Docking and Bioinformatics Tools to Guide Enzyme Engineering

Thesis submitted to University College London for the degree of Doctor of Philosophy

John Strafford
Department of Biochemical Engineering
2010
One general law, leading to the advancement of all organic beings, namely, multiply, vary, let the strongest live and the weakest die.

-Charles Darwin, The Origin of Species (1859)
Declaration

I, John Strafford confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.
Acknowledgements

I wish to thank my supervisors, Paul Dalby and David Jones, for their continued support throughout my PhD. I would also like to extend my thanks to all those in the Biochemical Engineering Department of UCL, in particularly Nigel Titchener-Hooker for his words of encouragement at difficult times.

I would not have been able to complete this work without the unending support of my girlfriend, Hannah Summers, who has kept me motivated and very happy throughout the last six years. I also need to give a massive thanks to my mum and my close family, who have always been there for me in the good, and the difficult times.

Last, but not least, I would like to thank the Biotechnology and Biological Sciences Research Council for funding my PhD.
Abstract

The carbon-carbon bond forming ability of transketolase (TK), along with its broad substrate specificity, makes it very attractive as a biocatalyst in industrial organic synthesis. Through the production of saturation mutagenesis libraries focused on individual active site residues, several variants of TK have been discovered with enhanced activities on non-natural substrates. We have used computational and bioinformatics tools to increase our understanding of TK and to guide engineering of the enzyme for further improvements in activity.

Computational automated docking is a powerful technique with the potential to identify transient structures along an enzyme reaction pathway that are difficult to obtain by experimental structure determination. We have used the AutoDock algorithm to dock a series of known ketol donor and aldehyde acceptor substrates into the active site of E. coli TK, both in the presence and the absence of reactive intermediates. Comparison of docked conformations with available crystal structure complexes allows us to propose a more complete mechanism at a level of detail not currently possible by experimental structure determination alone.

Statistical coupling analysis (SCA) utilises evolutionary sequence data present within multiple sequence alignments to identify energetically coupled networks of residues within protein structures. Using this technique we have identified several coupled networks within the TK enzyme which we have targeted for mutagenesis in multiple mutant variant libraries. Screening of these libraries for increased activity on the non-natural substrate propionaldehyde (PA) has identified combinations of mutations that act synergistically on enzyme activity. Notably, a double variant has
been discovered with a 20-fold improvement in $k_{\text{cat}}$ relative to wild type on the PA reaction, this is higher than any other TK variant discovered to date.
# Table of Contents

1 Introduction .......................................................................................... 14
1.1 Transketolase structure and mechanism .............................................. 17
  1.1.1 Transketolase structure ................................................................. 17
  1.1.2 Cofactor binding ........................................................................... 21
  1.1.3 Substrate binding and recognition ................................................. 23
  1.1.4 Molecular mechanism of Transketolase ........................................ 28
1.2 Transketolase as a biocatalyst ............................................................ 31
  1.2.1 Use of Transketolase in enzymatic syntheses ............................... 33
  1.2.2 Optimisation of process for industrial application ......................... 35
  1.2.3 Transketolase enzyme engineering .............................................. 37
1.3 Current methods in enzyme engineering ............................................. 42
  1.3.1 Error-prone polymerase chain reaction ........................................ 43
  1.3.2 Saturation mutagenesis ................................................................. 44
  1.3.3 Limitations in enzyme engineering .............................................. 45
1.4 Computational methods for enhanced of enzyme engineering .......... 48
  1.4.1 Computational structural modelling .............................................. 48
  1.4.2 Statistical coupling analysis ......................................................... 52
1.5 Conclusions ........................................................................................ 57

2 Mechanistic analysis of Escherichia coli transketolase by in silico docking of substrates in the active site .......................................................... 60
2.1 Introduction ....................................................................................... 60
2.2 Materials and methods ...................................................................... 64
  2.2.1 AutoDock 3.0.5 .......................................................................... 64
2.2.2 Docking of D-erythrose 4-phosphate in yeast TK .................................................64
2.2.3 Docking of D-erythrose 4-phosphate in E. coli TK...........................................64
2.2.4 Creation of a model of the ThDP-enamine intermediate in E. coli TK.65
2.2.5 Docking DE4P and glycolaldehyde in ThDP-enamine complexed forms
of yeast and E. coli TK.........................................................................................65
2.2.6 Docking of natural and non-natural aldehyde substrates into E. coli TK
65
2.2.7 PyMol Molecular Graphics System .................................................................66

2.3 Results and discussion.........................................................................................67

2.3.1 Automated docking of D-erythrose-4-phosphate in the active site of
yeast TK67
2.3.2 Automated docking of D-erythrose-4-phosphate in the active site of E.
coli TK 69
2.3.3 Modelling the enamine intermediate in E. coli TK and docking of D-
erythrose-4-phosphate into the yeast and E. coli TK-ThDP-enamine complexes
71
2.3.4 Docking of other substrates in E. coli holo-TK...............................................74
2.3.5 Comparison of docked and crystallized DR5P conformations...............79
2.3.6 Glycolaldehyde docking .................................................................................81
2.3.7 Comparison of docking for the ketol donors ..............................................83
2.4 Conclusions..........................................................................................................87

3 Statistical Coupling Analysis of Escherichia coli transketolase .......................89
3.1 Introduction..........................................................................................................89
3.2 Materials and Methods .......................................................................................93
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.1</td>
<td>Multiple sequence alignments</td>
<td>93</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Statistical coupling energy calculation</td>
<td>94</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Matrix assembly and cluster analysis</td>
<td>95</td>
</tr>
<tr>
<td>3.3</td>
<td>Results and Discussion</td>
<td>96</td>
</tr>
<tr>
<td>3.3.1</td>
<td>TPP-dependent multiple sequence alignment validation and perturbation of position 469</td>
<td>96</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Identification of residues statistically coupled to Asp 469</td>
<td>99</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Global SCA analysis of TPP-dependent enzymes</td>
<td>103</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Single PYR domain global SCA analysis</td>
<td>114</td>
</tr>
<tr>
<td>3.4</td>
<td>Conclusions</td>
<td>119</td>
</tr>
<tr>
<td>4</td>
<td>Statistical coupling guided library design for enzyme engineering of Transketolase</td>
<td>122</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>122</td>
</tr>
<tr>
<td>4.2</td>
<td>Materials and methods</td>
<td>126</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Materials</td>
<td>126</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Standard procedures</td>
<td>128</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Mutant library construction</td>
<td>133</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Colorimetric screening of libraries for activity</td>
<td>135</td>
</tr>
<tr>
<td>4.2.5</td>
<td>Solubilisation and refolding of insoluble variants</td>
<td>137</td>
</tr>
<tr>
<td>4.2.6</td>
<td>Activity of purified enzymes</td>
<td>138</td>
</tr>
<tr>
<td>4.2.7</td>
<td>Enantioselectivity of purified enzymes</td>
<td>139</td>
</tr>
<tr>
<td>4.3</td>
<td>Results and Discussion</td>
<td>140</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Library design strategy</td>
<td>140</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Construction and screening of SCA directed single mutant library</td>
<td>142</td>
</tr>
</tbody>
</table>
Index of Figures

Figure 1.1 Kinetic Scheme of a Bi Bi Ping Pong Mechanism .................................................. 15
Figure 1.2 In vivo and In vitro reactions catalysed by transketolase ........................................ 16
Figure 1.3 E. coli transketolase structure ................................................................................. 20
Figure 1.4 ThDP bound in E. coli TK ....................................................................................... 23
Figure 1.5 DE4P bound by yeast TK ......................................................................................... 26
Figure 1.6 Covalent complex of X5P and ThDP in E. coli TK .................................................. 28
Figure 1.7 Reaction mechanism of transketolase ................................................................. 29
Figure 1.8 Covalent complex of X5P and ThDP in E. coli TK .................................................. 31
Figure 2.1 DE4P binding in yeast and E. coli TK ................................................................. 69
Figure 2.2 ThDP-enamine intermediate in yeast and E. coli TK .............................................. 73
Figure 2.3 Experimental versus calculated log(Km) values .................................................. 76
Figure 2.4 Binding of phosphorylated and non-phosphorylated substrates ............................ 78
Figure 2.5 DR5P docking in E. coli TK .................................................................................. 81
Figure 2.6 Glycolaldehyde and hydroxypyruvate docked in E. coli TK ................................. 83
Figure 3.1 Conservation and validation of the MSA for SCA ............................................... 98
Figure 3.2 Coupling to residue Asp 469 of E. coli TK ......................................................... 102
Figure 3.3 Validation of the MSA for global SCA ............................................................... 104
Figure 3.4 Unclustered matrix following global SCA of PP and Pyr domains ...................... 105
Figure 3.5 Clustered matrix following global SCA of PP and PYR domains ....................... 106
Figure 3.6 Focused matrix following global SCA of PP and PYR domains .......................... 107
Figure 3.7 PP-Pyr network mapped onto the surface of TK ............................................. 111
Figure 3.8 PP-Pyr network mapped onto the structure of TK ........................................... 113
Figure 3.9 Matrices following global SCA of Pyr domains .................................................. 116
Figure 3.10 Pyr network mapped onto the structure of TK .............................................. 118
Figure 4.1 Validation of colorimetric assay ........................................................................... 137
Figure 4.2 Pyr coupled network mapped onto the structure of TK ...................................... 141
Figure 4.3 Distribution of cell growth across the microplate ............................................... 142
Figure 4.4 Relative lysate activities of single mutants ....................................................... 144
Figure 4.5 Relative lysate activities of variants in the triple mutant cycle .......................... 150
Figure 4.6 Concentrations of TK variants in mutant cycle ............................................... 152
Figure 4.7 k_cat, K_m and k_cat/K_m for the double mutant cycle ........................................... 155
Index of Tables

Table 1.1 Key residues in yeast and *E. coli* TK.................................................................21
Table 2.1 Experimental versus calculated TK $K_m$...............................................................76
Table 3.1 Hierarchical clusters of coupled positions. ............................................................109
Table 4.1 Primers used in site directed mtagenesis. ...............................................................135
Table 4.2 Frequency of residues at selected positions within the MSA .........................141
Table 4.3 Sample sequencing results of variant libraries. .....................................................147
Table 4.4 Experimental versus expected library activities....................................................147
Table 4.5 Initial velocities of TK double mutant cycle. .........................................................154
Table 4.6 Kinetic parameters and enantioselectivities of triple mutant cycle. ....155
Abbreviations

2OXO: 2-oxoisovalerate dehydrogenase
AA: Acetaldehyde
ALS: Acetolactone synthase
BAL: Benzaldehyde lyase
BFDC: Benzoylformate decarboxylase
DE: D-erythrose
DE4P: D-erythrose-4-phosphate
DG: D-glyceraldehyde
DG3P: D-glyceraldehyde-3-phosphate
DHAS: Dihydroxyacetone synthase
DHP: 1,3-dihydroxypentan-2-one
DR: D-ribose
DR5P: D-ribose-5-phosphate
DX5P: D-xylulose-5-phosphate
DXPS: D-xylulose-5-phosphate synthase
E4P: Erythrulose-4-phosphate
ePCR: Error-prone polymerase chain reaction
GA: Glycolaldehyde
GXC: Glyoxylate carboligase
HPA: Hydroxypyruvic acid
HPLC: High performance liquid chromatography
IEMR: Immobilised enzyme microreactor
IPDC: Indolepyruvate decarboxylase
ISPR: In situ product removal
MSA: Multiple sequence alignment
NMR: Nuclear magnetic resonance
PA: Propionaldehyde
PCR: Polymerase chain reaction
PDC: Pyruvate decarboxylase
PFRD: Pyruvate ferrodoxin reductase
PhPDC: Phenylpyruvate decarboxylase
PKL: Phosphoketolase
PO: Pyruvate oxidase
PPDC: Phosphopyruvate decarboxylase
R5P: Ribose-5-phosphate
SCA: Statistical coupling analysis
SPDC: Sulfopyruvate decarboxylase
ThDP: Thiamine diphosphate
TK: Transketolase
1 Introduction

Transketolase (TK) is a key constitutive enzyme in metabolic regulation, providing a link between the pentose phosphate pathway and glycolysis through the production of 3 and 6 carbon sugars (glyceraldehyde-3-phosphate and fructose-6-phosphate respectively) (Figure 1.2 a and b) [1]. Found in the non-oxidative branch of the pentose phosphate pathway, TK catalyses the reversible transfer of two carbon ketol groups between several donor and acceptor substrates. In addition to supplying substrates for glycolysis, TK controls the supply of ribose-5-phosphate (R5P), essential for biosynthesis of nucleotides and nucleic acids, and catalyses the production of erythrulose-4-phosphate (E4P) which is utilised by microorganisms in the shikimate pathway for the biosynthesis of aromatic amino acids.

