Computational Modeling of the Mechanisms and Selectivity of Organophosphate Hydrolases

MIHA PURG
Abstract


Computational modeling is becoming an increasingly integral part of (bio)chemistry, providing a powerful complementary view into the dynamics, binding, and reactivity of biochemical systems. In particular, molecular simulations based on multiscale models are now regularly employed in studies of enzymatic reactions, offering invaluable mechanistic insight through the lens of molecular energy landscapes. In this thesis, I used the empirical valence bond (EVB) and related methods to study the mechanisms and selectivity of organophosphate hydrolases.

Organophosphate hydrolases are a diverse class of enzymes capable of degrading some of the most toxic compounds known to mankind, including pesticides and chemical warfare agents. They are particularly interesting from a mechanistic and evolutionary point of view, having evolved the ability to catalyze the hydrolysis of compounds which were introduced to nature less than a century ago. Moreover, they show promise as effective organophosphate decontamination agents and a thorough understanding of their function is fundamental to the future design of efficient and selective biocatalysts.

As organophosphate hydrolases are metal-dependent enzymes, a reliable metal model was a prerequisite to our simulations. First, I present the development of force-field independent parameters for several alkaline-earth and transition-metal ions described using the nonbonded cationic dummy model. The model was subsequently employed in EVB simulations to probe the origin of metal-ion activity and selectivity patterns observed in methyl parathion hydrolase (MPH) and to provide mechanistic insight into its paraoxonase and promiscuous arylesterase activities. I further set out to resolve open mechanistic questions surrounding diisopropyl fluorophosphatase (DFPase) by performing extensive simulations of two mechanistic pathways proposed in literature, including calculating the effects of mutations, temperature, and protonation states on the rate of hydrolysis. Using this knowledge, I address the origin of cross-selectivity between DFPase and a structurally similar enzyme serum paraoxonase 1 (PON1). Finally, I present the latest developments in the software used to perform the simulations.

Keywords: Organophosphate Hydrolase, Computational enzymology, MPH, DFPase, PON1, Empirical Valence Bond, EVB

Miha Purg, Department of Chemistry - BMC, Biochemistry, Box 576, Uppsala University, SE-75123 Uppsala, Sweden.

© Miha Purg 2018

ISSN 1651-6214
urn:nbn:se:uu:diva-360518 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-360518)
All models are wrong, but some are useful.

George E.P. Box
List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Reprints were made with permission from the publishers.
Additional Publications


XII Purg, M. qtools: A Python toolset for the Q molecular simulation package. *Submitted to the Journal of Open Source Software*
Contribution report

Contributions to the articles that are part of this thesis are listed here:

Paper I Performed part of the parameterization.

Paper II Performed part of the simulations and analysis.

Paper III Performed all of the simulations and analysis.

Paper IV Contributed to the development of the code. Performed and analysed part of the simulations.
# Contents

1 Introduction ........................................................................................................... 13

2 Enzymes ................................................................................................................. 14
   2.1 Kinetics of Enzymatic Reactions ..................................................................... 15
      2.1.1 The Steady-State Approximation .............................................................. 15
      2.1.2 The Connection to Thermodynamics ....................................................... 18
      2.1.3 Factors Affecting the Rate ..................................................................... 19
   2.2 Enzyme Catalysis ............................................................................................. 21
      2.2.1 Transition State Stabilization ................................................................. 21
      2.2.2 Ground State Destabilization ................................................................. 22
      2.2.3 Dynamic Effects ..................................................................................... 23
      2.2.4 Metal Ion Catalysis ................................................................................ 23
   2.3 Selectivity and Promiscuity ............................................................................ 24

3 Computational Modeling of Enzymatic Reactions ............................................... 26
   3.1 Defining the Potential Energy Landscape ....................................................... 26
      3.1.1 Quantum Mechanics ............................................................................. 28
      3.1.2 Molecular Mechanics ............................................................................ 32
      3.1.3 Multiscale Models .................................................................................. 35
      3.1.4 Empirical Valence Bond ........................................................... 36
   3.2 Exploring the Configurational Space ............................................................... 41
      3.2.1 Molecular Dynamics ............................................................................. 43
      3.2.2 Free-Energy Perturbation .................................................................... 44
      3.2.3 Umbrella Sampling and Potential of Mean Force ................................. 46
      3.2.4 EVB FEP/US ......................................................................................... 48
      3.2.5 Linear Response Approximation ........................................................... 52

4 Studied Systems ..................................................................................................... 54
   4.1 Organophosphates ......................................................................................... 54
      4.1.1 Non-enzymatic Hydrolysis .................................................................... 55
   4.2 Organophosphate Hydrolases ......................................................................... 58
      4.2.1 Methyl Parathion Hydrolase ................................................................. 60
      4.2.2 Squid DFPase ....................................................................................... 63
      4.2.3 Serum Paraoxonase 1 ........................................................................... 68

5 Summary of Publications ....................................................................................... 73
   5.1 Paper I - Force Field Independent Metal Parameters Using a Nonbonded Dummy Model .............................................................. 73
5.2 Paper II - Probing the Mechanisms for the Selectivity and Promiscuity of Methyl Parathion Hydrolase ........................................ 74
5.3 Paper III - Similar Active Sites and Mechanisms Do Not Lead to Cross-Promiscuity in Organophosphate Hydrolysis: Implications for Biotherapeutic Engineering ............................. 75
5.4 Paper IV - Q6 - A Comprehensive Toolkit for Empirical Valence Bond and Related Free Energy Calculations .................. 77

6 Concluding Remarks ........................................................................................................ 78

7 Sammanfattning .............................................................................................................. 79

References .......................................................................................................................... 81
Abbreviations

DFP     Diisopropyl Fluorophosphate
DFPase  Diisopropyl Fluorophosphatase
EVB     Empirical Valence Bond
FEP     Free Energy Perturbation
KIE     Kinetic Isotope Effect
LRA     Linear Response Approximation
MM      Molecular Mechanics
MPH     Methyl Parathion Hydrolase
OP      Organophosphate
OPH     Organophosphate Hydrolase
PES     Potential Energy Surface
PON1    Paraoxonase 1
PXN     Paraoxon
QM      Quantum Mechanics
TS      Transition State
TST     Transition State Theory
US      Umbrella Sampling
1. Introduction

Organophosphate hydrolases are a diverse class of enzymes capable of degrading some of the most toxic compounds known to man. Introduced in the 1930s, these toxic man-made chemicals found widespread application as pesticides, herbicides, and chemical warfare agents. The fact that these compounds are normally not found in nature raises the question of how these enzymes evolved their functionality in such a short time-span.

Significant progress on this topic has been made in the last decades, revealing that their functionality most likely emerged as a promiscuous activity in enzymes where the native function is chemically similar to organophosphate hydrolysis, for example the hydrolysis of lactones. Substantial insight into the structure, function and evolution of these enzymes has been obtained from both experimental and computational studies, however, certain questions regarding the mechanisms and selectivity remain unanswered.

In this thesis, I address some of the open questions using a computational approach, that is, through modeling and simulation of chemical reactions in organophosphate hydrolases. Molecular simulations find tremendous utility in mechanistic studies, providing a direct view into the atomistic world of an enzymatic reaction, as well as giving valuable insight into various forces and interactions, which govern the reactivity and selectivity. The publications presented here span a range of topics from the development of the software used for the calculations and analysis, the derivation and parameterization of specific metal ion models required for later studies, and finally, to addressing several questions on mechanistic pathways, substrate specificity, and promiscuity in several members of the enzyme family, specifically, methyl parathion hydrolase (MPH), diisopropyl fluorophosphatase (DFPase), and paraoxonase 1 (PON1).

The content of the thesis is structured as follows. Chapter two provides a general overview of enzymes, paying special attention to kinetics, catalysis, as well as the connection between experimental and computational enzymology. The next chapter provides a thorough theoretical background on the computational methods used in this thesis, in particular the empirical valence bond (EVB) method. It progresses from building the EVB model from theories of quantum and molecular mechanics, to calculating a reaction free energy profile from statistical ensembles. The fourth chapter is an overview of the structure, function and evolution of organophosphate hydrolases and their substrates. Publications included in the thesis are summarized in the fifth and final chapter, followed by the concluding remarks and the reprints of the publications.
2. Enzymes

Essential for life as we know it, the macromolecules called enzymes are present in every living organism, mediating nearly all biologically relevant reactions, from fermentation of sugar and metabolism of alcohol to replication of DNA. The term enzyme finds its origins in the late 19th century studies on alcoholic fermentation in yeast extracts\(^1\) which, via demonstration of cell-free fermentation, put an end to the notion of vitalism and marked the beginning of modern biochemistry. \(^1\) Enzymes carry a tremendous capacity for accelerating the rates of chemical reactions, allowing the vital biochemical transformations to occur on a life-sustaining timescale with typical reaction rates of around 1 per second. \(^2\) Their remarkable nature becomes apparent when one considers the fact that the respective uncatalyzed rates in solution are up to \(10^{19}\) times slower, corresponding to half-lives of billions of years. \(^3\) This makes enzymes typically far more efficient at catalyzing chemical reactions than synthetic or inorganic catalysts. In addition to their efficiency, enzymes exhibit significant specificity towards their substrates, and are conveniently optimized to operate in very mild conditions of temperature and pH. \(^1\)

With the exception of a few catalytic RNAs, the majority of enzymes are proteins - natural polymers made of amino acids. Their three-dimensional structure or fold is defined by their respective and unique amino-acid sequence, and in turn defines the enzyme’s function. However, a direct and accurate prediction of either structure or function from the sequence, is for the time being limited to identification of homologous structures. The elucidation of protein structures is thus an indispensible part of enzymology, and in fact a prerequisite in the field of computational enzymology\(^2\). Yet, the structure-function relation is by no means a unique prescription. A structure alone can rarely be used to elucidate the enzyme’s function, while functionally related enzymes can adopt radically different structural folds. Furthermore, a single structure can facilitate several functions. \(^1\) This promiscuous character of enzymes, that is, the simultaneous adoption of many functionalities, seems contradictory to the previously mentioned concept of specificity. However, as discussed later in this section, such behavior is assumed to play a major role in evolution.

Interestingly, while the glue which holds and shapes the protein structure in-place is well understood to be the weak non-covalent interactions, the origin of catalysis, and in particular, why enzymes are so good at catalysis, has

---

\(^1\) Enzyme in Greek translates to ‘in yeast’

\(^2\) I hereby express my sincere gratitude to all structural biologists for their much-appreciated contributions to science.
been a subject of heated debate in the literature. Here, computational models prove to be invaluable in the analysis. Complementary to the ‘outside-in’ perspective of experiments, the computational techniques allow us to analyze the enzymatic reactions at an atomic level, to describe the mechanistic pathway, and finally to rationalize the various aspects of chemical reactions such as pH dependence, metal-ion dependence, temperature dependence, substrate specificity, as demonstrated in this thesis.

In addition I mention that while the search for knowledge is in itself a reward, the understanding of enzyme function carries an enormous potential in applications of drug-discovery and enzyme design, for example via identification of transition state analogues and mutagenesis hot-spots.

2.1 Kinetics of Enzymatic Reactions

The study of chemical kinetics is fundamental to understanding the nature of chemical reactions and thus by extension, the function of enzymes. The main goal of kinetics is to establish a rate equation (also rate law), which describes the relation between the concentration of reactive species and the rate of a given chemical reaction. The obtained model carries information on which moieties are involved in the reaction mechanism leading up to the rate-limiting step, and what is the nature of the kinetic rate constants, which characterize the individual steps in the mechanism. The kinetic rate constants are also most often the link between experimental and computational enzymology.

While kinetic analysis is an essential tool in the enzymologist’s tool-belt, it cannot provide sufficient evidence to confirm a particular mechanism on its own, although it can be used to reject the mechanisms that do not agree with the established rate law. Its usefulness becomes more evident when used in conjunction with other analyses such as isotope-effects, linear free energy relationships (LFER) and if dealing with enzymatic reactions - site-directed mutagenesis. These, as opposed to raw kinetics, directly identify bonds, which are forming and breaking, and provide invaluable insight into the nature of intermediate and transition states. [4] Furthermore, the dependence of rate constants on varying conditions such as pH, temperature and salt concentration often provides a great deal of mechanistic insight, especially in enzymatic environments. [5]

2.1.1 The Steady-State Approximation

Kinetic analysis of enzymes dates back to the pioneering work of scholars such as Brown and Henri in the early 20th century, who first realized the saturating effect of the enzyme-substrate complex and derived rudimentary rate laws governing a enzymatic reactions. [6] It was Michaelis and Menten who followed up with meticulous analysis of the problem by controlling for pH
and measuring only initial reaction rates, allowing them to accurately derive, in 1913, [7, 8] a rate law which to a good degree describes the majority of single-substrate enzymatic reactions. The equation was later slightly modified by Briggs and Haldane by including the steady state assumption [9] resulting in what is now known as the Michaelis-Menten equation.

The proposed mechanism for a single-substrate enzymatic reaction without any intermediate states was written as,

$$E + S \overset{k_1}{\underset{k_-1}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} E + P$$

where E, S, ES and P are the enzyme, substrate, the enzyme-substrate complex, also known as the Michaelis complex, and product, respectively. The formation of the ES complex is governed by diffusion ($k_1 \approx 10^9$ s$^{-1}$) and is in equilibrium with the breakdown to free enzyme and substrate, while the path to products is assumed to be irreversible. The rate of product formation and change in complex concentration is thus given as,

$$\frac{d[P]}{dt} = k_2[ES]$$  \hspace{1cm} (2.1.1)

$$\frac{d[ES]}{dt} = k_1[E][S] - k_-1[ES] - k_2[ES]$$  \hspace{1cm} (2.1.2)

Since the above expressions are not easily solvable, we turn to the steady-state assumption$^3$, which assumes that the concentration of [ES] is constant, i.e. the complex is in fast equilibrium with the free enzyme and substrate:

$$\frac{d[ES]}{dt} = 0$$  \hspace{1cm} (2.1.3)

By expressing the enzyme concentration in terms of the initial concentration $[E_0] = [E] + [ES]$ and combining the above relations, one can, with some algebraic shuffling, arrive at the equation,

$$v = \frac{d[P]}{dt} = \frac{k_2[E_0][S]}{k_-1 + \frac{k_2}{k_1} + [S]}$$  \hspace{1cm} (2.1.4)

which is typically written in a more general form, not specific to the two-step mechanism proposed above,

$$v = \frac{k_{cat}[E_0][S]}{K_m + [S]}$$  \hspace{1cm} (2.1.5)

where $k_{cat}$ and $K_m$ are the Michaelis-Menten parameters discussed below. A schematic representation of the dependence of rate $v$ on substrate concentration $[S]$ for a reaction obeying this rate law, is shown in Figure 2.1.

$^3$The equations are in fact directly integrable with lesser approximations. See for example Chapter 2.5 in ref [6].
Figure 2.1. A schematic representation of a profile of rate versus substrate concentration, for an enzyme reaction exhibiting Michaelis-Menten kinetics. The equations corresponding to the two limiting cases in substrate concentration are described in the text.

While the rate equation governing steady-state kinetics holds true for many complicated reaction pathways, it should be noted that it provides only an indirect, abstract view into the possible chaotic complexity of the actual mechanistic pathway. In order to resolve intermediate kinetic steps, one should preferably resort to more elaborate analysis, such as transient state kinetics.

**Michaelis-Menten parameters**

$k_{cat}$

$k_{cat}$ is a first-order rate constant, defining the rate of the reaction following the formation of the enzyme-substrate complex. While often attributed to a chemical step ($k_2$ in the above analysis), it strictly speaking provides only a lower limit on the rate constant of the rate-limiting step following the binding. [5] This can include conformational changes, chemical reactions, product release, or a complex combination of several such steps. [11] At saturating substrate concentrations ($[S] \gg K_m$, Figure 2.1), Equation 2.1.5 becomes,

$$\lim_{[S] \to \infty} v = v_{max} = k_{cat}[E_0]$$

showcasing the significance of $k_{cat}$ as the maximum possible rate for the reaction. In the field of computational enzymology, the kinetic constant has significant value as it is directly comparable with the effective free energy of activation via transition state theory. The majority of the work presented in this thesis in fact, relies heavily on experimentally obtained $k_{cat}$ values.

$K_m$

$K_m$, the *Michaelis constant*, is in its most strict definition: the concentration of the substrate, which yields a reaction rate that is half of the maximum

---

4More precisely, the rate constant corresponding to the highest *effective* energy barrier. [10]
rate (Figure 2.1). It equals to \((k_1 + k_2)/k_1\) in the simple two-step mechanism above and is generally, like \(k_{cat}\), a function of all the rate constants in the pathway. [11] It is often over-interpreted as the measure of binding \((K_m \approx K_d^{-1} = k_{-1}/k_1)\), however, this is reasonably valid only in a somewhat specific scenario of a simple two-step reaction mechanism \((k_{cat} = k_2)\) and slow chemical step with regards to dissociation \((k_2 \gg k_{-1})\). [11]

\[
k_{cat}/K_m
\]

\(k_{cat}/K_m\) is usually understood as a measure of the catalytic efficiency of an enzyme towards a particular substrate. It is an apparent second-order rate constant, describing productive substrate binding and is particularly relevant for conditions where \([S] \ll K_m\) (Figure 2.1), e.g. in vivo, where the specificity is governed both by binding and by catalysis:

\[
 v \approx \frac{k_{cat}}{K_m}[E_0][S]
\] (2.1.7)

In special cases where the catalytic step is actually faster than the dissociation into free reactants \((k_{cat} \gg k_{-1})\), the rate of the reaction is limited only by diffusion and the enzyme is said to have achieved ‘kinetic perfection’. [5]

### 2.1.2 The Connection to Thermodynamics

The thermodynamic potential of interest is the Gibbs free energy \(G\). The reaction free energy \(\Delta G_0\), can be obtained directly from the equilibrium constant,

\[
\Delta G_0 = -RT \ln K_{eq}
\] (2.1.8)

which is equal to the reaction free energy of the uncatalyzed reaction in solution (as per definition of a catalyst).

