, *FEMS Microbiol. Lett.*, 1993, **110**, 239–242.

33 D. C. Alexander and S. E. Jensen, *J. Bacteriol.*, 1998, **180**, 4068–4079.

34 M. Laws, A. Shaaban and K. M. Rahman, *FEMS Microbiol. Rev.*, 2019, **43**, 490–516.

35 J. L. Hansen, P. B. Moore and T. A. Steitz, *J. Mol. Biol.*, 2003, **330**, 1061–1075.

36 T. P. Ko, Y. K. Chen, H. Robinson, P. C. Tsai, Y. G. Gao, A. P. Chen, A. H. Wang and P. H. Liang, *J. Biol. Chem.*, 2001, **276**, 47474–82.

37 W. Sinko, Y. Wang, W. Zhu, Y. Zhang, F. Feixas, C. L. Cox, D. A. Mitchell, E. Oldfield and J. A. McCammon, *J. Med. Chem.*, 2014, **57**, 5693–5701.

38 K. Fujikura, Y.-W. Zhang, M. Fujihashi, Kunio Miki and T. Koyama, *Biochemistry*, 2003, **42**, 4035–41.

39 J.-J. Pan, L.-W. Yang and P.-H. Liang, *Biochemistry*, 2000, **39**, 13856–13861.

40 Y. Kharel, Y. W. Zhang, M. Fujihashi, K. Miki and T. Koyama, *J. Biol. Chem.*, 2001, **276**, 28459–64.

41 R.-T. Guo, T.-P. Ko, A. P.-C. Chen, C.-J. Kuo, A. H.-J. Wang and P.-H. Liang, *J. Biol. Chem.*, 2005, **280**, 20762–74.

42 S.-Y. Chang, T. Ko, A. P. -C. Chen, A. H. -J. Wang and P. Liang, *Protein Sci.*, 2004, **13**, 971–978.

43 Y.-H. Chen, A. P.-C. Chen, C.-T. Chen, A. H.-J. Wang and P.-H. Liang, *J. Biol. Chem.*, 2002, **277**, 7369–7376.

44 S. Y. Chang, Y. K. Chen, A. H. J. Wang and P. H. Liang, *Biochemistry*, 2003, **42**, 14452–14459.

45 T.-P. Ko, Y.-K. Chen, H. Robinson, P.-C. Tsai, Y.-G. Gao, A. P.-C. Chen, A. H.-J. Wang and P.-H. Liang, *J. Biol. Chem.*, 2001, **276**, 47474–47482.

46 S.-Y. Chang, T.-P. Ko, P.-H. Liang and A. H.-J. Wang, *J. Biol. Chem.*, 2003, **278**, 29298–307.

47 R.-T. Guo, R. Cao, P.-H. Liang, T.-P. Ko, T.-H. Chang, M. P. Hudock, W.-Y. Jeng, C. K.-M. Chen, Y. Zhang, Y. Song, C.-J. Kuo, F. Yin, E. Oldfield and A. H.-J. Wang, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 10022–7.

48 J. D. Durrant, R. Cao, A. A. Gorfe, W. Zhu, J. Li, A. Sankovsky, E. Oldfield and J. A. McCammon, *Chem. Biol. Drug Des.*, 2011, **78**, 323–32.

49 Y. Zhang, F.-Y. Lin, K. Li, W. Zhu, Y.-L. Liu, R. Cao, R. Pang, E. Lee, J. Axelson, M. Hensler, K. Wang, K. J. Molohon, Y. Wang, D. A. Mitchell, V. Nizet and E. Oldfield, *ACS Med. Chem. Lett.*, 2012, **3**, 402–406.

50 W. Zhu, Y. Zhang, W. Sinko, M. E. Hensler, J. Olson, K. J. Molohon, S. Lindert, R. Cao, K. Li, K. Wang, Y. Wang, Y.-L. Liu, A. Sankovsky, C. A. F. de Oliveira, D. A. Mitchell, V. Nizet, J. A. McCammon and E. Oldfield, *Proc. Natl. Acad. Sci.*, 2013, **110**, 123–128.

51 Y. Wang, J. Desai, Y. Zhang, S. R. Malwal, C. J. Shin, X. Feng, H. Sun, G. Liu, R.-T. Guo and E. Oldfield, *ChemMedChem*, 2016, **11**, 2311–2319.

52 M. Jukič, K. Rožman, M. Sova, H. Barreteau and S. Gobec, *Front. Microbiol.*, 2019, **9**, 3322.

53 M. A. Farha, T. L. Czarny, C. L. Myers, L. J. Worrall, S. French, D. G. Conrady, Y. Wang, E. Oldfield, N. C. J. Strynadka and E. D. Brown, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 11048–11053.

54 D. E. Danley, E. T. Baima, M. Mansour, K. F. Fennell, B. A. Chrunyk, J. P. Mueller, S. Liu and X. Qiu, *Protein Sci.*, 2015, **24**, 20–26.

55 N. Concha, J. Huang, X. Bai, A. Benowitz, P. Brady, L. C. Grady, L. H. Kryn, D. Holmes, K. Ingraham, Q. Jin, L. Pothier Kaushansky, L. McCloskey, J. A. Messer, H. O'Keefe, A. Patel, A. L. Satz, R. H. Sinnamon, J. Schneck, S. R. Skinner, J. Summerfield, A. Taylor, J. D. Taylor, G. Evindar and R. A. Stavenger, *J. Med. Chem.*, 2016, **59**, 7299–7304.

56 S. Peukert, Y. Sun, R. Zhang, B. Hurley, M. Sabio, X. Shen, C. Gray, J. Dzink-Fox, J. Tao, R. Cebula and S. Wattanasin, *Bioorg. Med. Chem. Lett.*, 2008, **18**, 1840–1844.

57 S. Peukert, Y. Sun, R. Zhang, B. Hurley, M. Sabio, X. Shen, C. Gray, J. Dzink-Fox, J. Tao, R. Cebula and S. Wattanasin, *Bioorg. Med. Chem. Lett.*, 2008, **18**, 1840–1844.

58 D. Temiakov, N. Zenkin, M. N. Vassylyeva, A. Perederina, T. H. Tahirov, E. Kashkina, M. Savkina, S. Zorov, V. Nikiforov, N. Igarashi, N. Matsugaki, S. Wakatsuki, K. Severinov and D. G. Vassylyev, *Mol. Cell*, 2005, **19**, 655–666.

59 S. Tuske, S. G. Sarafianos, X. Wang, B. Hudson, E. Sineva, J. Mukhopadhyay, J. J. Birktoft, O. Leroy, S. Ismail, A. D. Clark, C. Dharia, A. Napoli, O. Laptenko, J. Lee, S. Borukhov, R. H. Ebright, E. Arnold and E. Arnold, *Cell*, 2005, **122**, 541–52.

60 M. G. Gänzle, *Appl. Microbiol. Biotechnol.*, 2004, **64**, 326–332.

61 T. Watanabe, M. Igarashi, T. Okajima, E. Ishii, H. Kino, M. Hatano, R. Sawa, M. Umekita, T. Kimura, S. Okamoto, Y. Eguchi, Y. Akamatsu and R. Utsumi, *Antimicrob. Agents Chemother.*, 2012, **56**, 3657–3663.

62 Y.-C. Jeong, M. Anwar, Z. Bikadi, E. Hazai and M. G. Moloney, *Chem. Sci.*, 2013, **4**, 1008–1015.

63 E. Bakkeren, M. Diard and W. D. Hardt, *Nat. Rev. Microbiol.*, 2020, **18**, 479–490.

64 M. L. Nadimpalli, S. J. Marks, M. C. Montealegre, R. H. Gilman, M. J. Pajuelo, M. Saito, P. Tsukayama, S. M. Njenga, J. Kiiru, J. Swarthout, M. A. Islam, T. R. Julian and A. J. Pickering, *Nat. Microbiol.*, 2020, **5**, 787–795.

65 L. L. Silver, *Clin. Microbiol. Rev.*, 2011, **24**, 71–109.

66 M. N. Gwynn, A. Portnoy, S. F. Rittenhouse and D. J. Payne, *Ann. N. Y. Acad. Sci.*, 2010, **1213**, 5–19.

67 L. V. Lee, B. Granda, K. Dean, J. Tao, E. Liu, R. Zhang, S. Peukert, S. Wattanasin, X. Xie, N. S. Ryder, R. Tommasi and G. Deng, *Biochemistry*, 2010, **49**, 5366–5376.

68 M. Zaghouani and B. Nay, *Nat. Prod. Rep.*, 2016, **33**, 540–548.

69 J. Markopoulos, G. Athanasellis and O. Igglessi-Markopoulou, *Bioinorg. Chem. Appl.*, 2010, https://doi.org/10.1155/2010/315056.

70 Y.-C. Jeong and M. G. Moloney, *J. Org. Chem.*, 2011, **76**, 1342–1354.

71 B. J. L. Royles, *Chem. Rev.*, 1995, **95**, 1981–2001.

72 R. Schobert and A. Schlenk, *Bioorganic Med. Chem.*, 2008, **16**, 4203–4221.

73 M. H. Lebrun, P. Duvert, F. Gaudemer, A. Gaudemer, C. Deballon and P. Boucly, *J. Inorg. Biochem.*, 1985, **24**, 167–181.

74 R. T. Gallagher, J. L. Richard, H. M. Stahr and R. J. Cole, *Mycopathologia*, 1978, **66**, 31–6.

75 N. Imamura, K. Adachi and H. Sano, *J. Antibiot. (Tokyo).*, 1994, **47**, 257–261.

76 Z. Cvetnić and S. Pepeljnjak, *Nahrung - Food*, 1998, **42**, 321–323.

77 M. Laursen, M. Bublitz, K. Moncoq, C. Olesen, J. V. Møller, H. S. Young, P. Nissen and J. P. Morth, *J. Biol. Chem.*, 2009, **284**, 13513–13518.

78 B. Biersack, R. Diestel, C. Jagusch, F. Sasse and R. Schobert, *J. Inorg. Biochem.*, 2009, **103**, 72–76.

79 J. Jia, F. Zhu, X. Ma, Z. W. Cao, Y. X. Li and Y. Z. Chen, *Nat. Rev. Drug Discov.*, 2009, **8**, 111–128.

80 L. L. Silver, *Clin. Microbiol. Rev.*, 2011, **24**, 71–109.

81 M. Jukic, K. Rozman and S. Gobec, *Curr. Med. Chem.*, 2016, **23**, 464–482.

82 P. Atkins and R. Friedman, *Molecular Quantum Mechanics*, Oxford

University Press, Oxford, United Kingdom, 5th Editio., 2010.

83 L. H. Thomas, *Math. Proc. Cambridge Philos. Soc.*, 1927, **23**, 542–548.

84 E. Fermi, *Zeitschrift für Phys.*, 1928, **48**, 73–79.

85 P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864.

86 W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133.

87 J. P. Perdew, in *AIP Conference Proceedings*, AIP Publishing, 2003, vol. 577, pp. 1–20.

88 C. L. Wennberg, T. Murtola, S. Páll, M. J. Abraham, B. Hess and E. Lindahl, *J. Chem. Theory Comput.*, 2015, **11**, 5737–5746.

89 P. P. Ewald, *Ann. Phys.*, 1921, **64**, 253.

90 S. Haider, G. N. Parkinson and S. Neidle, *Biophys. J.*, 2008, **95**, 296–311.

91 A. Amadei, A. B. M. Linssen and H. J. C. Berendsen, *Proteins Struct. Funct. Bioinforma.*, 1993, **17**, 412–425.

92 A. Amadei, A. B. M. Linssen, B. L. de Groot, D. M. F. van Aalten and H. J. C. Berendsen, *J. Biomol. Struct. Dyn.*, 1996, **13**, 615–625.

93 H. Yamaguchi, D. M. F. Van Aalten, M. Pinak, A. Furukawa and R. Osman, *Nucleic Acids Res.*, 1998, **26**, 1939–1946.

94 C. C. David and D. J. Jacobs, *Methods Mol. Biol.*, 2014, **1084**, 193–226.

95 M. S. Nair and A. Shukla, *J. Biomol. Struct. Dyn.*, 2019, **38**, 3087–3097.

96 D. Sindhikara, Y. Meng and A. E. Roitberg, *J. Chem. Phys.*, 2008, **128**, 024103.

97 Y. I. Yang, Q. Shao, J. Zhang, L. Yang and Y. Q. Gao, *J. Chem. Phys.*, 2019, **151**, 70902.

98 W. You, Z. Tang and C. E. A. Chang, *J. Chem. Theory Comput.*, 2019, **15**, 2433–2443.

99 C. Domene and S. Furini, *Methods Enzymol.*, 2009, **466**, 155–177.

100 J. Kästner, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2011, **1**, 932–942.

101 K. Murata, Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, 2004, **385**, 1–7.

102 Y. Sugita, A. Kitao and Y. Okamoto, *J. Chem. Phys.*, 2000, **113**, 6042–6051.

103 D. Sabri Dashti and A. E. Roitberg, *J. Chem. Theory Comput.*, 2013, **9**, 4692–4699.

104 S. Jo, D. Suh, Z. He, C. Chipot and B. Roux, *J. Phys. Chem. B*, 2016, **120**, 8733–8742.

105 S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, *J. Comput. Chem.*, 1992, **13**, 1011–1021.

106 J. S. Hub, B. L. De Groot and D. Van Der Spoel, *J. Chem. Theory Comput.*, 2010, **6**, 3713–3720.

107 M. N. Kobrak, *J. Comput. Chem.*, 2003, **24**, 1437–1446.

108 Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, 1999, **314**, 141–151.

109 R. Qi, G. Wei, B. Ma and R. Nussinov, in *Methods in Molecular Biology*, Humana Press Inc., 2018, vol. 1777, pp. 101–119.

110 D. J. Earl and M. W. Deem, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3910–3916.

111 M. J. Abraham and J. E. Gready, *J. Chem. Theory Comput.*, 2008, **4**, 1119–1128.

112 X. Periole and A. E. Mark, *J. Chem. Phys.*, 2007, **126**, 014903.

113 D. J. Sindhikara, D. J. Emerson and A. E. Roitberg, *J. Chem. Theory Comput.*, 2010, **6**, 2804–2808.

114 H. Fukunishi, O. Watanabe and S. Takada, *J. Chem. Phys.*, 2002, **116**, 9058–9067.

115 R. Affentranger, I. Tavernelli and E. E. Di Iorio, *J. Chem. Theory Comput.*, 2006, **2**, 217–228.

116 S. K. Mishra, M. Kara, M. Zacharias and J. Koča, *Glycobiology*, 2014, **24**, 70–84.

117 G. Bussi, *Mol. Phys.*, 2014, **112**, 379–384.

118 P. Liu, B. Kim, R. A. Friesner and B. J. Berne, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 13749–13754.

119 M. R. Shirts and D. L. Mobley, in *Biomolecular Simulations*, Humana Press, Totowa, NJ, 2013, pp. 271–311.

120 C. Chipot and A. Pohorille, *Springer Ser. Chem. Phys.*, 2007, **86**, 33–75.

121 M. K. Gilson, J. A. Given, B. L. Bush and J. A. Mccammon, *The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review*, 1997, vol. 72.

122 M. Aldeghi, J. P. Bluck and P. C. Biggin, in *Methods in Molecular Biology*, Humana Press Inc., 2018, vol. 1762, pp. 199–232.