Transketolase was first purified from Saccharomyces Cerevisiae [2] and requires divalent cations and thiamine diphosphate (ThDP) for its activity [3]. There is a high level of sequence identity between the TK proteins of different organisms with many residues displaying complete invariance [4]. A second TK encoding gene was identified in Escherichia Coli in 1993 [5], this gene was named tktB to distinguish it from tktA. tktA and tktB share high sequence identity (74%) but tktA encodes the major TK activity in E. coli. All future references to E. coli TK refer to tktA encoded transketolase. In all structures solved to date, TK exists as a homodimer with two identical active sites positioned at the interface between the subunits.

The reaction catalysed by TK proceeds via a Ping Pong Bi Bi mechanism: two substrates are converted into two products as the ThDP cofactor within the enzyme active site shuttles between a free and a substrate modified intermediate state [6].
The Ping Pong Bi Bi model describes a specific type of Bi Bi mechanism in which substrates and products are bound and released sequentially and the enzyme shuttles between a free and a substrate modified intermediate state (Figure 1.1).

Enzymatic thiamine catalysis is recognised as proceeding through two intermediate states: the ylide of ThDP in which the C2 proton of the thiazolium ring is abstracted, and the 2-α carbanion which is formed following nucleophilic attack by the ylide C2 on the donor substrate [6]. The α-carbanion is stabilised by the thiazolium ring which acts as an electron sink. Further stabilisation is provided by interconversion of the α-carbanion into a neutral enamine, creating a resonance hybrid. Following formation of the intermediate, the two carbon unit is transferred from the carbanion to the acceptor substrate forming a ketose with an extended carbon skeleton through nucleophilic attack. The ThDP in TK is bound in a V conformation which brings the 4-amino group of the pyrimidine ring into close proximity with the C2 carbon atom of the thiazolium ring, this conformation is essential for catalysis [7].
Broad substrate specificity, stereospecificity and stereoselectivity have made TK an attractive target for applications in organic synthesis [8]. The reversible reaction catalysed by TK in vivo has been utilised in a synthetic manner, but if the natural ketol donor is replaced with hydroxypyruvic acid (HPA), carbon dioxide is released as a by-product rendering the reaction irreversible and far more industrially useful (Figure 1.2 c). The industrial applicability of TK has been further adapted by engineering the protein sequence of the enzyme to improve attributes such as substrate specificity and enantioselectivity [9-11]. This thesis explores the potential of computational applications to direct and optimise the engineering of this enzyme towards even greater improvements.

Figure 1.2 In vivo and In vitro reactions catalysed by transketolase. Reactions (a) and (b) occur in vivo in the non-oxidative branch of the pentose phosphate pathway and are reversible. (c) In vitro the ketol donor is generally replaced with β-HPA, rendering the reaction irreversible through elimination of CO₂. Various aldehyde acceptors are accepted by TK but TK preferentially accepts α-hydroxylated aldehydes with the (R)-configuration.
1.1 Transketolase structure and mechanism

1.1.1 Transketolase structure

The first transketolase structure was solved for yeast TK in 1992 [12] this was refined to 2Å in 1994 [13] (1TRK). Since then, several other transketolase structures have been solved for TK. The *E. coli* TK structure was solved in 1994 [14] (1QGD), Maize TK in 2003 [15] (1ITZ) and *Leishmania Mexicana* TK in 2004 [16] (1R9J). Several further yeast TK structures have also been solved in the quest to refine our functional understanding of this catalyst. These include Apo TK [17], D-erythrose-mutants of TK [19] (1AYO) and several complexes of TK with ThDP analogues [20] (1TKA, 1TKB, 1TKC). More recently, *E. coli* TK structures were determined in covalent complexes with DX5P (2R8O) and DF6P (2R8P), and in non-covalent complex with DR5P (2R5N) [21].

The majority of detailed structural analysis and functional studies have been carried out on the yeast TK protein. Yeast and *E. coli* TK share a very high level of sequence identity and all homologous TK structures defined show near identical conformations of functional residue side chains. The vast majority of functional residues identified in yeast TK are 100% conserved in all TK proteins sequenced so far. We can therefore utilise the data and information gathered for yeast TK and apply it in our study of *E. coli* TK. Throughout this thesis, except where scientific evidence is based solely on yeast TK, numbering refers to *E. coli* TK and is based on the PDB structure 1QGD. Where yeast TK numbering is used, residues are underlined and numbering is based on the PDB structure 1TRK (in these situations *E. coli* numbering is also reported in brackets). Key functional residues are listed in
1 - Introduction

Table 1.1 together with the corresponding *E. coli* and yeast numbering for reference.

In all TK structures solved to date, TK is a homodimer consisting of two subunits each of 70-74 kDa (Figure 1.3 a). ThDP binds along with the metal ion cofactor at the interface between these two subunits. Each subunit of TK is made up of three domains, the N-terminal domain or the PP domain, the middle domain or Pyr domain, and the C-terminal domain (Figure 1.3 b). Each of the domains is of α/β type and the PP and Pyr domains are structurally similar when superimposed upon each other. The following structural analysis of transketolase is based on the structure for *E. coli* TK [14], but could equally be applied to the structure of yeast TK due to the high degree of structural equivalence.

The PP domain of *E. coli* includes residues 2-317 and comprises of a five stranded parallel β sheet with several helices on either side and some on top of the sheet. The α/β connection after the third strand contains a hairpin loop 187-191, this loop is involved in binding the cofactor and has been shown to be mobile in the apo-transketolase of yeast [17]. In the holo-enzyme of yeast TK, Asp 192 (190) and Ile 191 (189) are in contact with the metal ion and the cofactor. These interactions keep the hairpin in a closed conformation, enclosing the cofactor and shielding it from solvent [14].

The Pyr domain includes residues 318-527. This domain is made up of a parallel β-sheet of six strands. As mentioned above, the Pyr domain is structurally similar to the PP domain. The similarity between these two structural motifs is most pronounced in the last four α/β units of the two domains. Like the PP domain, the
Pyr domain forms interactions with the THDP cofactor through the loops at the carboxy ends of the β-sheet [14].

The C-terminal domain of TK consists of a mixed β-sheet with one antiparallel strand followed by four parallel strands. This domain is not involved in binding the ThDP cofactor and contributes less to the dimer interface interactions than do the other domains. Recent results [22] have demonstrated that the C-terminal domain is not essential for catalysis. The function of this domain remains unknown but it has been suggested that it may have a regulatory or a cellular localisation role.

The interface between the subunits of TK consists of a buried region representing approximately 18% of the accessible surface area of one free monomer [13]. Interactions between the two equivalent PP domains consist of tight packing interactions between the equivalent helices that link β-strands two and three and the equivalent helices that link strands three and four. The main interactions between the Pyr domains of each subunit are limited to the equivalent helices that link strands four and five of the Pyr β-sheet. This dimeric configuration positions the loops at the carboxy ends of the PP β-sheet facing the loops at the carboxy ends of the Pyr β-sheet of the other subunit. The region where these loops come together constructs the ThDP binding site and the active site of TK.
Figure 1.3 (a) *E. coli* transketolase homodimeric structure (1QGD) coloured by chain. (b) Chain A of *E. coli* transketolase coloured by domain.
1 - Introduction

1.1.2 Cofactor binding

ThDP binds in a deep cleft at the interface of the two TK subunits. Bound ThDP is totally isolated from the surrounding solvent apart from the reactive C2 carbon atom of the thiazolium ring. Unlike the structure of free ThDP, the bound cofactor is strained into a V-conformation. This brings the pyrimidine ring 4’-NH2 group into close proximity with the reactive C2 carbon and contributes to the catalytic mechanism of all ThDP dependent enzymes (Figure 1.4).

Conserved E. coli TK residues His 66 and His 261 form hydrogen bonds with the diphosphate group of ThDP. Two oxygen atoms of diphosphate together with Asp
155, Asn 185 and the main chain oxygen of Ile 187 are ligands of Ca$^{2+}$ and create further indirect interactions [14].

Whilst the diphosphate of ThDP forms interactions with one subunit of TK, the thiazolium and pyrimidine rings of the cofactor are bound in a cleft between the subunits. The thiazolium ring forms hydrophobic interactions with conserved residues Leu 116 and Ile 189. The C4 methyl group of the thiazolium ring interacts with the side chains of Leu 382 and Val 409 [14].

The pyrimidine ring is stacked with the ring system of Phe 437 and forms further interactions with conserved residues Phe 434 and Tyr 440. Main chain atoms of Gly 114 and Leu 116 form H-bonds with the pyrimidine ring 4’-NH2 group and the N3’ nitrogen atom respectively. The N1’ nitrogen atom of the pyrimidine ring forms a H-bond with Glu 411 [14]. This interaction is very important in the molecular mechanism of enzymatic thiamine catalysis [23].
1.1.3 Substrate binding and recognition

The active site binding cleft of TK is a deep funnel leading towards the exposed reactive C2 of ThDP. Conserved loops make up the walls of this binding funnel. Two conserved arginine residues are positioned at the entrance to the binding funnel. The middle of the binding channel contains several conserved residues including Asp 469, Ser 385 and His 461. Towards the base of the binding cleft, near the thiazolium ring of ThDP, there is a cluster of histidine residues on one side and there are also several conserved hydrophobic residues [14].
The crystal structure of the acceptor substrate, DE4P, bound in yeast TK gives insight into the residues involved in substrate binding and recognition [18]. The phosphate group of the acceptor substrate forms interactions with several conserved residues near the entrance of the binding funnel. Arg 359 (358), Arg 528 (520), Ser 386 (385) and His 469 (461) form interactions with the substrate phosphate group. These interactions position the substrate in the binding channel in the correct orientation, the long side chains of the arginine residues may also provide some flexibility to allow the substrate to move towards the reactive C2 of ThDP [18]. Asp 477 (469) forms polar interactions with the C2-hydroxyl group of the substrate and the aldehyde oxygen atom is within H-bonding distance of His 30 (26) and His 263 (261) (Figure 1.5) [18].