The kinetic rate constant is directly related to the activation free energy of a process via Transition State Theory (TST): [12]

\[
 k(T) = \frac{k_B T}{h} e^{-\Delta G^\ddagger/RT}
\] (2.1.9)

where the activation free energy \(\Delta G^\ddagger\) is defined at standard conditions and \(k_B\), \(h\), and \(R\) are the Boltzmann, Planck, and the ideal gas constant, respectively. The expression above is commonly referred to as the Eyring-Polanyi equation.

TST is based on the assumption that there exists a so-called transition state, a hypersurface in phase space, which divides the phase space into a reactant region and a product region. It is further assumed that the state of reactants and the metastable transition state are in a quasi-equilibrium, and that the crossing of the barrier will always lead to the product state (as opposed to forming the reactants, which is referred to as recrossing). [13] Corrections to TST
include recrossing events, non-equilibrium effects, and quantum tunneling, where non-equilibrium effects are assumed to be small for enzymatic reaction [14] and recrossing has been shown to be insignificant although slightly dependent on choice of reaction coordinate. [15] Quantum tunneling can have a significant impact on the rate, however, it has been demonstrated to be equally relevant in solution and in the enzyme, thus not contributing to catalysis. [16]

A visual representation of a free energy profile of a reaction in solution and in the enzyme is shown in Figure 2.2.

![Free energy profile](image)

**Figure 2.2.** A schematic free energy profile of an uncatalyzed reaction in solution (blue) and in the enzyme (green).

### 2.1.3 Factors Affecting the Rate

The rates of enzymatic reactions are in general sensitive to a number of solution and environment conditions, such as pH, ionic strength, viscosity, and temperature, as well as more direct perturbations such as isotope substitutions, a fact frequently exploited to gain mechanistic insight. This section gives a brief overview over pH and temperature effects, as these were directly employed in the thesis.

**pH Dependence**

The first measurements of pH dependence date back to 1911, when Michaelis and Davidsohn observed the characteristic bell-shaped plots of activity versus pH, which they attributed to the amphoteric nature of the enzyme. [17, 18] Although enzymes contain a myriad of possible protonation states, it is typical to find such single or double ionization curves. This is because the ionizations of residues which are either directly involved in the chemical reaction or are
responsible for maintaining the active conformation of the enzyme are significantly more important to the rate. [19] As ionizable amino-acid residues are frequently involved in enzymatic catalysis, either as proton donors and/or acceptors thereby directly by modifying the chemical mechanism, or by providing the appropriate electrostatic environment for catalysis, pH dependence analysis provides clues about the nature of the titratable residues which are suspected to be directly involved in the catalytic process.

However, any assignment of the apparent pKa values obtained from such experiments to specific residues (or groups of residues) should be done with caution due to the possible underlying complexity of the reaction mechanism, [20] and the fact that pKa values of protein residues are likely to be perturbed. [21] For definite conclusions, further analysis using methods like spectroscopy or site-directed mutagenesis are required. [22]

In a study relevant to the work done in Paper III, the authors measured pH-rate profiles of the wild-type as well as mutant enzyme, and observed a loss of pH dependence upon a specific mutation. This observation led them to conclude that the mutated histidine residue is involved in the mechanism, acting as a general base. [23] The conclusion was later largely dismissed, however the observed effect was not explained. In Paper III, we show the effect can be attributed instead to a change in the electrostatic environment upon ionization of the histidine residue.

Temperature Dependence
Temperature effects on rate (ignoring the extreme case of protein denaturation) can be used to resolve the contributions of enthalpy and entropy to the activation free energy. The activation free energy is defined as,

$$\Delta G^\ddagger = \Delta H^\ddagger - T \Delta S^\ddagger$$  \hspace{1cm} (2.1.10)

where the individual terms can be determined computationally by simply calculating $\Delta G^\ddagger$ along a certain temperature range and performing linear regression analysis.

Experimentally, it is reasonable to instead use kinetic rate constants, and plotting $\ln \frac{k}{T}$ versus $T$,

$$\ln \frac{k}{T} = -\frac{\Delta H^\ddagger}{RT} + \frac{\Delta S^\ddagger}{R} + \ln \frac{k_B}{h}$$  \hspace{1cm} (2.1.11)

where the above relation is a rearranged version of Equation 2.1.9.

The utility of temperature dependence analysis is for example demonstrated in a study by Wolfenden and coworkers, where they examined the differences in temperature dependence between enzyme-catalyzed and uncatalyzed reactions in solutions and revealed that the rate enhancement, that is catalysis, is in most cases largely enthalpic in nature. The authors claim this is consistent with the formation of new electrostatic and hydrogen bonds whilst proceeding from the ground state to the transition state. [24]
In addition, despite the challenges associated with addressing temperature dependence computationally such as poor precision and insufficient sampling, is has nevertheless been successfully employed to elucidate mechanistic insight in a number of recent studies, [25–27] including the work presented in Paper III.

2.2 Enzyme Catalysis

Enzymes are remarkably efficient and selective catalysts, increasing the rates of chemical reactions up to $10^{19}$ fold. [3] The exact source of their catalytic power has been a topic of heated debate in the literature, with various theories proposed based on evidence from both experimental and computational studies. One strategy employed by enzymes is the use of specific catalytic functional groups and/or cofactors, altering the mechanism and thus allowing the reaction to proceed through a pathway with lower effective activation energy. Examples of these include general-base catalysis, nucleophilic catalysis and metal-ion catalysis. [1] This type of ‘chemical’ catalysis has long been understood and well characterized by physical-organic chemistry experiments in solution. It has been demonstrated that these effects alone cannot account for a significant part of the catalytic power observed in enzymes. Early extrapolations from these experiments by Koshland suggested that the missing ‘secret of enzyme catalysis’ accounted for $10^{12}$-fold rate-acceleration. He hypothesized this effect to be due to precise orientation, strain, push-pull and/or microscopic environment factors. [28]

2.2.1 Transition State Stabilization

A key contribution to the understanding of enzyme function is Pauling’s 1946 hypothesis of transition state (TS) stabilization, where he, before knowing what an enzyme looks like, proposed that an enzyme’s active-site has evolved to be complementary to the transition state of a reaction, leading to its preferential binding relative to the Michaelis complex. [29] This theory gained further support with the discovery of transition state analogues by Wolfenden in 1969, revealing their remarkable binding affinities. [30, 31] With the use of computational methods, Warshel showed in 1978 that the electrostatic stabilization of the TS in enzyme is similar in magnitude than that of solution, thus cannot directly account for the stabilization effect. This puzzling observation was resolved by realizing that in order to achieve the same solvation, the dipoles in solutions require additional work for reorientation. This theory of preorganized electrostatics [32] states that the protein provides a preorganized polar environment that stabilizes the transition state, while the reaction in solution has to overcome an additional penalty of reorganizing the solvent.
molecules. [13] A schematic depiction is seen in Figure 2.3. A great deal of computational studies have since eluded to the fact that preorganization is indeed the dominant contribution to enzyme catalysis. [33]

**Figure 2.3.** Schematic depiction of preorganized electrostatics. The dipoles in the enzyme are preorganized to accommodate the transition state of the reaction, while work is required to reorganize the dipoles in water solution. Adapted ref. [33]

### 2.2.2 Ground State Destabilization

Among other more prominent theories are those relating the catalytic effect of enzymes to *ground-state destabilization*. In 1960, Bruice and Pandit showed that intramolecular-nucleophilic catalysis of ester hydrolysis accelerated the rate by up to $10^8$ [34] which gave rise to the idea that enzymes fix the reactive species in a steric conformation closely resembling the transition state. [35] The theory was later refined by introducing the term *near-attack conformations (NAC)* to describe the restriction of configurational space of the substrate in the reactants state. [36] Warshel and coworkers showed that the contribution from NAC in haloalkane halogenase is small in comparison to TS stabilization [37], while in the case of chlorismate mutase they concluded that it is only a side effect of transition state stabilization and not the reason behind catalysis. [38] Moliner and coworkers reached a similar conclusion, acknowledging that the equilibrium of substrate conformers is displaced towards the reactive (closer to TS) conformations, and concluding that both effects (TS stabilization and MC preorganization) share the same origin in the enzyme structure. [39, 40]

An influential theory emerged from Page and Jencks’ rationalization of the intramolecular catalysis observed by Bruice. [35, 41] They reasoned that the freezing out of the translational and rotational motions in a bimolecular reaction is a major contribution to the observed rate-acceleration. The so-called *entropy trap* theory proposes that the entropy of aligning the reacting species constitutes a significant part of the activation free energy in solution, while in the enzyme the cost of aligning is paid-for in the formation of the Michaelis complex. While conceptually similar to the NAC theory, the formation of NAC is proposed to be enthalpic in nature. The theory has been addressed by Warshel and coworkers via a computational approach, concluding that en-
tropic contributions are ‘much smaller than previous estimates’. In addition, Murphy showed that the analysis used by Page and Jencks highly overestimates the entropic contributions and Wolfenden demonstrated via temperature dependence analysis that the rate enhancement in enzymes is largely enthalpic in origin. Moreover, recent computational work by Kazemi and Åqvist casts some doubt into the several established explanations of enzyme function via entropy traps.

2.2.3 Dynamic Effects

In recent years, several theories attributing the catalytic power of enzymes to their dynamic motions have emerged. It is well understood that proteins are dynamic and fluctuating entities, adopting many different conformations, only a few of which might be relevant for catalysis. However, here the proposition is that the dynamics itself is responsible for rate enhancement. Examples of such proposals include ‘molecular memory’ as revealed by single-molecule experiments, dynamically promoted tunneling, correlated motions, and promoting vibrations. The attributed role of dynamics in catalysis has been heavily scrutinized in literature, chiefly by computational studies.

2.2.4 Metal Ion Catalysis

Nearly a third of all known enzymes require one or more metal ions for catalysis, where the highest proportion of enzymes, which employ metal ions, belong to the oxidoreductase, transferase and hydrolase functional classes. The metal ions can take on several roles in catalysis, acting either as redox reagents, providing electrostatic stabilization, or acting as Lewis acids in the activation of a reacting species. They play a central role in this thesis, from the parameterization of a physical model (Paper I) to utilization of the model in molecular simulations (Papers II and III).

All organophosphate hydrolases are metal-dependent enzymes, with either binuclear or mononuclear metal active sites, where the metal ions aid in nucleophile activation, electrostatic stabilization and/or substrate binding and positioning. They display similar levels of activity with a variety of different metals (Section 4.2), suggesting the mechanism of action is not specific to a particular metal, as is the case, for example, in the oxidoreductase class of binuclear manganoenzymes. Distinct metal-ion selectivity patterns were recently observed in methyl parathion hydrolase (MPH) and were linked to the functional evolution of its organophosphatase activity. These variations in activities and in particular their origin were addressed in the work presented in Paper II.
2.3 Selectivity and Promiscuity

Enzymes are often portrayed as very selective catalysts, readily discriminating between different types of reactions, substrates with quite similar structures, and even enantiomers. [1] It was this observation that led Fischer to describe the complementary nature of an enzyme and its substrate using the Lock and Key analogy. [58] Also referred to as specificity, the role of enzyme selectivity is essential in areas like DNA replication and protein biosynthesis where such high fidelity was guided by evolutionary pressures. [19] The defining measure relevant to biological systems is the specificity constant $k_{cat}/K_m$, combining both the effects of the ‘chemical’ rate as well as binding, which can vary significantly and often compensate. This constant is dependent solely on the free-energy binding of the transition state by the enzyme. [19]

In spite of the traditional picture of enzymes as highly selective, Jensen in 1976 proposed a hypothesis that ancient primitive enzymes possessed very broad specificities, permitting them to react with a wide range of related substrates and thus maximizing the catalytic versatility of the cell restrained by limited resources. [59] The subsequent divergence into specialized enzymes, via gene duplication, mutation and selection, led to the current diversity of enzymes and to increased metabolic efficiency. [60] It was later recognized that enzymes possess the ability to not only accept different substrates, but also to catalyze different reactions. O’Brien and Hershlag introduced the term catalytic promiscuity to describe this phenomena, and suggested that a low level of promiscuity could be a common characteristic of enzymes, allowing the evolution of new physiological functions via gene duplication followed by advantageous mutations. Furthermore, they demonstrate that even single mutations can substantially improve the promiscuous activities. [61]

From a mechanistic point of view, the facilitation of promiscuous activities can be achieved by conformational diversity, accommodation of alternative substrates, different protonation states, different subsites within the same active site, water assisted promiscuity, and alternative cofactors. [60] The latter is of particular relevance to this thesis. Studies on metalloenzymes have shown that the substitution of metal cofactors can significantly alter the activity towards the native substrate and promote catalysis of non-native promiscuous activities. [57, 62–64] A recent study on five functionally diverged enzymes from the metallo-β-lactamase superfamily, each reconstituted with six different metals and assayed against eight catalytically distinct hydrolytic reactions, revealed each metal isoform to have distinct activity profiles for the native and promiscuous activities. The authors suggest that the enzymes exist in vivo as ensembles of various metal isoforms and that the observed metal isoform heterogeneity expands the repertoire of promiscuous activities, enhancing the evolvability of enzymes. [57] Our work presented in Paper II provides focuses on one of these enzymes (MPH) in an attempt to explain the origin of the ob-
served metal-ion activity patterns for both the ‘native’ paraoxonase activity as well as (presumably) ancestral arylesterase activity.
3. Computational Modeling of Enzymatic Reactions

This chapter describes the theoretical concepts behind the computational tools employed in this thesis. It is divided into two conceptually distinct sections, the first of which deals with the concept of the internal energy of molecules at rest, providing a short overview of the theoretical frameworks of quantum mechanics and molecular mechanics, and finally, building on the two theories, derives the empirical valence bond (EVB) model. The second section deals with the concepts of molecular simulations, configurational sampling, statistical mechanics, and calculation of reaction free energy profiles.

Throughout the chapter, special attention is placed on describing the empirical valence bond approach in a verbose fashion, as the literature on EVB methodology can often be difficult to approach. By doing so I hope to alleviate the learning process for any aspiring EVB practitioner who happens to stumble upon this thesis. A more practical take on setting-up, analyzing, and troubleshooting EVB simulations was recently presented in ref. [65]

3.1 Defining the Potential Energy Landscape

The most important quantity in computational chemistry is arguably that of energy. Since energy comes in many forms (e.g. kinetic, electrostatic, Gibbs free energy, etc.), I should emphasize that in the following section I consider only the potential energy of a system, where the system is defined as a stationary arrangement of atoms. Such a depiction corresponds to a molecular system at complete rest, essentially at 0 K, although excluding even the zero-point vibrations arising from Heisenberg’s uncertainty principle. Additionally, I will limit the discussion to cases without any external electric or magnetic fields. We may now write this in a form of a function that returns the energy of the system and is depended solely on the coordinates of the nuclei of the atoms:

\[ E = f(r_1, r_2, r_3, \ldots, r_N) \]  

(3.1.1)

where \( r_i = (x_i, y_i, z_i) \) is the coordinate vector for atom \( i \) and \( N \) is the number of atoms. In terms of Cartesian coordinates, the energy function is a 3N-dimensional hypersurface. However, since we are interested only in the relative positions of the nuclei with respect to each other (i.e. molecular geometry)

26
and since we do not have external influences, this description is somewhat redundant. By removing the translational and rotational degrees of freedom, we can instead express the function in terms of internal coordinates (inter-atomic distances and angles), reducing the dimensionality to $3N - 6$ ($3N - 5$ for linear systems). This function, which relates the internal geometry of a chemical system to its energy is fundamental to computational chemistry, and is referred to as the Potential Energy Surface (PES). It allows us to visualize the relationship between molecular geometry and its energy, find energetically stable geometries (minima) or transition states (saddle points), determine the relative stability of conformations, visualize a dihedral profile, calculate a reaction’s enthalpy of activation, etc. By exploring the PES using sampling techniques (see Section 3.2), one can also obtain macroscopic thermodynamic quantities such as free-energy and entropy. An example of a PES is depicted in Figure 3.1.

![Figure 3.1](image)

**Figure 3.1.** An example of a potential energy surface. Shown is a calculated two-dimensional PES of an $S_N2$ reaction between chloride and methyl bromide. The stationary points are marked with R (reactant state), ‡ (transition state), and P (product state).

The concept of an energy function that is dependent only on coordinates of the nuclei is made possible only by invoking the Born-Oppenheimer approximation. [66] It states that the motions of electrons can be decoupled from the nuclei due to the considerable difference in their velocities. In other words, since electrons are much faster, they ‘relax’ instantaneously to any
change in nuclear geometry, resulting in a unique (and non-changing) electron distribution for each geometry of the nuclei.

The following chapter gives a brief introduction into the two main computational models for calculating potential energy surfaces - Quantum Mechanics (QM) and Molecular Mechanics (MM), and building on this knowledge, the last part of this chapter, addresses the derivation of the model used extensively in this thesis - the Empirical Valence Bond (EVB).

3.1.1 Quantum Mechanics

At the turn of the 20th century, following some of the greatest advances in understanding the fundamental laws of nature, including J.C. Maxwell’s work on electromagnetism and J. Gibbs’ work on thermodynamics, the wide-held belief that most of fundamental discoveries of science had already been made was shattered with the discovery of the theory of Quantum Mechanics (QM). Max Planck’s quantum hypothesis of blackbody radiation and Albert Einstein’s explanation of the photoelectric effect implied that light is made of particle-like packets with discrete quantized energy. This wave-particle duality of light was later extended by Louis de Broglie to include all matter. [12] Shortly after, in 1926, a fundamental equation of quantum mechanics describing the wave-like behavior of the (non-relativistic) quantum world, was conceived by Erwin Schrödinger: [67]

$$\hat{H} \psi(r) = E \psi(r)$$

The solutions to the time-independent Schrödinger equation shown above are the total energy of the system, and the so-called stationary-state wave function $\psi(r)$.