123 C. Chipot, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2014, **4**, 71–89.

124 T. Steinbrecher and A. Labahn, *Curr. Med. Chem.*, 2010, **17**, 767–785.

125 M. Almlöf, A. Jens Carlsson and J. Åqvist, *J. Chem. Theory Comput.*, 2007, **3**, 2162–2175.

126 T. Hansson, J. Marelius and J. Åqvist, *J. Comput. Aided. Mol. Des.*, 1998, **12**, 27–35.

127 H. Gutiérrez-de-Terán and J. Åqvist, *Methods Mol. Biol.*, 2012, **819**, 305–323.

128 T. Hou, J. Wang, Y. Li and W. Wang, *J. Comput. Chem.*, 2011, **32**, 866–877.

129 R. Kumari, R. Kumar and A. Lynn, *J. Chem. Inf. Model.*, 2014, **54**, 1951–1962.

130 L. Xu, H. Sun, Y. Li, J. Wang and T. Hou, *J. Phys. Chem. B*, 2013, **117**, 8408–8421.

131 M. S. Lee and M. A. Olson, *Biophys. J.*, 2006, **90**, 864–877.

132 H. J. Woo and B. Roux, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 6825–6830.

133 L. Sutto, S. Marsili and F. L. Gervasio, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2012, **2**, 771–779.

134 A. Cavalli, A. Spitaleri, G. Saladino and F. L. Gervasio, *Acc. Chem. Res.*, 2015, **48**, 277–285.

135 S. Makino and I. D. Kuntz, *J. Comput. Chem.*, 1997, **18**, 1812–1825.

136 M. Rarey, B. Kramer, T. Lengauer and G. Klebe, *J. Mol. Biol.*, 1996, **261**, 470–489.

137 D. S. Goodsell and A. J. Olson, *Proteins Struct. Funct. Bioinforma.*, 1990, **8**, 195–202.

138 O. Trott and A. J. Olson, *J. Comput. Chem.*, 2009, **31**, 455–461.

139 G. Jones, P. Willett and R. C. Glen, *J. Mol. Biol.*, 1995, **245**, 43–53.

140 G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *J. Mol. Biol.*, 1997, **267**, 727–748.

141 W. T. M. Mooij and M. L. Verdonk, *Proteins Struct. Funct. Genet.*, 2005, **61**, 272–287.

142 B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, 1983, **4**, 187–217.

143 W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *J. Am.*

Chem. Soc., 1996, **118**, 2309–2309.

144 W. L. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.*, 1988, **110**, 1657–1666.

145 M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini and R. P. Mee, *J. Comput. Aided. Mol. Des.*, 1997, **11**, 425–445.

146 H. J. Böhm, *J. Comput. Aided. Mol. Des.*, 1994, **8**, 243–256.

147 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.

148 E. A. Rifai, M. van Dijk and D. P. Geerke, *Front. Mol. Biosci.*, 2020, **7**, 114.

149 J. Aqvist and J. Marelius, *Comb. Chem. High Throughput Screen.*, 2001, **4**, 613–626.

150 E. Stjernschantz and C. Oostenbrink, *Biophys. J.*, 2010, **98**, 2682–2691.

151 J. Åqvist, C. Medina and J. E. Samuelsson, *Protein Eng. Des. Sel.*, 1994, **7**, 385–391.

152 J. Aqvist, V. B. Luzhkov and B. O. Brandsdal, *Acc. Chem. Res.*, 2002, **35**, 358–365.

153 R. Baron, *Computational Drug Discovery and Design*, Springer, New York, NY, 2011.

154 J. Å. and and T. Hansson, *J. Phys. Chem.*, 1996, **100**, 9512–9521.

155 M. van Dijk, A. M. ter Laak, J. D. Wichard, L. Capoferri, N. P. E. Vermeulen and D. P. Geerke, *J. Chem. Inf. Model.*, 2017, **57**, 2294–2308.

156 L. Perić-Hassler, E. Stjernschantz, C. Oostenbrink and D. P. Geerke, *Int. J. Mol. Sci.*, 2013, **14**, 24514–24530.

157 C. R. Vosmeer, R. Pool, M. F. Van Stee, L. Peric-Hassler, N. P. E. Vermeulen and D. P. Geerke, *Int. J. Mol. Sci.*, 2014, **15**, 798–816.

158 A. Nunes-Alves and G. M. Arantes, *J. Chem. Inf. Model.*, 2014, **54**, 2309–2319.

159 L. Capoferri, M. van Dijk, A. S. Rustenburg, T. A. Wassenaar, D. P. Kooi, E. A. Rifai, N. P. E. Vermeulen and D. P. Geerke, *J. Cheminform.*, 2017, **9**, 58.

160 Y. Meng, D. Sabri Dashti and A. E. Roitberg, *J. Chem. Theory Comput.*, 2011, **7**, 2721–2727.

161 W. Jiang and B. Roux, *J. Chem. Theory Comput.*, 2010, **6**, 2559–2565.

162 P. V. Klimovich, M. R. Shirts and D. L. Mobley, *J. Comput. Aided. Mol. Des.*,

2015, **29**, 397–411.

163 M. Aldeghi, A. Heifetz, M. J. Bodkin, S. Knapp and P. C. Biggin, *Chem. Sci.*, 2016, **7**, 207–218.

164 R. W. Zwanzig, J. G. Kirkwood, I. Oppenheim and B. J. Alder, *J. Chem. Phys.*, 1954, **22**, 783–790.

165 C. H. Bennett, *J. Comput. Phys.*, 1976, **22**, 245–268.

166 M. R. Shirts, D. L. Mobley and S. P. Brown, in *Drug Design*, Cambridge University Press, 2010, pp. 61–86.

167 M. R. Shirts, E. Bair, G. Hooker and V. S. Pande, *Phys. Rev. Lett.*, 2003, **91**, 140601.

168 D. L. Mobley, J. D. Chodera and K. A. Dill, *J. Chem. Phys.*, 2006, **125**, 84902.

169 I. J. General, *J. Chem. Theory Comput.*, 2010, **6**, 2520–2524.

170 S. Boresch, F. Tettinger, M. Leitgeb and M. Karplus, *J. Phys. Chem. B*, 2003, **107**, 9535–9551.

171 C. J. Woods, J. W. Essex and M. A. King, *J. Phys. Chem. B*, 2003, **107**, 13711–13718.

172 C. J. Woods, J. W. Essex and M. A. King, *J. Phys. Chem. B*, 2003, **107**, 13703–13710.

173 R. B. Mariutti, J. E. Hernández-González, A. F. Z. Nascimento, M. A. B. de Morais, M. T. Murakami, C. M. A. Carareto and R. K. Arni, *Biochim. Biophys. Acta - Gen. Subj.*, 2020, **1864**, 129597.

174 A. M. Potshangbam, R. S. Rathore and P. Nongdam, *Heliyon*, 2020, **6**, e03378.

175 X. Chen and S. D. Schwartz, *ACS Catal.*, 2019, **9**, 9870–9879.

176 J. Ricci-López, A. Vidal-Limon, M. Zunñiga, V. A. Jimènez, J. B. Alderete, C. A. Brizuela and S. Aguila, *PLoS One*, 2019, **14**, e0213028.

177 J. E. Hernández González, L. Hernández Alvarez, P. G. Pascutti and V. B. P. Leite, *J. Phys. Chem. B*, 2019, **123**, 7327–7342.

178 R. Creutznacher, E. Schulze, G. Wallmann, T. Peters, M. Stein and A. Mallagaray, *ChemBioChem*, 2020, **21**, 1007–1021.

179 S. Pérot, O. Sperandio, M. A. Miteva, A. C. Camproux and B. O. Villoutreix, *Drug Discov. Today*, 2010, **15**, 656–667.

180 X. Zheng, L. Gan, E. Wang and J. Wang, *AAPS J.*, 2013, **15**, 228–241.

181 J. R. Wagner, J. Sørensen, N. Hensley, C. Wong, C. Zhu, T. Perison and R. E. Amaro, *J. Chem. Theory Comput.*, 2017, **13**, 4584–4592.

182 J. D. Durrant, L. Votapka, J. Sørensen and R. E. Amaro, *J. Chem. Theory Comput.*, 2014, **10**, 5047–5056.

183 S. R. Ellingson, Y. Miao, J. Baudry and J. C. Smith, *J. Phys. Chem. B*, 2015, **119**, 1026–1034.

184 S. Y. Huang and X. Zou, *Proteins Struct. Funct. Genet.*, 2007, **66**, 399–421.

185 M. Totrov and R. Abagyan, *Curr. Opin. Struct. Biol.*, 2008, **18**, 178–184.

186 W. Sinko, C. de Oliveira, S. Williams, A. Van Wynsberghe, J. D. Durrant, R. Cao, E. Oldfield and J. A. McCammon, *Chem. Biol. Drug Des.*, 2011, **77**, 412–420.

187 V. Le Guilloux, P. Schmidtke and P. Tuffery, *BMC Bioinformatics*, 2009, **10**, 168.

188 P. Schmidtke, V. Le Guilloux, J. Maupetit and P. Tufféry, *Nucleic Acids Res.*, 2010, **38**, W582–W589.

189 D. B. Kokh, S. Richter, S. Henrich, P. Czodrowski, F. Rippmann and R. C. Wade, *J. Chem. Inf. Model.*, 2013, **53**, 1235–1252.

190 D. B. Kokh, P. Czodrowski, F. Rippmann and R. C. Wade, *J. Chem. Theory Comput.*, 2016, **12**, 4100–4113.

191 I. R. Craig, C. Pfleger, H. Gohlke, J. W. Essex and K. Spiegel, *J. Chem. Inf. Model.*, 2011, **51**, 2666–2679.

192 T. Paramo, A. East, D. Garzón, M. B. Ulmschneider and P. J. Bond, *J. Chem. Theory Comput.*, 2014, **10**, 2151–2164.

193 B. Laurent, M. Chavent, T. Cragnolini, A. C. E. Dahl, S. Pasquali, P. Derreumaux, M. S. P. Sansom and M. Baaden, *Bioinformatics*, 2015, **31**, 1478–1480.

194 J. Desaphy, K. Azdimousa, E. Kellenberger and D. Rognan, *J. Chem. Inf. Model.*, 2012, **52**, 2287–2299.

195 J. D. Durrant, C. A. F. De Oliveira and J. A. McCammon, *J. Mol. Graph. Model.*, 2011, **29**, 773–776.

196 S. G. Akl and G. T. Toussaint, in *4th. Int. Joint Conf. on Pattern Recognition (Kyoto, Japan)*, Kyoto, Japan, 1978, pp. 483–487.

197 E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt,

E. C. Meng and T. E. Ferrin, *J. Comput. Chem.*, 2004, **25**, 1605–1612.

198 V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson and D. C. Richardson, *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 2010, **66**, 12–21.

199 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1**–**2**, 19–25.

200 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.

201 H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, A. Dinola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684–3690.

202 B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *J. Comput. Chem.*, 1997, **18**, 1463–1472.

203 M. H. M. Olsson, C. R. SØndergaard, M. Rostkowski and J. H. Jensen, *J. Chem. Theory Comput.*, 2011, **7**, 525–537.

204 X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, A. E. Mark, A. Rigault, J. Siegel, J. Harrowfield, B. Chevrier, D. Moras, J. Lehn, M. Garrett, U. Koert, D. Meyer and J. Fischer, *Angew. Chem. Int. Ed. Engl*, 1998, **31**, 1387–1404.

205 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graph.*, 1996, **14**, 33–38.

206 J. Uranga, P. Mikulskis, S. Genheden and U. Ryde, *Comput. Theor. Chem.*, 2012, **1000**, 75–84.

207 C. R. Søndergaard, M. H. M. Olsson, M. Rostkowski and J. H. Jensen, *J. Chem. Theory Comput.*, 2011, **7**, 2284–2295.

208 X. Mo, Q. Li and J. Ju, *Rsc Adv.*, 2014, **4**, 50566–50593.

209 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, M. C. X. Li, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. J. Montgomery, J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin,

V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16, Revis. C.01, Gaussian, Inc., Wallingford CT*, 2016.

210 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–6170.

211 W. J. Hehre, C. F. Pau, R. W. Taft, A. D. Headley and R. D. Topsom, *J. Am. Chem. Soc.*, 1986, **108**, 1711–1712.

212 S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, **32**, 1456–1465.

213 A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2009, **113**, 6378–6396.

214 A. W. Sousa Da Silva and W. F. Vranken, *BMC Res. Notes*, 2012, **5**, 367.

215 M. R. Shirts and J. D. Chodera, *J. Chem. Phys.*, 2008, **129**, 124105.

216 J. D. Chodera and M. R. Shirts, *J. Chem. Phys.*, 2011, **135**, 194110.

217 P. V Klimovich, M. R. Shirts and D. L. Mobley, *J Comput Aided Mol Des*, 2015, **29**, 397–411.

218 M. A. Olsson, A. T. García-Sosa and U. Ryde, *J. Comput. Aided. Mol. Des.*, 2018, **32**, 211–224.

219 G. J. Rocklin, D. L. Mobley, K. A. Dill and P. H. Hünenberger, *J. Chem. Phys.*, 2013, **139**, https://doi.org/10.1063/1.4826261.

220 N. A. Baker, D. Sept, S. Joseph, M. J. Holst and J. A. McCammon, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 10037–10041.

221 Y.-C. Jeong and M. Moloney, *Synlett*, 2009, **2009**, 2487–2491.

222 V. V. Gromak, V. G. Avakyan and O. F. Lakhvich, *J. Appl. Spectrosc.*, 2000, **67**, 205–215.

223 Y.-C. Jeong, M. Anwar and M. G. Moloney, *Bioorg. Med. Chem. Lett.*, 2014, **24**, 1901–1906.

224 T. D. Panduwawala, S. Iqbal, A. L. Thompson, M. Genov, A. Pretsch, D. Pretsch, S. Liu, R. H. Ebright, A. Howells, A. Maxwell and M. G. Moloney, *Org. Biomol. Chem.*, 2019, **17**, 5615–5632.