In order to further elucidate the role played by the arginine and histidine residues at the entrance to the binding funnel, these residues have been subjected to site directed mutagenesis in yeast TK [18]. Substitution of residues Arg359 (358), Arg528 (520) and His469 (461) for alanine did not have a great effect on catalytic activity (residual catalytic activities were 31%, 17% and 77% respectively) but did increase $K_m$ values for donor substrate, and in particular acceptor substrates, significantly. Consistent with the crystal structure of DE4P bound in the TK active site, these results support a role for these residues in binding the phosphate group of substrates.

The pattern of H-bonds formed by Asp 477 (469), His 30 (26) and His 263 (261) with the acceptor substrate is consistent with the enantiosensitivity TK displays towards D-threo configured donor substrates. Inversion of the stereocentres in the favoured
configuration would disrupt this H-bond network and reduce enzyme affinity for substrate. The potential of forming a H-bond with Asp 477 (469) also explains the preference for α-hydroxylated acceptor substrates. Replacement of Asp 477 (469) with alanine in yeast TK resulted in an enzyme with severely impaired catalytic activity [24]. Kcat/Km for this variant is reduced relative to wild type TK for D-α-hydroxyaldehydes (DE4P, DR5P) and this reduction is equivalent to the reduction in Kcat/Km for the wild type enzyme with 2-deoxyaldoses or L-α-hydroxyaldehydes [24].

Yeast TK residues His 30 (26) and His 263 (261) have also been mutated to alanine [19]. These residues are within H-bonding distance to the carbonyl oxygen of the acceptor substrate and their mutation to alanine has a large effect on kcat [19].
A cluster of histidine residues is located towards the base of the binding funnel close to the reactive C2 of ThDP. Based on the structures of TK and acceptor substrate (DE4P) bound TK, histidine residues His 69 (66) and His 103 (100) were predicted to form H-bonds with the C1-hydroxyl group of the donor sugar phosphate [18]. In support of this role, replacement of His 69 (66) or His 103 (100) with alanine had little effect on the $K_m$ values for acceptor substrates but significantly increased those of donor substrates [19, 25]. These mutants also displayed a significant decrease in catalytic activity. Although hydroxyphenylacetate is a donor substrate for TK, pyruvate is not. The recognition of the C1-hydroxyl group of the donor by these two histidine residues might explain this molecular selectivity. In
a crystal structure of the covalent enamine intermediate bound in yeast TK [7], His 103 (100) formed a H-bond with the β-hydroxyl oxygen of the intermediate. The β-hydroxyl oxygen also interacted with His 69 (66), indirectly through a water molecule. These interactions further support a role for His 69 (66) and His 103 (100) in discrimination between hydroxypyruvate and pyruvate.

Recent structures determined of covalent intermediates in *E. coli* TK support the structural studies above and the conserved function of residues between yeast and *E. coli* TK [21]. DX5P and DF6P covalent intermediates adopted very similar extended conformations in the active site of *E. coli* TK, forming at least 11 well defined hydrogen bonds with the side chains of active site residues. The C1-hydroxyl and C2-hydroxyl groups of both substrates formed interactions with His 473 and the 4’-amino group of the ThDP pyrimidine ring. The C1-hydroxyl group also formed a hydrogen bond with His 100. The C3-hydroxyl groups interact with the two histidine residues His 261 and His 26 and the C4-hydroxyl group interacts with Asp 469 and His 26 (Figure 1.6). Phosphate interactions were mediated by residues His 461, Ser 385 and Arg 358 with the phosphate group of DF6P slightly closer to these residues due to the longer carbon chain. The additional C5-hydroxyl group of DF6P interacts with Ser 385.
1.1.4 Molecular mechanism of Transketolase

Transketolase catalyses the transfer of a two carbon unit from a ketose donor to an aldose acceptor, the reaction proceeds through two major steps. In the first step, the donor substrate is cleaved to produce an aldose and a covalent intermediate, ThDP α-carbanion. The second step is initiated by nucleophilic attack by the α-carbanion on the acceptor substrate, resulting in a ketose product with an extended carbon skeleton. This reaction mechanism, described in further detail below, was proposed by Schneider and Lindqvist (1993) [6] (Figure 1.1Figure 1.7).
Prior to the first step in catalysis, the C2 carbon of the thiazolium ring must be deprotonated in order to create an ylide that can attack the donor substrate. Evidence suggests that the deprotonation of C2 is catalysed by the cofactor itself [23]. The transketolase molecule contributes to cofactor deprotonation by maintaining the V-conformation of ThDP, which brings the 4’-NH2 group into close proximity with the C2 carbon, and through protonation of the N1’ nitrogen of the pyrimidine ring. The N1’ nitrogen is protonated by a H-bond with Glu 411, this interaction alters the pKa of the 4’-NH2 group and leads to the production of a 4’-imino group which is sufficiently basic to deprotonate the C2 carbon of the thiazolium ring [6].

Once the C2 carbon has been deprotonated, the carbanion formed attacks the carbonyl oxygen of the donor substrate to create a high energy intermediate. During covalent bond formation between ThDP and the donor substrate, a proton...
donor is required to stabilise the negative charge forming at the carbonyl oxygen. His 473 and the charged 4’-imino group of ThDP are possible proton donors at this step in the reaction. While site directed mutagenesis of the His 481 (His 473) in yeast TK suggested a role in transition state stabilisation, the lack of conservation of this histidine residue across other TK enzymes suggests that the 4’-imino group may be responsible for the majority of transition state stabilisation.

The final steps in TK catalysis require an acid/base catalyst that can deprotonate the hydroxyl group of the substrate at C3, catalysing the cleavage that produces the α-carbanion intermediate, and act as a proton donor to the carbonyl oxygen of the acceptor substrate as it is attacked by the α-carbanion. Both His 26 and His 261 are within H-bonding distance from the C3-hydroxyl group of the reaction intermediate and the carbonyl oxygen of the acceptor substrate. Replacement in yeast of either of these residues by alanine severely impairs catalytic activity and they may act together as the acid/base catalyst in the reaction [19].

Crystal structures of the covalent high energy intermediates formed with DX5P and DF6P in E. coli TK reveal a conformation in which the newly formed C2-Cα bond is out-of-plane with the thiazolium ring by 25-30° (Figure 1.8) [21]. This strained conformation will be relieved upon product elimination conceivably providing the driving force for the reaction. Density functional theory (DFT) calculations supported the above out-of-plane conformation, demonstrating that this conformation is energetically favourable relative to a model with a co-planar C2-Cα bond [21]. No structural rearrangements were seen in the active site following intermediate formation, suggesting that the active site is poised for catalysis such
that the substrate binding energy and the enthalpic energy gain following covalent bond formation between ThDP and the donor substrate can be channelled directly into the formation of the high energy, strained intermediate.

![Covalent complex of X5P and ThDP isolated in the active site of E. coli TK. The strained angle of the out-of-plane C2-C\(\alpha\) bond is displayed relative to the thiazolium ring of ThDP.](image)

**1.2 Transketolase as a biocatalyst**

*In vivo*, transketolase catalyses the transfer of a 2-carbon ketol unit from D-xylulose-5-phosphate to D-ribose-5-phosphate, generating D-sedulose-7-phosphate and D-glyceraldehyde-3-phosphate. The carbon-carbon bond formation catalysed by TK is both stereospecific and stereoselective, making this a very attractive enzyme for the industrial production of complex organic structures. Unlike traditional chemical methods for organic synthesis, enzymatic catalysis does not require complex protection and de-protection steps, additional advantages include greater chiral control and milder reaction conditions.
The reversible reaction catalysed by TK in vivo has been utilised for industrial applications but the usefulness of the biocatalyst can be enhanced by replacing the D-xylulose-5-phosphate (X5P) ketol donor with β-hydroxypyruvate (HPA). Use of HPA results in the elimination of CO₂ thereby rendering the reaction irreversible [26].

In addition to varying the ketol donor, much work has been carried out exploring the substrate specificity of TK for aldehyde acceptors. Although TK favours aldehydes containing an α-hydroxylated group in an (R)-configuration [27], specificity with respect to the aldehyde is relatively broad with both non-phosphorylated and phosphorylated aldehydes of varying sizes being accepted [28]. Steric hindrance appears to impact on relative activity with cyclic aldehydes and aldehydes containing bulky groups displaying lower activities.

The broad nature of aldehyde substrate specificity is advantageous for the application of TK in organic synthesis as it provides flexibility in the nature of structures that can be produced. The broad specificity also provides a good starting point for engineering TK to improve activity on non-natural substrates.

Early work in the development of TK as a biocatalyst utilised commercially available S. cerevisiae TK, and spinach TK extracted from spinach leaves. These sources are readily available but do not offer the scale required for an industrial process. In 1993, Hobbs et al developed an efficient and reliable source of TK by introducing the previously cloned E. coli TK gene fragment into a high copy number plasmid and transforming this into E. coli [29]. The resulting transformants overexpressed TK with superior specific activity to that obtained previously from E. coli TK. Evaluation
of the substrate specificity of *E. coli* TK demonstrated a similar profile to yeast TK and the same preference for α-hydroxylated aldehydes with an (R)-configuration [29]. An additional benefit of *E. coli* TK was its increased activity with HPA (60 U/mg protein) relative to yeast (8.6 U/mg protein) or spinach (2 U/mg protein) [1, 28, 30].

1.2.1 Use of Transketolase in enzymatic syntheses

Transketolase has been used successfully in various organic syntheses of both natural and unnatural complex, chiral compounds. The compounds produced using transketolase are expensive and would require complex multi-step synthesis if produced using traditional chemical methods. Some of the compounds synthesised by transketolase cannot be produced chemically.

In an early example, D-[1,2-\(^{13}\)C\(_2\)]-xylulose was produced from [2,3-\(^{13}\)C\(_2\)]-hydroxypyruvate and D-glyceraldehyde using spinach transketolase [31]. The isotopic labelling of sugars is useful for the study of metabolism and structure. Introduction of two adjacent \(^{13}\)C labels is particularly useful due to \(^{13}\)C-\(^{13}\)C coupling. Transketolase has also been utilised to produce glycosidase inhibitors fagomine [32] and 1,4-dideoxy-1,4-imino-D-arabinitol [33], which have applications as agrochemicals and therapeutic agents.

Commercially available TK was used together with chemical steps to produce α-exo brevicamycin, a beetle pheromone with applications in pest control. In this example TK was used to convert a racemic mixture of 2-hydroxy butyraldehyde and HPA into the key tri-hydroxy ketone intermediate in the process [34].
Production of expensive food additives has also utilised TK. Spinach TK was used to produce 6-deoxy-L-sorbose, a precursor for synthetic furaneol, an aromatic product with a caramel like flavour. In the above synthesis, serine glyoxylate amino transferase was used to prepare HPA and 4-deoxy-L-threose was obtained by microbial isomerisation of 4-deoxy-L-erythulose [35].

Multi-enzyme approaches have been successfully applied to the production of sugars 4-deoxy-D-fructose 6-phosphate and D-xylulose 5-phosphate. In the first example, 4-deoxy-D-fructose 6-phosphate was synthesised in a process including four steps and two enzymatic reactions. Epoxide hydrolase was first used to obtain (3S)-1,1-Diethoxy-3,4-epoxybutane which was subsequently opened by inorganic phosphate to produce 2-deoxy-D-erythrose 4-phosphate. This aldehyde was then reacted with L-erythulose in the presence of yeast TK to introduce the second chiral centre [36].