The fundamental postulate of quantum mechanics is that for any system, there exists a wave function which completely describes its behavior and from which, all the information about this system can be derived. The wave function has no classical analogy; however, its square represents the probability density for the system (e.g. electron) in space. By acting on this wave function via a quantum mechanical operator (e.g. the Hamiltonian $\hat{H}$, we get as a result the same wave function (eigenfunction), multiplied by a scalar quantity (eigenvalue) - an observable corresponding to the operator. In Equation 3.1.2, the eigenvalue is the energy $E$ of the system, corresponding to the Hamiltonian operator $\hat{H}$, which is comprised of terms for the kinetic and potential energies of electrons and nuclei:

$$\hat{H} = -\sum_i \frac{\hbar}{2m_e} \nabla_i^2 - \sum_k \frac{\hbar}{2m_k} \nabla_k^2 - \sum_i \sum_k \frac{e^2 Z_k}{r_{ik}} + \sum_{i<j} \frac{e^2}{r_{ij}} + \sum_{k<l} \frac{e^2 Z_k Z_l}{r_{kl}}$$

where indices $i$ and $j$ denote electrons, $k$ and $l$ denote nuclei, $\hbar$ is the Planck constant $\hbar$ divided by $2\pi$, $m_e$ is the mass of the electron, $m_k$ is the mass of the
nucleus $k$, $\nabla^2$ is the Laplace operator, $e$ is the charge of the electron, $Z$ is the atomic number and $r$ is distance. [68]

Due to the correlated motions of electrons and nuclei, Equation 3.1.2 can only be solved analytically for systems with a maximum of two particles (e.g. H, He$^+$), demanding the use of approximate methods and numerical approaches. The Born-Oppenheimer approximation, already mentioned in Section 3.1, remedies this problem somewhat, with little cost in accuracy. By decoupling the electronic and nuclear motion, the nuclei are treated as stationary. The second term in Equation 3.1.3 thus vanishes, while the fifth term becomes constant. A more cumbersome issue, namely electron correlation, however, still persists, and the accuracy of the quantum chemical method is largely determined by how well it is able to describe this phenomena. The prototypical ab initio (lat. from first principles) method for solving the Schrödinger equation - the Hartree-Fock (HF) method solves the issue by neglecting electron correlation altogether, with each electron experiencing only an effective average potential. Post-HF methods such as Möller-Plesset perturbation theory (MPn), configuration interaction (CI) and coupled cluster theory (CC) try to account for correlation in varying degrees, resulting in significant improvement in accuracy compared to HF, although doing so at a considerable computational cost. This becomes especially noticeable in larger systems, since the scaling of calculations increases from $N^4$ (HF), to $N^5$ (MP2) and $N^7$ (CCSD(T)). [69]

The Variational Principle
The energy of a system can be expressed from Equation 3.1.2 by multiplying both sides with the complex conjugate wave function $\psi^*$ and integrating over all space:

$$E_{\text{exact}} = \frac{\int \psi^* \hat{H} \psi \, dr}{\int \psi^* \psi \, dr}$$

(3.1.4)

Direct evaluation of this equation is generally not possible as the wave function $\psi$ is not known. However, it turns out that if we substitute $\psi$ with any other arbitrary trial function $\Phi$, the resulting energy will always be greater-than or equal to the exact energy, according to the variational principle:

$$E \geq \frac{\int \Phi^* \hat{H} \Phi \, dr}{\int \Phi^* \Phi \, dr} = E_{\text{exact}}$$

(3.1.5)

This provides us with a lower limit for the energy, suggesting that the lower the trial energy is, the better the trial wave function. [12]
If the trial wave function is dependent on some arbitrary parameters, the lowest-energy solution is obtained by minimizing the energy with respect to those parameters. Often, wave functions are expressed as linear combinations of the type,

$$\Phi = \sum_{i=1}^{N} c_i \phi_i$$  \hspace{1cm} (3.1.6)

where $\phi_i$ is for the time being unknown, and $c_i$ are the so-called expansion coefficients. Here, the minimum in energy is obtained by,

$$\frac{\partial E}{\partial c_1} = \frac{\partial E}{\partial c_2} = \ldots = \frac{\partial E}{\partial c_N} = 0$$  \hspace{1cm} (3.1.7)

which results in a set of linear equations,

$$\sum_{i=1}^{N} c_i (H_{ki} - ES_{ki}) = 0 \quad \forall \, k$$  \hspace{1cm} (3.1.8)

often expressed in the form of a secular equation,

$$\begin{bmatrix}
H_{11} - ES_{11} & H_{12} - ES_{12} & \ldots & H_{1N} - ES_{1N} \\
H_{21} - ES_{21} & H_{22} - ES_{22} & \ldots & H_{2N} - ES_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
H_{N1} - ES_{N1} & H_{N2} - ES_{N2} & \ldots & H_{NN} - ES_{NN}
\end{bmatrix} = 0$$  \hspace{1cm} (3.1.9)

where $H_{ki} = \int \phi_k \hat{H} \phi_i \, d\mathbf{r}$ and $S_{ki} = \int \phi_k \phi_i \, d\mathbf{r}$. The energy of this system can now be obtained by solving the above secular equation, where the lowest energy solution is termed the ‘ground state’. The rest are deemed the ‘excited states’. [68]

The method above assumes that we have a vague idea of what $\phi_i$ in Equation 3.1.6 looks like in order to build a valid wave function. Two distinct approaches to solve this problem are presented in the following section.

**Molecular Orbital and Valence Bond Theory**

Only a year after Schrödinger presented his wave equation, Heitler and London used it to unravel the mystery of chemical bonding between non-polar atoms. They acknowledged the indistinguishable nature of electrons and wrote the wave function of the H$_2$ molecule as a linear combination of states H$_a^{(1)}$H$_b^{(2)}$ and H$_a^{(2)}$H$_b^{(1)}$,

$$\psi_{HL} = \frac{1}{\sqrt{2}} \left[ \phi_a(1)\phi_b(2) + \phi_a(2)\phi_b(1) \right] \cdot \chi$$  \hspace{1cm} (3.1.10)

where $\phi_a(1)$ denotes electron ‘1’ in atomic orbital 1s of atom H$_a$, and $\chi$ is the (antisymmetric) spin function $[\alpha(1)\beta(2) - \alpha(2)\beta(1)]$. 

30
The above formulation accounts for around 75% of the bonding energy, where the attractive force between the shared valence electrons emerges due to electron exchange, a characteristically quantum-mechanical effect. Known as the Heitler-London wave function, this covalent description of bonding forms the basis of the quantum Valence Bond (VB) theory. [70]

The remainder of the bonding character is obtained by considering also the ionic states, where the electrons are both located on either one of the hydrogen AO-s. [70] The full expression for a valence bond wave function is thus written as a linear combination of all possible covalent and ionic states (described via localized AO or fragment orbitals):

$$\psi_{\text{VB-full}} = \sum_{i=1}^{N} c_i \psi_i$$

which is in the case of the H\textsubscript{2} molecule simply,

$$\psi_{\text{VB-full}} = \lambda \psi_{\text{HL}} + \mu (\psi_{\text{ionic}}^a + \psi_{\text{ionic}}^b)$$

where $\psi_{\text{ionic}}^a = \frac{1}{\sqrt{2}} \phi_a(1) \phi_a(2)$, and $\lambda$ and $\mu$ are covalent and ionic coefficients, respectively. A more complex example of a valence bond wave function describing an S\textsubscript{N}2 reaction is shown in Figure 3.2.

The appeal of VB theory is that it ‘resonates’ with chemists due to its inherent connection to the practical conceptual models we use on a daily basis to describe molecular geometry (hybridization) and bonding patterns (resonance structures). However, as it was historically more difficult to implement into computer algorithms, as well as due to some apparent failures of the theory, its use in computational chemistry has declined since its inception, giving way to Molecular Orbital (MO) theory. [70, 72]

As opposed to VB, molecular orbital theory treats electrons as completely delocalized, occupying the so-called molecular orbitals (MO). The theory was conceived in the early 1930s by Hund and Mulliken, gaining significant momentum through the descriptive power of the Hückel MO theory, as well as practical application of Fukui’s frontier MO theory for describing HOMO/LUMO interactions, and finally, achieving mainstream adoption with the first appearance of dedicated software tools. [70]

The complete (full CI) wave function is in MO constructed as a linear combination of Slater determinants,

$$\Phi_{\text{CI}} = a_0 \Phi_{\text{HF}} + \sum_s a_s \Phi_S + \sum_D a_D \Phi_D + \cdots = \sum_i a_i \Phi_i$$

where HF, S, and D denote the Hartree-Fock, Singly excited and Doubly excited determinants, respectively. [73] The individual determinants ensure antisymmetry of the wave function, and are populated using single-electron molecular orbitals (MO), which are in practice approximated as linear combinations of atomic orbitals (LCAO), also known as basis sets. The mathematical representation of a basis set is the same as that of Equation 3.1.6, where $c_i$
Figure 3.2. Valence bond states of an SN2 halogen exchange reaction $X + CH_3−Y → X−CH_3 + Y$ obtained by employing a 4e/3c description (4 valence electrons shared on 3 localized atomic orbitals centered on atoms X, C, and Y). The wave functions shown on the right side use the shorthand notation $|\bar{x}\bar{y}\rangle = 1/\sqrt{2}[\phi_x(1)\phi_y(2)\alpha(1)\beta(2)−\phi_x(2)\phi_y(1)\alpha(2)\beta(1)]$. The Heitler-London wave function is used to describe covalent bonds (e.g. $\psi_{\text{HL}}(C−Y) = |c\bar{y}−\bar{c}y\rangle$). [70] Relative energies of the states are displayed on the diagram on the left. The figure is adapted from refs. [71, 72].

is the molecular orbital expansion coefficient, and $\phi_i$ is an AO-like function - typically a Gaussian-type orbital (GTO). [68]

While the two theories take on the challenge from completely opposite sides (localized electron pairs vs delocalized electron clouds), they converge to the same solution when taking into account all possible ionic states in VB theory, and considering full CI in MO theory: [70]

$$\psi_{\text{VB−full}} \equiv \psi_{\text{MO−CI}}$$

(3.1.14)

3.1.2 Molecular Mechanics

The mathematical model used to obtain the PES could in principle be any arbitrary mathematical formulation which adequately maps the geometry of a molecule to its energy. For example, we might consider using Hook’s law to describe bonds between atoms even though we are well aware that bonds are not material, let alone look like mechanical springs (section 3.1.1). It turns out, however, that the harmonic approximation provides an accurate description of the energy profile of a bond stretch, so long as we stay close to the equilibrium value. [12]
This simplified representation is the basis of Molecular Mechanics (MM), a model describing atoms as completely classical objects bonded by mechanical springs (“ball and springs” model). The fundamental building blocks are thus atoms, with the electron density modeled in abstract form (e.g. using partial atomic charges). An important consequence is that the bonding information cannot be obtained from the model itself as in QM but must be provided explicitly. [73] In addition to the Born-Oppenheimer approximation, the model assumes additivity of contributions (bonds, angles, etc.) and transferability of parameters beyond the training set. [74]

A set of parameters and potential energy functions used in a molecular mechanics model is commonly referred to as a force-field. Modern biomolecular force fields (e.g. OPLS-AA/M [75], AMBER-ff14SB [76], and CHARMM36m [77]), find their roots in the pioneering work of Lifson and coworkers [78] in late 1960s, sharing common features such as the Harmonic potential for describing stretching and bending vibrations, Coulomb’s potential for electrostatic interactions, and Lennard-Jones’s potential for Pauli repulsion and dispersion interactions. Additional periodic terms are usually introduced to correct for the torsional interactions due to rotation around a bond (dihedral angle), as well as out-of-plane bending (improper dihedral). Parameterization of these functional terms is typically performed by fitting to experimental data such as X-ray structures, vibrational spectra, densities, dipole moments, heat of vaporization, heat capacity, solvation free-energies and/or by using QM to calculate the PES and other properties. [74]

The full functional used in the OPLS-AA family of force fields has the following form,

\[
U_{MM}(r) = \sum_{\text{bonds}} \frac{1}{2} k_b (r - r_0)^2 + \sum_{\text{angles}} \frac{1}{2} k_\theta (\theta - \theta_0)^2 +
\]

\[
\sum_{\text{dihedrals}} \sum_{n=1}^{4} \frac{1}{2} K_n \left[ 1 + \cos(n \phi - \gamma_n) \right] + \sum_{\text{impropers}} \frac{1}{2} K_\xi (1 - \cos 2 \xi) +
\]

\[
\sum_{\text{nonbonded pairs} i,j} k_e \frac{q_i q_j}{r_{ij}} + \sum_{\text{nonbonded pairs} i,j} 4 \varepsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \quad (3.1.15)
\]

where \( k_b \) is the bond stretch force constant \( r \) is the bond distance, \( r_0 \) is the equilibrium bond distance, \( k_\theta \) the angle bend force constant, \( \theta \) is the bending angle, \( \theta_0 \) is the equilibrium bending angle, \( K_n \) is the barrier height of the \( n \)-th term of the Fourier series describing the dihedral potential, \( \phi \) is the dihedral angle, \( \gamma_n \) is the phase shift angle, \( K_\xi \) is the barrier height of the improper dihedral potential, \( \xi \) is the out-of-plane bending angle, \( q \) is partial atomic charge, \( k_e \) is the Coulomb’s constant, \( r_{ij} \) is the distance between atoms \( i \) and \( j \), and \( \varepsilon_{ij} \) and \( \sigma_{ij} \) are Lennard-Jones parameters. A visual depiction of the interactions described by a force field is shown in Figure 3.3.
The benefit of using such a model lies in its simplicity - the calculation of energy on modern hardware is very fast for even large biomolecular systems comprised of several thousand atoms, allowing the use of sampling techniques as described in Section 3.2. A significant downside of molecular mechanics is the inherent inability of describing chemical reactions due to fixed bonding patterns, although there exist specialized force fields such as ReaxFF, which allow for changes in bonding. [79] An additional drawback of traditional force fields with fixed partial atomic charges is the lack of electronic polarizability, which can reportedly account for around $10 - 20\%$ of the interaction energy between molecules at the van der Waals distance. [80] Force fields which account for polarizability via strategies such as induced point dipole (IPD) model, shell model, and fluctuating charge (FQ) model, are however gaining significant momentum in recent years. [80, 81] Finally, the simplistic and heavily parameterized nature of the molecular mechanics model brings with it
the requirement of constant validation, to ensure that the use of said models in practice is justifiable, [77, 82–84] and its use outside of the training set is always subject to errors which are difficult to quantify.

3.1.3 Multiscale Models
The enormous computational cost of modeling large condensed-phase systems with high-level QM methods deems their use in studying reactions in full enzyme environments, at least for time being, unfeasible. One option is to truncate the system to include only the reactive species and surrounding active site residues, thus allowing the use of very accurate high-level QM. The so-called QM-cluster model has been shown to be very effective for elucidating reaction mechanisms, particularly in the case of redox-active metalloenzymes. [85, 86] The downside of this model is that the complex heterogenous enzyme environment is largely neglected and described using a polarizable continuum model. Additionally, it has been demonstrated that the reliability of QM-cluster models can be heavily dependent on the size of the included system, with results in some situations differing up to 60 kJmol$^{-1}$. [87] While it is nowadays possible to use quite large clusters (>300 atoms) [88] which should in principle alleviate such issues, this presents another so-called local-minima problem (more in Chapter 3.2).

On the other side of the spectrum, molecular mechanics presents a too simplistic physical model, as it generally does not allow for forming of breaking of bonds due to its fixed pre-assigned bonding patterns. However, it can be combined with the more accurate QM, forming the so-called multiscale model. 

Multiscale models date back to as early as 1976, to the pioneering work done by Warshel and Levitt on lysozymes. [89] The authors introduced the idea of partitioning the system into two regions, where the reactive region is treated with a quantum mechanical model, and the surrounding region is treated with molecular mechanics force fields. The total potential energy was thus expressed as a sum of the classical potential, the quantum potential, and the coupling of the two.

\[
U = U_{\text{classical}} + U_{\text{quantum}} + U_{\text{quantum/classical}} \quad (3.1.16)
\]

This so-called QM/MM hybrid model takes the best from both worlds - the accuracy of quantum mechanics for a proper description of chemical reactivity, and the simplicity of molecular mechanics for an efficient depiction of the environment and its effect on the quantum region.

The QM/MM approach has now matured into a well established tool for studying enzymatic reactions and their mechanisms, elucidating the origin of catalysis, selectivity, promiscuity, etc., and ultimately, to aid in the design of novel biocatalysts. Reviews on the topic include [90–92]. A particular type of multiscale model, which was used extensively in this thesis, is presented in the next chapter.
3.1.4 Empirical Valence Bond

I will conclude the chapter on potential energy surfaces by combining the ideas presented in previous sections on quantum mechanics, molecular mechanics and multiscale modeling into a discussion on the empirical valence bond (EVB) model, which was used throughout this thesis.