225 H. B. Broughton and P. R. Woodward, *J. Comput. Aided. Mol. Des.*, 1990, **4**, 147–153.

226 Y.-C. Jeong, Z. Bikadi, E. Hazai and M. G. Moloney, *ChemMedChem*, 2014, **9**, 1826–1837.

227 Y.-C. Jeong and M. G. Moloney, *Beilstein J. Org. Chem.*, 2013, **9**, 1899–1906.

228 M. Gutiérrez, G. A. Vallejos, M. P. Cortés and C. Bustos, *Chem. Biol. Drug Des.*, 2019, **93**, 1117–1128.

229 A. Pohorille, C. Jarzynski and C. Chipot, *J. Phys. Chem. B*, 2010, **114**, 10235–10253.

Supporting Information

Table S1: Propka pKa predictions of H43 (H50) in various UPPS complexes

Structure	Residue	Buried	Propka	State
apo-UPPS	H43	46	5.46	HID
UPPS-FPP	H50	100	4.43	HID
UPPS-1A	H50	73	5.86	HID
UPPS-2A	H50	68	5.01	HID
UPPS-3D	H50	69	5.03	HID
UPPS-5aD	H50	65	4.56	HID
UPPS-5bD	H50	52	5.18	HID

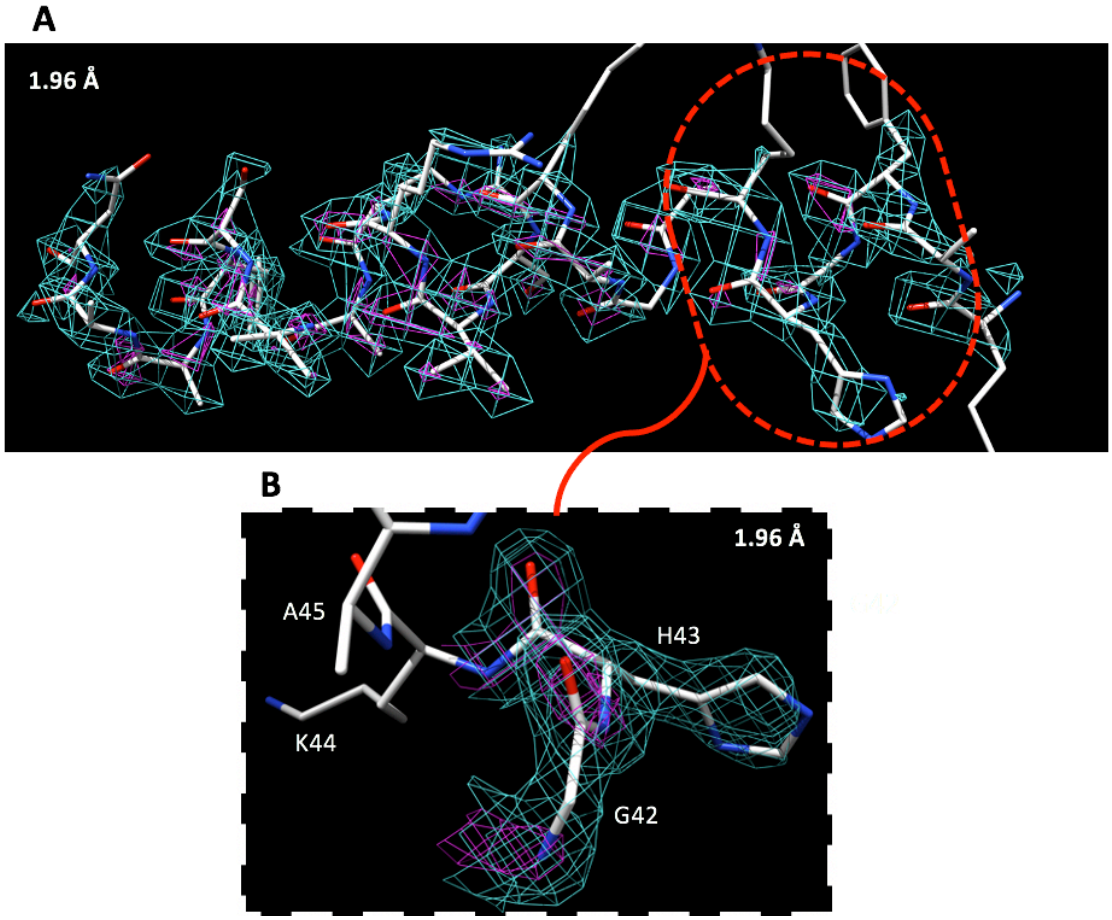


Figure S1: **A**) The electron density map of residues 39-59 (subunit A) of *E. coli* UPPS (PDB IS: **3QAS**), generated using Chimera. **B)** The electron density of residues 42-45 of *E. coli* UPPS (PDB ID: **3QAS**).

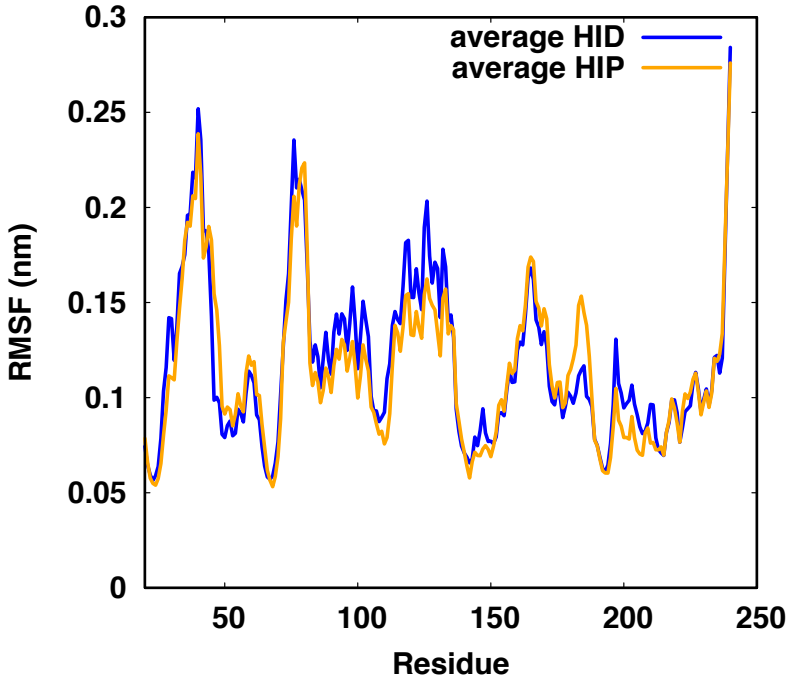


Figure S2: The mean RMSF of subunit A of apo-HIP and apo-HID UPPS, calculated from the concatenated 500 ns MD simulation for each protonation state.

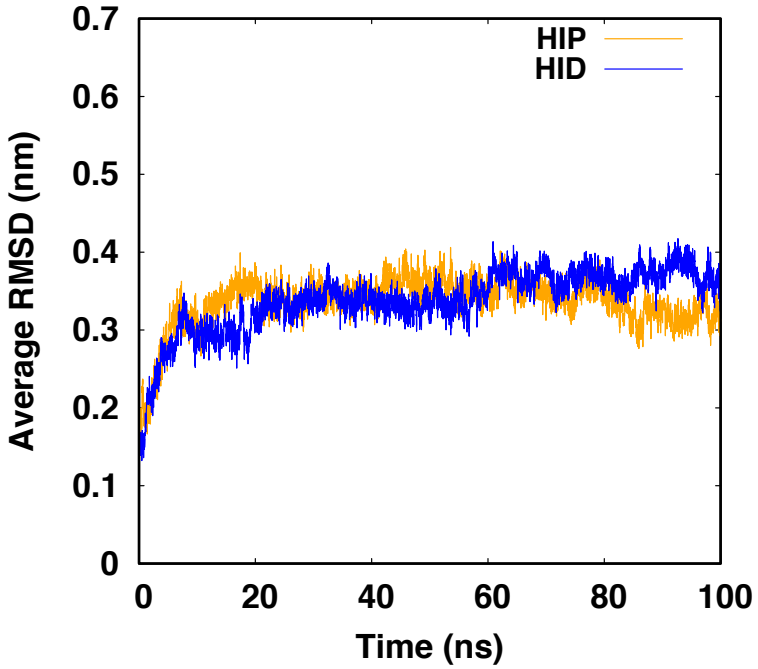


Figure S3: The mean RMSD of R77 in apo-HIP and apo-HID UPPS, calculated from the concatenated 500 ns MD simulation for each protonation state.

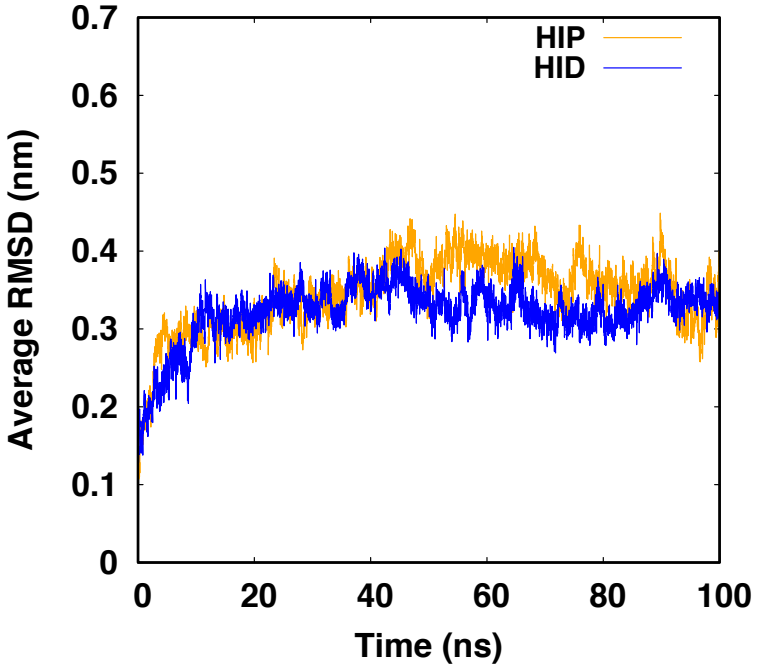


Figure S4: The mean RMSD of R39 in apo-HIP and apo-HID UPPS, calculated from the concatenated 500 ns MD simulation for each protonation state.

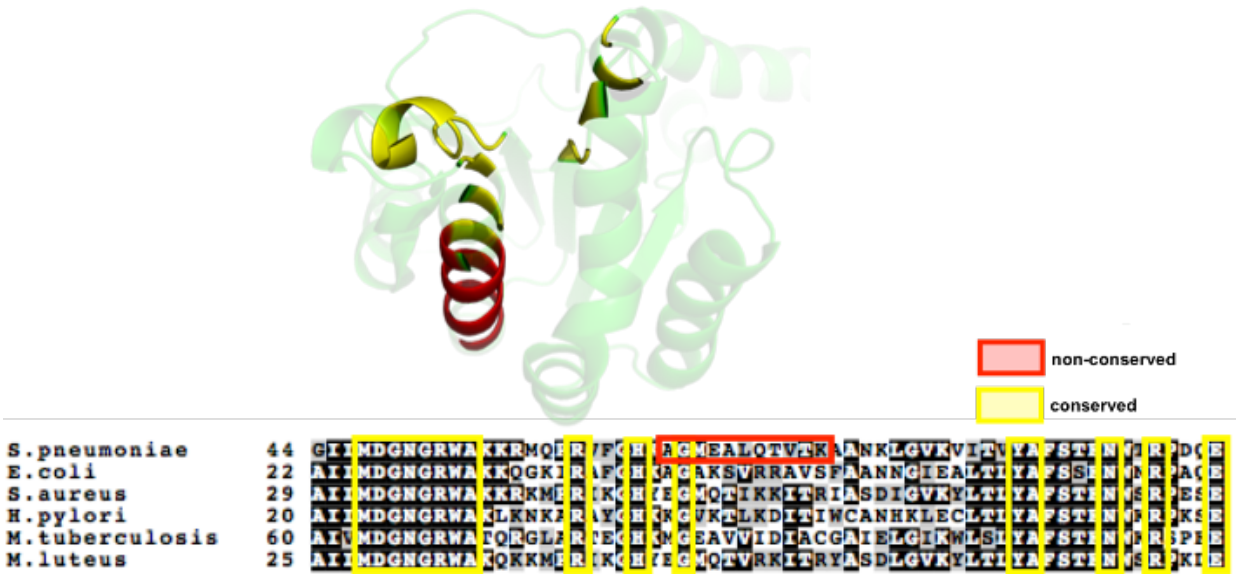


Figure S5: Conserved residues (yellow) of the catalytic site in the bacterial species *S. pneumoniae*, *E. coli*, *S. aureus*, *H. pylori*, *M. tuberculosis*, and *M. luteus*, and the non-conserved sequence (red) encompassing the allosteric site in *S. pneumoniae* UPPS. Multiple Sequence Alignment (MSA) was performed using Clustal, accessible from the MPI Bioinformatics Toolkit, and the results were presented using the Boxshade tool of SIB Bioinformatics Resource Portal.[63,64]

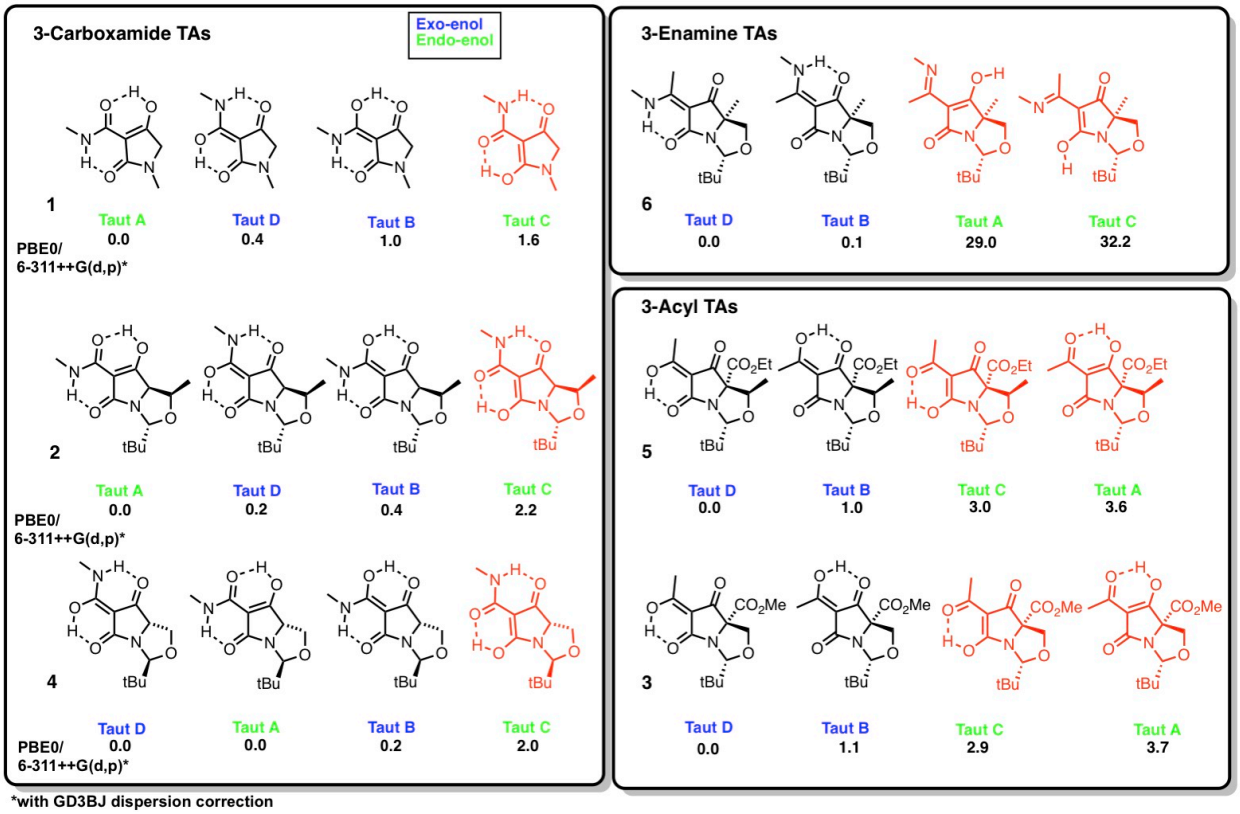


Figure S6: DFT results for the simplified structures of TAs **1-6**, categorised according to the 3-carboxamide, 3-acyl and 3-enamine classes. All single point energy calculations (PBE0/6-311++G(d,p)) were carried out with GD3BJ dispersion correction, employing SMD implicit water solvation, in Gaussian 16 (refer to **Section 4.2.1**).