In the second example, D-xylulose-5-phosphate, a valuable substrate required for enzymatic assays, was prepared in gram quantities through a one-pot procedure incorporating fructose 1,6 bisphosphate aldolase followed by *E. coli* TK [37]. The aldolase was initially used to produce D-glyceraldehyde-3-phosphate and dihydroxyacetone phosphate (DHAP) through retro-aldolization of D-fructose-1,6-bisphosphate. The D-glyceraldehyde-3-phosphate produced was then coupled with HPA by *E. coli* TK to produce X5P. Zimmermann *et al* also incorporated triosephosphate isomerase (TPI) to equilibrate the two products of the retro-aldolization thereby increasing the overall yield. Thanks to recent advances in
synthesis [38], DHAP is now available as a starting point, opening up a new route to the synthesis of X5P using just TPI and TK [39].

Finally, transketolase was utilised to catalyse the key asymmetric step in a process to produce N-hydroxypyrrolidine, a glucosidase inhibitor. In this synthesis TK was used to couple (+)-3-O-benzylglyceraldehyde with HPA to yield 5-O-benzyl-D-xylulose in multigram quantities [40].

1.2.2 Optimisation of process for industrial application

To utilise TK in large scale industrial enzymatic syntheses, limitations in the process must be identified and overcome to ensure product purity and maximise yield and efficiency. Limitations in TK biocatalysis include stability of the enzyme (reactive α-hydroxy aldehydes inactivate TK [41]), and product inhibition. More general process options also need to be optimised during process development to ensure maximum yield, minimum cost and scalability.

One strategy to overcome the problem of substrate deactivation of TK is to minimise substrate concentration by carrying out the reaction in an enzyme membrane reactor. Applied to the production of L-erythrulose from GA and HPA this strategy was able to increase the half-life of TK from 5.6 hours (repetitive batch reactor) to 106 hours (enzyme membrane reactor). This improvement in stability resulted in an increase in space time yield from 28 g/L/d to 45 g/L/d [41].

An alternative approach to increasing enzyme stability is to immobilise the enzyme on a support. GA is believed to deactivate TK by forming Schiff bases with amino acid side chains on the surface of the enzyme. Formation of Schiff bases can alter
the three dimensional structure of proteins. Stabilisation of TK by immobilisation can be explained by the prevention of the formation of Schiff bases, or by rigidifying the three dimensional structure of TK. Applied to the production of L-erythulose from GA and HPA, immobilisation of TK on commercially available supports (Amberlite XAD-7 and Eupergit C) increased the stability of TK by 80- and 100-fold respectively. Immobilisation was unable to prevent inactivation by oxidation but this could be reduced by the inclusion of a stabilising solute such as mercaptoethanol [42]. Immobilisation confers additional benefits on process productivity: allowing the enzyme to be retained in the bioreactor, extracted from the product stream, and/or reused.

In situ product removal (ISPR) has been explored as a means to overcome limitations introduced by product inhibition. L-erythulose was successfully removed using an immobilised phenylboronate resin, however there were also considerable levels of nonspecific substrate binding to the resin which reduced the actual yield. To overcome the problem of nonspecific binding, a fed batch system was utilised. The fed batch mode gave the added benefit of reducing the deactivation of TK by GA. The rate of deactivation of TK by substrate is much higher than by the synthesised product [41], therefore in practice the reduction in aldehyde toxicity overcame the benefits of product removal and negated the need for ISPR [43].

Work has also been carried out to explore more general process development strategies to optimise yield, cost and scalability. The synthesis of X5P using TK and TPI has been utilised as a model reaction to investigate the potential of
Introduction

The strategy successfully identified new biocatalytic routes and processes for further investigation [44].

As an alternative to reducing the number of process options to screen, high-throughput microwell based methods can be used to screen multiple process options in parallel. For TK process characterisation a more efficient alternative to microwells has recently been developed in the form of an immobilised enzyme microreactor (IEMR). The microreactor developed is composed of a 25 cm long fused silica capillary with a 200 μm internal diameter. His-tagged TK is reversibly immobilised inside the capillary via Ni-NTA linkages. His-tagged TK is expected to be kinetically identical to un-tagged protein based on the location of the His tag and on previous kinetic characterisation. The reactor can be operated in stop flow or continuous flow mode and product is analysed by HPLC. For high throughput screening of different process options or different enzyme variants the IEMR has several advantages over traditional microwell approaches, these include reduction in reactant volume, enhanced productivity, reduced reaction time and increased reusability [45].

1.2.3 Transketolase enzyme engineering

Although TK displays broad substrate specificity, non-natural aldehyde substrates are converted at far lower rates than the phosphorylated, α-hydroxylated natural substrates. The production of non-phosphorylated chiral products is one of the
major advantages of TK over other enzymes such as aldolases, it would therefore be advantageous to increase the activity of TK on these non-natural substrates.

In addition to process engineering, enzyme engineering can be applied to TK to further optimise the biocatalyst for industrial applications. The good structural and mechanistic understanding of TK and its broad substrate specificity make the wild type enzyme a good candidate for enzyme engineering. However, the two-substrate mechanism of TK complicates engineering as any change can result in positive or detrimental effects to the binding of either substrate. In recent years TK has been engineered to improve its activity on non-phosphorylated and non-hydroxylated substrates. The enzyme has also been engineered to improve or reverse its stereoselectivity with these non-natural substrates.

Early work on TK engineering focussed on improving activity in the model reaction of GA and HPA to produce L-erythulose. Variants with up to 5-fold improvements in activity against GA were identified by screening a library of single point mutants generated by saturation mutagenesis of nineteen positions. Residues were selected for mutagenesis based on structural and phylogenetic criteria. Two sets were included: residues within 4 Å of bound substrates, and phylogenetically varied residues within 10 Å of TPP. Following the screen, twelve variants were identified with enhanced specific activity on GA relative to wild type TK. Six of the nineteen residue libraries yielded variants with improved activity.

The greatest improvements in activity against GA were associated with variations at residues His 461, Arg 520 and Ala 29. His 461 and Arg 520 interact with the phosphate group of natural substrates, mutation of these residues could improve
GA accessibility to the active site by removing a charged group or by reducing the steric bulk around the entrance to the active site. The third position identified, Ala 29, is harder to rationalise as it is in the active site second shell and is in direct contact with the terminal phosphate group of TPP. Overall, residues with high sequence entropy were more likely to confer enhanced activity on GA following mutagenesis. The important exceptions to this rule were the three residues known to interact with the phosphate group of natural substrates. Saturation libraries of His 461, Arg 358 and Arg 520 all yielded at least one mutant with increased specific activity on GA despite the low sequence entropy at these positions. Interestingly the most successful mutations were non-natural variants [9].

In addition to increasing activity on non-phosphorylated substrates it would be advantageous to extend the applicability of TK by enhancing activity on aliphatic, non-hydroxylated aldehydes. Typically, activity of wild type TK on these substrates is only 5%-35% of that for α-hydroxylated aldehydes such as GA. The active site variant libraries described above were screened for TK catalysed production of 1,3-dihydroxypentan-2-one (DHP) using the aldehyde substrate propioanldehyde (PA).

Twenty-six distinct mutants were identified with increased specific activity on PA relative to wild type. These variants represented eight of the nineteen residue libraries. Five of the eight libraries yielding enhanced activity on PA had been previously identified through screening for enhanced activity on GA. These included the group of residues which interacts with the phosphate group of natural substrates, Arg 358, His 461, and Arg 520. The phylogenetically variant residues in the active site second shell, Ala 29 and Asp 259, were also identified. A third group
of residues not previously identified yielded variants with the greatest increase in activity on PA. This group was made up of conserved residues His 26, Asp 469 and His 100. These residues form a pocket in the active site and directly interact with the hydroxyl group at C2 of erythrose-4-phosphate. The greatest increase in activity on PA was observed with D469T which demonstrated a 5-fold improvement in specific activity over wild type [10].

Variation of phylogenetically variant residues or those interacting with the phosphate group of natural substrates led to enhanced activity on both GA and PA [10, 46]. In contrast, mutation of hydroxyl interacting residues produced variants with enhanced substrate specificity for PA over GA. The D469Y mutant displayed the greatest substrate specificity, with a 64-fold higher activity on PA relative to GA [10].

Wild type TK catalyses the production of L-erythrulose from GA and HPA with 95% ee, but 3S-DHP is only produced in 58% ee using PA and HPA. To identify variants with increased stereoselectivity for the production of 3S-DHP, Smith et al screened the three variant libraries of residues that interact with the hydroxyl group of natural substrates, His 26, Asp 469 and His 100. Significant increases in stereoselectivity were obtained with the D469E variant which produced 3S-DHP in 90% ee. Interestingly the majority of His 26 variants lead to the formation of 3R-DHP and H26Y produced 3R-DHP in 88% ee [11].

This work was extended to establish the enantioselectivity of wild type TK and selected variants on linear aliphatic aldehydes of increasing length and cyclic aldehydes. Compared to wild type TK, D469E TK produced products in greater yields.
for the longer chain and cyclic aliphatic aldehydes although yield did decrease as chain length increased. Although yields were lower, D469E displayed enhanced stereoselectivity with longer chain aldehydes and over 99% ee for cyclic aldehydes cyclopropanecarbaldehyde and cyclopentanecarbaldehyde. H26Y gave product with lower yields but the reversal of ee was maintained across all the aldehydes tested, with the highest ee noted for butanal at 92% [47].

Although work to date has been limited to single point variants and residues in close proximity to the active site, considerable success has been achieved in the engineering of TK. Activity on both non-phosphorylated and non-hydroxylated aldehydes has been improved by 5-fold relative to wild type, and enantioselectivity has also been improved for the non-natural substrates. Surprisingly, a single point mutation was also able to reverse the enantioselectivity of the enzyme. The variants identified further expand the potential applications of TK for industrial synthesis applications. However, there is significant potential to further engineer TK to the point where non-natural substrates convert at the same rate and with the same exquisite stereoselectivity as the natural substrates.
1.3 Current methods in enzyme engineering

Directed evolution can be used to engineer the catalytic properties of enzymes and has been used successfully to modify properties such as specificity, selectivity and enantioselectivity. The general process of directed evolution can be broken down into three steps; generation of a variant library, screening for a desired property and selection of positive variants for the next cycle. Although most directed evolution experiments follow this overall process, there are many different ways to carry out each step. The main limitation in directed evolution is the size of the library that can be screened. Typically, libraries of $10^3$ to $10^6$ variants can be screened using high-throughput techniques and robotic equipment. In some cases a desired attribute can be linked to a growth advantage in bacteria, allowing significantly larger libraries (up to $10^{9-13}$) to be screened. Unfortunately for most directed evolution experiments this cannot be applied and we are limited to libraries containing thousands rather than millions of variants. In some situations we are limited to even smaller libraries, for example where a suitable colorimetric or fluorogenic assay is not available.

The limitation in the size of libraries means we can only sample a tiny fraction of the possible sequence space. A protein with 300 amino acids has $20^{300}$ distinct possible sequences so even with a large screening effort of $10^6$ variants we can only sample a minute fraction of the potential sequence space. The strategy chosen to produce the variant library must therefore be selected very carefully in order to target the best section of sequence space for the desired property.
1.3.1 Error-prone polymerase chain reaction

The most commonly used method to produce a library is error-prone polymerase chain reaction (epPCR). Error-prone PCR involves the introduction of random copying errors through imperfect reaction conditions (e.g. by adding Mn\(^{2+}\) or Mg\(^{2+}\) to the PCR reaction mixture) usually with the aim of introducing approximately one mutation each time the gene is copied [48]. Such a technique applied to a 300 amino acid protein will produce a library of 5,700 potential variants, easily screenable even accounting for the oversampling required. In such examples it is usually possible to find a variant with improved properties although several cycles may be required as multiple variations are usually needed to generate the required level of improvement in a particular property. Previous experiments have demonstrated that on average 30% - 50% of random mutations are deleterious, 50% - 70% are neutral, and just 0.01% - 0.5% are beneficial [49]. In our example of a 300 amino acid protein we could expect to find 1 – 30 beneficial mutations in the 5,700 variant library.