Conceived in 1980 by Warshel and Weiss, EVB is presented as a simple and reliable approach for comparing reactions in solution and in enzymes. [93] The model builds on the valence bond theory, describing the system (e.g. chemical reaction) as a mixture of valence states. However, instead of explicitly describing the electron structure, it approximates the valence states with classical force-field functions (although these are still coupled in a QM-like formalism). While the model has received heavy criticism in the past due to its apparent simplicity, [94] it has nevertheless proven itself over the years to be an invaluable tool for studying enzymatic (and other [95]) reactions, [96] perhaps most prominently by providing invaluable insight into the electrostatic origins of catalysis. [33]

**Derivation of a Two-State EVB Model**

We will derive here a two-state EVB model as used in the thesis, although it should be evident that the derivation can be easily extended to three or more states. As a pedagogical system, we will make use of an abstract S\textsubscript{N}2 reaction analogous to that depicted in Figure 3.2:

\[
A + BC \rightarrow AB + C \tag{3.1.17}
\]

The total wave function of the reaction in Figure 3.2 is comprised of six VB states, where four valence electrons were chosen to be active and are shared between three orbitals centered on three atoms. [72] Here, we consider only the two lowest energy VB states, as these have significantly larger contributions to the overall energy. In accordance to Equation 3.1.11, we write out the wave function of the system as,

\[
\psi = c_1 \phi_1 + c_2 \phi_2 \tag{3.1.18}
\]

Here we note that the states described by \( \phi_1 \) and \( \phi_2 \) have a direct physical representation, namely, they correspond to the diabatic states of the reactants \((A + BC)\) and products \((AB + C)\), respectively.
The total energy of the system is written as (Equation 3.1.5):¹,

\[ E = \int \psi \hat{H} \psi \, dr \]  
\[ \int \psi^2 \, dr \]  
\[ = \int (c_1 \varphi_1 + c_2 \varphi_2) \hat{H} (c_1 \varphi_1 + c_2 \varphi_2) \, dr \]  
\[ \int (c_1 \varphi_1 + c_2 \varphi_2)^2 \, dr \]  
\[ = \frac{c_1^2 H_{11} + 2 c_1 c_2 H_{12} + c_2^2 H_{22}}{c_1^2 S_{11} + 2 c_1 c_2 S_{12} + c_2^2 S_{22}} \]  
(3.1.20)

We now add some constraints on the wave function, defining the valence states as orthogonal and normalized (orthonormal),

\[ S_{ij} = \delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \]  
(3.1.21)

acknowledging that some of the error introduced by neglecting the overlap between valence states (i.e. \( S_{12} = 0 \)) will be compensated by \( H_{12} \) through calibration. [93] By applying the variational principle (\( \frac{\partial E}{\partial c_1} = \frac{\partial E}{\partial c_2} = 0 \); see Equation 3.1.7), we obtain the following set of equations,

\[ c_1 (H_{11} - E) - c_2 H_{12} = 0 \]  
(3.1.22a)

\[ c_2 (H_{22} - E) - c_1 H_{12} = 0 \]  
(3.1.22b)

for which the non-trivial solution can be obtained from the secular equation:

\[ \begin{vmatrix} H_{11} - E & H_{12} \\ H_{12} & H_{22} - E \end{vmatrix} = 0 \]  
(3.1.23)

By diagonalizing the matrix and solving for \( E \), we obtain an analytical expression for the EVB ground state potential energy function as the lowest-energy solution of the secular equation,

\[ E_g = \frac{1}{2} (H_{11} + H_{22}) - \frac{1}{2} \sqrt{(H_{11} - H_{22})^2 + 4 H_{12}^2} \]  
(3.1.24)

where the high-energy solution is the potential describing the excited-state. A schematic depiction of the ground state potential is shown in Figure 3.4.

The expansion coefficients (corresponding to a particular energy state), can be expressed from equations 3.1.34 using the imposed constraint on orthonormality of valence states,

\[ \int \psi^2 \, dr = 1 = \sum c_i c_j S_{ij} = \sum c_i c_j \delta_{ij} = \sum c_i^2 \]  
(3.1.25)

¹\( H_{12} = H_{21} \) is a consequence of the Hamiltonian being a Hermitian operator.
Figure 3.4. Schematic representation of the ground state EVB potential $E_g$ and the potentials $H_{11}$ and $H_{22}$ for a reaction $A + B – C \rightarrow A – B + C$ (Equation 3.1.24), described with a two-state, constant $H_{12}$ EVB model. Adapted from ref. [65].

leading to,

$$c_1^2 + c_2^2 = 1$$

(3.1.26)

and

$$c_1^2 = \frac{1}{1 + \frac{H_{12}^2}{(H_{22} - E_g)^2}}$$

(3.1.27)

The weights of each valence structure can be directly obtained from the coefficients, for example according to the Coulson–Chirgwin scheme [72]

$$w_i = c_i^2 + \sum_{i \neq j}^N c_i c_j S_{ij} = c_i^2$$

(3.1.28)

As of yet, we have not made any empirical approximations, and the Equation 3.1.24 is still considered to be pure VB, albeit approximated to an extent where the resulting energies are unlikely to be very accurate. It is at this point that we introduce the empirical part of Empirical Valence Bond, by first realizing that $H_{11}$ and $H_{22}$ have a clear physical meaning - they represent the energies of the two diabatic, non-coupled, valence states:

$$H_{11} = \int \phi_1 \hat{H} \phi_1 \, d\mathbf{r} = E_1$$

(3.1.29a)

$$H_{22} = \int \phi_2 \hat{H} \phi_2 \, d\mathbf{r} = E_2$$

(3.1.29b)
We then approximate the complex, many-electron integrals with analytical molecular mechanics (MM) potential functions:

\[ H_{11} = U_{MM,RS}(r) \]  
\[ H_{22} = U_{MM,PS}(r) + \alpha_{12} \]

where \( \alpha_{12} \) is the energy difference between the individual states at infinite separation and is included to correct for the fact that both \( U_{MM,RS} \) and \( U_{MM,PS} \) are potential energy functions with arbitrary reference points. It is also known as the gas-shift and is obtained via calibration of the EVB potential.

With the exception of reactive bonds, the interatomic interactions (in \( H_{11} \) and \( H_{22} \)) are described with typical force-field functions as described in Chapter 3.1.2. For bonds that are involved in the reaction, that is, the bonds that are forming or breaking, the harmonic approximation is no longer justifiable. The Morse potential is used instead as it provides a more accurate representation of the potential energy of a bond stretch (Figure 3.3),

\[ U_{Morse}(r) = D_e \left(1 - e^{-a(r-r_0)}\right)^2 \]

where \( D_e \) is the depth of the potential energy minimum, \( a \) is the ‘steepness’ of the potential, and \( r_0 \) is the equilibrium distance for the bond.

While the diagonal elements of the hamiltonian are easily obtain via force-field parameterization of the respective valence states, the off-diagonal \( H_{12} \) requires a different approach.

**The Resonance Integral \( H_{12} \)**

The off-diagonal element \( H_{12} \), also known as the resonance integral, represents the quantum coupling of states, but unlike its diagonal counterparts, does not have a classical analogy. It was initially approximated by Warshel and coworkers by using gas-phase experimental data. [93, 97] In later studies, an analytical, distance-dependent exponential form was introduced,

\[ H_{12} = A_{12} e^{-\mu_{12}(r_{12}-r_{12}^0)} \]

and the unknown parameters were obtained by fitting \( E_g \) to reproduce an *ab initio*-calculated potential energy surface. [98, 99] Due to the challenges associated with obtaining a reliable potential energy surface at the time, Åqvist and Warshel [100] made a further approximation by neglecting the geometric dependence of \( H_{12} \) (\( \mu = 0 \)),

\[ H_{12} = \text{const}. \]

pointing out that the difference between energy profiles in solution and enzyme is not very sensitive to the exact value of the resonance integral. The calibration of the ‘constant \( H_{12} \)’ EVB model is significantly less demanding.
as it requires only one observable\(^2\) (instead of many points on the PES) - the experimental (or calculated) activation energy. This makes it particularly convenient for studying enzymatic reactions as we will see later. The approach, albeit simple, has been successfully employed in a variety of studies, including the work presented in Papers II, III, and IV.

While the simplified treatment of the resonance integral presented above yields good agreement for reproducing the energy of activation, the overall shape of the potential energy surface has been shown to be distorted and unable to accurately reproduce transition state frequencies. [101] More elaborate functions have been introduced, where a more accurate representation of the PES was required. Examples include the generalized gaussian by Chang and Miller, [102] the distributed gaussian by Sonnenberg and coworkers, [101, 103] or a simpler single gaussian function by Glowacki et al. [95]. For a more thorough review, see Chapter 2 in ref. [96].

**Measuring Catalysis**

The convenience of the EVB model, insofar as enzymology is concerned, lies in the fact that the off-diagonal element \(H_{12}\) is assumed to be transferable between different environments (e.g. gas-phase, solution, enzyme, mutant-enzyme, etc.). Interestingly, the assumption was for a long time considered somewhat *ad hoc*, rationalized mainly by its consistency with empirically observed linear free energy relationships, until finally validated using Constrained DFT. [104]

By partitioning the system into a reactive region (e.g. substrate and catalytic residue) and its surroundings (e.g. enzyme and/or solvent) \(^3\), and taking into consideration the transferability of \(H_{12}\), we may write the matrix elements of the multiscale model as,

\[
H_{11} = H_{11}^r + H_{11}^{rs} + H_{11}^s \\
H_{22} = H_{22}^r + H_{22}^{rs} + H_{22}^s \\
H_{12} = H_{12}^r
\]

where superscripts \(r\) and \(s\) denote the internal interactions of the reactive region and the surroundings, respectively, and \(rs\) denotes the interactions between the two regions.

This scheme allows us to calibrate the unknown parameters found in the reactive-region hamiltonian (\(H_{12}\) and \(\alpha_{12}\)) using any environment of our choosing (provided the reaction in this environment is sufficiently characterized). The calibrated parameters can then be transferred to the environment of interest simply by changing the \(H^{rs}\) and \(H^s\) terms in \(H_{11}\) and \(H_{22}\) (thus in \(E_g\)). This

\(^2\)In addition to one observable required to obtain the gas-shift \(\alpha_{12}\).

\(^3\) Analogous to procedure in Section 3.1.3, with the exception that the two regions are not described using different models.
allows one to capture the change in environments, or in the specific case of moving from water solution to an enzyme - catalysis\(^4\).

A natural choice of reference for studying enzymatic reactions is the corresponding reaction in water solution (see Section 2.2), and in typical a scenario (employing a two-state, constant \(H_{12}\), EVB potential), the calibration is done such as to reproduce the solution activation free energy (\(\Delta G^{\ddagger}\)) and the reaction free energy (\(\Delta G_0\)). The calculation of free energies (as opposed to potential energies) is the topic of the next chapter.

3.2 Exploring the Configurational Space

As the complexity of the system increases, that is, as the system grows in size, an additional challenge is presented, beside the obvious greater computational cost of calculating potential energies. The large number of loose degrees of freedom in a typical condense-phase system results in a highly-dimensional and rugged potential energy surface, with many local minima. In order to describe such a system at some finite temperature, a single configuration of the system is not enough. All thermally accessible microscopic states must be considered to describe experimentally obtained macroscopic observables, most notably, free energy.

The commonly used minimization (also optimization) approach for studying chemical reactions, in which activation and reaction enthalpies are obtained by locating stationary points along a reaction coordinate, should be used with care when studying enzymatic reactions. [105] Conformational changes of the complex enzyme environment can significantly affect the result - the formation of a single hydrogen bond, for example, can shift the activation enthalpy by \(\sim 20kJ\). [106] In studies employing this approach, the local-minima problem is often addressed by repeating the calculation using snapshots from classical molecular dynamics; however, it is at the moment still unclear how the individual calculations should be averaged, with some evidence suggesting exponential averaging to be preferred over arithmetic. [107] Moreover, it was recently demonstrated that the number of optimizations required for converged results is substantially larger than what is typically done in practice. [106]

A more rigorous approach involves the calculation of free energies by explicit (and extensive) configurational sampling, where the link between the individual microstates obtained from a simulation and experimental macroscopic observables, such as free energy and entropy, is obtained through statistical mechanics. Unfortunately, for the time being, \textit{ab initio} and Density Functional Theory (DFT) QM/MM studies are limited to a few hundred picoseconds of simulation time, [108] with proper sampling typically requiring

\(^4\)Not accounting for acid/base, nucleophilic, metal ion, etc. catalysis.
the use of semi-empirical methods [109–111] or force-field based approaches such as EVB and others. [112, 113]

The main focus of this chapter is to provide a comprehensive derivation of the EVB MD/FEP/US methodology as is typical used in EVB studies of enzymatic reactions and in the work presented in this thesis.

**Essential Statistical Mechanics**

The central concept of statistical mechanics is the so-called *ensemble*. An ensemble is a collection of microscopic states of the system that make up the macroscopic whole. It is constrained by a set of variables that define a particular thermodynamic state, for example, constant number of particles, constant volume, and constant temperature (NVT) in the case of a closed, isothermal system. The ensemble corresponding to these conditions is the *canonical ensemble* discussed below.

A fundamental postulate of statistical mechanics is that an experimentally observed value of any quantity $f$ is equal to the *ensemble average* of that quantity,

$$f_{\text{obs}} = \langle f \rangle_{\text{ens}} = \sum_i p_i f_i$$  \hspace{1cm} (3.2.1)

where $p_j$ is the probability (weight) of the $j$-th discrete microstate. The probability of states is, in the canonical ensemble, proportional to the Boltzmann factor $e^{-\beta E}$ and written as,

$$p_i = \frac{e^{-\beta E_i}}{\sum_j e^{-\beta E_j}} = \frac{e^{-\beta E_i}}{Q_{\text{NVT}}}$$  \hspace{1cm} (3.2.2)

where $\beta = (k_b T)^{-1}$. $Q$ is the sum over all states, also known as the *partition function*. The partition function describes the system completely and from it, one can derive all macroscopic quantities. The bridge to phenomenological thermodynamics is given by the relationship,

$$A = \frac{1}{\beta} \ln Q_{\text{NVT}}$$  \hspace{1cm} (3.2.3)

where $A$ is Helmholtz free energy, the characteristic function for the NVT ensemble. [12, 114]

In the classical limit, the partition function is expressed instead as an integral over continuous position and momentum,

$$Q_{\text{NVT}} = c \int e^{-\beta H(r,p)} dr dp$$  \hspace{1cm} (3.2.4)

where $H$ is the classical Hamiltonian and $(r,p)$ defines a point in the $6N$-dimensional position-momentum space termed the *phase space*. $c$ is the classical proportionality constant. By separating the kinetic and the potential contributions to the partition function we can write,

$$Q_{\text{NVT}} \propto Z_{\text{NVT}} = \int e^{-\beta U(r)} dr$$  \hspace{1cm} (3.2.5)
$Z_{NVT}$ is the excess (also potential or non-ideal) contribution to the partition function dependent only on the position of the particles and is known as the configuration integral or configuration partition function. [114, 115] $Z_{NVT}$ is the operational element of statistical mechanics in the determination of free energy. [116]

The expression for calculating an ensemble average from a classical continuous distribution is written as:

$$
\langle f \rangle = \int f(\mathbf{r}) e^{-\beta U(\mathbf{r})} d\mathbf{r} = \frac{\int f(\mathbf{r}) e^{-\beta U(\mathbf{r})} d\mathbf{r}}{Z_{NVT}} \tag{3.2.6}
$$

Generally, a direct evaluation of the integrals in the above equation is not possible in practice. The ensemble average is instead evaluated indirectly by exploring (sampling) the thermally accessible configurations on the potential energy landscape via Monte Carlo (MC) or Molecular Dynamics (MD) simulations. [115]

### 3.2.1 Molecular Dynamics

While difficult to prove, it is strongly believed that almost all many-body systems obey the so-called ergodic hypothesis, which states that an ensemble average of a system is equal to its time-average: [115]

$$
\langle f \rangle = \lim_{t \to \infty} \frac{1}{t} \int_0^t f(t) \, dt \tag{3.2.7}
$$

The evolution of a system through time is the province of Molecular dynamics (MD), a simulation method where a (classical) time-dependent trajectory of particles is obtained by numerically integrating Newtonian equations of motion. The only requirement is knowing the interatomic forces, which can be readily obtained by calculating the gradient of the potential energy surface. The forces yield the acceleration ($\mathbf{F} = m \mathbf{a} = -\nabla U$), the change in velocity ($\mathbf{a} = \partial \mathbf{v} / \partial t$), and finally, the change in position ($\mathbf{v} = \partial \mathbf{r} / \partial t$). The equations of motion are in practice solved numerically using a finite-difference approach, where the change in position and momentum is re-evaluated at regular time intervals $\delta t$, for example, using the following scheme:

$$
\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \mathbf{v}(t) \delta t + \frac{1}{2} \mathbf{a}(t) \delta t^2 \tag{3.2.8}
$$

$$
\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \frac{1}{2} \left[ \mathbf{a}(t + \delta t) + \mathbf{a}(t) \right] \delta t \tag{3.2.9}
$$

The example above is the commonly used velocity Verlet algorithm for integrating the equations of motion. [117] When using a finite-difference approach, the time-step $\delta t$ should be at least an order of magnitude smaller than 43.
the characteristic time of the fastest motion in the system (e.g. O–H stretch has a frequency of around 10 fs, thus $\delta t$ should be 1 fs or lower).

The ensemble average from a (sufficiently long, i.e. equilibrated) MD simulation is obtained as a simple arithmetic average over a finite set of configurations sampled in the simulation,

$$\langle f \rangle \approx \frac{1}{n} \sum_{i} f_i$$  \hspace{1cm} (3.2.10)

where $n$ is the number of configurations. [115]

Direct integration of Newton’s equations of motion results in a constant total energy. The sampled NVE ensemble (microcanonical ensemble) has limited applicability in biochemistry as experiments are typically performed in NVT or NPT conditions. The canonical ensemble can be sampled instead by artificially regulating the temperature of the system with a thermostat. These can be roughly divided into stochastic (Andersen, Langevin, Parinello-Bussi) and deterministic (Nose-Hoover, Berendsen weak-coupling) [117, 118]. Additionally, a barostat can be introduced if constant pressure (and thus density) is required (isobaric-isothermal ensemble). However, the near incompressibility of liquids such as water means that the difference between the two ensembles, that is, the $PV$ (pressure-volume) work contribution to free energy, is negligible as long as the simulation is conducted at the average volume for that state. [119]

3.2.2 Free-Energy Perturbation

Free energy is one of the most fundamental concepts in physical (bio)chemistry, describing the spontaneity and equilibria of natural processes. It deals with some of the most interesting questions in biochemistry such as binding affinities, protein folding, and (enzyme) reaction kinetics, as well as guiding application driven sciences of drug-discovery and enzyme-design.

Our interests (both experimental and computational) lie typically in the evaluation of free energy differences, as opposed to their absolute values. Using Equation 3.2.3 we write (omitting NVT for clarity),

$$\Delta A = -\frac{1}{\beta} \ln \frac{Q_1}{Q_0} = -\frac{1}{\beta} \ln \frac{Z_1}{Z_0}$$ \hspace{1cm} (3.2.11)

$$\approx \Delta G$$  \hspace{1cm} (3.2.12)

where indices 0 and 1 denote two physical states (e.g. bound and unbound ligand; reactant and transition state). Note that the approximation to Gibbs free energy is valid due to the negligible difference between NVT and NPT ensembles in terms of free energy (see previous section).