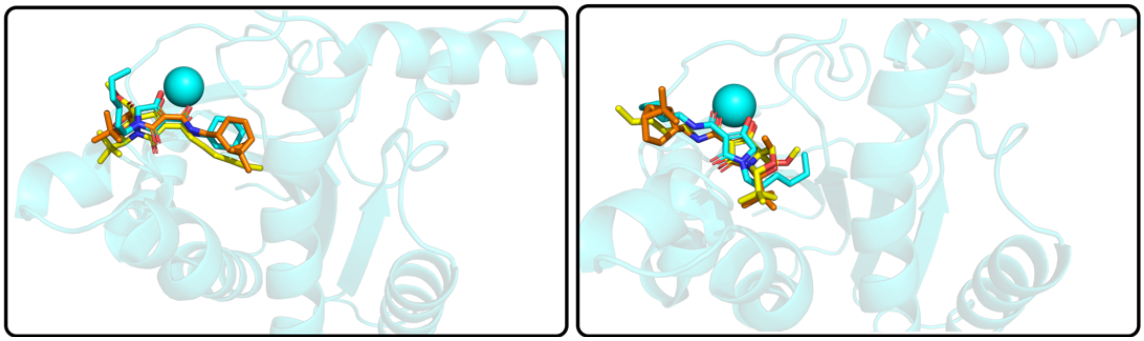


Figure S7: Poses 1 and 2 generated by manual alignment of the charged TA inhibitors such that all were chelated to Mg^{2+} in site 1.

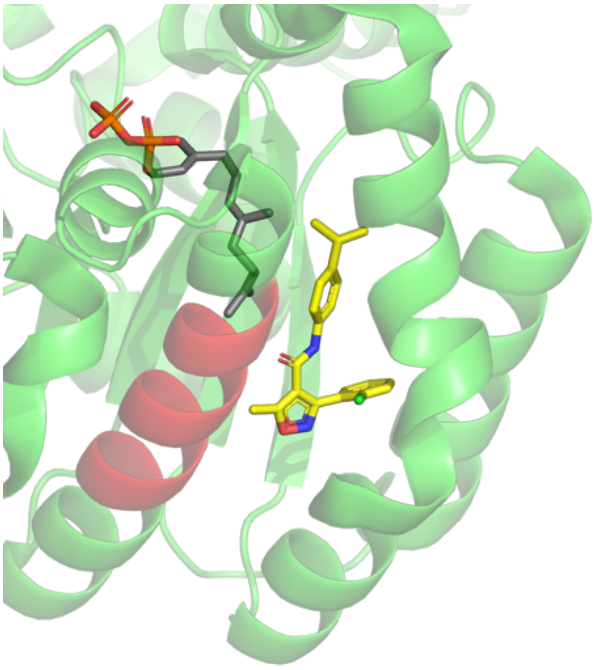


Figure S8: Crystal structure (PDB ID: **4Q9M**) of *S. pneumoniae* UPPS in complex with the allosteric inhibitor, **UK-106051** (yellow sticks), and FPP (grey sticks). The reported site of attachment (**AGMEALQTVTK**) of the TA-based probe is indicated (red ribbon).[34]

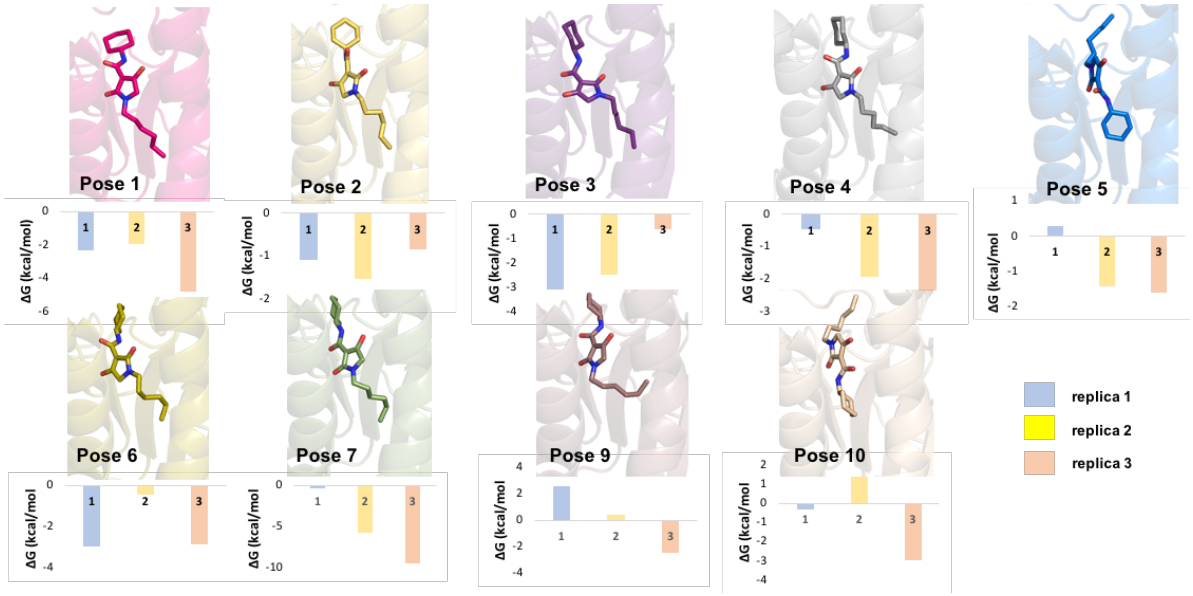


Figure S9: Iterative LIE results for TA **1A** in site 2 of closed *E. coli* UPPS (PDB ID: **4H8E**).

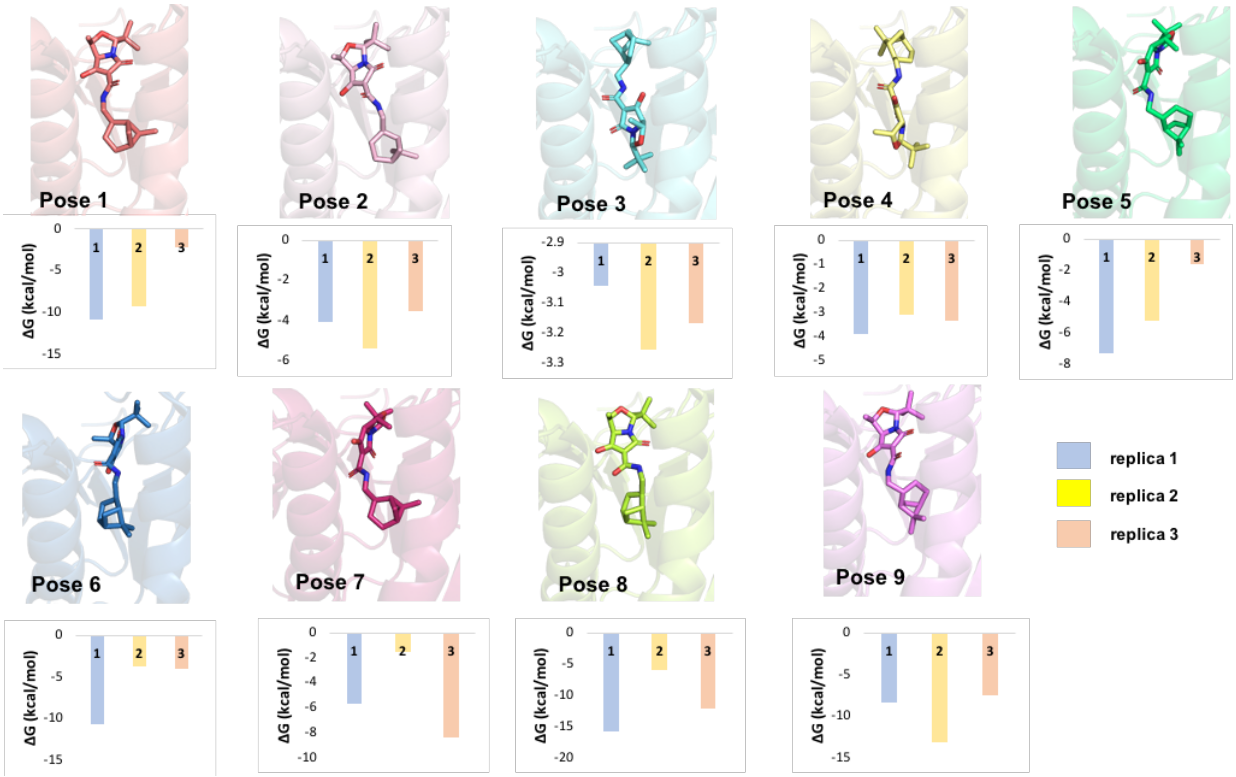


Figure S10: Iterative LIE results for TA **2A** in site 2 of closed *E. coli* UPPS (PDB ID: **4H8E**).

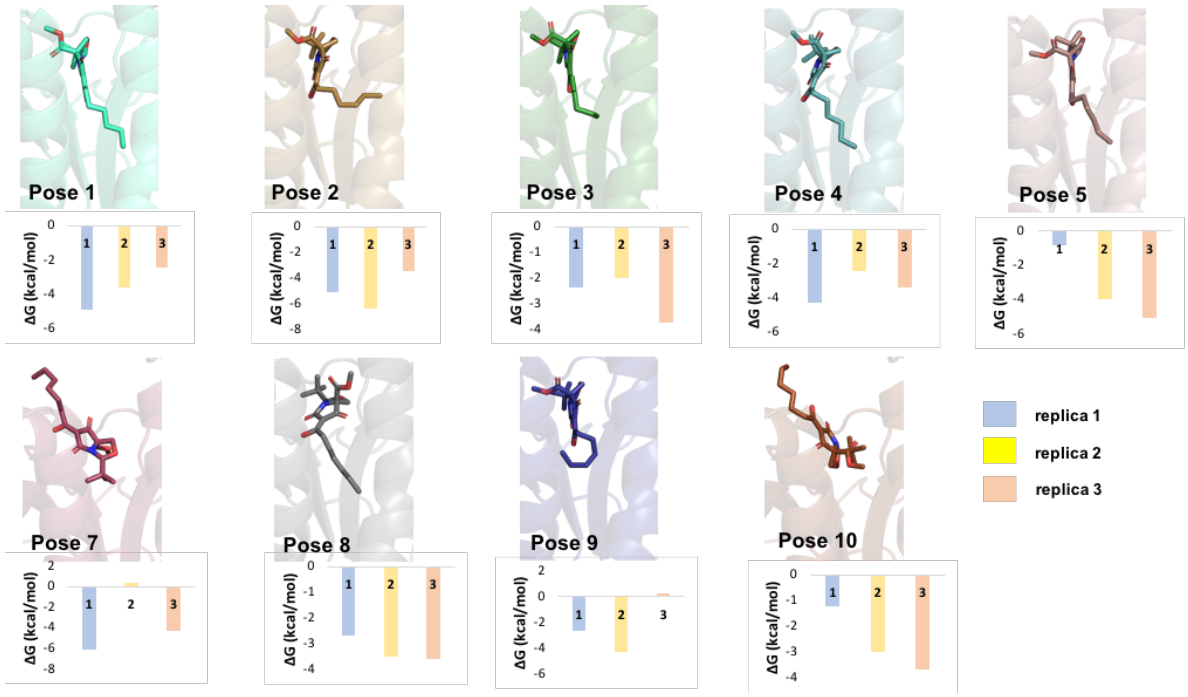


Figure S11: Iterative LIE results for TA **3D** in site 2 of closed *E. coli* UPPS (PDB ID: **4H8E**).

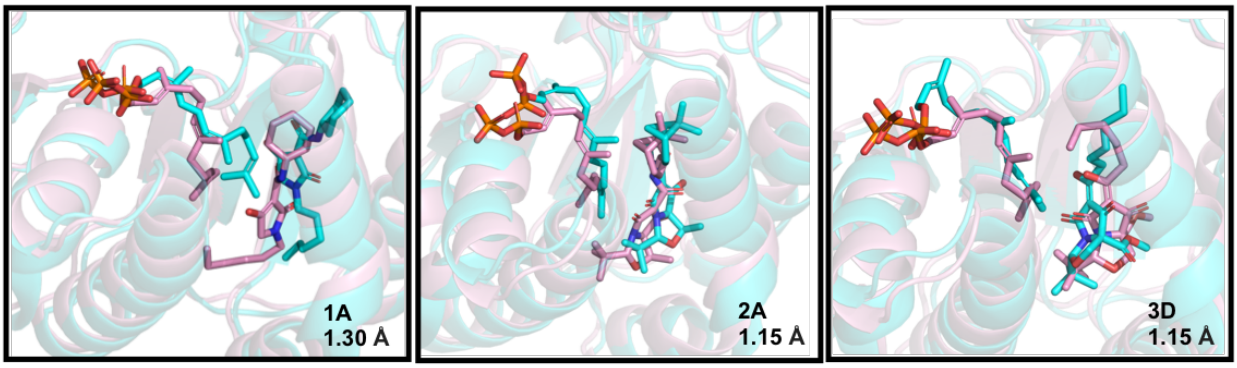


Figure S12: Backbone RMSD between the docked input complex (cyan ribbon) and the complex after REMD (pink ribbon) for site 3.