Using epPCR the full length of the protein sequence can be probed; but the cycling nature of directed evolution, and the introduction of one change at a time, means that an evolutionary trajectory is entered once the first variant has been selected. This trajectory theoretically limits the potential optimal sequence that can be achieved. If all the individual variations that constitute an improved activity are independently beneficial and additive in nature then this is not a problem because all trajectories should arrive at the same optimal solution. Unfortunately, variation
at one site in a protein sequence often requires the simultaneous variation at another site to be beneficial.

Following several rounds of selection directed evolution experiments often plateau with further rounds failing to generate improvements. This has been suggested to represent protein sequences becoming stuck on fitness peaks in sequence space. Usually this can be overcome by one or two rounds of selection for stabilising mutations, it is hypothesised that the accumulation of beneficial mutations gradually reduces stability until no further mutations can be tolerated [49]; introduction of stabilising mutations allows the protein to tolerate further mutations. We can envisage a situation where an incredibly beneficial variant is so destabilising that it cannot be introduced without the simultaneous introduction of a specific stabilising variation, which by itself may have no impact on the stability of the wild type sequence. In such a situation, it would be impossible to identify the beneficial variant even following pre-stabilisation of the protein fold.

1.3.2 Saturation mutagenesis

Another common method of library generation is saturation mutagenesis. This approach requires the selection of one, or a small number of sites in the protein sequence; randomised codons are then used in PCR primers to generate all the possible variants at the individual sites. Single site libraries created in this manner are small enough to allow the use of conventional GC or HPLC methods for screening and therefore allow selection for properties that are intractable with high throughput techniques. However, as more sites are added to the library, the size quickly becomes intractable to screening. Complete saturation at three sites would
create 7,999 potential variant sequences (including 57 single point mutants and 1,083 double mutants), four sites results in approximately $1.6 \times 10^5$ possible sequences, representing the approximate upper limit even for high throughput screening techniques. Such a mutagenesis strategy generally requires structural information and a good understanding of the functionally important residues. Sites are normally selected based on their proximity to the active site and on their hypothesised role in specific elements of the reaction being catalysed.

Saturation mutagenesis overcomes some of the issues which limit epPCR, allowing multiple amino acids to be varied simultaneously means we can identify combinations of variations that may be deleterious or neutral in isolation. Reetz et al identified multiple beneficial variants by saturated mutagenesis of three positions in the epoxide hydrolase enzyme from *Aspergillus niger* (ANEH), variants were screened for activity on a new substrate. 5000 variants were screened resulting in 26 unique hits. Twenty two (85%) of the resulting hits were triple mutants and four (15%) were double mutants; none were single mutants [50]. The authors did not create double and triple mutant cycles to investigate whether the multiple beneficial variants were synergistic or additive in nature but it is interesting that no single variants were identified by the screen.

1.3.3 Limitations in enzyme engineering

Many variations and combinations of the above methods have been used in directed evolution experiments. A further method, DNA shuffling, represents another technique used for library generation but this is not covered in detail here. For each enzyme engineering experiment a choice of technique is made in an effort
to maximise the proportion of beneficial variants in the limited library size. However, no one method has proved universally superior to the others and each has its own benefits and limitations.

In examples where mutagenesis is focused on residues likely to confer beneficial properties, positions are usually selected based on their proximity to the active site. Where an attempt is made to introduce multiple mutations, such sites are often selected based on their proximity to each other [51]. Such a simple selection strategy illustrates the limitations of current enzyme engineering approaches. Enzyme properties such as substrate specificity are sometimes determined by sites distant from the active site. Hedstrom et al successfully engineered Trypsin serine protease to accept Chymotrypsin substrates, but to do so required mutation of both the binding pocket and distributed surface loops which don’t interact directly with the substrate. A particular residue (172) was identified as a determinant of substrate specificity through interaction with both the binding pocket residues and surface loops [52, 53]. Using proximity to the active site to select residues for mutagenesis, it would not be possible to engineer a Trypsin enzyme to accept Chymotrypsin substrates.

Modern computational techniques such as structural modelling and statistical coupling analysis provide a new resource to refine our choice of enzyme residues to target in enzyme engineering experiments. In addition to supporting library design, structural modelling can be used to rationalise positive hits identified in a successful enzyme engineering campaign. This new information can be cycled around for further potential benefit in later rounds of design. Statistical coupling analysis has
the potential to identify relationships between residues that are not apparent from examination of the structure. Such information could lead to the production of multiple variant libraries that are not limited to a proximal shell around the active site. In the following sections these techniques are discussed in further detail.
1.4 Computational methods for enhanced of enzyme engineering

1.4.1 Computational structural modelling

Crystal structures provide a great deal of information that can be utilised in the generation of hypotheses on enzyme mechanism and to guide the design of variant libraries for enzyme engineering. Protein crystallisation does however have various limitations- the time required to produce crystals is often limiting; the nature of the crystallisation process renders it impossible to derive reactive structures that only exist transiently; this process also limits crystal structures to static structure solutions, crystal structures often fail to indicate the dynamic nature of a particular protein structure. Computational docking of ligands in protein active sites allows us to address some of the issues that limit crystal structures of proteins and extends the use of structure data. With a protein structure as a starting point, computational docking allows us to generate multiple structures representing the likely conformations of different ligands bound in the active site of the protein in a fraction of the time required to achieve this experimentally. We can also model the structures of intermediate, transient, structures within the protein active site. This would not be possible experimentally. Computational modelling has even been extended to design de novo functional enzymes, Baker et al have utilised computational enzyme design to produce both kemp elimination catalysts [54] and a Diels-Alderase [55].

Computational automated docking involves searching for a conformation of a ligand in an active site that has minimal energy. The energy of the ligand in the context of
the active site is calculated using a molecular mechanics forcefield with parameters for all the different types of interaction that contribute to the bound energy of the ligand. There are two main categories of automated docking methods: matching methods and docking simulation methods. In matching methods, a model of the active site is created and rigid ligands are docked into this model. Dock is a good example of an automated docking algorithm using a matching method. In contrast to matching methods, docking simulation methods involve exploration of flexible ligand translations and orientations until an ideal conformation is found within the protein active site. Docking simulation methods are more computationally intensive than matching methods but this is not a problem unless a large chemical database of lead compounds needs to be screened against an active site. Docking simulation methods allow the docking of a flexible ligand and the use of a more detailed molecular mechanics forcefield which can more accurately calculate the binding energy of the ligand.

Although other programs such as DOCK are available for ligand-protein docking, Autodock is the best known example of a docking simulation method. This program couples a well optimised empirical molecular mechanics forcefield with an efficient search algorithm. These attributes, together with the fact that Autodock is freely available, have led to a good support network for this program and many publications utilising it. Docking simulation aims to identify the minimum energy binding conformation in a huge energy landscape, resulting in a very computationally intensive problem requiring sophisticated search algorithms to reduce the search space to a tractable size. Autodock (Version 3.0.5) uses a
Lamarckian genetic algorithm and an empirical free energy function to find the minimal energy binding conformation

A genetic algorithm is a search technique that utilises the principles of biological evolution to find a solution which exhibits maximum fitness. In the genetic algorithm used by Autodock, the state of the ligand in the context of the protein is defined by a set of state variables which describe the translation, orientation and conformation of the ligand. Each state variable corresponds to a gene in the genetic algorithm. The state of the whole ligand corresponds to the genotype, and the atomic coordinates of the ligand correspond to the phenotype. Finally, the fitness of the ligand state is defined by the total interaction energy of the ligand with the protein.

The genetic algorithm initiates with a random set of genotypes which makes up the population. Individuals with better fitness in the initial population are allowed to reproduce whereas others die. Reproduction involves the mating of random pairs of individuals, during this process crossover takes place with new individuals inheriting genes from either parent. Some of the offspring also undergo random mutation where one gene changes by a random amount.

Autodock 3.0.5 combines a genetic algorithm with a local search method which performs energy minimisation. The genetic algorithm stage is a global search of the energy landscape and allows transitions over energy barriers which may separate energy valleys. The local search method uses the same forcefield as the genetic algorithm to make fine adjustments and find the energetic minimum within the energy trough. The step size of the local search method is adaptive, becoming
smaller in response to a series of consecutive successes. Following a local search the individual can be replaced by the result of the local search. As the local search is carried out at the phenotypic level, this is an example of an inverse mapping function, a genotype can be derived from a given phenotype. The term “Lamarckian genetic algorithm” refers to Jean Baptiste de Lamarck’s belief that phenotypic characteristics acquired during an individual’s lifetime can become inheritable traits [56].

The scoring function used by Autodock to represent fitness is based on an empirical free energy function that can reproduce experimentally derived binding constants of ligands. The energy function consists of five entropic terms which represent Van der Waals forces, H-bonding, electrostatic forces, entropy of the ligand and solvation. Coefficients for each of these terms have been determined using linear regression from a set of protein ligand complexes with known binding constants. Autodock utilises a fast grid based method for energy evaluation in which ligand-protein pairwise interaction energies are precalculated and used as a look up table during the simulation. Summations are performed for all ligand (i) and protein (j) atom pairs as well as all ligand atom pairs three or more bonds apart (Equation 1.1).

\[
\Delta G = \Delta G_{\text{vdw}} + \Delta G_{\text{Hbond}} + \Delta G_{\text{elec}} + \Delta G_{\text{tor}} + \Delta G_{\text{sol}}
\]

\[
\Delta G_{\text{vdw}} = \sum_{i,j} \left( \frac{A_{ij} B_{ij}}{r_{ij}^6} \right)
\]

\[
\Delta G_{\text{Hbond}} = \sum_{i,j} E(t) \left( \frac{C_{ij} D_{ij}}{r_{ij}^12} \frac{r_{ij}^{10} E_{\text{Hbond}}}{r_{ij}} \right)
\]

\[
\Delta G_{\text{elec}} = \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}}
\]

\[
\Delta G_{\text{tor}} = N_{\text{tor}}
\]

Equation 1.1 Energy function utilised by Autodock 3.0.5 calculations of binding energy.
The three terms that describe the interaction energies of atom pairs include a Lennard Jones 12-6 van der waals term, a directional 12-10 H-bonding term in which \( E(t) \) is a directional weight based on the angle between the H-bond donator and the H-bond acceptor atom, and a coulombic electrostatic potential with a distance cut off. Ligand binding is accompanied by unfavourable entropy as the ligands conformational degrees of freedom are reduced. This contribution to the total binding energy is proportional to the number of SP3 bonds in the ligand and is represented by \( N_{tor} \). The grid based method for energy evaluation used by Autodock limits the choice of solvation terms to use as most of these methods are based on surface area calculations. Autodock uses the pairwise volume based method of Stouten et al., in which the percentage of volume around a ligand atom that is occupied by protein atoms is weighted against the atomic solvation parameter of the ligand atom [57]. This gives the desolvation energy contribution from the ligand atom upon binding.