While the individual configurational integrals $Z_0$ and $Z_1$ can in principle be obtained from a simulation, they will in practice be inadequately sampled in
high-energy regions which contribute significantly to the ensemble average. [116] We thus instead use some clever rearranging,

\[
\Delta G = -\frac{1}{\beta} \ln \frac{\int e^{-\beta U_1} \, dr}{\int e^{-\beta U_0} \, dr} = -\frac{1}{\beta} \ln \frac{\int e^{-\beta U_1} e^{-\beta U_0} \, dr}{\int e^{-\beta U_0} \, dr} \\
= -\frac{1}{\beta} \ln \frac{\int e^{-\beta (U_1 - U_0)} e^{-\beta U_0} \, dr}{\int e^{-\beta U_0} \, dr} \\
= -\frac{1}{\beta} \ln \langle e^{-\beta (U_1 - U_0)} \rangle_0
\]

(3.2.13)

where \( \langle \ldots \rangle_0 \) denotes an average over the configurational space sampled on the potential \( U_0 \) and the dependence of \( U_1(\mathbf{r}) \) and \( U_2(\mathbf{r}) \) on \( \mathbf{r} \) has been omitted for clarity. The configurations sampled on \( U_1 \) are not used, although they can be used to reduce the error by calculate the reverse process (1 \( \rightarrow \) 0) and taking the average of the two. [116]

The relation described in Equation 3.2.13 was originally derived by Zwanzig in 1954 [120] and is known as Free Energy Perturbation (FEP). Though exact in its derivation, it is accurate only for very small perturbations, that is, when the configurational spaces of the two states overlap significantly. More precisely, when the microstates representative of state 1 have been sampled in state 0. [121]

A solution to this problem is stratification, also known as multistage sampling. A new potential is devised as a linear combination of the two potentials,

\[
U(\lambda) = \lambda U_0 + (1 - \lambda) U_1
\]

(3.2.14)

where \( \lambda \) is the coupling parameter taking on a range of values between 1 \((U = U_0)\) and 0 \((U = U_1)\). The free energy difference can then be expressed as a sum over all intermediate perturbations:

\[
\Delta G_{0 \rightarrow 1} = -\frac{1}{\beta} \sum_{i=0}^{n-1} \ln \langle e^{-\beta [U(\lambda_{i+1}) - U(\lambda_i)]} \rangle_i
\]

(3.2.15)

While the method is typically associated with calculations of ligand binding and solvation free energies via alchemical transformations [122], it has significant value in the context of studying enzymatic reactions. It is used either in the context of alchemical transformations or for calculating the perturbation of ‘switching’ from lower- to higher-level potentials. FEP is for example used in the EVB FEP/US methodology (as described in Section 3.2.4), in the QM/MM-FE approach introduced by Yang and coworkers, in the QTCP approach by Rod and Ryde [113], in the DH-FEP approach by Thiel and coworkers [111] and others (see refs in [90, 123]).
A popular alternative for estimating free energies from two ensembles is the Bennett’s Acceptance Ratio (BAR) [124] as well as its multi-state extension MBAR. [125]

3.2.3 Umbrella Sampling and Potential of Mean Force
Typical enzymatic reactions are associated with relatively high energy barriers ($k_{cat} = 1\text{s}^{-1} \Rightarrow \Delta G^\ddagger = 17.6\text{kalmol}^{-1}$) and thus occur on timescales far greater than those obtainable by present day simulations (in the range of milliseconds). Since it is not possible to simply measure the reaction rate by observing the turnover in the simulation\footnote{Unless the energy barrier is sufficiently low.}, we instead resort to using various computational techniques for calculating the free-energy profile of a reaction (thus indirectly the rate). One of the more common techniques is presented below and put into the context of the EVB methodology in the next chapter.

We start by defining the probability distribution sampled on potential $U_u$ as a function of some reaction coordinate $\xi$,

$$P_u(\xi) = \frac{Z_{u,\xi}}{Z_u} = \frac{\int \delta[\xi(\mathbf{r}) - \xi]e^{-\beta U_u(\mathbf{r})}d\mathbf{r}}{\int e^{-\beta U_u(\mathbf{r})}d\mathbf{r}} = \langle \delta[\xi(\mathbf{r}) - \xi] \rangle_u \quad (3.2.16)$$

where the probability distribution is obtained by integrating out all other degrees of freedom. [126] The Dirac delta function $\delta[\xi(\mathbf{r}) - \xi]$ is equal to 1 for configurations where $\xi(\mathbf{r}) = \xi$ and is equal to 0 for others.

From the probabilities along $\xi$, we can readily calculate the differences in free energy:

$$\Delta G_u(\xi_0 \rightarrow \xi_1) = -\frac{1}{\beta} \ln \frac{Z_{u,\xi_1}}{Z_{u,\xi_0}} = -\frac{1}{\beta} \ln \frac{P_u(\xi_1)}{P_u(\xi_0)} \quad (3.2.17)$$

The probabilities along the coordinate can be obtained from the simulation by counting the frames when $\xi = \xi_0$ and $\xi = \xi_1$, i.e., by constructing a histogram. The obtained free-energy profile along a reaction coordinate is commonly referred to as the Potential of Mean Force (PMF). [127]

However, the computational cost for simulating rare events such as enzymatic reactions is at the moment still beyond reach, [128] which means that the obtained probability distributions will in practice be poorly-defined (or more likely non-existent) in the high-energy regions of the configurational space.

To address the issue of insufficient sampling, Torrie and Valleau introduced in 1976 an ingenious work-around known as Umbrella Sampling (US). [129] Essentially, the idea is that one can modify the sampling potential by adding an artificial biasing potential $U_{bias}$, in order to sample more (i.e. neighboring regions) of the configuration space,

$$U_b = U_u + U_{bias} \quad (3.2.18)$$
and then retrieve unbiased properties (in this case the probability distribution) using the following relationship,

\[ P_u(\xi) = \frac{\langle \delta[\xi(r) - \xi] \rangle_{b}}{\langle e^{+\beta U_{bias}} \rangle_{b}} \]  

(3.2.19)

where \( \langle \ldots \rangle_{b} \) are ensemble averages obtained using the modified potential \( U_{b} \).

The term ‘Umbrella’ originates from the fact that the new potential is sampling (‘covering’) more of configurational space.

This methodology is most commonly applied to studying processes of binding and chemical reactions in a multi-stage approach. Due to the fact the a single biasing potential is unlikely to span across the whole reaction coordinate, several windows \( \xi_j \) are defined at discrete points on the coordinate \( \xi \) and for each window, a separate simulation is performed to sample the respective configurational space. The windows are associated with biasing potentials \( U_{bias}^j \), which force the system to sample the region of configurational space around \( \xi_j \) (Figure 3.5 A).

The traditional approach (distinct from the EVB FEP/US methodology described in the next section) is to employ a harmonic biasing potential of the type \( U_{bias}^j(\xi(r)) = k_j[\xi(r) - \xi_j]^2 \), in which case Equation 3.2.19 can be rearranged to

\[ P_u^j(\xi) = e^{+\beta U_{bias}^j} \cdot \langle \delta[\xi(r) - \xi]\rangle_{b,j} \cdot \langle e^{-\beta U_{bias}^j} \rangle_{b,j} \]  

(3.2.20)

leading to the equation for calculating the free-energy profile along \( \xi \), within the particular window described by the biasing potential \( U_{bias}^j \) (Figure 3.5 B):

\[ G_u^j(\xi) - K^j = -\frac{1}{\beta} \ln P_u^j(\xi) \]  

(3.2.21)

\[ = -\frac{1}{\beta} \ln P_b^j(\xi) - U_{bias}^j(\xi) + f^j \]  

(3.2.22)

where \( K^j \) is an arbitrary reference point (e.g. \( K^j = G_u^j(\xi_x) - \frac{1}{\beta} \ln P_u^j(\xi_x) \)).

The last term on the r.h.s. of the Equation 3.2.22 is a window-specific constant, which must be determined to construct to full PMF \( G_u(\xi) \) along all windows (Figure 3.5 C). It represents the free energy associated with introducing the window potential,

\[ f^j = -\frac{1}{\beta} \ln \langle e^{-\beta U_{bias}^j}\rangle_{b,j} \]

\[ = -\frac{1}{\beta} \ln \frac{\int e^{-\beta U_{bias}^j} e^{-\beta U_{bias}^j} dr}{\int e^{-\beta U_{bias}^j} dr} = -\frac{1}{\beta} \ln \frac{\int e^{-\beta U_{bias}^j} dr}{\int e^{-\beta U_{bias}^j} dr} = -\frac{1}{\beta} \ln \frac{Z_{b,j}}{Z_u} = \Delta G(u \to b_j) \]  

(3.2.23)

\( ^{6}U_{bias} \) is dependent only on \( \xi(r) \), therefore the second part in the numerator can be integrated out directly.
Figure 3.5. Schematic representation of the umbrella sampling approach. (A) Use of biased potentials to sample high-energy regions of phase space. (B) Calculating unbiased free-energy profiles in each window - Equation 3.2.22. (C) Constructing the full potential of mean force by calculating the window offsets $f_j$.

and cannot be directly calculated from the simulations [126]. The $f_j$ values are obtained by either adjusting the values such that the PMF’s from neighboring windows overlap, either by calculating the relative window offsets $(f_{j+1} - f_j)$ via FEP, or as is most commonly done, to use the Weighted Histogram Analysis Method (WHAM) by Kumar et al. [130, 131]. Additionally, the MBAR method by Shirts and Chodera can be used to get the most optimal estimate for the free energy profile. [125]

3.2.4 EVB FEP/US

The EVB ground-state potential $E_g$ can in principle be used in conjunction with any of the more traditional enhanced sampling techniques such as the ‘harmonic bias’ umbrella sampling presented in the previous chapter, metadynamics [132], or Yang’s minimum free energy path (MFEP) approach [112] and its derivatives.

However, one of the strengths of EVB in studying enzymatic reactions is actually the sampling methodology which naturally extends the EVB model, referred to as FEP/US (also mapping approach).

Energy-gap reaction coordinate

The FEP/US methodology employs the energy gap reaction coordinate, which has been shown to allow for more efficient sampling and better localization of TS compared to the traditional geometric coordinates. [133]. The energy gap is defined as the difference between the two Hamiltonians:

$$\Delta \varepsilon = H_{11} - H_{22} = \xi \tag{3.2.24}$$

The method does not require any assignment of reaction coordinate (or other) variables, which is often a non-trivial task due to the many degrees of freedom involved. [134] The energy gap reaction coordinate traces a path from reactants to products (Figure 3.2.4 A) where all of the coordinates of the system are relaxed, and crosses the transition state (as defined by $E_g$) exactly (at $\Delta \varepsilon = 0$).
The Mapping Potential

Following the umbrella sampling procedure described in the previous section, we define a biased potential as a sum of the ground-state EVB potential $E_g$ and a biasing potential $U_{bias}$,

$$U_m = E_g + U_{bias} \quad (3.2.25)$$

where $U_m$ is the so-called mapping potential. Instead of defining an explicit biasing potential $U_{bias}$, the mapping potential is defined as a linear combination of potentials $U_{11}$ and $U_{22}$,

$$U_m(\lambda) = \lambda U_{11} + (1 - \lambda) U_{22} \quad (3.2.26)$$

where $U_{11}$ and $U_{22}$ are the potential energy functions describing the EVB diabatic states 1 and 2. Gradually changing the coupling parameter $\lambda$ from 1 to 0 alchemically transforms the reactive subsystem from diabatic state 1 to diabatic state 2, allowing the environment to polarize itself (to match the reactive subsystem) along the effective reaction coordinate $\lambda$. [135] In each mapping window, the configurational space is sampled via molecular dynamics (Figure 3.2.4 B).

Such a mapping potential is by design not a function of the unknown EVB parameters $\alpha_{12}$ and $H_{12}$. By combining Equations 3.1.24, 3.1.30 and 3.2.26, we can write out the implicit biasing potential,

$$U_{bias} = U_m - E_g$$

$$= \alpha_{12}(\lambda - 1) + \xi \left( \lambda - \frac{1}{2} \right) + \frac{1}{2} \sqrt{\xi^2 + 4H_{12}^2} \quad (3.2.27)$$

which includes the unknown parameters in the bias. The practical implication of this is that it is possible to run two-state, constant-$H_{12}$, EVB trajectories without prior knowledge of the unknown EVB parameters. This fact is especially useful for performing simulations of reference reactions, which are used to calibrate the missing parameters. The calibration can thus be performed in post-processing, as explained below.

Constructing the Free Energy Profile

In order to obtain the unbiased $E_g$ probability distributions along the $\xi$ reaction coordinate from the biased simulations on the mapping potential, we employ the umbrella sampling approach discussed in Section 3.2.3,

$$P_g^j(\xi) = \frac{\langle \delta[\xi(r) - \xi] \cdot e^{+\beta U_{bias}^j} \rangle_{m,j}}{\langle e^{+\beta U_{bias}^j} \rangle_{m,j}} \quad (3.2.28)$$

$$= \langle \delta[\xi(r) - \xi] \cdot e^{-\beta(E_g - U_m^j)} \rangle_{m,j} \cdot \frac{Z_m^j}{Z_g} \quad (3.2.29)$$
where $j$ is the index of the simulation window defined by $\lambda_j$.

The PMF in the $j$-th simulation window is then obtained analogous to Equation 3.2.22,

\[
G^j_g(\xi) - K^j = -\frac{1}{\beta} \ln P^j_g(\xi)
\]

\[
= -\beta^{-1} \ln \langle \delta[\xi(r) - \xi] \cdot e^{-\beta(E_g - U^j_m)} \rangle_{m,j} - \frac{1}{\beta} \ln \frac{Z^j_m}{Z_g}
\]

(3.2.30)

(3.2.31)

where $K^j$ is an arbitrary reference point (e.g. $K^j = G^j_u(\xi_x) - \frac{1}{\beta} \ln P^j_u(\xi_x)$). The question now is - how to calculate the last term on r.h.s. in order to ‘stitch’ the mapping windows into a full free energy profile?

Due to the convenient nature of the mapping potential $U_m(\lambda)$, the free energy difference between neighboring mapping windows can be readily obtained via the FEP approach. Since $Z_g$ is constant and is common to all windows, we can substitute it with a more convenient value, for example $Z^0_m$, corresponding to the first mapping window. The evaluation of the final term then becomes a straightforward multi-stage FEP calculation (Figure 3.6 C):

\[
\Delta G^j_m = \sum_{i=0}^{j-1} -\frac{1}{\beta} \ln \frac{Z^{i+1}_m}{Z^i_m}
\]

\[
= \sum_{i=0}^{j-1} -\frac{1}{\beta} \ln \langle e^{-\beta(U_m(\lambda_{i+1}) - U_m(\lambda_i))} \rangle_{m,i}
\]

(3.2.32)

The final expression for constructing the full ($\lambda$- and $\xi$-dependent) PMF can now be obtained as (Figure 3.6 D):

\[
G^j_g(\xi) - K^j = \Delta G^j_m - \frac{1}{\beta} \ln \langle \delta[\xi(r) - \xi] \cdot e^{-\beta(E_g - U^j_m)} \rangle_{m,j}
\]

(3.2.33)

However, as different $\lambda_j$ windows can contribute points to one $\xi$ bin, the values are additionally combined into a ‘smoothed’ profile by performing a weighted average (Figure 3.6 E),

\[
G_g(\xi) - K = \sum_{j=0}^{\infty} G^j_g(\xi) \cdot \frac{n(\xi, \lambda_j)}{n(\xi)}
\]

(3.2.34)

where $n(\xi, \lambda_j)$ is the number of points from window $\lambda_j$ that fall into bin $\xi$, and $n(\xi)$ is the total number of points in this bin. The same technique is also reported by Woolf and Roux [136], and is in later work referred to as weighted-PMF, or W-PMF. [131]

Alternatively one can use the weighted histogram analysis method (WHAM) in concert with $U_{bias}$ (Equation 3.2.27) to extract a statistically optimal PMF. However, the resulting free-energy profiles have been demonstrated to be largely
Figure 3.6. A visualization of the FEP/US approach discussed in this chapter based on the model reaction shown in Figure 3.4. (A) Energy gap contour levels overlayed on the two dimensional $E_g$ ground-state potential surface. Dashed lines denote negative values. (B) Configurations sampled on the mapping potential $E_m(\lambda)$ (color-coded by $\lambda$-value) in relation to the actual ground-state potential energy surface $E_g$. (C) Free-energy profile for the alchemical transformation of State 1 ($\lambda = 1$) to State 2 ($\lambda = 0$), obtained with Equation 3.2.32. (D) Free-energy profiles obtained using Equation 3.2.33. Points on the plot are color-coded based on $\lambda$ as in B. Note that many mapping windows can contribute to a single energy-gap bin. (E) Bin-averaged free-energy profile obtained by weight-averaging the data in D using Equation 3.2.34. The figure is adapted from ref. [65].

equal to those obtained with the above FEP/US procedure (Figure 3.7). While it is conceivable that WHAM would provide a better estimate for severely poorly sampled simulations, such issues should instead be addressed at the simulation level. We discuss some of the practical aspects of performing EVB simulations including convergence criteria and troubleshooting in a recent volume of Methods in Enzymology. [65].
Figure 3.7. Comparison of free-energy profiles obtained using FEP/US and WHAM. The simulation data is from simulations of DFPase hydrolysis of DFP presented in Paper III. The difference in free-energy of activation is less than 0.1 kcal mol⁻¹.