Table S2: Average electrostatic $\langle V^{Ele} \rangle$ and van der Waals $\langle V^{VdW} \rangle$ interaction energies for **1A, 2A and 3A** (site 1, 100 ns)

	Solvent			**Protein-bound**			
TA	**V^{Ele} (kJ/mol)**	**V^{VdW} (kJ/mol)**	**Orient.**	**V^{Ele} (kJ/mol)**	**V^{VdW} (kJ/mol)**		**ΔG (kcal/mol)**
1A	-276.79 ± 0.78	-95.64 ± 0.31	1	-28.92 ± 12.00	-121.30 ± 4.20		28.51 ± 1.45
			2	-23.85 ± 0.81	-122.96 ± 1.00		29.04 ± 0.14
2A	-285.62 ± 0.74	-106.77 ± 0.55	1	109.00 ± 2.30	-120.47 ± 3.20		46.57 ± 0.32
			2	86.72 ± 3.90	-144.19 ± 3.00		42.88 ± 0.49
3A	-286.79 ± 1.70	-96.00 ± 0.37	1	84.92 ± 2.30	-106.39 ± 4.10		43.97 ± 0.39
			2	56.60 ± 1.90	-158.33 ± 1.60		38.34 ± 0.31

Table S3: Average electrostatic $\langle V^{Ele} \rangle$ and van der Waals $\langle V^{VdW} \rangle$ interaction energies for **1A** (site 2, 30 ns)

	Solvent			**Protein-bound**						
Pose	**V^{Ele} (kJ/mol)**	**V^{VdW} (kJ/mol)**	**Replica**	**V^{Ele} (kJ/mol)**			**V^{VdW} (kJ/mol)**			**W_i**
1	169.70 ± 6.00	-101.50 ± 0.86	1	-69.38	±	12.00	-414.45	±	2.10	0.00
			2	-58.97	±	5.00	-417.31	±	2.60	0.00
			3	-134.81	±	6.20	-395.14	±	1.60	0.00
2			1	-47.47	±	2.20	-401.99	±	1.10	0.00
			2	-49.51	±	2.80	-417.89	±	2.00	0.00
			3	-69.19	±	1.10	-346.22	±	3.10	0.00
3			1	-95.84	±	4.30	-396.16	±	2.50	0.00
			2	-79.73	±	23.00	-399.98	±	4.50	0.00
			3	-34.46	±	1.80	-406.10	±	2.00	0.00
4			1	-44.61	±	3.70	-378.07	±	4.40	0.00
			2	-63.16	±	1.20	-408.64	±	1.60	0.00
			3	-69.67	±	2.80	-414.14	±	0.90	0.00
5			1	-26.39	±	1.10	-381.34	±	1.80	0.00
			2	-59.84	±	3.70	-392.52	±	1.50	0.00
			3	-70.53	±	2.80	-378.86	±	0.80	0.00
6			1	-89.41	±	3.20	-403.80	±	2.00	0.00
			2	-43.45	±	0.73	-381.47	±	0.34	0.00
			3	-75.04	±	1.50	-427.19	±	1.10	0.00
7			1	-34.07	±	13.00	-398.78	±	2.90	0.00
			2	-151.41	±	2.20	-403.86	±	0.84	0.00
			3	**-254.26**	**±**	**7.30**	**-365.66**	**±**	**1.10**	**1.00**
9			1	-11.07	±	3.40	-307.64	±	7.90	0.00
			2	-24.15	±	0.77	-380.86	±	2.00	0.00
			3	-103.53	±	2.90	-349.72	±	1.00	0.00
10			1	-57.96	±	5.00	-344.98	±	0.67	0.00
			2	-30.90	±	5.00	-321.19	±	3.60	0.00
			3	-114.40	±	7.60	-350.51	±	1.90	0.00

Table S4: Average electrostatic $\langle V^{Ele} \rangle$ and van der Waals $\langle V^{VdW} \rangle$ interaction energies for **2A** (site 2, 30 ns)

	Solvent			**Protein-bound**						
Pose	**V^{Ele} (kJ/mol)**	**V^{VdW} (kJ/mol)**	**Replica**	**V^{Ele} (kJ/mol)**			**V^{VdW} (kJ/mol)**			**W_i**
1	155.96 ± 4.60	-117.10 ± 0.76	1	-257.12	±	6.00	-410.75	±	1.70	0.00
			2	-231.48	±	5.80	-388.87	±	2.90	0.00
			3	-69.49	±	0.61	-392.46	±	3.60	0.00
2			1	-96.50	±	12.00	-426.88	±	3.50	0.00
			2	-127.92	±	2.50	-423.64	±	3.30	0.00
			3	-76.77	±	0.90	-441.73	±	3.40	0.00
3			1	-55.08	±	1.00	-463.92	±	1.90	0.00
			2	-58.41	±	1.20	-466.96	±	0.87	0.00
			3	-66.95	±	2.30	-445.37	±	1.60	0.00
4			1	-72.69	±	0.99	-466.88	±	3.00	0.00
			2	-65.72	±	3.80	-445.22	±	2.70	0.00
			3	-70.62	±	2.30	-445.00	±	1.40	0.00
5			1	-179.86	±	7.20	-404.63	±	5.40	0.00
			2	-121.30	±	22.00	-427.47	±	4.90	0.00
			3	-25.84	±	0.45	-456.02	±	1.10	0.00
6			1	-271.72	±	15.00	-374.66	±	5.60	0.00
			2	-100.20	±	1.10	-403.32	±	1.30	0.00
			3	-107.20	±	9.60	-400.91	±	1.30	0.00
7			1	-165.56	±	4.10	-358.56	±	2.00	0.00
			2	-30.52	±	0.86	-444.38	±	0.84	0.00
			1	-191.57	±	7.90	-428.42	±	0.86	0.00
8			**1**	**-374.64**	**±**	**11.00**	**-397.26**	**±**	**4.00**	**1.00**
			2	-164.25	±	22.00	-378.03	±	6.90	0.00
			3	-292.56	±	0.99	-399.41	±	1.70	0.00
9			1	-204.90	±	2.00	-402.74	±	3.20	0.00
			2	-325.07	±	1.80	-378.32	±	2.10	0.00
			3	-192.243	±	4.30	-389.23	±	2.60	0.00

Table S5: Average electrostatic $\langle V^{Ele} \rangle$ and van der Waals $\langle V^{VdW} \rangle$ interaction energies for **3D** (site 2, 30 ns)

	Solvent			**Protein-bound**						
Pose	**V^{Ele} (kJ/mol)**	**V^{VdW} (kJ/mol)**	**Replica**	**V^{Ele} (kJ/mol)**			**V^{VdW} (kJ/mol)**			**W_i**
1	-132.60 ± 2.30	-112.40 ± 0.44	1	-83.89	±	5.00	-438.51	±	1.40	0.00
			2	-75.81	±	4.00	-392.75	±	2.70	0.00
			3	-39.10	±	2.90	-416.10	±	4.00	0.00
2			1	-97.69	±	6.70	-417.55	±	3.50	0.00
			2	**-110.00**	**±**	**2.50**	**-452.13**	**±**	**2.50**	**0.88**
			3	-58.14	±	1.80	-422.68	±	1.80	0.00
3			1	-46.67	±	1.20	-397.65	±	2.00	0.00
			2	-47.72	±	3.00	-379.17	±	3.80	0.00
			3	-64.69	±	1.60	-422.35	±	0.97	0.00
4			1	-78.97	±	1.20	-417.70	±	1.20	0.00
			2	-31.15	±	1.30	-430.00	±	2.70	0.00
			3	-55.09	±	1.20	-425.68	±	0.84	0.00
5			1	-25.14	±	0.99	-370.93	±	3.70	0.00
			2	-58.89	±	2.20	-445.67	±	2.20	0.00
			3	-98.16	±	2.80	-415.85	±	2.70	0.00
7			**1**	**-185.02**	**±**	**17.00**	**-287.59**	**±**	**5.40**	**0.12**
			2	-64.33	±	7.70	-231.77	±	4.70	0.00
			3	-148.12	±	5.30	-276.75	±	4.30	0.00
8			1	-41.33	±	1.00	-421.66	±	3.40	0.00
			2	-47.57	±	0.88	-446.85	±	1.30	0.00
			3	-50.58	±	4.90	-444.39	±	3.30	0.00
9			1	-65.74	±	3.50	-370.63	±	1.90	0.00
			2	-69.38	±	0.93	-440.38	±	2.60	0.00
			3	-24.14	±	0.87	-323.26	±	5.90	0.00
10			1	-53.37	±	3.80	-330.83	±	3.40	0.00
			2	-87.78	±	2.00	-341.40	±	1.10	0.00
			3	-125.44	±	2.70	-295.75	±	1.30	0.00

Table S6: Average electrostatic $\langle V^{Ele} \rangle$ and van der Waals $\langle V^{VdW} \rangle$ interaction energies for **5aD** (site 2, 30 ns)

	Solvent			Protein-bound						
Pose	**V^{Ele} (kJ/mol)**	**V^{VdW} (kJ/mol)**	**Replica**	**V^{Ele} (kJ/mol)**			**V^{VdW} (kJ/mol)**			**W_i**
1	-157.51 ± 0.49	-150.13 ± 1.30	1	-24.35	±	5.19	-271.45	±	7.28	0.00
			2	**-65.44**	**±**	**5.40**	**-252.50**	**±**	**11.00**	**0.98**
			3	-34.06	±	2.59	-264.51	±	1.72	0.00
			4	-44.73	±	2.01	-273.63	±	2.22	0.02
			5	-33.72	±	4.69	-226.18	±	4.31	0.00
			6	-30.38	±	5.48	-255.93	±	14.02	0.00

Table S7: Average electrostatic $\langle V^{Ele} \rangle$ and van der Waals $\langle V^{VdW} \rangle$ interaction energies for **5bD** (site 2, 30 ns)

	Solvent			Protein-bound						
Pose	**V^{Ele} (kJ/mol)**	**V^{VdW} (kJ/mol)**	**Replica**	**V^{Ele} (kJ/mol)**			**V^{VdW} (kJ/mol)**			**W_i**
1	-144.70 ± 1.40	-161.38 ± 2.60	1	-36.29	±	2.60	-281.58	±	2.70	0.00
			2	-54.03	±	1.70	-284.48	±	4.60	0.00
			3	-11.08	±	4.30	-289.24	±	2.30	0.00
			4	-47.26	±	1.80	-296.15	±	2.80	0.00
			5	-25.86	±	1.60	-280.65	±	7.00	0.00
			6	-30.42	±	2.30	-277.47	±	1.90	0.00
2			1	-20.36	±	2.00	-263.50	±	2.10	0.00
			2	-43.03	±	2.10	-287.40	±	3.40	0.00
			3	**-115.88**	**±**	**4.70**	**-200.19**	**±**	**2.70**	**1.00**
			4	-41.45	±	1.70	-277.43	±	2.40	0.00
			5	-24.67	±	3.30	-222.29	±	16.00	0.00
			6	-40.03	±	1.70	-298.55	±	1.30	0.00

Table S8: LIE results for poses with most favourable ΔG_{calc} (site 2, 100 ns)

TA	**ΔV^{Ele} (kcal/mol)**	**ΔV^{VdW} (kcal/mol)**	**ΔG_{calc} (kcal/mol)**
1A	-28.09 ± 2.87	-46.10 ± 1.10	0.78 ± 1.50
2A	-42.99 ± 2.39	-47.08 ± 0.36	-4.73 ± 1.80
3D	-7.68 ± 0.74	-47.37 ± 0.41	4.89 ± 0.30
5aD	-8.65 ± 0.60	-73.72 ± 1.24	3.88 ± 0.30
5bD	-32.58 ± 0.48	-51.41 ± 0.33	-1.58 ± 0.20

Table S9: Average electrostatic $\langle V^{Ele} \rangle$ and van der Waals $\langle V^{VdW} \rangle$ interaction energies for **1A, 2A and 3D** (site 3, 100 ns)

	Solvent			Protein-bound						
TA	**V^{Ele} (kJ/mol)**	**V^{VdW} (kJ/mol)**	**Orient.**	**V^{Ele} (kJ/mol)**			**V^{VdW} (kJ/mol)**			**ΔG (kcal/mol)**
1A	-170.89 ± 12.00	-101.75 ± 1.40	1	-26.69	±	0.88	-165.25	±	1.90	10.00 ± 1.07
2A	-166.32 ± 18.00	-115.33 ± 2.40	1	-3.47	±	0.79	-201.73	±	0.67	10.66 ± 1.60
3A	-128.84 ± 2.10	-113.50 ± 0.25	1	-11.98	±	0.79	-202.08	±	1.50	6.50 ± 0.21

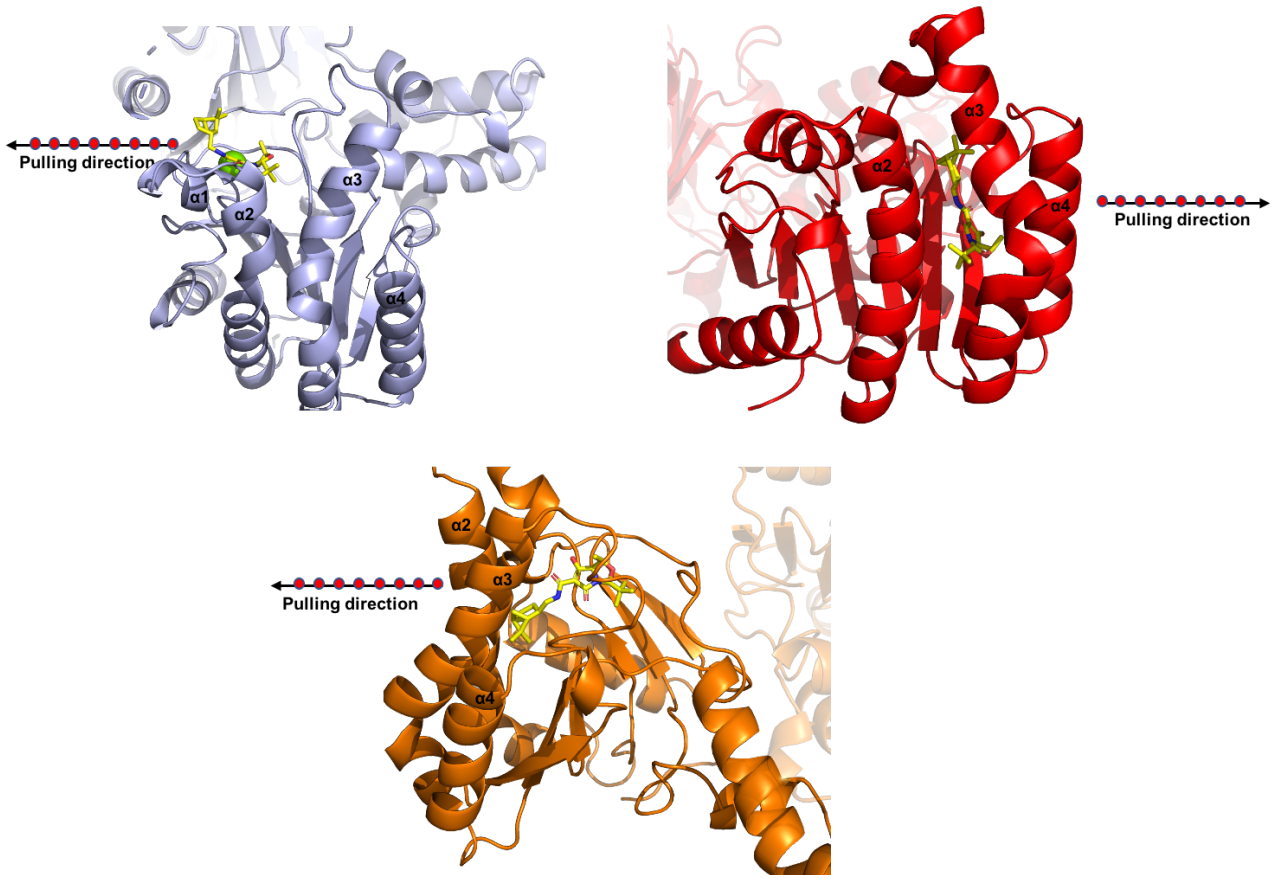


Figure S13: Representation of the initial simulation setup: the ligand is pulled along the z axis using a harmonic force with a spring constant of 1000 kJ mol^{-1} nm^{-2} and a pull rate of 0.01 nm ps^{-1}.

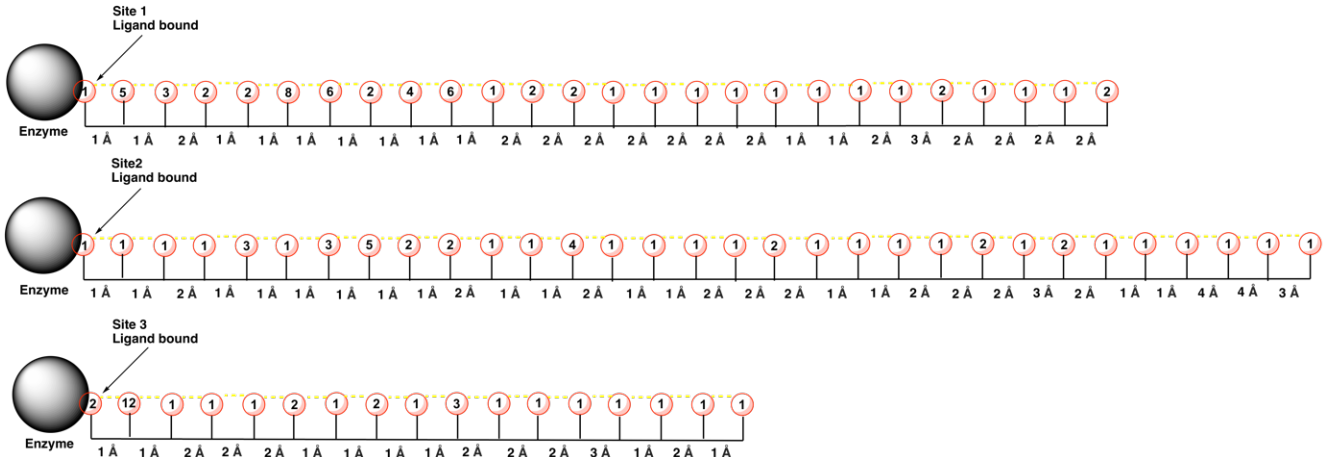


Figure S14: Asymmetric distribution of the umbrella sampling windows for UPPS-**2A** complex of sites 1-3, separated by COM distance. Numbers within the red circles represent the number of US replica windows which were run at each COM distance.