1.4.2 Statistical coupling analysis

There are many examples of epistatic coupling within proteins. Signalling proteins such as GPCRs rely on information transfer between distant residues [58, 59], the exquisite specificity of antibodies generated through B-cell maturation is often determined by residues distant from the antigen binding site [60], co-operative binding of oxygen in haemoglobin is mediated by networks of interacting residues [61-64]. In all of these examples, energy transduction mechanisms have evolved which make possible the highly adapted functions of these varied proteins.
Identification of these interacting networks of residues could enable enhanced engineering of new properties into protein scaffolds.

A new method termed Statistical Coupling Analysis (SCA) has been developed to identify epistatically coupled networks of residues [65]. SCA utilises evolutionary data contained within multiple sequence alignments (MSA’s) to identify co-evolved positions within a protein sequence. The method is based on two hypotheses, firstly that without evolutionary constraint, the amino acids at a specific position in an MSA will approach their mean distribution in all proteins. Secondly, that functional coupling of two positions in a protein should mutually constrain their evolution. If two positions are functionally coupled, alteration of the distribution of amino acids at one site (by the removal of sequences from the MSA) should result in a change in the distribution of amino acids at the other site. Importantly, this does not require that the level of conservation change at the second site, just that the distribution of amino acids be altered [65].

In the first application of the SCA technique, Lockless et al applied the method to the identification of coupled sites within the PDZ domain, a small protein binding motif [65]. PDZ domains can be divided into two classes based on their target sequence specificity. The identity of the residue at position 76 in the PDZ domain is known to be an important determinant of this property. Lockless et al constructed a multiple sequence alignment consisting of 274 eukaryotic PDZ domains, including 4 PDZ domains with known structures. This MSA was then perturbed by removing all the sequences apart from those with a histidine at position 76. In response to this perturbation, the distributions of amino acids at several other positions in the MSA
were found to be altered. These sites are statistically coupled to position 76. Positions identified included both sites in close proximity to position 76 and other surface residues which are involved in sequence recognition. Coupling to proximal cooperative surface sites can be explained by energy propagation through the bound substrate. A third, unexpected, class of residues were also identified by SCA; these sites were a long distance from position 76 and were found in the core or on the opposite surface of the PDZ domain. Although the mechanism and function of this coupling is unknown, pathways of sterically connected, coupled residues were identified that connect position 76 to these distant residues. These pathways may represent routes of signal transduction through the tertiary structure of the protein. Lockless et al went on to verify the coupling interactions identified through thermodynamic mutant cycle analysis. This verification demonstrated good correlation between the statistically coupled sites and the thermodynamically coupled sites including those both proximal and distant from position 76.

Since this early demonstration of statistical coupling analysis, the technique has been applied to several different protein folds including G-protein coupled receptors, haemoglobin and serine proteases [66]. In these examples, Suel et al hypothesised that if networks of coupled residues exist and are conserved, perturbations at positions within the network should redundantly identify each other. In each of the examples above, Suel et al carried out global perturbation analysis and displayed the resulting statistical coupling energies on a matrix with perturbations represented by columns and positions represented by rows. Using
two-dimensional cluster analysis, global patterns of statistical coupling could be identified in the protein folds.

Of particular interest is the analysis of statistical coupling in serine proteases. A multiple sequence alignment was constructed consisting of 616 chymotrypsin serine proteases. Global SCA was carried out involving 69 site specific perturbations. Iterative two-dimensional clustering of the resulting matrix identified two distinct clusters, each containing positions that demonstrate similar patterns of coupling. One of the clusters was found to encompass both the S1 binding pocket and the surface loops known to determine substrate specificity. Residue 172 was also present in this cluster. Although distant from the active site and the binding pocket, SCA was able to identify positions known to determine substrate specificity.

Substrate specificity of the transketolase enzyme has previously been modified using saturation mutagenesis. Improved specific activity towards non-natural substrates such as glycoaldehyde [9] and propionaldehyde [10] has been engineered into transketolase by targeting residues in close proximity to the TPP cofactor. Phylogenetic information was also used to select residues but in both cases no residues more than 10Å from the ThDP cofactor were modified. In the examples above, single point variant libraries were constructed and screened using a colorimetric assay or conventional HPLC. Although hits were identified for different non-natural substrates within these libraries, the screening process limited the number of residues that could be probed and multiple simultaneous mutations could not be assessed. Statistical coupling analysis was able to identify
sites in chymotrypsin that appear to have co-evolved to determine substrate specificity. Identification of such networks in the transketolase enzyme may allow screening to be directed towards areas of sequence space unidentifiable by traditional structure and phylogeny directed selection methods.
1.5 Conclusions

The natural activity of transketolase can be applied to many applications where regio- and stereo-specific carbon-carbon bonds need to be created. This activity is hugely sort after in the synthesis of fine chemicals and pharmaceuticals. In addition to the added specificity afforded by transketolase over more traditional chemical processes, the use of a biocatalyst reduces the need for harsh reaction conditions, organic solvents, and multistep processes. To date, all biotransformations using transketolase have used the wild type molecule, an enzyme which has evolved over millions of years to catalyse two specific reactions in vivo. Through protein engineering, the tools exist to tailor-make transketolase variants with improved properties desirable to the fine chemical and pharmaceutical industries.

Arguably the most powerful method in enzyme engineering is directed evolution. However, limitations in library size impose limits on the sequence space that can be searched using this technique. Effort is being made to reduce these size limitations through advancements in high throughput screening and improved ligation steps. But, as long as there is any limit at all in library size, it will be necessary to target the sequence space to be searched to areas most likely to lead to functional improvements. Various strategies can be adopted to select residues to target; from simple spatial constraints to more complex phylogenetic strategies, such as common ancestor rebuilding. We can also take a rational approach to choose changes that are likely to improve characteristics. Most of these selection methods utilise one branch of knowledge and data. Here we attempt to utilise sophisticated computational techniques to merge information sources, creating a more
sophisticated knowledge base that can be used to direct the creation of more intelligent variant libraries.

Crystal structures of proteins represent one of the richest data sources available to us as enzyme engineers. But, these structures are limited to long-lived, static targets through the nature of crystallisation. Our understanding of the nature of non-covalent interatomic interactions and data from ligand binding experiments allow us to computationally model the energy landscape of substrate-protein interactions, and predict conformations in which substrates are likely to bind. This extends our understanding of function and structure beyond what is possible with X-ray crystallography alone. Computational docking of substrates can identify residues directly involved in substrate interactions as well as rationalising the results from previous library screens.

The second great data source available to protein engineers is the vast, ever growing, collection of sequence data. Using statistical coupling analysis we can delve into this data and discover energetic coupling between sites within proteins. The true power of this technique becomes apparent when the networks of energetic coupling are superimposed onto the three-dimensional structure of the protein. Using the results from computational docking of substrates together with the knowledge of coupled networks of sites in the protein structure we can start to target our variant library very efficiently.

Structural and sequence data represents the culmination of a 3.5 billion year experiment in evolution. It is evolution which ties each of these data sets inextricably together. Random mutation of sequence affects the structure of
proteins, altering the chemical properties of active sites, leading to changes in the function of enzymes. These changes in enzyme function drive evolution. Only when each class of data is viewed in the context of the others does its true potential become apparent. Using the most modern computational methods, we hope to develop rich, combinatorial, information which can help direct the production of variant libraries to those regions which hold the most potential for improved activities.
2 Mechanistic analysis of *Escherichia coli* transketolase by *in silico* docking of substrates in the active site

2.1 Introduction

The carbon-carbon bond forming ability of TK, along with its broad substrate specificity, makes it very attractive as a biocatalyst in industrial organic synthesis [8, 39, 67]. If the ketol donor in the reaction is replaced by hydroxypyruvic acid (HPA) the reaction is rendered irreversible by the release of carbon dioxide. The use of α-hydroxyaldehydes as acceptors together with HPA as the ketol donor allows the creation of enantiomerically pure chiral triols. The potential for producing non-phosphorylated products simplifies their isolation and avoids the requirement to remove phosphate from the product [8, 67]. The TK enzyme from *E. coli* is a preferable biocatalyst to that from yeast due to the higher specific activity of *E. coli* TK towards HPA [68]. Therefore, it is useful to establish that the structural and mechanistic information gained for yeast TK is equally applicable to TK from *E. coli*, for which there is also a crystal structure available [14]. Considerable mechanistic detail has been obtained for *S. cerevisiae* and *E. coli* TK from crystal structures and NMR experiments [7, 18, 21]. Crystal structures have been obtained for the DE4P acceptor substrate bound to the yeast-TK active-site [18], and also for the enamine intermediate formed upon reaction of the yeast enzyme with the DX5P donor substrate [7]. More recently, *E. coli* TK structures were obtained in covalent complexes with DF6P (2R8P.pdb) and DX5P (2R8O.pdb) prior to enamine formation, as well as a non-covalent complex with the cyclic form of DR5P (2R5N) [21].
Despite these impressive structural studies, there is still little information regarding the mode of binding for the donor substrate before it reacts to form the covalent complex and subsequent enamine intermediate, or for the aldehyde acceptor binding in the presence of the enamine intermediate. While crystal structures offer valuable insights into how ligands interact with protein binding-pockets, the binding of enzyme substrates is more challenging as the substrate will usually only bind transiently in the correct conformation before reaction occurs. Notably, for the crystal structure of DE4P bound to the TK active-site [18], the original aim was to solve the structure with DF6P, but the electron density of the resulting structure revealed that the donor substrate had been cleaved into DE4P by TK during the crystal formation process. This exemplifies the difficulties in obtaining structures of substrates bound in active sites.

There is considerable interest in the further development of TK as an efficient biocatalyst, with rational mutagenesis and directed evolution approaches previously having resulted in mutants with altered or improved activity [9], substrate specificity [10, 69] and enantioselectivity [11]. However, further protein engineering to accept an even broader range of substrates would benefit from methods to rationalise the behaviour of existing mutants in structural terms, and to understand how non-natural substrates bind to the active site of this enzyme. Unlike DE4P, many of the non-natural aldehyde acceptor substrates so far examined for biocatalysis with TK, do not contain a phosphate group or an α-hydroxyl group, which are both known to have an important role in substrate recognition [18]. Furthermore, the engineering of TK variants that are less
Computational substrate docking

susceptible to substrate or product inhibition will require a better understanding of the roles played by various residues within the enzyme active-site.

Automated computational docking presents an alternative and complementary means to x-ray crystallography for probing the binding of reactive substrates in short-lived conformations, and also for studying the many non-natural substrates and products, for which crystallography would be time-consuming. Computational automated docking involves searching for the conformation of a ligand bound within in an enzyme active-site that has minimal energy. AutoDock is the best known example of a docking simulation method in which the active site is created and ligands are docked into an enzyme active-site model with an accurate calculation of the binding energy. Flexible ligand translations and orientations are explored until an ideal conformation is found within the protein active site [70]. The observation that very little structural change occurs in the TK active site upon formation of covalent complexes with substrates [21], suggests that it would be an excellent system for docking different substrate complexes without requiring the modelling of amino-acid sidechains for induced fit.