3.2.5 Linear Response Approximation

The free-energy difference between two states, A and B, can be also estimated using the linear response approximation (LRA),

$$\Delta G(A \rightarrow B) = \frac{1}{2} \left[ \langle U_B - U_A \rangle_A + \langle U_B - U_A \rangle_B \right]$$  \hspace{1cm} (3.2.35)

where $\langle \ldots \rangle_A$ is the ensemble average sampled on $U_A$, and $\langle \ldots \rangle_B$ is the ensemble average sampled on $U_B$. The above relationship is obtained by assuming that the free energy functionals describing the two states are parabolas of equal curvature. \[137\] A more rigorous derivation involves taking a cumulative expansion of the Zwanzig equation (3.2.13) and truncating the series after the second term. \[138\]

The method has been used extensively in the past for estimating electrostatic free energies of ligand binding, while also forming the basis for Åqvist's linear interaction energy (LIE) method. \[137, 139\] Additionally, it provides a convenient and computationally inexpensive way for estimating small perturbations to free-energy processes. For example, it has been used to estimate the contributions of individual amino-acid residues to ligand binding (group contributions). \[140, 141\]

Of special interest to enzymology and protein design is the contribution of individual amino-acids to catalysis. These are obtained via virtual deletion of the residue in question while moving from the reactant state to the transition state:

$$\Delta \Delta G^\dagger(p \rightarrow p') = \frac{1}{2} \left[ \langle U_{TS}^{p'} - U_{R'}^{p'} \rangle_R + \langle U_{TS}^{p'} - U_{R'}^{p'} \rangle_{TS} \right]$$

$$- \frac{1}{2} \left[ \langle U_{TS}^p - U_R^p \rangle_R + \langle U_{TS}^p - U_R^p \rangle_{TS} \right]$$

$$\approx - \frac{1}{2} \left[ \langle U_{TS}^jR - U_R^j \rangle_R + \langle U_{TS}^j \rangle_{TS} - \langle U_R^j \rangle_{TS} \right]$$  \hspace{1cm} (3.2.36)
Note that the configurational space of the system $p'$, where a residue has been ‘deleted’, is approximated to be equal to that of the full system $p$.

The main advantage of this method is that the calculation is computationally inexpensive and can be performed as post-processing on existing EVB trajectories (corresponding to the reactant and transition state), therefore not requiring additional simulations.
4. Studied Systems

4.1 Organophosphates

Organophosphates (OP) are a large class of diverse organic chemicals, derivatives of phosphoric, phosphonic, and phosphinic acid, sharing a characteristic phosphoryl or thiophosphoryl bond (P=O, P=S), and differing in varying side-chain groups comprising of oxygen, carbon, sulfur, nitrogen and halogen elements. [142]

The term organophosphate is commonly associated with a particular subset of these compounds which were synthesized since the 1930s for use as pesticides (e.g. parathion, paraoxon), nerve agents (e.g. sarin, tabun, VX), flame retardants, parasiticides, and herbicides. [142] Several of the more prominent compounds belonging to this class are displayed in Figure 4.1.

![Chemical structures of several common organophosphate pesticides and chemical warfare agents.](image)

*Figure 4.1. Chemical structures of several common organophosphate pesticides and chemical warfare agents.*

These man-made *phosphotriester* derivatives are highly toxic, acting as irreversible inhibitors of acetyl choline esterase, causing overstimulation of the
nervous system, ultimately leading to paralysis and death. [143] The compounds are very effective phosphorylating agents, reacting readily with a serine residue in the AChE active site, forming a stable phosphorylated complex and inactivating the enzyme. [144] Beside the contamination and associated negative impacts on terrestrial and aquatic ecosystems due to widespread use of OP pesticides, there are between 750,000 to 3 million human poisonings reported annually due to exposure, either because of accidental spillage, terrorist attacks or suicide attempts. [145, 146] Effective methods for degradation of organophosphates are thus required, with OP-hydrolyzing enzymes showing promise as environmentally friendly, non stoichiometric bioremediation agents or as bioscavengers for treating OP poisoning. [145, 147, 148]

4.1.1 Non-enzymatic Hydrolysis

Hydrolysis of phosphate esters can proceed through either: (1) a dissociative step-wise mechanism (elimination-addition; $D_N + A_N$) which involves the formation of a metaphosphate intermediate; (2) a concerted $S_N2$-like mechanism ($A_N D_N$) proceeding through a single transition state; (3) an associative step-wise mechanism (addition-elimination; $A_N + D_N$) involving the formation of a pentavalent phosphorane intermediate. These are however only limiting cases, where the continuum from the dissociative to associative case is described by the tightness of the transition state. A tight TS has higher bonding character than the ground state, a loose TS has lower bonding than the ground state, and a synchronous TS has similar bonding as the ground state. [149]

The nature of the transition state gradually moves from loose in monoesters to tight in triesters as evident from kinetic isotope effect (KIE) studies. A larger KIE for the non-bridging oxygen atom observed in triester hydrolysis suggests a tighter transition state and weakening of the $P-O_{(nonbridge)}$ bond. The KIE’s for the leaving group oxygen in triester hydrolysis are highly variable depending on the nature of the leaving group, but are in general smaller than mono- and diesters, indicating larger bonding character and thus also a tighter transition state. [149]

Linear free energy relationships (LFER), particularly Brønsted correlations for nucleophiles ($\beta_{NUC}$) and leaving groups ($\beta_{LG}$) have been used extensively in the past in order to elucidate the nature of phosphate hydrolysis. In general, a loose transition state is associated with a negative and larger $\beta_{LG}$ due to a more pronounced substituent effect on the more polar transition state, while a positive and large $\beta_{NUC}$ is indicative of a tight transition state. [149] Additionally, breaks in Brønsted plots are indicative of either a change in mechanism (upward deviation with increasing nucleophile $pK_a$) or a change in the rate-limiting step and thus the existence of a phosphorane intermediate (downward deviation with increasing nucleophile $pK_a$). [150]
Phosphotriesters are readily hydrolyzed in alkaline solution, while the background rates of hydrolysis in neutral solutions are around $10^5$ fold slower, as shown in Table 4.1.1. These reactions often display good Brønsted correlations without breaks indicative of intermediates or alternate pathways. The $\beta_{\text{nuc}}$ values in Table 4.1.1 fall in the range of synchronous to tight transition states. However, the precise mechanistic pathway has been shown to be strongly dependent on the nature of the nucleophile and leaving group, modulating not only the relation between the kinetically equivalent nucleophilic and general-base catalysis, but also suggesting the existence of phosphorane intermediates in certain situations. Based on the data in literature, which is described in the following sections, we modeled the alkaline hydrolysis of paraoxon in Paper II as a concerted nucleophilic substitution. In Paper III, we modeled the reactions of DFP and paraoxon hydrolysis in acetate buffer as concerted processes, general-base catalyzed in case of paraoxon, and both general-base catalyzed and nucleophile catalyzed in case of DFP. The second-order rate constant for hydrolysis of DFP in acetate buffer solution was calculated from Table 1 in ref. [151] to be $k_B = 2.84 \times 10^{-5} \text{M}^{-1}\text{s}^{-1} (25^\circ\text{C}, 0.4\text{M NaCl})$.

### Table 4.1.1

<table>
<thead>
<tr>
<th>Substituents</th>
<th>$k_{\text{OH}}$ (M$^{-1}\text{s}^{-1}$)</th>
<th>$k_w$ (10$^{-4}$ s$^{-1}$)</th>
<th>$\beta_{\text{NUC}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paraoxon</td>
<td>EtO EtO PNP</td>
<td>0.075 [152]</td>
<td>0.0021 [153]</td>
</tr>
<tr>
<td>DFP</td>
<td>i-PrO i-PrO F</td>
<td>0.3 [154]</td>
<td>0.017 [155]</td>
</tr>
<tr>
<td>Sarin</td>
<td>i-PrO Me F</td>
<td>25.8 [156]</td>
<td>0.5 [154]</td>
</tr>
<tr>
<td>DNPDEP</td>
<td>EtO EtO DNP</td>
<td>0.29 [158]</td>
<td>0.01 [158]</td>
</tr>
<tr>
<td>DNPPEMP</td>
<td>EtO Me DNP</td>
<td>24.2 [157]</td>
<td>1 [157]</td>
</tr>
</tbody>
</table>

*a Comprehensive compilations of values can be found in the review by Cox and Ramsay [159] and review by Lassila et al. (Sup. Information) [149]

b PNP: p-nitrophenol; DNP: 2,4-dinitrophenol

### Concerted versus Associative

Experimental data suggests that phosphate mono-, di- and triesters generally react via a concerted pathway. [149] The absence of a pentavalent phosphorane intermediate and a single transition state pathway was demonstrated for the reaction of diphenyl p-nitrophenyl phosphate with substituted phenolates. [160] Furthermore, $^{18}$O-exchange study revealed no stable intermediates in hydrolysis of DFP and analogues. [155]

However, extensive LFER and stereochemical studies on cyclic phosphotriesters have suggested a continuum of mechanisms from concerted to associative, where the later is followed at least to some extent by strongly basic
nucleophiles and poor leaving groups. Large $\beta_{lg}$ values observed for weak nucleophiles like water, methoxyacetate and acetate indicate significant bond breakage at the TS and thus a fully concerted mechanism. [161, 162]

Several theoretical studies have demonstrated a shift towards the step-wise associative mechanism with respect to the stability of the leaving group. Tarrat et al. demonstrated that leaving groups with $pK_a > 8$ lead to a formation of a pentavalent intermediate, with its stability increasing with $pK_a$ of the leaving group. [163, 164] A similar finding was later reported by Maxwell et al. who concluded that methanolysis of phosphate and phosphorotioate triesters is concerted when the $pK_a$ of the leaving group is less than 12.2 and associative otherwise. [165] Xia et al. showed that the alcoholysis of triesters with aryl leaving groups proceed through a concerted mechanism or through an associative mechanism with a very low energy barrier for the second step, but is distinctly step-wise with alkyl leaving groups. [166] Lopez et al. demonstrated for paraoxon that the alkaline hydrolysis follows a concerted mechanism. [167]

**Nucleophilic versus General-base Catalysis**

The hydrolysis of DFP was shown to be catalyzed by acids and bases, with the observed rate constant for hydrolysis in acetate buffer written as,

$$k_{obsd} = k_W + k_H[H^+] + k_{OH}[OH^-] + [B]\{k_B + Rk_A\}$$

where $R = [A]/[B] = [CH_3COOH]/[CH_3COO^-]$, and the first three terms are constant. A small salt effect was noted as being indicative of general acid/base catalysis, and the catalytic effect of the acid in the diethanolamine buffer was shown to be negligibly small. [151] Solvent KIE’s for hydrolysis of DFP analogues revealed that proton transfer occurs as part of the rate-determining step. [155] General-base catalysis was demonstrated for the base $n$-butylamine. [168]

Solvent KIE of hydrolysis of cyclic phosphotriesters with aromatic oxygen leaving groups reveals predominantly nucleophilic catalysis, although contributions from general-base catalysis are significant in the reactions of less basic nucleophiles (e.g. water, acetate, pyridine) with triesters with poorer leaving groups (e.g. 4-nitrophenol, 3-nitrophenol, phenol). [161] A notable exception is the reaction between the acetate ion and the 2,4-dinitrophenyl ester, which is characterized with a $k_H/k_D = 1.0$ and an entropy of activation of -17.3 e.u., in the usual region for a bimolecular reaction, much higher than that measured for the neutral general-base catalyzed hydrolysis, which is -34.5 e.u., in the region of other termolecular reactions. [161] On the other hand, the hydrolysis of 2,4-dinitrophenyl ethylmethylphosphonate (DNPEMP), a sarin analogue with the same leaving group as above, was suggested to proceed via general-base catalysis for the base Et$_3$N, based on solvent KIE of 1.4. Additionally, it was shown for DNPEMP that bulky bases do not deviate from the Brønsted plot as would be expected in nucleophilic catalysis. [169]
4.2 Organophosphate Hydrolases

The first reported case of enzymatic organophosphate hydrolysis dates back to 1946 when Mazur observed the degradation of diisopropyl fluorophosphate (DFP) in rabbit serum. [170] In the following decades, a number of enzymes capable of hydrolyzing various OPs were found in mammals, invertebrates and bacteria. The more prominent members of OP degrading enzymes include the mammalian serum paraoxonase 1 (PON1) [171, 172] and senescence marker protein 30 (SMP-30), [173, 174] squid diisopropyl fluorophosphatase (DFPase), [175] and bacterial methyl parathion hydrolase (MPH), [176] phosphotriesterase (PTE), [152] and organophosphate acid anhydrase (OPAA). [177]

Overview of Structure and Function

While similar in function, the tertiary structures of these enzymes are highly diverse, encompassing a range of different protein folds including the TIM-barrel fold (PTE), β-propeller fold (PON1, SMP-30, DFPase), β-lactamase fold (MPH), and pita-bread fold (OPAA), as shown in Figure 4.2. [55] Interestingly, several key characteristics are shared among the functional family. The binding sites are generally big and hydrophobic so as to accommodate bulky substrates and consist of three binding pockets, one for each phosphate side-chain. In addition, all OP hydrolyzing enzymes are metal-dependent hydrolases, requiring at least one divalent metal which directly ligates the substrate. [55]

PTE, MPH, and OPAA possess a bimetallic active site, where the catalytic metal ions separated on average by 3.6 Å, with a bridging γ-hydroxide ion located between the two metal centers in case of PTE and MPH. The native metal found in bacterial PTE is Zn$^{2+}$, although the enzyme is active with other ions such as Mn$^{2+}$, Ni$^{2+}$, Co$^{2+}$. [178] Similarly, methyl parathion hydrolase’s native metal ion is thought to be Zn$^{2+}$, [179] but can be substituted with a variety of ions, including Cd$^{2+}$, Co$^{2+}$, Mn$^{2+}$, Ni$^{2+}$, and Fe$^{2+}$ - a common trait in the metallo-β-lactamase family. [57] On the other hand, the native ion in OPAA is Mn$^{2+}$, with activity for DFP hydrolysis observed also with, Co$^{2+}$ ions. while Zn$^{2+}$ was demonstrated to have a particularly inactivating effect. [177] DFPase and PON1 are calcium dependent hydrolases. While they also require two ions, these are spatially separated (7.4 Å in PON1 and 9.4 Å in DFPase [180]), with the more buried structural calcium maintaining the fold of the enzyme, and the surface exposed catalytic calcium facilitating the binding of the substrate and catalysis. [23, 181] Minor increase in activity has been reported for Mg$^{2+}$ and Ba$^{2+}$ ions in DFPase, [182] while the effect was opposite in PON1, with Mg$^{2+}$, Ba$^{2+}$ and Cu$^{2+}$ having an inhibitory effect. [183] Interestingly, the structurally related SMP-30 is completely inactive towards DFP with Ca$^{2+}$, but shows activity with Mg$^{2+}$, Mn$^{2+}$, Co$^{2+}$ and to a lesser degree with Cd$^{2+}$ ions. [174] It is active towards its native gluconolactone substrate with bound Ca$^{2+}$, Mn$^{2+}$, Mg$^{2+}$, and Zn$^{2+}$ ions. [184]
Figure 4.2. Tertiary structures of several representative organophosphate hydrolases, demonstrating high structural diversity. Shown are TIM-barrel (PTE), β-propeller (top-down view: DFPase; side view: PON1, SMP-30), β-lactamase (MPH), and pita-bread (OPAA/prolidase) folds.
It is unlikely that the ability to degrade organophosphates has evolved as a completely novel function, considering that human-made organophosphate substrates were introduced less than a century ago. Instead, it is proposed to have arisen as a promiscuous activity due to physical similarities between the phosphoryl center and the transition state of the enzyme’s natural reaction. [55]

Most OP-degrading enzymes are functionally classified as phosphoric-triester hydrolases and further divided into aryldialkylphosphatases (EC 3.1.8.1) and diisopropyl-fluorophosphatase (EC 3.1.8.2), where the members of the former sub-class act on phosphate ester bonds, and the latter on phosphate anhydride bonds (including halide and cyanide).