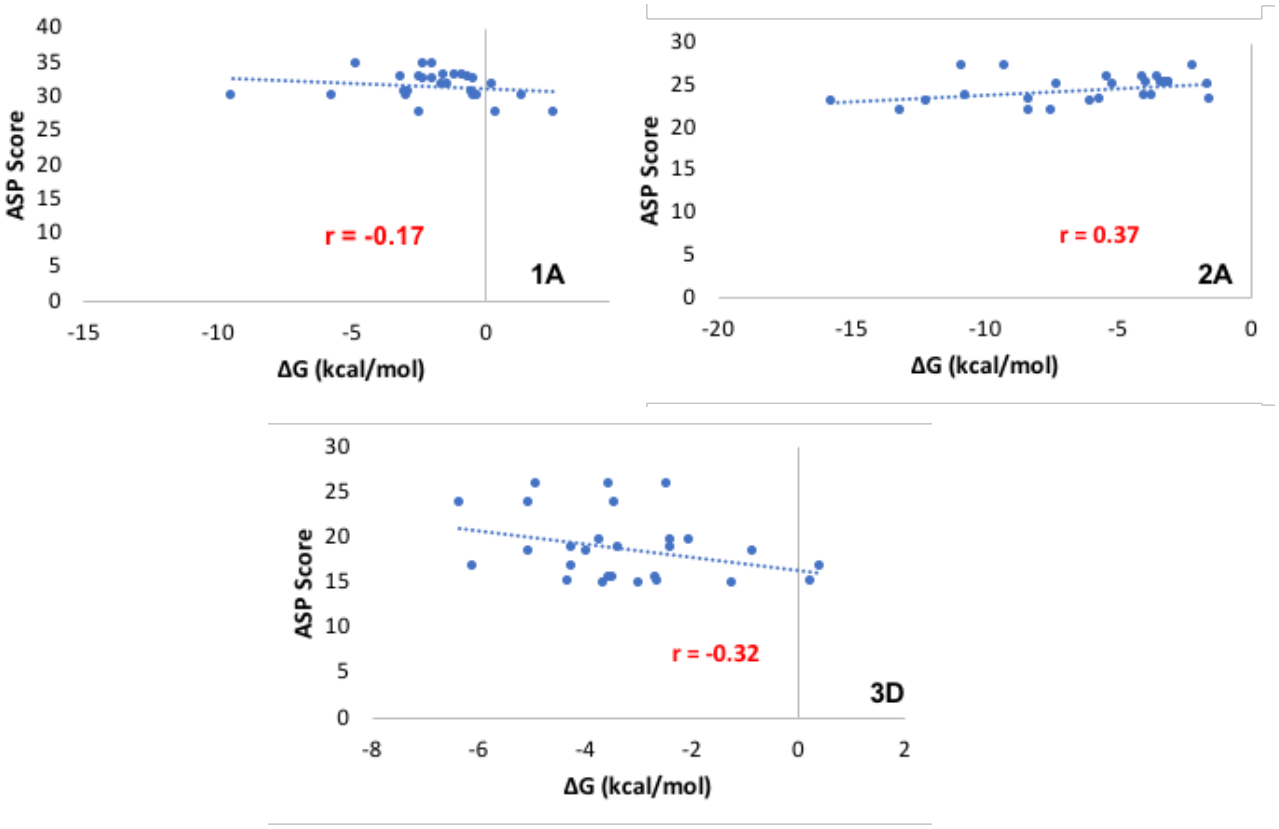


Figure S15: (Lack of) correlation between the ASP docking score and the LIE-determined free energies of binding of compounds **1A**, **2A** and **3D** in site 2.

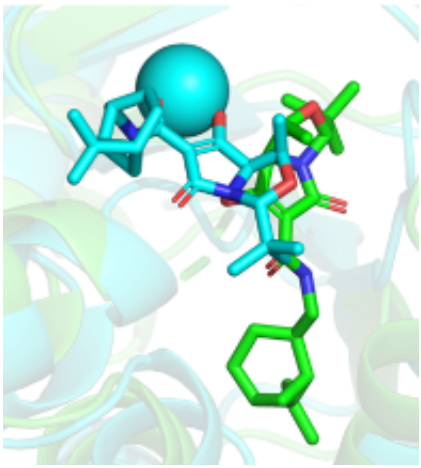


Figure S16: Compound 2A in site 1 (cyan sticks) and site 2 (green sticks).

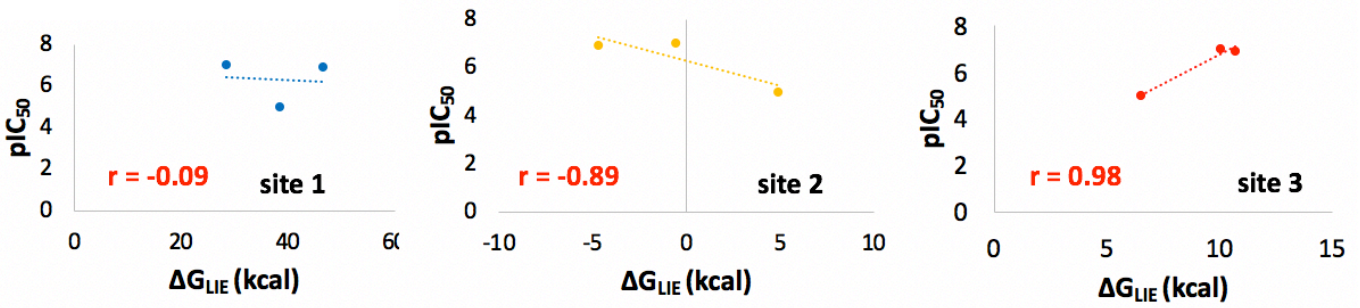


Figure S17: Correlations between *in vitro* binding data and ΔG_{LIE} for compounds **1-3** in sites 1-3.

Table S10: IC_{50} binding data for TAs 1, 2, 3, 5a and 5b.[52]

TA	IC_{50} (μM)
1	0.10
2	0.12
3	>10
5a	1.9
5b	>10

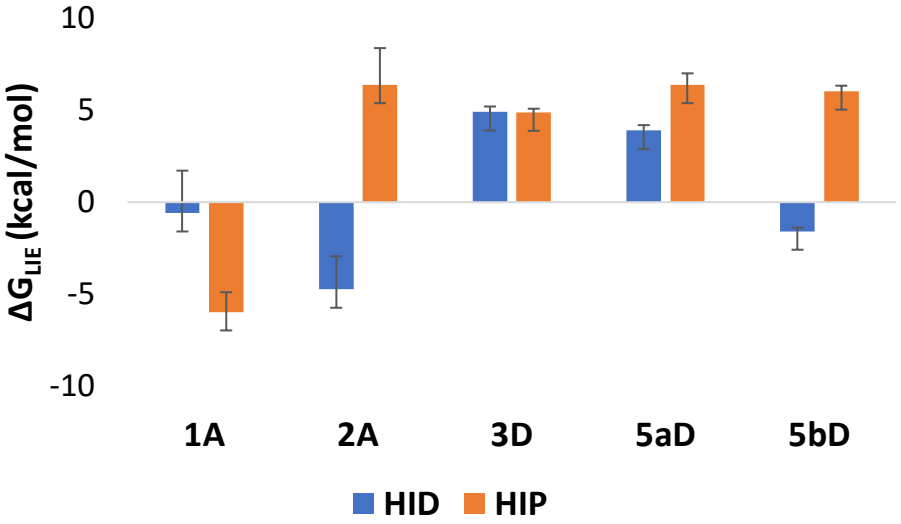


Figure S18: LIE results for the HID and HIP trajectories of compounds 1-3 in site 2.

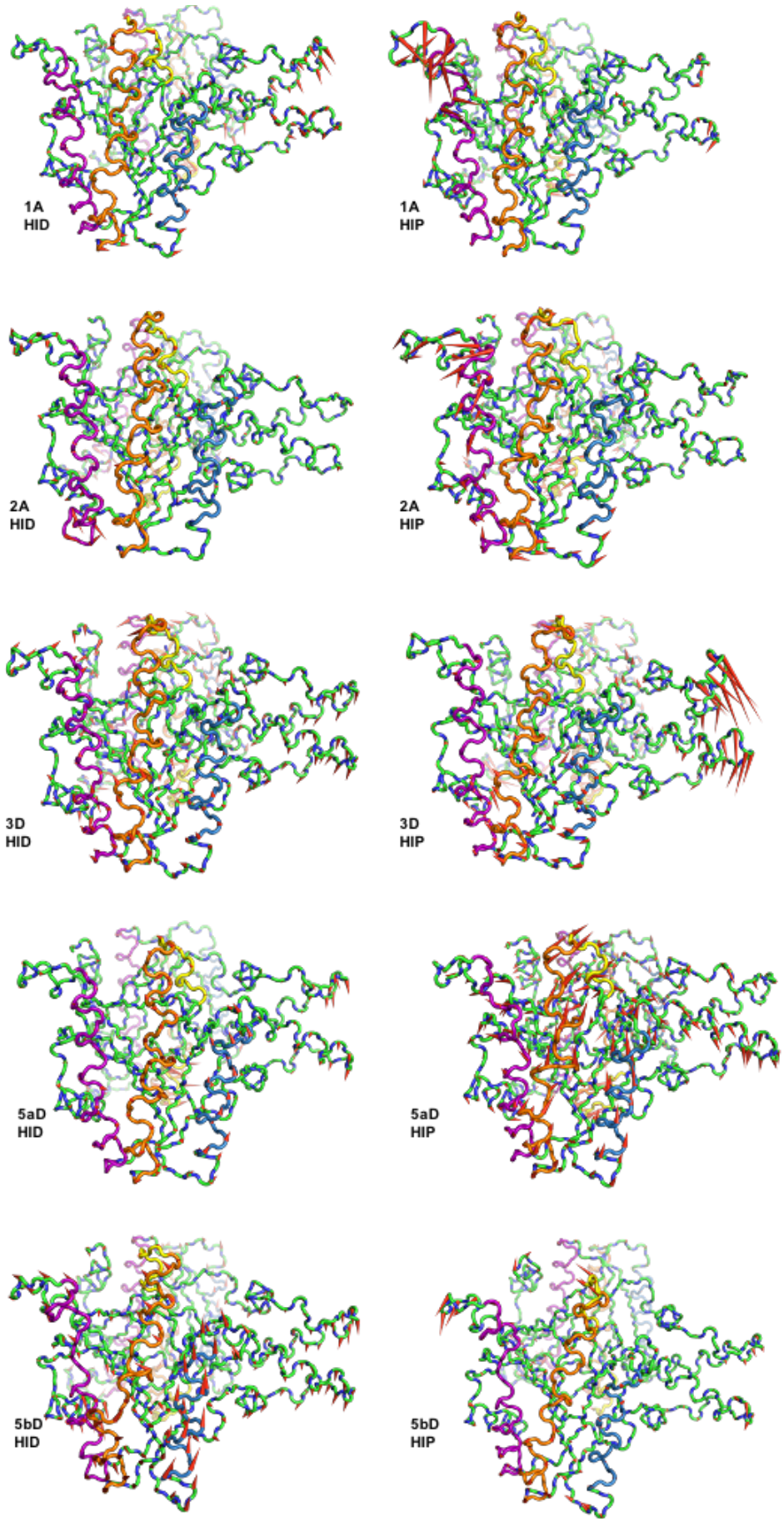


Figure S19: Porcupine plots generated from the extreme projections of PC1 in the HID and HIP simulations of each of the UPPS-TA complexes. Structural elements which exhibited the highest movement were highlighted in yellow.

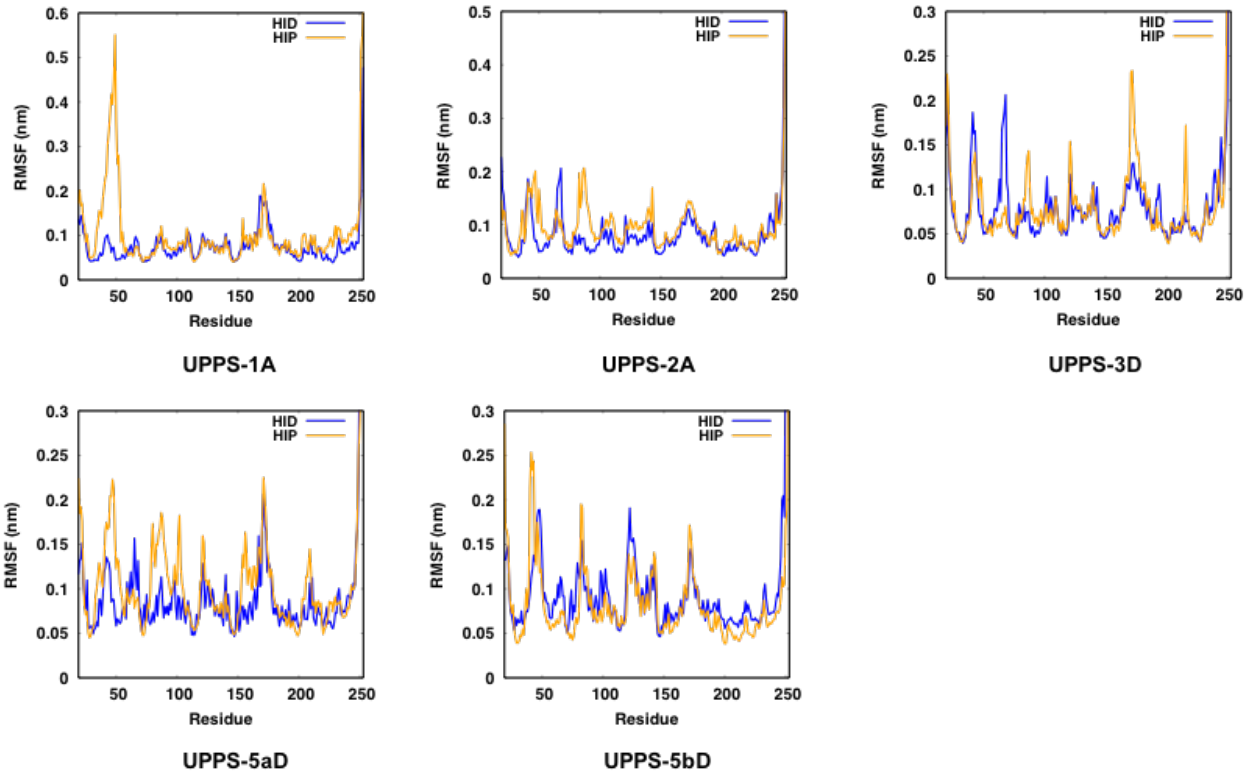


Figure S20: RMSF of chain A of UPPS in the HID and the HIP trajectories of the UPPS-TA complexes.