Here we show that automated docking can produce accurate models of substrates bound in the active site of TK. The accuracy of our results is demonstrated by comparison of a computationally derived structure with the crystal structure of DE4P in yeast TK. Further validation is provided by a correlation of experimentally derived Km values for yeast TK, with those calculated from computationally derived docking energies in AutoDock. Having demonstrated the accuracy of the approach we explored the differences and similarities between the binding of DE4P in the
active-sites of *E. coli* TK and yeast TK, and the impact this may have on function. We then examined the binding of natural and non-natural substrates in *E. coli* TK in non-covalently associated complexes that are too reactive to be obtained by crystallography. We also discuss the implications on the potential nucleophilic attack of the deprotonated ThDP cofactor upon the ketol substrate at an unusual Bergi-Dunitz angle, and also the mechanism for ring opening of the cyclic form of D-ribose-5-phosphate. These results will have a significant bearing on attempts to further engineer TK as a biocatalyst for organic synthesis, as well as generating useful hypotheses for future experimental studies to understand the enzyme mechanism of TK.
2.2 Materials and methods

2.2.1 AutoDock 3.0.5

The open source AutoDock software version 3.0.5 was used for all the automated docking reported. AutoDock combines a Lamarckian Genetic algorithm with an empirical free energy function to obtain ligand docked conformations [70]. Substrate docking models were obtained using the *E. coli* TK structure 1qgd.pdb with a cubic grid in the active site of sides 80 Å. Defaults were used for docking each substrate except for the following: the maximum number of energy evaluations was increased to 1 million, the number of genetic algorithm runs was increased from 10 to 200, and the grid spacing used was 0.375 Å. AutoDock performed a cluster analysis to each final conformation obtained from the 200 GA runs such that two conformations with an RMSD less than 0.5 Å are stored in the same cluster. Clusters are output in ranked order of increasing energy following completion of analysis. Manual visual analysis of docked conformations and further analysis of the docked conformations was carried out with Pymol and Ligplot.

2.2.2 Docking of D-erythrose 4-phosphate in yeast TK

D-erythrose 4-phosphate (DE4P) was removed from the yeast TK PDB file 1NGS. AutoDock was used to re-dock the substrate back into the binding site. Grid centre and size used for AutoDock run: (-12.645, 56.02, 19.419) 80 Åx80 Åx80 Å.

2.2.3 Docking of D-erythrose 4-phosphate in *E. coli* TK

DE4P was docked into the binding site of *E. coli* TK (1QGD). Grid centre and size used for AutoDock run: (-10.6, 27.6, 36.4) 80 Åx80 Åx80 Å.
2.2.4 Creation of a model of the ThDP-enamine intermediate in *E. coli* TK

The ThDP-enamine intermediate was docked into *E. coli* TK (1QGD). Grid centre and size for AutoDock run: (-10.0, 28.1, 36.0) 80 Åx80 Åx80 Å.

2.2.5 Docking DE4P and glycolaldehyde in ThDP-enamine complexed forms of yeast and *E. coli* TK

DE4P was docked into the yeast ThDP-enamine-TK complex (1GPU) and in the modelled *E. coli* ThDP-enamine-TK complex. Glycolaldehyde (GA) was docked into the modelled *E. coli* ThDP-enamine-TK complex. Grid centres and sizes were (-6.6, 56.7, 18.4) 60 Åx60 Åx60 Å for DE4P in yeast ThDP-enamine-TK, and (-11.4, 26.3, 36.4) 60 Åx60 Åx60 Å for DE4P and GA in *E. coli* ThDP-enamine-TK.

2.2.6 Docking of natural and non-natural aldehyde substrates into *E. coli* TK

PDB files for the ten TK substrates for which there are published $K_m$ values, and also fluoropyruvate, found to be a potential inhibitor (unpublished data), were generated using the Dundee PRODRG server [71]. Each substrate was docked into the active site of *E. coli* TK. Preliminary docking identified two docking regions within the binding funnel of *E. coli* TK for some of these substrates. Grid sizes and positions were altered to obtain docked conformations for each substrate in the binding region closest to the ThDP cofactor. For some substrates the grid centres were adapted to avoid inaccessible pocket “traps” within the protein. Grid centres and sizes were as follows (grid centres in brackets):
### 2 – Computational substrate docking

Docking energies ($\Delta G$) were converted to a $K_m$ values using $\Delta G = -RT\ln(K_m)$, where $R$ is the gas constant and $T$ is the temperature in Kelvin.

#### 2.2.7 PyMol Molecular Graphics System

All visualisations of docked conformations were produced using PyMol, available from http://www.pymol.org [72].

<table>
<thead>
<tr>
<th>Compound</th>
<th>Coordinates (Å)</th>
<th>Size (Å x Å x Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydroxypyruvate</td>
<td>(-15.991 21.945 37.096)</td>
<td>60 x 60 x 60 Å</td>
</tr>
<tr>
<td>Acetaldehyde</td>
<td>(-18.344 24.016 40.547)</td>
<td>60 x 60 x 60 Å</td>
</tr>
<tr>
<td>D-erythrose 4-phosphate</td>
<td>(-18.344 24.016 40.547)</td>
<td>40 x 40 x 40 Å</td>
</tr>
<tr>
<td>D-erythrose</td>
<td>(-18.344 24.016 40.547)</td>
<td>40 x 40 x 40 Å</td>
</tr>
<tr>
<td>D-glyceraldehyde 3-phosphate</td>
<td>(-10.586 27.153 35.586)</td>
<td>80 x 80 x 80 Å</td>
</tr>
<tr>
<td>D-glyceraldehyde</td>
<td>(-18.344 24.016 40.547)</td>
<td>40 x 40 x 40 Å</td>
</tr>
<tr>
<td>D-ribose 5-phosphate</td>
<td>(-10.586 27.153 35.586)</td>
<td>80 x 80 x 80 Å</td>
</tr>
<tr>
<td>D-ribose</td>
<td>(-10.586 27.153 35.586)</td>
<td>80 x 80 x 80 Å</td>
</tr>
<tr>
<td>Glycolaldehyde</td>
<td>(-15.991 21.945 37.096)</td>
<td>60 x 60 x 60 Å</td>
</tr>
<tr>
<td>Xylulose 5-phosphate</td>
<td>(-10.586 27.153 35.586)</td>
<td>80 x 80 x 80 Å</td>
</tr>
<tr>
<td>Fluoropyruvate</td>
<td>(-15.991 21.945 37.096)</td>
<td>60 x 60 x 60 Å</td>
</tr>
</tbody>
</table>
2.3 Results and discussion

2.3.1 Automated docking of D-erythrose-4-phosphate in the active site of yeast TK

Nilsson et al (1997) previously solved the crystal structure for a substrate protein complex of DE4P bound in the active site of yeast TK [18]. The original aim of their crystallisation was to solve the structure with DF6P but the electron density of the resulting structure would not fit a six-carbon chain and the conclusion was drawn that the donor substrate had been cleaved into DE4P by TK during the long period necessary for crystal formation. This cleavage would also yield the $\alpha,\beta$-dihydroxyethyl thiamine diphosphate intermediate but this would degrade in a few hours into ThDP and glycolaldehyde explaining the lack of an intermediate in the electron density map.

To assess the accuracy and potential of automated computational docking on the substrates of TK we initially docked DE4P into the empty active site of yeast holo-TK, to recreate the substrate-holoenzyme complex. AutoDock accurately predicts the binding conformation of DE4P producing a docked structure within 1.65 Å RMSD of the crystal structure (Figure 2.1). The hydrogen bonding network of the docked substrate is accurately predicted by AutoDock, supporting the role and evolution of this network in determining the stereospecificity of TK. In the crystal structure the C1 aldo carbon atom of the acceptor substrate is positioned 4.16 Å away from the reactive C2 carbon of the thiazolium ring of ThDP. This distance is sufficient to allow the presence of an enamine intermediate in a reactive conformation with the acceptor substrate. The computationally docked
conformation positioned the C1 aldo carbon 4.81 Å away from the thiazolium ring of ThDP. Although AutoDock docked DE4P slightly further away from the ThDP reactive centre, it was still within an acceptable distance to react in the presence of the intermediate.

The main discrepancy between the experimentally determined structure and that created by computational docking was the position of the phosphate group. In the computational docking the phosphate group of DE4P was pulled closer to the side chains of the Arg 358, Arg 520, Ser 385 and His 461 residues (E. coli numbering) at the entrance to the binding funnel (by just over 1 Å). This has the effect of pulling DE4P slightly out of the funnel away from the ThDP cofactor (Figure 2.1). The interaction between the phosphate group and the positively charged arginine residues of yeast TK is strongly affected by the electrostatic interactions between these groups. The force field of AutoDock includes a term to model electrostatic interactions but the discrepancy between the modelled and the crystal structures may be explained by errors in the electrostatic term of the forcefield. While all of the thirty ligand-protein complexes used to calibrate the AutoDock forcefield included H-bond interactions, only a small proportion involved electrostatic interactions of explicitly charged groups [70]. Electrostatic interactions between these groups would also be strongly influenced by pH which may further explain the slight discrepancy. However, the TK structure itself was resolved to 2.4 Å and so a 1 Å shift is acceptable within error. Despite the possible error in the modelling of phosphate binding, the results show that AutoDock is capable of reproducing accurate docked conformations of substrates in the active site of TK.
2.3.2 Automated docking of D-erythrose-4-phosphate in the active site of *E. coli* TK

The active sites of yeast and *E. coli* TK, including the orientation of conserved residues, are nearly identical. For *E. coli* TK, structures are available in covalent complexes with DF6P (2R8P.pdb) and DX5P (2R8O.pdb), as well as a non-covalent complex with the cyclic form of DR5P (2R5N) [21]. However, no crystal structure exists for the non-covalent complex of *E. coli* TK with DE4P. If *E. coli* TK binds DE4P in the same conformation as for yeast TK then mechanistic insights derived from studies of yeast TK could be used with confidence to infer equivalent mechanistic details in the *E. coli* protein. This would be of great value as much of the work on the function of TK to date has been carried out on the yeast enzyme.
DE4P was docked into the active site of *E. coli* TK (1QGD) using AutoDock. Many of the docked conformations predicted by AutoDock involved an inversion of the DE4P, with the phosphate group oriented towards the ThDP and the carbonyl active centre of DE4P pointing out of the active site in an un-reactive conformation. This could be explained by the error in the handling of phosphate interactions. Like the arginine residues at the entrance to the active site normally involved in phosphate binding, the ThDP molecule carries an explicit positive charge. Hypothetically, the DE4P may actually be able to bind in this orientation and form an un-reactive inhibitory complex. If this is not an artefact of computational docking it may therefore have implications for substrate inhibition of TK at high concentrations.

AutoDock also docked DE4P in the reactive orientation in *E. coli* TK (Figure 2.1). This conformation is very similar to that of DE4P bound in yeast TK (compared in Figure 2.1). The phosphate group is positioned close to the entrance of the active site and interacts with residues Arg 358, Arg 520, Ser 385 and His 461. These interactions are equivalent to those maintaining the position of the DE4P phosphate group in yeast TK. The carbon chain of DE4P extends down the active site of *E. coli* TK forming a hydrogen-bonding network with the side chains of several conserved residues. The C3 hydroxyl group forms an interaction with the backbone oxygen of Gly 262 that is not seen in the yeast TK bound structure. Like yeast TK, the conserved Asp 469 residue of the *E. coli* enzyme interacts with the C2 hydroxyl group of DE4P but an additional interaction is formed through His 26. The C1 aldo oxygen atom interacts with residues His 261 and His 26 in *E. coli* as
observed in yeast TK, to position the C1 aldo carbon at 4.89 Å away from the C2 atom of the ThDP thiazolium ring.

The general binding conformation for DE4P in *E. coli* TK is the same as that for yeast TK. The minor differences in the hydrogen bonding network do not change the favoured stereospecificity of the recognition and these interactions could be transiently present in the yeast structure with only a small degree of dynamic movement. The conformation of DE4P docked in *E. coli* TK supports the hypothesis that the conserved residues of the TK active sites of *E. coli* and yeast TK have the same roles in substrate binding.