Beside OPAA, whose native activity has been determined to be prolidase (EC 3.4.13.9), [185] the native activity has in most cases been proposed to be hydrolysis of lactones (EC 3.1.1.25). [55] PON1’s natural substrates have been suggested to be dihydrocoumarin (DHC), long chain fatty acid lactones and acylhomoserine lactones (AHLs), [186–188] while SMP-30 was shown to be a gluconolactonase (EC 3.1.1.17) involved in L-ascorbic acid (AA) biosynthesis. [189]

A particularly interesting case is that of bacterial PTE, which is thought to have evolved during the past several decades specifically to degrade synthetic organophosphate compounds. [145] Homologous PTE-like lactonases (PLL) sharing key sequence and active site features and exhibiting much lower promiscuous PTE activities were identified, suggesting that PTE could have diverged from a member of the PLL family, utilizing its latent promiscuous paraoxonase activity. [190] Further support came from later studies revealing a key evolutionary link between PLL and PTE to be an insertion of a loop. [191, 192] Moreover, Jacquet et al. recently reported transforming a PLL (SsoPox) into a broad spectrum PTE via rational engineering. [193]

MPH displays phosphotriesterase, arylesterase and to a lesser extent lactonase activity, with distinct metal ion dependencies. [57] Analogous to PLL, a ‘quorum-quenching’ lactonase AiiA from the β-lactamase superfamily has been identified to exhibit some phosphotriesterase activity, suggesting that MPH could have evolved from a lactonase as well. [145, 190]

4.2.1 Methyl Parathion Hydrolase

Methyl parathion hydrolase (MPH) has been identified in several soil-dwelling bacteria able to degrade the insecticide methyl parathion (MPS) and use it as their sole source of carbon and nitrogen. [55, 176, 194, 195] It is suggested that its evolution was affected by the environment, since all known mpd (methyl parathion degrading) genes have been isolated from one country. [145] Due to its ability to degrade various organophosphate pesticides, MPH has been suggested as a good candidate in the development of bioremedia-
tion agents, and has already been used to develop and evaluate biosensors for detecting OP contamination. [145]

![Active site of methyl parathion hydrolase. Adapted from Figure 3 in ref. [196]](image)

**Figure 4.3.** Active site of methyl parathion hydrolase. Adapted from Figure 3 in ref. [196]

**Structural features**
The crystal structure of MPH from *Pseudomononas sp. WBC-3* has been solved to 2.4 Å resolution (PDB: 1P9E), revealing it to be a member of the metallo-\(\beta\)-lactamase superfamily, [179] and not homologous to any other OPH as initially suspected. [145] It forms an intimate, symmetric 2-fold dimer, with each monomeric unit described as an \(\alpha\beta\beta\alpha\) sandwich typical of the \(\beta\)-lactamase fold. [55, 179] Each subunit includes a binuclear metal center, located between the two \(\beta\)-sheets and surrounded by two \(\alpha\beta\)-loops. [179]

The two ions in the metal center are separated by 3.5 Å, and are coordinated by seven protein residues and two water molecules. The more buried \(\text{Zn}^{2+}\) ion is coordinated with residues Asp151, His152, Asp255, His302 and the bridging hydroxide (\(\gamma\)-OH). [179] The more solvent-exposed ion has been determined to be \(\text{Cd}^{2+}\); however, the nature of the ion is presumably due to high Cadmium concentrations in the crystallization conditions. [179] Its coordination sphere includes residues His147, His234, His149, Asp255 and the bridging hydroxide. The binuclear active site motif is quite similar to PTE despite sharing no sequence or structural homology. Both display octahedral
coordination in the mixed Zn\(^{2+}\)Cd\(^{2+}\) metal isoform and a hydrophobic surrounding with three distinct pockets. [55, 179]

Activity and Selectivity

The enzyme displays particularly high activity for methyl parathion \((k_{cat}/K_m \sim 10^6 \text{M}^{-1}\text{s}^{-1}; k_{cat} = 37.1 \text{s}^{-1})\), [179] while also possessing moderate paraoxonase activity \((\text{Ni}^{2+}: k_{cat}/K_m = 3.6 \times 10^3 \text{M}^{-1}\text{s}^{-1}; k_{cat} = 4.8 \text{s}^{-1})\), arylesterase activity \((\text{Ni}^{2+}: k_{cat}/K_m = 1.7 \times 10^3 \text{M}^{-1}\text{s}^{-1}; k_{cat} = 3.7 \text{s}^{-1})\), and very limited phosphodiesterase and lactonase activity. [57] A recent extensive substrate profiling study revealed the activity extends towards other pesticides such as methyl paraoxon, ethyl parathion, methyl chlorpyrifos, and ethyl chlorpyrifos. [197]

MPH exhibits a distinct metal-ion specificity pattern, retaining paraoxonase and esterase activity to varying degree with metal ions including Cd\(^{2+}\), Co\(^{2+}\), Mn\(^{2+}\), Ni\(^{2+}\), and Fe\(^{2+}\). [57] This feature was shown to be a common trait of this superfamily, and suggests an underlying mechanism for evolution and functional divergence through an expanded activity profile of the many metal isoforms. [57]

Figure 4.4. Plausible mechanisms for the MPH-catalyzed hydrolyses of (A,B) paraoxon and (C) \(p\)-nitrophenyl butyrate. OR denotes the \(p\)-nitrophenyl leaving group. The figure was taken with permission from ref. [196]
Reaction Mechanisms

Limited mechanistic work has been done on MPH, although close active-site similarity with PTE has led workers to propose a similar mechanism (via the bridging hydroxide) [55] An alternative mechanism employing the terminal hydroxide as the nucleophile was suggested from structural studies on a bacterial phosphotriesterase (OpdA). [198, 199] These revealed the substrate is bound monodentately to the α metal, while the product is bound bidentately with the bridging hydroxide in-tact in the product state. Both mechanistic proposals are shown in Figure 4.4. The work presented in Paper II challenges these proposals and additionally aims to rationalize the observed metal-ion selectivity patterns. [196]

4.2.2 Squid DFPase

In 1966, while bathing the squid giant axon in diisopropyl fluorophosphate (DFP), Hoskin noticed significantly less AChE inhibition than was expected and attributed the effect to enzymatic degradation of DFP. [175] The term squid-type DFPase was given to the enzyme to signify its unique characteristics when compared to the previously reported ‘Mazur’-type DFPase1 (found in mammalian tissues), [170, 201] specifically, its stability, inhibition by Mn2+ ions and substantial preference for hydrolysis of DFP compared to Sarin, Soman and Tabun. [202] A detailed structural and kinetic characterization began in the early 2000, when Rüterjans and coworkers successfully cloned the DFPase gene from Loligo Vulgaris into E. coli. [23, 147, 203]

Structural features

At the time of writing, there were an impressive 24 structures of squid DFPase deposited to the Protein Data Bank, [204] out of which, one structure has sub-Ångstrom resolution, one was obtained using joint neutron and X-ray diffraction, one structure with a bound inhibitor (Figure 4.5), and 17 structures of DFPase variants.

DFPase has a molecular weight of 35 kDa, consists of 314 amino acid residues and folds into 6-bladed β-propeller with a central water tunnel containing two calcium ions. The buried ‘structural’ calcium has higher affinity and was shown to be important for maintaining the proper fold. The solvent exposed ‘catalytic’ calcium at the base of the active site has lower affinity and was shown to be crucial for binding and catalysis. [205] The catalytic calcium is coordinated by seven ligands: Asn120, Asp229, Asn175, Glu21, and three water molecules. The hydrophobic cleft that is the active site, is limited to one side by Trp244, Thr195, and Phe173, and to the other by His287, Pro36, Ala74, and Met90 (Figure 4.5).

1Although a definite connection has not been established in literature, it appears that the early reports of DFPase activity in mammals could be attributed to SMP-30. [173, 188, 200]
Figure 4.5. Top-down view of squid DFPase tertiary structure and active site (PDB ID: 2GVV [182]). The co-crystallized inhibitor $O,O$-dicyclopentylphosphoramoide (DcPPA) is shown in yellow.
The overall structure of the enzyme is quite resilient to mutations, with crystal structures of several DFPase variants having virtually identical tertiary structures. Moreover, prolonged incubations of the enzyme at pH ranging from 5.5 to 11.0 and temperatures up to 55°C showed no significant loss of activity or changes in the CD spectrum, indicating its stability under these conditions. [23] Together with the fact that the enzyme is readily available in bulk amounts, these properties make DFPase a good candidate for use as an industrial biodecontamination agent. [147]

No homology or evolutionary relationships were found when comparing to other β-propellers. [203] The sequence identities with similar β-propeller OP-degrading enzymes SMP-30 and PON1 are around only 20%, although high sequence diversity is a common trait of β-propeller folds. [206]

Activity and Selectivity
Squid DFPase displays high specificity towards compounds containing fluoride and cyanide leaving groups including DFP and G-type nerve agents Sarin, Soman, Cyclosarin, and Tabun, and is generally inert towards P-O and P-S bonds, such as VX and Paraoxon. [55] An exception are a few fluorogenic analogues of nerve agents which were reported to be hydrolyzed with high efficiency. [207] The enzyme has been thoroughly characterized with respect to DFP, while kinetic data for other more toxic substrates is mostly limited to $k_{cat}/K_m$ values or specific activities.

The Michaelis-Menten parameters for the hydrolysis of DFP were determined at various conditions of temperature, pH and ionic strength, with $k_{cat} = 130 \text{s}^{-1}$ and $k_{cat}/K_m = 2.41 \times 10^5 \text{M}^{-1}\text{s}^{-1}$ at standard conditions (pH = 7.5, 25°C, 10mM NaCl), and approaching $k_{cat} = 526 \text{s}^{-1}$ when the conditions were optimal (pH ≥ 8, 35°C, 500mM NaCl). [23] The rate increases with pH to a limiting value, indicating the titration of a single base, with $pK_a = 6.97$. [23]

Like PON1, DFPase is stereoselective and showing preference for the less toxic $R_P$ enantiomers of the G-type nerve agents. [208] Using docking and rational design, Melzer et al. successfully reversed the enantioselectivity for Sarin (GB) and Cyclosarin (GF), increasing the observed $S$-GB/$R$-GB ratio of $k_{cat}/K_m$ values from 0.89 to 5.11 and $S$-GF/$R$-GF from 0.02 to 3.77. [208] Three mutations were chosen to reduce the steric constraints in the vicinity of the leaving group (E37A/Y144A/R146A), and one to restrict the size of the cavity encompassing the sidechain of the $R_P$ enantiomer (T195M). All four mutations are reportedly required to achieve the change in selectivity, with the single and double mutants E37A and T144A/R146A not displaying any change in preference, and mutant T195M showing only slightly less pronounced preference for the less toxic $R_P$ enantiomer of GF. Interestingly, another reported quadruple mutant, where E37 was instead mutated to an aspartate (E37D/Y144A/R146A/T195M), did not yield reversed selectivity for GF.
Reaction Mechanisms
Several mechanistic pathways have been proposed for the hydrolysis of DFP in DFPase in literature and are discussed below.

Hartlieb et al. initially suggested that a His residue acts as a general base, activating a water nucleophile. The attack of water and subsequent displacement of the fluoride leaving group was suggested to proceed via the formation of a pentavalent intermediate. [23] The justification came from pH-rate analysis, revealing the titration of a single base with $pK_a = 6.96$. Follow-up site-directed mutagenesis revealed H287N to have less than 4% WT activity, and virtually no pH dependence. Residual activity was attributed to water activation via hydrogen bond between the Asn287 or possible change of catalytic base. The authors also noted a positive effect of increasing ionic strength on both $k_{cat}$ (up to 500 mM) and $K_m$, which they rationalized to be due to the strengthening of hydrophobic interactions between the enzyme and the substrate, as well as stabilization of the developing negative charge on the fluoride leaving group. [23] Further, temperature dependence indicates that the rate-limiting step does not change with temperature and is product formation. [23]

In a thorough mutational study, Katsemi et al. found some of the H287 mutants to retain up to 79% activity relative to WT, casting doubt on the involvement of H287 as a general base. [209] Specifically, residues Phe, Trp, and Leu led to a retention of 64%, 66%, and 79% specific activity, respectively, suggesting the requirement of a large hydrophobic residue at position 287. Furthermore, substitution with Ala, Asn, and Asp resulted in only 10%, 4%, and 1% residual activity. Loss of activity due to conformational change was ruled out after solving the crystal structure of the H287A variant, which showed little conformational differences compared to WT. [209]

While the polar nature of His287 should according to the observations provide a detrimental effect, its polarity was suggested to be reduced due to the hydrogen bond network Trp244-His287-Ala20. [209] This also provides a possible explanation for the high relative activity of H287Q (46%) compared to H287N (4%), as only the former can form the hydrogen bonds. Interestingly, using NMR pH-titration experiments they demonstrated that His287 does not titrate within the scanned pH range (5.76 - 9.45), which is somewhat conflicting with Hartleib’s reported pH dependence (and lack thereof in H287N). [23] The importance of larger hydrophobic residues was observed also at position 173, where activity decreased in the order Phe/Trp > Leu > Val/Tyr > Ser > Ala, [209] which is surprising considering the densely charged fluoride leaving group is presumed to be in the immediate vicinity. A similar effect was observed with mutating T195, a residue located in one of the sidechain pockets. Mutation into a smaller alanine resulted in a 60% loss in activity; however, more hydrophobic substitutions T195V and T195L did not affect the activity significantly. [209] Finally, the authors reported a double mutation D229N/N120D in the coordination sphere of the catalytic calcium,
which resulted in a completely inactive enzyme, with the structure shown to be unaffected, containing two bound calciums and a similar \(^1\)H NMR spectrum to that of the WT protein. [209] While the authors conclude that the resulting loss of activity is due to relative position of the charge in the active site, a more probable reason is the loss of the catalytic residue aspartate in position 229.

Soon after, Blum et al. solved a crystal structure of WT DFPase with a bound substrate analogue \(O, O\)-dicyclopentylphosphoroamidate (DcPPA) and proposed a new catalytic mechanism involving the residue Asp229\(^2\). The inhibitor is shown to directly ligate the catalytic calcium via the phosphoryl oxygen and is positioned with the amino group pointing toward the proposed catalytic residue Asp229 as can be seen in Figure 4.5. The authors deemed the pose as unproductive, and provided docking evidence for an additional productive binding pose for DFP with the fluoride leaving group pointing away from the proposed catalytic base and thus allowing for in-line attack. [182] Double mutants D229N/N120D, D229N/N175D, E21Q/N120D, E21Q/N175D were shown to inactivate the enzyme, suggesting that the correct electrostatic topology of the active site is necessary for binding and catalysis. [182] To exclude any conformational changes or missing calcium ions, the crystal structures of D229N/N120D and D229N/N175D were solved and revealed virtually identical active site geometries as WT DFPase. [182] The Asp229 was proposed to act as a nucleophile, generating a covalent phosphoenzyme intermediate which is subsequently hydrolyzed by water at the Asp229 carboxylate carbon (scheme B in Figure 4.6). Evidence for the existence of such a phosphoenzyme intermediate was provided by \(^18\)O incorporation experiments. [182]

Wymore et al. performed QM/MM umbrella sampling simulations to investigate the mechanisms of DFP and (S)-sarin hydrolysis, providing additional support for the mechanism where Asp229 acts as a nucleophile. [210] The rate-limiting step for DFP hydrolysis was shown to be a step-wise formation of the phosphoenzyme intermediate via a pentavalent phosphoenzyme species, with an effective barrier of 14.2 kcal mol\(^{-1}\), in very good agreement with experimental \(k_{\text{cat}} = 130\) s\(^{-1}\) [23] (via TST: \(\Delta G^\ddagger = 14.6\) kcal mol\(^{-1}\)). The subsequent hydrolysis of the intermediate involves a calcium-bound water molecule activated by Asp121. [210] The same mechanism for (S)-sarin yielded a highly unfavorable barrier of around 28 kcal mol\(^{-1}\), which prompted the authors to propose that the hydrolysis of (S)-sarin follows a different mechanism, where the Asp229 acts instead as a general base, activating an active site water nucleophile, although the hypothesis was not tested. [210]

Elias et al. performed a detailed investigation and refinement of the available DFPase sub-Ångstrom resolution crystal structure, and concluded that a calcium-coordinated water molecule \(\text{HOH}524\) is partially activated into a

---

\(\text{HOH}524\) is chemically significantly more similar to the hydroxyl group of the product than the substrate's fluoride leaving group, thus the term substrate analogue could be misleading (see Figure 4.7).
hydroxide ion by Asp229. This suggests an alternative role for Asp229 is possibly similar to the analogous Asp269 in PON1 which was demonstrated to act as a general base. [211]

Xu et al. used a DFT cluster-model approach to investigate the role of Asp229 in the reaction mechanism, modeling both the nucleophilic and the general-base roles of Asp229. [212] With dimethyl fluorophosphate as the substrate, they obtained a barrier of 14.8 kcal mol\(^{-1}\) for the formation of the phosphoenzyme intermediate (i.e. the nucleophilic mechanism), in good agreement with the value reported by Wymore et al. [210] Their analysis however did not reveal any stable pentavalent species, possibly due to the smaller inductive stabilization of the methyl sidechains. The authors also reported an alternative pathway with a lower barrier of 10.8 kcal mol\(^{-1}\), where the calcium is hexa- instead of hepta-coordinated, suggesting the coordination to have an important role, [212] however, such a coordination has to my knowledge not been found in any crystal structures or has been observed in our simulations. The subsequent hydrolysis of the phosphoenzyme intermediate was modeled via the activation of a water molecule by Glu21, and was found to have a similar activation energy of around 15.2 kcal mol\(^{-1}\). Finally, the authors demonstrated the general-base pathway to have a significantly more favorable effective barrier of 6.0 kcal mol\(^{-1}\), suggesting it is more plausible than the nucleophilic mechanism. [212] The pathway, although shown to be step-wise, can be considered as concerted, since the most stable intermediate state has a reverse barrier of only 0.4 kcal mol\(^{-1}\).

The two plausible mechanisms are shown in Figure 4.6, with the proposed stable pentavalent species omitted. The work presented in this thesis tries to resolve the apparent conflicts using a computational study that heavily relies on available experimental data for validation.

### 4.2.3 Serum Paraoxonase 1

Enzymatic degradation of paraoxon (diethyl \(p\)-nitrophenylphosphate) was first reported by Aldridge in 1953 and was attributed to an \(A\)-esterase enzyme found in mammalian sera. [171] The enzyme was initially classified as an arylesterase (EC 3.1.1.2) due to its activity towards \(p\)-nitrophenyl acetate, propionate and butyrate. The reclassification to paraoxonase (EC 3.1.8.2) was accepted based on evidence that two different enzymes are responsible for these activities, [213] although this was later proven to be false. [214].

Out of the three mammalian members of the paraoxonase family (PON) which share 60-70% amino acid identity, PON1 is the most investigated due to its ability to degrade various organophosphates. [187] It has gained particular notice for its potential use as a biotherapeutic for OP poisoning, since it will, unlike other OPHs, not induce a response from the immune system. [147] \(In vivo\), it is associated with the ‘good cholesterol’ HDL (high-density lipopro-
Figure 4.6. Plausible mechanisms for the hydrolysis of DFPase by DFP. (A) General-base mechanism, in which Asp229 acts as a general base to activate the nucleophilic water molecule and the reaction proceeds via a single, concerted transition state. (B) Nucleophilic substitution mechanism involving direct nucleophilic attack by the carboxylate side chain of Asp229, proceeding via a covalent intermediate that is hydrolyzed by a water molecule. Adapted from Figure 2 in Paper III.

Figure 4.7. Overlay of the crystal structure with bound inhibitor DcPPa (green, PDB code 2gvv [182]) and representative snapshot of the product state (cyan) as obtained from simulations of the general-base mechanism (Paper III). The inhibitor DcPPA is bound in a similar pose as the hydrolyzed DFP (diisopropyl phosphate) and is chemically analogous, with the amino group mimicking the hydroxyl phosphate group and presumably also forming two hydrogen bonds with residues Glu21 and Asp229. This suggests that the general-base mechanism proceeds through a low-energy pathway resembling the bound inhibitor.

...tein) particles, contributing to their antioxidant properties, [215] while at the same time making its purification, crystallization and characterization challenging. [172, 216] As the human PON1 is unstable and tends to aggregate in the absence of detergents, Tawfik and coworkers used directed evolution in order to identify and express variants with increased bacterial expression and
solubility. One of the variants RePON1-G2E6 was found to have similar kinetic properties as WT enzyme as well as a completely conserved active site and was successfully expressed in *E. coli* and crystalized. 