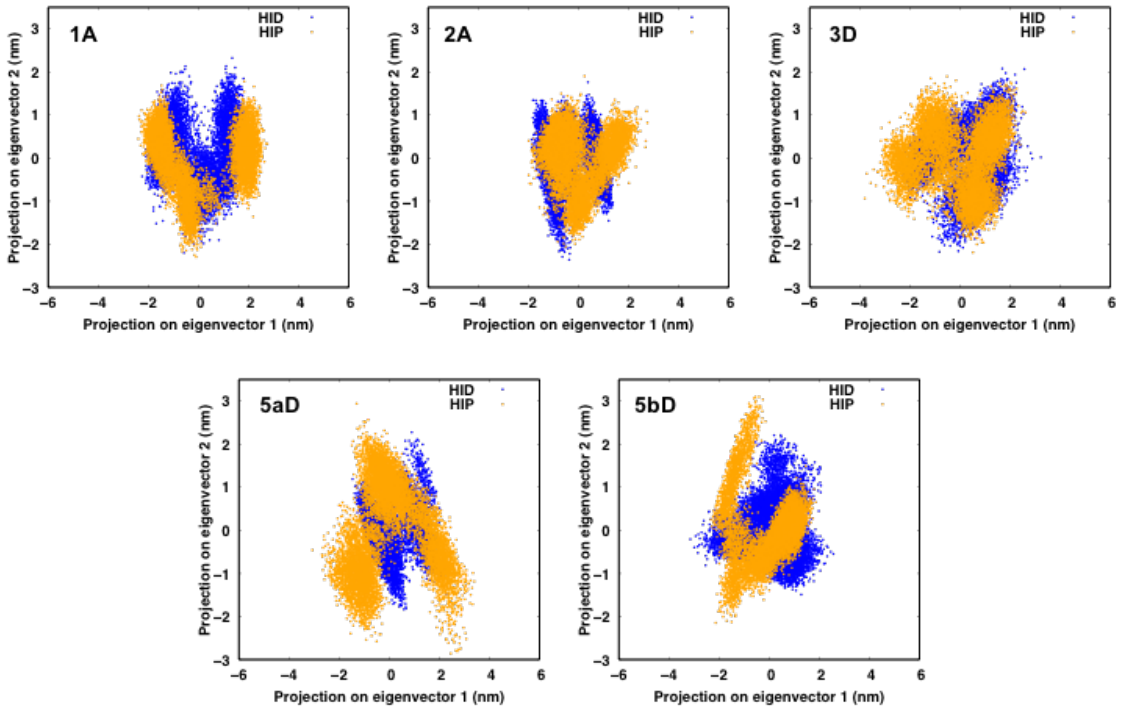


Figure S21: Two-dimensional projection plots of PC1 and PC2 of the HID and HIP trajectories of the UPPS-TA complexes, in which the TAs occupied site 2.

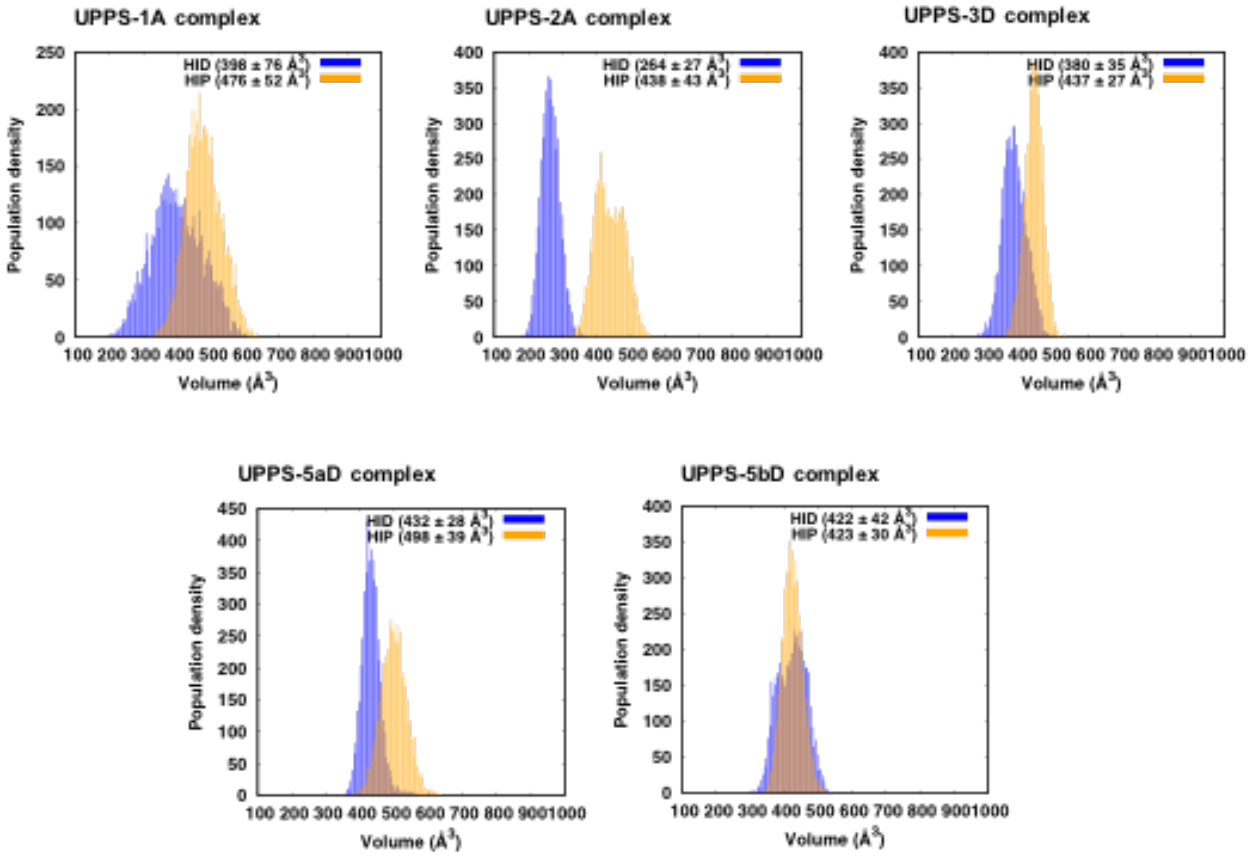


Figure S22: Volumes of the hydrophobic tunnel are presented in histogram form. All volumes were calculated from the pre-aligned 100 ns trajectories of the respective HIP and HID simulations for each UPPS-TA complex. Volumes of the binding pocket in apo-HID and apo-HIP were calculated from the average of all 5 independent 100 ns simulations for each protonation state.

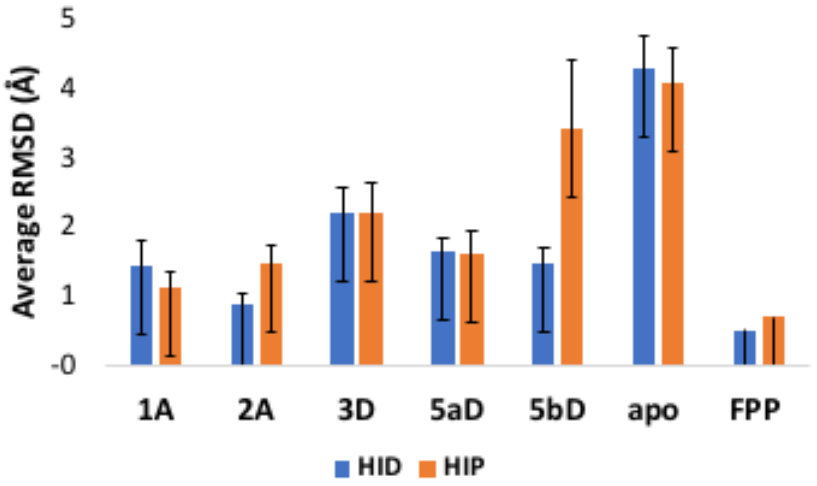


Figure S23: Average RMSD of the flexible loop in the HID and HIP trajectories of apo-UPPS and the UPPS-TA complexes.

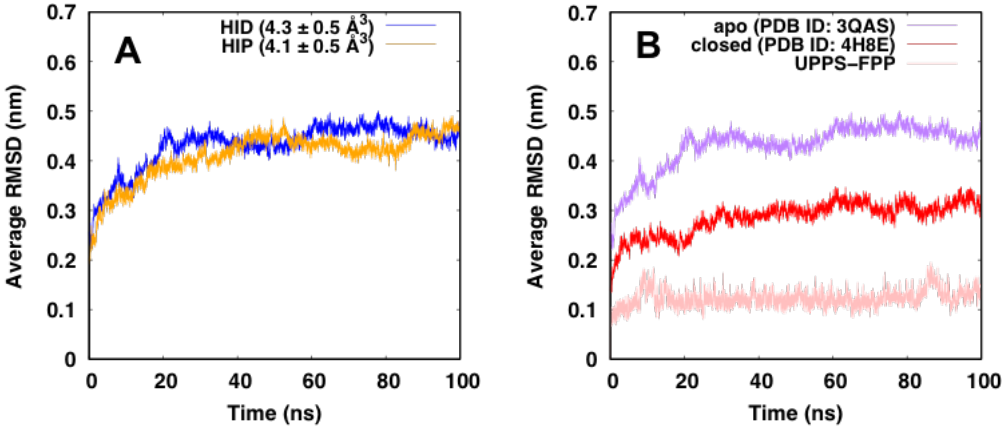


Figure S24: Average RMSD of the flexible loop (residues 72-82) in the HID and HIP simulations of the apoenzyme (PDB ID: **3QAS**). The average was calculated over 5 independent 100 ns MD simulations for each protonation state (**A**). Average RMSD of the flexible loop in apo (PDB ID: 3QAS), closed (PDB ID: **4H8E**, ligand removed), and FPP-bound (PDB ID: **4H8E**) UPPS, each with **H43**/**H50** in the HID protonation state. The average for the apoenzyme was calculated over 5 independent 100 ns simulations, while that for the closed

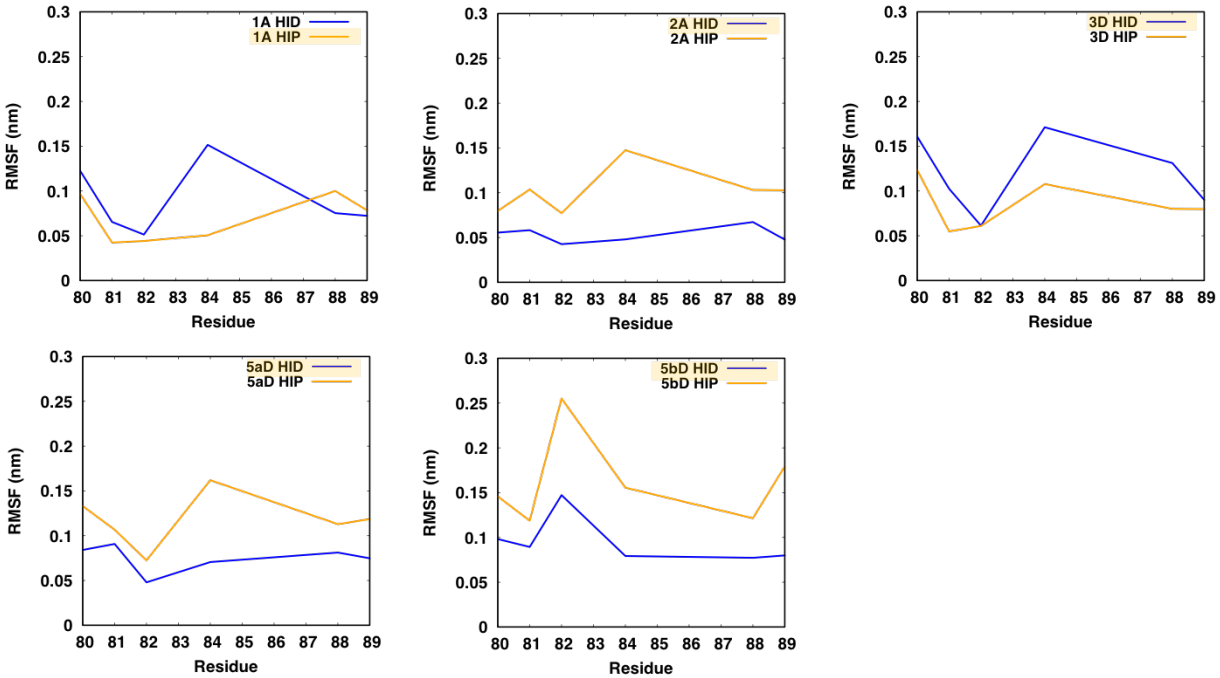


Figure S25: RMSF of the conserved residues of the flexible loop (residues 82-92) for the HID and HIP trajectories of all UPPS-TA site 2 complexes.

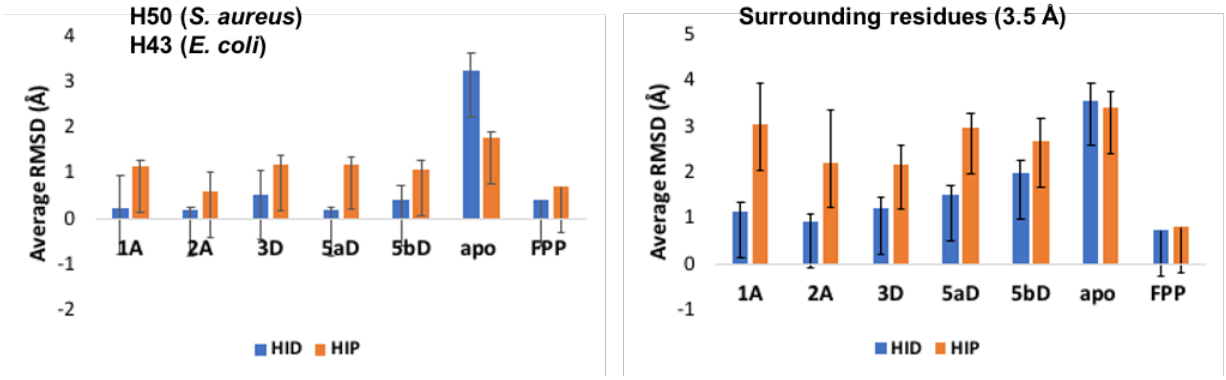


Figure S26: Comparisons of the RMSD of active site histidine, surrounding residues, and the flexible loop in the HID (blue) and HIP (orange) states. For the UPPS-TA complexes, all results were presented as averages and standard deviations over single 100 ns MD simulations for each protonation state. For the apoenzyme, the results were presented as averages and standard deviations over 5 independent 100 ns MD simulations for each protonation state.