**Structural features**

The crystal structure of rePON1 was co-crystalized with a bound competitive inhibitor 2-hydroxyquinoline (Figure 4.8). The structure is very similar to DFPase, adopting a β-propeller fold, with a central water tunnel and two calcium ions. The coordination shell of the catalytic calcium is almost identical, with the catalytically important residues Asp269, Glu53, Asn224 and Asn168 superimposable with DFPase’s Asp229, Glu21, Asn175 and Asn120. One difference between the two is the coordination of Asp270 in PON1, which is replaced by a water molecule in DFPase, although these are located inside the water tunnel, opposite to the substrate. An overlay of the two active sites is shown in Figure 1 in Paper III. 

Several notable differences between the enzymes become apparent further away from the catalytic calcium, including the His115 in position of Ala74 which was shown to be the catalytic base in hydrolysis of lactones, as well as His285 which is replacing the bulky hydrophobic Trp244 and stabilizing the catalytic Asp229 via a hydrogen bond. Additionally, Tyr71 was shown to be crucial for maintaining active site hydrophobicity and in turn paraoxonase activity, by closing one of the active site loops via the formation of a hydrogen bond with Asp183. 

A major difference between the two enzymes are the two additional α-helixes sitting above the active site, which were suggested to play a role in anchoring PON1 to HDL. The membrane binding has been shown to immobilize a crucial hydrogen-bond network reaching from one of the helixes to the core of the active site (> 15 Å away), stabilizing a key catalytic residue Asn168. 

The enzymes have only around 22% sequence similarity, although high sequence diversity is a common trait of β-propeller folds.

**Activity and Selectivity**

PON1 has a significantly broader substrate specificity than DFPase, hydrolyzing not only organophosphates but also arylesters and γ- and δ-lactones, which have been proposed to be its native substrates. Organophosphates hydrolyzed by PON1 include the nerve agents Sarin, Soman, Cyclosarin, Tabun, VX, the insecticide paraoxon and others. Interestingly, RePON1 displays no activity towards a structural analogue of DFP, while DFP itself was shown to be a competitive inhibitor for paraoxonase and arylesterase activity in HuPON1. The origin of the apparent lack of DFPase activity in PON1 is addressed in Paper III.
Figure 4.8. Top-down view of the tertiary structure and active site of rePON1-G2E6 (PDB ID: 3SRG [218]). The bound inhibitor 2-hydroxyquinoline (2HQ) is shown in yellow.
Reaction Mechanism
Several experimental and computational studies have demonstrated that the paraoxonase activity is mediated by Asp269 acting as a general-base, deprotonating a water nucleophile. [187, 218–220] Importantly, the lack of observed burst kinetics for any substrate suggests the hydrolysis does not in fact proceed via a covalent phosphoryl-enzyme intermediate as was proposed for DF-Pase. [187] The hydrolysis of lactones and arylesterases was on the other hand shown to be mediated by a different general-base, namely His115. Its removal was demonstrated to abolish those activities, while not affecting paraoxonase activity in any way. [218]
5. Summary of Publications

5.1 Paper I - Force Field Independent Metal Parameters Using a Nonbonded Dummy Model

Metal ions are found in all functional classes of enzymes where they play a variety of catalytic roles, either activating nucleophiles, acting as redox centers, providing electrostatic stabilization, or facilitating optimal positioning. However, obtaining an adequate description of metal ions for use in classical biomolecular simulations is challenging. It is commonly done by imposing artificial bonds between the metal and the ligands (know as the bonded model) which is particularly stable, although cumbersome to parameterize, system-specific, and can become particularly problematic in studies where ligands are allowed to exchange. An alternative approach is to use the so-called non-bonded soft-sphere model which has the benefit of being simple and transferable. However, while it is applicable to alkali and alkaline-earth ions, it is unable to capture the solvation free-energy and metal-oxygen distances of transition metals. Another strategy is the cationic dummy model, where the metal ion is described with a set of cationic dummy particles, coordinated around a partially negatively charged central atom, characterized by a set of metal-ion-specific Lennard-Jones parameters.

In this publication we extend the range of dummy models to include three biologically relevant transition metals Ni^{2+}, Co^{2+}, Fe^{2+} and refine existing parameters for Mn^{2+}, Zn^{2+}, Mg^{2+}, Ca^{2+}. By adjusting the Lennard-Jones parameters we were able to accurately reproduce the experimental metal-oxygen distances and solvation free-energies in aqueous solution. We further demonstrate that the model is largely transferable between different water models (SPC and TIP3P). The ability of the model to maintain correct coordination in enzyme simulations was assessed by performing MD simulations of human and E. coli variants of Glyoxalase I which not only revealed a stable coordination but also captured subtle geometric changes that would be expected upon metal substitution.

The model was later successfully applied to simulations of chemical reactivity of organophosphate hydrolases in Papers II and III.
5.2 Paper II - Probing the Mechanisms for the Selectivity and Promiscuity of Methyl Parathion Hydrolase

Methyl parathion hydrolase (MPH) is an enzyme isolated from soil-dwelling bacteria which evolved to use the toxic organophosphate pesticide methyl parathion as their sole source of nitrogen and carbon. As these kinds of organophosphate compounds were introduced less than a century ago, this raises the question of how this functionality evolved in such a short timescale. A recent experimental study revealed intriguing metal-ion selectivity patterns in MPH which were linked to the evolution and divergence of novel functions, although the origin of the effect was not addressed.

The goal of the study presented here was to obtain a deeper understanding of the observed metal-ion selectivity patterns and, as this is the first mechanistic study on MPH, to elucidate a viable mechanistic pathway for the hydrolysis of organophosphates and arylesters by MPH, which has not been described before.

Using the empirical valence bond (EVB) approach we modeled the hydrolysis of paraoxon and $p$-nitrophenyl butyrate, where a terminal metal-bound hydroxide nucleophile attacks the substrate ligated to the second metal ion via the phosphoryl/carbonyl oxygen and in a concerted fashion displaces the $p$-nitrophenolate leaving group. Additionally, both reactions were simulated using five different transition metals Fe$^{2+}$, Zn$^{2+}$, Mn$^{2+}$, Co$^{2+}$ and Ni$^{2+}$, which were described using the cationic dummy model presented in Paper I.

The activation free energies were obtained in good agreement with experiment, providing justification for the proposed mechanistic scheme, as well as validation for using the dummy metal model to describe chemical reactivity in the EVB framework. From the analysis of our simulations it is evident that there are no major structural changes occurring during metal substitution, implying that the origin of the observed selectivity patterns is simply the difference in the inherent electrostatic properties of the metals themselves, coupled with very subtle changes in substrate and metal ion geometries.
5.3 Paper III - Similar Active Sites and Mechanisms Do Not Lead to Cross-Promiscuity in Organophosphate Hydrolysis: Implications for Biotherapeutic Engineering

The work presented in this publication aims to resolve open mechanistic questions as well as to provide insight into substrate specificity of two structurally very similar enzymes, squid diisopropyl fluorophosphate (DFPase) and mammalian serum paraoxonase 1 (PON1).

DFPase and PON1 share a $\beta$-propeller fold with a central water tunnel containing two calcium ions one of which has a catalytic and the other a structural function. The first coordination sphere of the catalytic calcium ion in DFPase and PON1 is almost identical, containing an Asp, Glu, two Asn residues, and two water molecules. One difference is in the more buried ligand, which is an Asn in PON1 and a water molecule in DFPase.

Despite having a similar tertiary structure and catalytic machinery, the two enzymes have very particular substrate specificities towards their ‘natural’ substrates diisopropyl fluorophosphate (DFP) and paraoxon, even though the two substrates are, apart from the leaving group, analogous. This peculiarity prompted us to study the two reactions using a theoretical approach to try to quantify the underlying reasons behind this selectivity.

The simulations of DFP and paraoxon hydrolysis, both in solution and in the enzymes were performed using the empirical valence bond (EVB) approach, which is described in detail in Chapter 3. A particular benefit in using this approach is its low computational cost which allows for extensive configurational sampling. This is required to obtain free energies of activation with reliable precision, particularly when calculating small free-energy differences associated with a change in temperature.

**Part I: DFPase Mechanism**

Extensive experimental and computational data suggests that PON1 hydrolyzes organophosphates via a general-base assisted mechanism, where Asp269 activates a water nucleophile which hydrolyzes the organophosphate in a concerted fashion. On the other hand, the literature has been divided in the case of DFPase. Several experimental and computational studies suggested Asp229 (analogous to PON1’s Asp269) acting as a nucleophile and forming a phosphoenzyme intermediate, while others proposed Asp229 to act as a general-base, analogous to PON1.

To resolve this controversy we modeled both of the proposed mechanisms, relying on a wealth of available experimental data. Beside the catalytic effect of the WT DFPase compared to a reference reaction in solution, we also cal-
culated the effects of various reported mutations, calculated the entropy and enthalpy of activation using temperature dependence analysis, and did a qualitative assessment of the pH dependence on the reaction rate.

Based on a significantly better agreement with experimental data, we concluded that the most likely mechanism is the general-base assisted mechanism, which was also previously suggested for PON1. Our analysis of the two mechanisms in DFPase suggests the preference for the general-base mechanism can partially be explained by the increased electrostatic stabilization by Glu21 and decreased differential solvation.

**Part II: Cross-Selectivity**

Based on the mechanistic insight obtained from the first part of the study, we then calculated the free energy profiles for hydrolysis of both substrates (DFP and paraoxon) in both enzymes (DFPase and PON1), with respect to the reference reactions in aqueous solution. While the calculated rate for DFP hydrolysis in PON1 is in qualitative agreement with experiment, the calculated rate for paraoxon hydrolysis in DFPase is well below the experimental estimate, implying that paraoxon is easily hydrolyzed by DFPase. This discrepancy with experiment was not resolved in the current study. However, we argue that there is a possibility of nonchemical effects governing the rate such as non-productive binding, which we are not able to capture in our simulations.

The observed increase in barrier for DFP hydrolysis in PON1 versus DFPase, is associated with a significant increase in endergonicity of the reaction, which we postulate to largely be coming from the product state due to its more polar nature. It is conceivable that the densely-charged fluoride ion is stabilized less in the more hydrophobic PON1 active site than in the solvent-exposed DFPase cleft. In the case of paraoxon hydrolysis, where the charge developing on the leaving group is delocalized, the difference in endergonicity is less pronounced, providing further support to the theory.

We used the linear response approximation (LRA) method to estimate the electrostatic effect of individual PON1 residues on the rate of hydrolysis of both substrates (DFP and paraoxon) in order to pinpoint the origin of the selectivity. In addition to Glu53 (analogous to DFPase Glu21) and Asn168 (analogous to DFPase Asn120), which have a notable but opposite effect on both substrates, we show Asp183 (no analogous residue in DFPase) to have a significantly larger detrimental effect on DFP hydrolysis than on paraoxon hydrolysis. The selective nature is proposed to originate from differential destabilization of the nearby leaving group which is in one case a densely charged fluoride ion, and in the other, a delocalized p-nitrophenolate ion.
5.4 Paper IV - Q6 - A Comprehensive Toolkit for Empirical Valence Bond and Related Free Energy Calculations

In this publication we present the latest version of Q, a generalized software package for empirical valence bond, linear interaction energy and free energy calculations. In addition to general technical improvements, new functionality includes the bisection quantum classical path (BQCP) method for calculating nuclear quantum corrections to classical energies, the Langevin and Nosé-Hoover thermostat, the LINCS constraint algorithm, and support for solvents such as methanol, ethanol, chloroform and dichloromethane. Additionally, to aid in the search for possible mutation and engineering sites in enzyme design, a ‘group exclusion’ approach was implemented. It allows for selected amino-acid residues to be excluded from the free-energy calculations, yielding a direct estimation of their contribution to the activation free energy of enzymatic reactions.

The BQCP approach was validated by calculating the kinetic isotope effect for the E2 elimination reaction of 2-phenylethylbromide in ethanol as well as for the transmethylation reaction catalyzed by catechol O-methyltransferase. Empirical valence bond simulations were used to obtain classical free energy profiles, and quantum effects were calculated with BQCP during post-processing. We were able to successfully reproduce both the normal and inverse kinetic isotope effects.

Calculations of group exclusions were tested on the epoxide ring opening reaction catalyzed by Solanum tuberosum epoxide hydrolase 1. The contributions of two catalytically important residues Y154 and Y235 were compared to the experimentally and computationally obtained values for the conservative truncations to phenylalanine, and were shown to give good estimates for the energy difference.

Finally, solvent models were tested in spherical boundary conditions to confirm the correct behavior concerning radial distribution functions and densities at different positions of the solvent sphere.
6. Concluding Remarks

The main topic of this thesis is the study of mechanisms and selectivity of organophosphate hydrolases, in particular methyl parathion hydrolase (MPH), diisopropyl fluorophosphatase (DFPase), and paraoxonase 1 (PON1). These enzymes have evolved to degrade toxic man-made compounds, which were introduced to nature less than century ago, and are thus of particular interest in the evolutionary as well as mechanistic context.

Computational methods were used to address several questions related to the mechanisms and selectivity of these enzymes. In particular, in Paper II, we used the empirical valence bond (EVB) approach to map the origin of metal-ion activity patterns observed in MPH to a simple electrostatic stabilization effect. In addition, we provided evidence for our proposed mechanism for the hydrolysis of paraoxon and p-nitrophenol butyrate, and further validated our dummy metal model which was derived and parameterized in our previous study presented in Paper I. This parameterization was an important prerequisite to our studies as all organophosphate hydrolases are metal-dependent enzymes. In Paper III, by reproducing a plethora of available experimental observables including mutational effects and temperature dependence, we were able to differentiate between two proposed mechanistic pathways for DFP hydrolysis in DFPase and rationalize the observed substrate selectivity in the related PON1. The majority of calculations and analyses were performed using the software package Q, the latest version of which is described in Paper IV.

The research presented in this thesis provides valuable insight into the mechanisms and selectivity of organophosphate hydrolases, and forms the foundations for ongoing research into the origins of stereoselectivity observed in DFPase. Moreover, the knowledge and models obtained here have a tremendous potential value in rational design of environmentally friendly, non stoichiometric, bioremediation agents and bioscavengers for treating organophosphate poisoning.
7. Sammanfattning

Organofosfathydrolaser är en klass av olika enzymer som kan nedbryta några av de giftigaste föreningarna man känner till. Deras substrat är en mängd olika konstgjorda, fosforhaltiga organiska kemikalier som introducerades första gången på 1930-talet, och som vanligen används som bekämpningsmedel, herbicider och kemiska krigsmedel. Eftersom dessa föreningar normalt inte förekommer i naturen ställs det frågan - hur utvecklade enzymerna denna funktionalitet på så kort tid. Under de senaste decennierna har gjorts betydande framsteg på detta område, som visar att funktionaliteten sannolikt framträdde som en promiskuös aktivitet i enzymer deras inhemska funktion kemiskt lik organisk fosfathydrolys är, till exempel hydrolys av laktoner. Väsentlig inblick i strukturen, funktionen och utvecklingen av dessa enzymer uppnådes genom både experimentella och beräkningsstudier; vissa frågor rörande mekanismerna och selektiviteten förblir dock obesvarade. Dessutom är enzymerna lovande som effektiva organofosfatdekontamineringssmedel. En djup kunskap om deras funktion är grundläggande för framtidens design av effektiva och selektiva biokatalysatorer inom detta användningsområde. I denna avhandling behandlar jag några av de öppna frågorna med hjälp av en beräkningsmetod, det vill säga genom datormodellering och simulering av kemiska reaktioner i organofosfathydrolaser.


De publikationer som presenteras i denna avhandling sträcker sig från utvecklingen av den programvara som används för beräkningar och analyser, till derivering och parametriserings av specifika metalljonmodeller som krävs för senare studier och slutligen för att ta itu med flera frågor om reaktionsmekanistiska vägar, substratspecificitet och promiskuitet hos flera medlemmar av enzymfamiljen, inklusive methylparathionhydrolas (MPH), diisopropylfluorofosfatas (DFPase) och paraoxonas 1 (PON1). Vi konstruerade och parametriserade
en molekylmekanisk modell för simulering av övergångsmetaller i enzymer, som beskriven i artikel I och använde modellen för att utreda ursprunget till metalljonaktivitetsmönster som observerades i MPH och ge mekanisk inblick i artikel II. Genom att reproducerera en mängd tillgängliga försöksobservatörer kunde vi i artikel III skilja mellan två föreslagna reaktionsmekanistiska vägar för DFP-hydrolys i DFPase och förklara den observerade substratselektiviteten i det besläktade PON1-enzymet. Slutligen presenteras de senaste framstegen i utvecklingen av simuleringssmjukvarapaketet Q, som användes för de flesta simuleringarna i denna avhandling, i artikel IV.

Forskningen som presenteras i denna avhandling ger teoretiska modeller som är tillämpliga i framtida beräkningsstudier på metallenzymer, liksom den senaste utvecklingen inom mjukvara för att utföra sådana molekylära simuleringar. Dessutom ger forskningen inblick i mekanismerna och selektiviteten hos organofosfathydrolaser och utgör grunden för pågående forskning om ur sprunget till stereoselektiviteten observerad i DFPase. Vidare har de kunskaper och modeller som erhållits här ett enormt potentiellt värde i rationell design av miljövänliga biologiska saneringsmedel och bioscavenger för behandling av organofosfatförgiftning.
References


(141) Muegge, I.; Tao, H.; Warshel, A. Protein engineering 1997, 10, 1363–1372.


(145) Singh, B. K. Nature Reviews Microbiology 2009, 7, 156.

(146) Balali-Mood, M.; Saber, H. Iranian journal of medical sciences 2012, 37, 74.


(171) Aldridge, W. *Biochemical journal* 1953, 53, 117.


(197) Ng, T.-K.; Gahan, L. R.; Schenk, G.; Ollis, D. L. *Archives of biochemistry and biophysics* 2015, 573, 59–68.


A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)