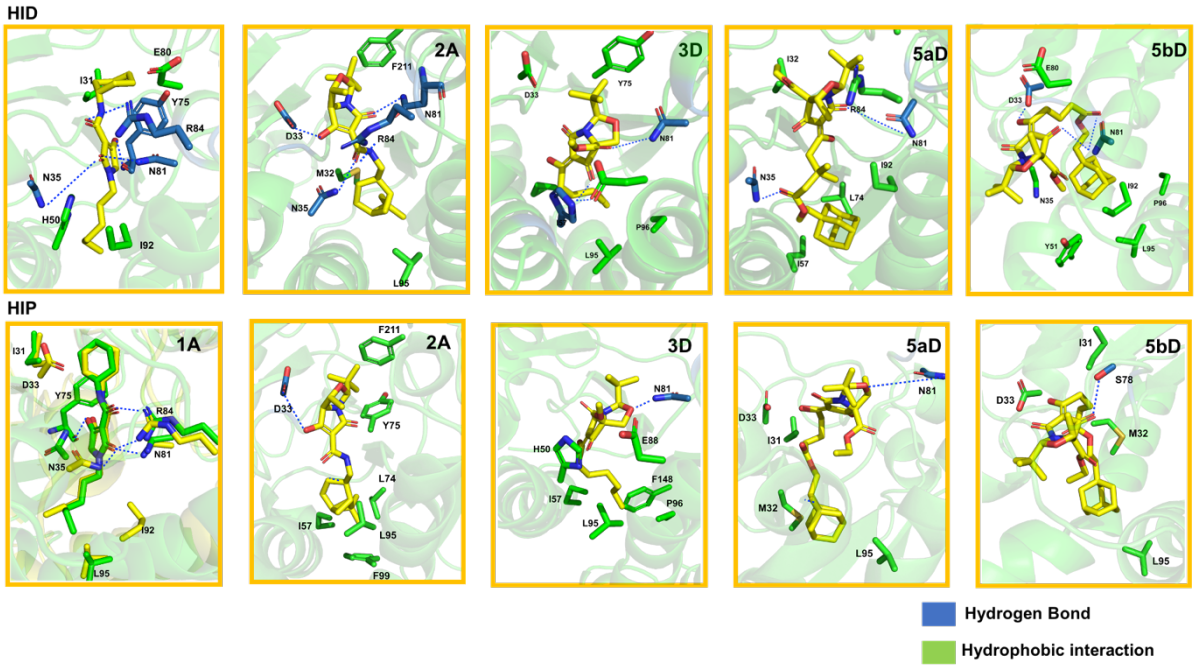


Figure S27: Comparison of the main interactions of the HID and HIP site 2 UPPS-TA complexes.

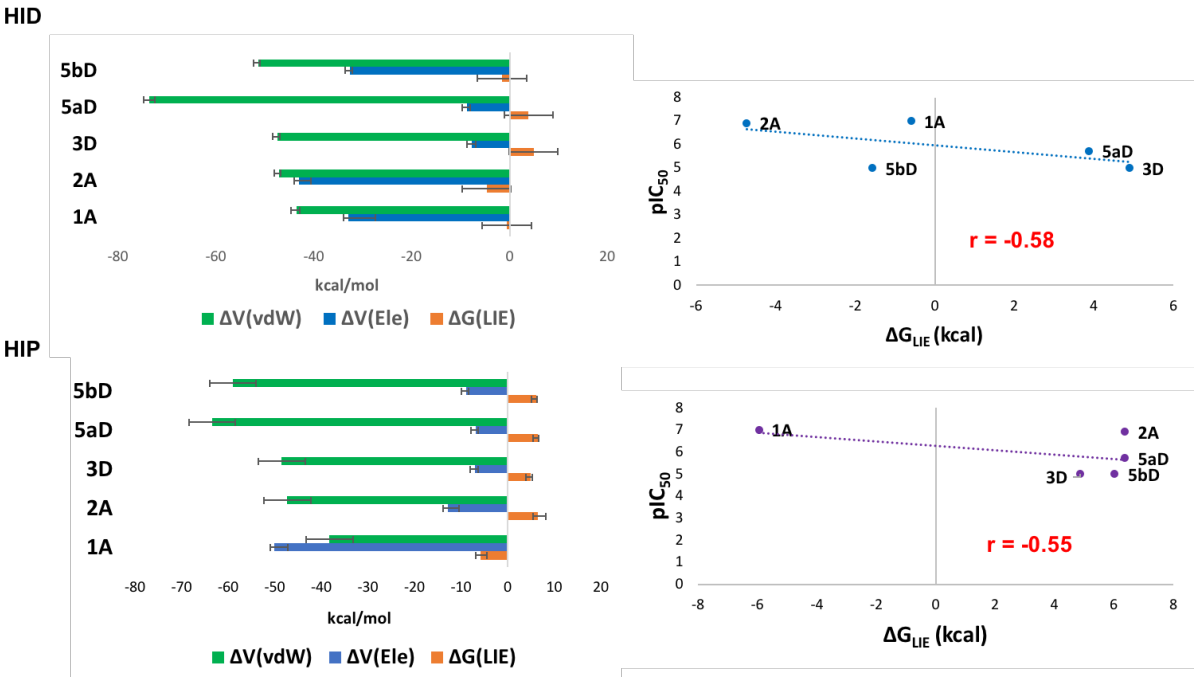


Figure S28: *Left:* Free energies of binding (ΔG(LIE)) and the differences in electrostatic (ΔV(Ele)) and van der Waals ΔV(vdW)) interaction energies and their standard errors, calculated using the LIE method for the HID and HIP simulations of all UPPS-TA complexes. *Right*: negative correlations between the negative logarithm of the IC_{50} (pIC_{50}) and the calculated free energies of binding ΔG_{LIE} of all UPPS-TA complexes.

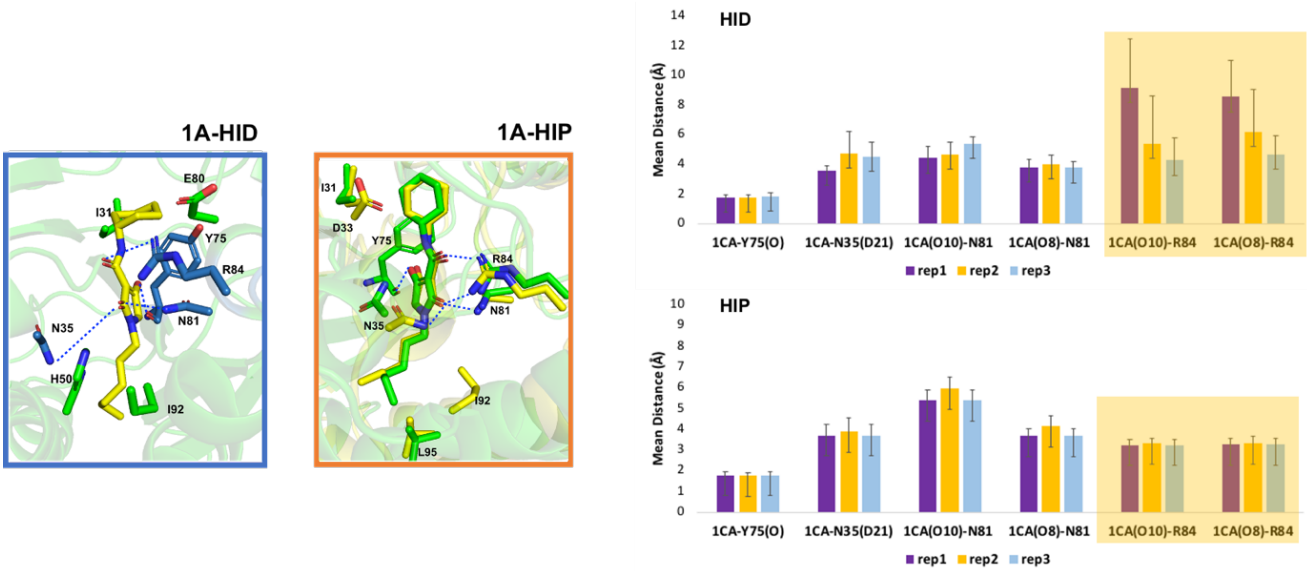


Figure S29: *Left*: Highest populated clusters of the HID and HIP trajectories of UPPS-1A associated with the most favourable ΔG_{LIE}. The first highest populated cluster (green ribbon) and the second highest populated cluster (yellow ribbon) of HIP UPPS-1A (i.e., accounting for >85% of the total trajectory) are shown. The first highest populated cluster (green ribbon) of HID UPPS-1A (i.e., accounting for >85% of the total trajectory) is shown. *Right:* the mean distances and standard deviations of critical electrostatic interactions measured for three 100 ns replicas of each of the HID and HIP trajectories.

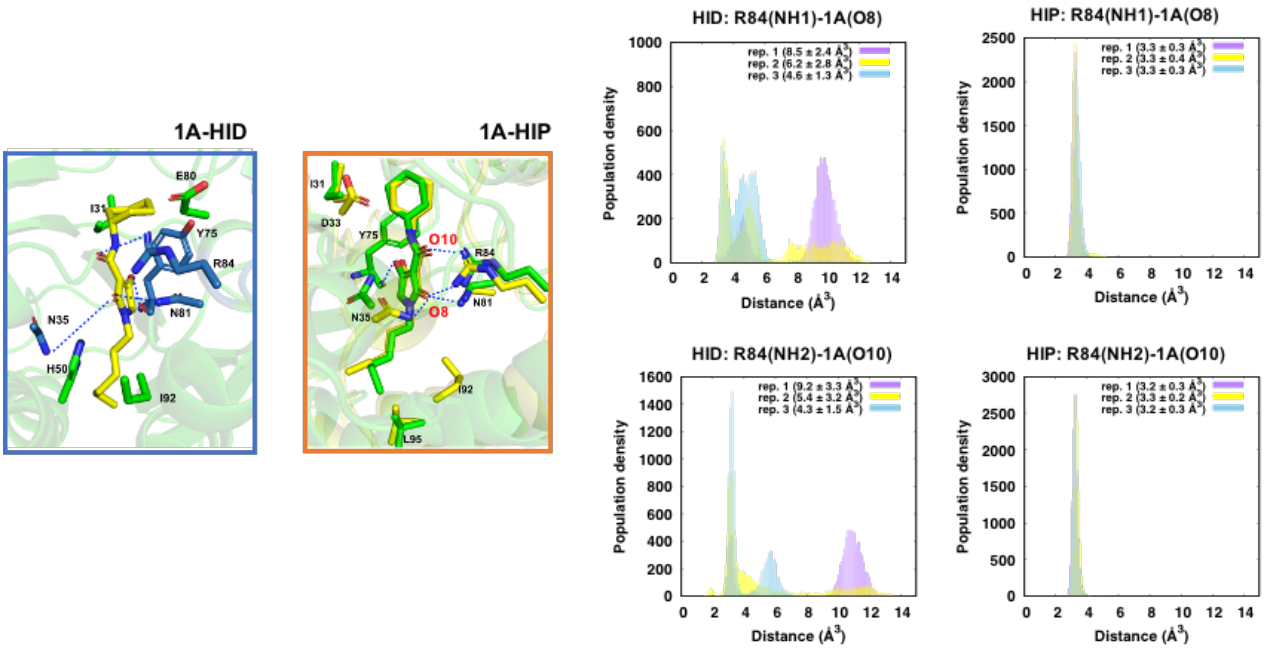


Figure S30: *Left*: Highest populated clusters of the HID and HIP trajectories of UPPS-1A associated with the most favourable ΔG_{LIE}, showing the two electrostatic interactions between the TA and **R84**. *Right*: Histograms of the average distances recorded for the latter in each of the three HID and three HIP trajectories.

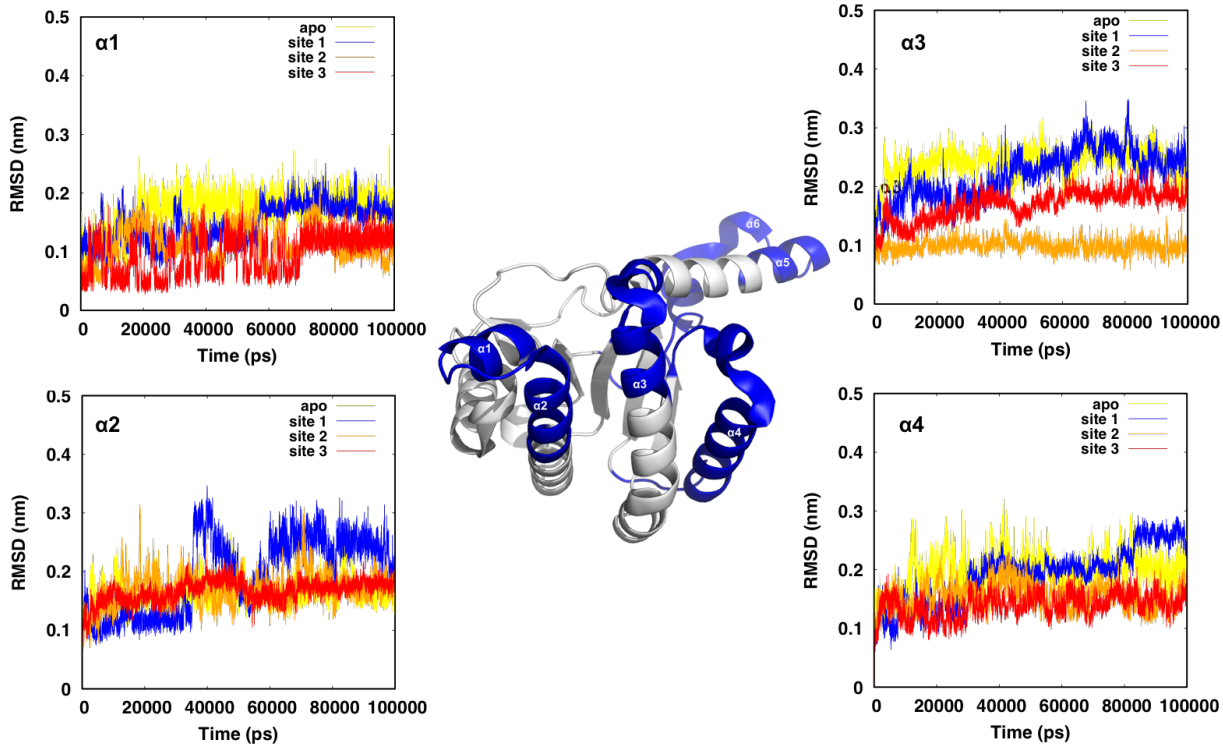


Figure S31: All-atom RMSD of helices α1, α2, α3, and α4 for the three UPPS-**2A** complexes (sites 1-3) over 100 ns of MD.

Table S11: Free energy results by FEP-MBAR

	$\Delta G^{solv}_{elec+vdw}$ **(kcal/mol)**	$\Delta G^{prot}_{elec+vdw+restr}$ **(kcal/mol)**	ΔG_{corr} [a] **(kcal/mol)**	ΔG^{solv}_{restr} **(kcal/mol)**
site 1	-73.75 $\pm$ 0.38	107.63 $\pm$ 0.31	5.82	-60.98
site 2	-10.40 $\pm$ 0.26	46.99 $\pm$ 0.23	N/A	-60.60
site 3	-10.40 $\pm$ 0.26	35.16 $\pm$ 0.18	N/A	-59.96

[a] $\Delta\Delta G_{elec\ corr}$ is the conbtribution to the binding free energy due to the correction for electrostatic finite size effects. It was applied to the charged ligand (i.e., 2A in site 1) only.