2.3.3 Modelling the enamine intermediate in *E. coli* TK and docking of D-erythrose-4-phosphate into the yeast and *E. coli* TK-ThDP-enamine complexes

We have demonstrated the ability of AutoDock to accurately model the binding conformation of DE4P in yeast TK and shown that the same binding conformation is formed in *E. coli* TK. However, in the TK catalysed reaction, DE4P cannot bind productively until an enamine intermediate has first been formed between the ThDP cofactor and the ketol donor substrate. It is possible that the binding conformation of DE4P in the TK active site is different in the presence of this intermediate. Due to the reactive nature of the intermediate, DE4P would not bind and exist in the presence of the enamine intermediate for long enough to obtain crystals for structural analysis. Currently, the only way to solve the structure for DE4P bound in the presence of ThDP-enamine intermediate is to model this structure computationally. In the yeast-TK crystal structure of the ThDP-enamine
intermediate there were no significant variations in the orientations or positions of active site residue side chains relative to the ThDP bound structure. The ThDP-enamine intermediate could therefore be confidently docked into the known *E. coli* holo-TK structure using AutoDock to obtain a model of the *E. coli* TK-ThDP-enamine complex.

As seen in Figure 2.2, the complex obtained for *E. coli* TK-ThDP-enamine was nearly identical to the solved structure of the yeast TK-ThDP-enamine complex (1GPU) [7]. All the major functional interactions were present in the modelled complex, including the Glu 411 (Glu 418 in yeast TK) interaction with N1' of the pyrimidine ring [73]. The ketol donor derived enamine intermediate is co-ordinated by hydrogen bonds to the conserved *E. coli* histidine residues His 100 and His 473. The only difference between the TK-enamine interactions of yeast and TK is that in the yeast-TK complex His 481 (His 473 in *E. coli*) interacts with both hydroxyl groups of the enamine whereas in the *E. coli* complex His 473 only interacts with the \( \alpha \)-hydroxyl group. Some other minor differences are present in the hydrogen-bonding network of the di-phosphate group but none of these differences significantly alter the position of ThDP-enamine relative to the TK molecule (Figure 2.2).
Figure 2.2 Comparison of the ThDP-enamine intermediate in yeast and *E. coli* transketolases. The ThDP-enamine intermediate in yeast TK (blue sticks) is from an available structure (1GPU), whereas that for *E. coli* TK was obtained by docking (green sticks) in 1QGD. The bound calcium ion is shown as a green sphere.

Following the creation of the *E. coli* TK-ThDP-enamine complex model, AutoDock was used to obtain docked conformations of DE4P in both the yeast (crystal structure) and the *E. coli* (modelled) TK-ThDP-enamine complexes. In each case the DE4P molecule docked in the same conformations as previously observed in Figure 2.1, in the absence of the enamine intermediate. The hydrogen bonding interactions between the TK and DE4P observed in the absence of the enamine are also preserved along with the few differences between yeast and *E. coli* TK described above. The DE4P carbonyl C1-atom is placed within 3.44 Å and 4.28 Å of the enamine α-carbon in the yeast TK and *E. coli* TK models respectively. As well as positioning the carbonyl group of DE4P in close proximity to the α-carbon of the enamine, the hydrogen-bonding network described orientates the carbonyl group
for the formation of the S-enantiomer product upon carbon-carbon bond formation as expected for TK (see Figure 2.4 a) [74].

### 2.3.4 Docking of other substrates in E. coli holo-TK

AutoDock has proven accuracy in predicting the binding affinities of ligand-protein complexes. Published kinetic data including $K_m$ values are available for several substrates in the reaction catalysed by *E. coli* TK [1]. Ten substrates, namely glycolaldehyde (GA), $\beta$-hydroxypyruvate (HPA), D-xylulose-5-phosphate (DX5P), D-ribose-5-phosphate (DR5P), D-glyceraldehyde-3-phosphate (DG3P), D-erythrose-4-phosphate (DE4P), acetaldehyde (AA), D-ribose (DR), D-glyceraldehyde (DG), D-erythrose (DE), were docked in the binding site of *E. coli* TK and their $K_m$ values were calculated from the binding affinities reported by AutoDock.

Initial docking using a grid that encompassed the entire binding funnel of *E. coli* TK revealed two regions in which substrates could bind in the active site of the enzyme. Following the identification of these separate binding sites, grid sizes and positions were altered to remove the non-productive and higher energy binding region, which was in the active-site funnel but further out from the ThDP molecule and the active centre of the enzyme. Such a site which involved mainly protein backbone interactions may prove to be physiologically relevant, for example resulting in substrate or product inhibition, but was not studied further here.

The calculated $K_m$ values for the docked substrates were compared against the experimentally determined values previously published [1] as shown in Figure 2.3. The calculated $K_m$ values we obtain from the binding energies of substrates in the
active site of TK were obtained only using the acyclic and monomeric forms of substrates. However, the values of $K_m$ can be affected in some cases by the equilibrium between dimeric, monomeric, cyclic and linear forms which decreases the availability of the reactive substrate. DR5P, DR and DE can all form ring structures and the proportion of acyclic monosaccharides in aqueous solution have been determined respectively as 0.6% [75], 0.05% [76] and 12.1% [77]. These equilibria were used to adjust the experimental Km data obtained by Sprenger and co-workers as presented in Table 2.1 and Figure 2.3. A recent crystal structure of *E. coli* TK bound to DR5P has shown that the substrate can bind directly in the cyclic form. Therefore, the adjustment of Km values above assumed that the ring opening of the sugars is not rate limiting, and therefore does not affect the observed $K_m$. Other substrates such as glycolaldehyde and D-glyceraldehyde form dimers at high concentrations but in dilute aqueous solutions they become essentially monomeric and so the experimental data for these substrates were not adjusted [77]. The calculated and experimental log($K_m$) values correlate well with a Pearson $R^2$ value of 0.82 indicating that the AutoDock algorithm is capable of predicting the conformations and binding energies of a broad range of mostly natural substrates. In all cases except DR5P, the calculated Km values are lower than the experimentally determined values indicating a systematic underestimation of Km values by AutoDock.
Figure 2.3 Comparison of log($K_m$) values obtained experimentally and calculated from the docking energies reported by AutoDock for ten substrates of TK. All values were taken from Sprenger et al [1], except the values for HPA and GA which were averaged with those from Hibbert et al [9]. Values for D-ribose, D-ribose-5-P and erythrose, were adjusted by experimentally determined equilibrium constants for the fraction of the acyclic forms [75-77].

<table>
<thead>
<tr>
<th>Substrate</th>
<th>$K_m$ (calc)</th>
<th>$K_m$ (exp) a</th>
<th>Distance to nucleophile g</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-Ribose</td>
<td>0.47</td>
<td>0.7 b</td>
<td>2.64</td>
</tr>
<tr>
<td>D-Ribose-5-phosphate</td>
<td>0.05</td>
<td>0.0084 c</td>
<td>3.13</td>
</tr>
<tr>
<td>D-Erythrose</td>
<td>0.67</td>
<td>18.15 d</td>
<td>3.93</td>
</tr>
<tr>
<td>D-Erythrose-4-phosphate</td>
<td>0.12</td>
<td>0.09</td>
<td>4.28</td>
</tr>
<tr>
<td>D-Glyceraldehyde</td>
<td>2.10</td>
<td>10 e</td>
<td>2.33</td>
</tr>
<tr>
<td>D-Glyceraldehyde-3-phosphate</td>
<td>0.23</td>
<td>2.1 e</td>
<td>5.01</td>
</tr>
<tr>
<td>Glycolaldehyde</td>
<td>9.14</td>
<td>24.5 f</td>
<td>2.66</td>
</tr>
<tr>
<td>Acetaldehyde</td>
<td>33.6</td>
<td>1200</td>
<td>2.69</td>
</tr>
<tr>
<td>β-Hydroxybutyrate</td>
<td>3.61</td>
<td>11.65 f</td>
<td>3.14</td>
</tr>
<tr>
<td>D-Xylosose-5-phosphate</td>
<td>0.022</td>
<td>0.16</td>
<td>4.73</td>
</tr>
</tbody>
</table>

Table 2.1 Comparison of *E. coli* TK $K_m$-values obtained by docking with those from experiment. a Experimental data obtained by Sprenger et al [1], adjusted by equilibrium for 0.05% acyclic form [76]. b adjusted by equilibrium for 0.6% acyclic form [75]. c adjusted by equilibrium for 12.1% acyclic form [77]. d Experimental data obtained with D,L racemate. e average of independent values from Sprenger et al [1] (glycolaldehyde 14 mM, β-Hydroxybutyrate 18 mM) and Hibbert et al [9] (glycolaldehyde 35 mM, β-Hydroxybutyrate 5.3 mM). g distances measured between either a) aldehyde carbon atom of aldot acceptor and the ThDP-enamine nucleophile, or b) carbonyl carbon atom of ketol donor and the ThDP thiazolium C2 nucleophile.

Examination of the structures of each substrate bound in the active site of *E. coli* TK reveals the same general binding conformation as seen for DE4P described above (Figure 2.4 a-b and Figure 2.5 a). Notably, all phosphorylated substrates docked with the phosphate group interacting with the conserved Arg 358, Arg 520, Ser 385 and His 461 residues at the entrance of the binding funnel as expected. This
conformation orientates the substrates so that their active centres are positioned within the binding funnel in an active conformation, but also creates small differences in their proximity to the ThDP/enamine depending on chain length which plays a role in substrate preference. The distance from the aldehyde carbon atom to the enamine nucleophile for phosphorylated substrates decreases progressively as the chain length increases from the three-carbon DG3P, and four-carbon DE4P, to the preferred five-carbon DR5P which has the shortest distance of 3.13 Å (Table 2.1). The phosphorylated substrates had generally lower experimental and predicted $K_m$ values (Table 2.1) which can be explained by this strong interaction at the entrance to the binding cleft. When overlaid with the docked enamine structure (Figure 2.4 a) the phosphorylated aldol acceptors are all positioned and oriented with the Re-face of their aldehydic carbonyls ready for nucleophilic attack by the nearby $\alpha$-carbon of the enamine to give products of the $S$-enantiomer.

Docking of the non-phosphorylated aldol acceptors was found to be dominated by hydrogen-bonding to His 26, His 261 and, with the exception of acetaldehyde, Asp 469 (Figure 2.4 b). Interestingly, D-ribose forms a hydrogen bond to Arg 520 via the C5-hydroxyl group in the absence of the phosphate, though this does not fully compensate for the much tighter binding achieved with the phosphate in DR5P. As for the phosphorylated aldehydes, an overlaid structure of the docked enamine (Figure 2.4 b) shows that the non-phosphorylated aldol acceptors are also all positioned and oriented with the Re-face of the aldehyde carbonyl prone to
nucleophilic attack by the nearby $\alpha$-carbon of the enamine to give the $S$-enantiomer products.

![Figure 2.4](image-url)

Figure 2.4 Comparison of the docking of phosphorylated and non-phosphorylated aldol acceptor and ketol donor substrates in *E. coli* TK. The enamine intermediate (not present during docking) is overlaid for reference in each panel (green sticks). a) Phosphorylated aldol acceptor substrates. DE4P (grey), G3P (dark blue) and DR5P (orange) are each oriented to present the $Re$ face of their *aldo* carbonyls to give the expected $S$-enantiomer products. The phosphate binding location is highly defined in the R520 and R358 pocket, and creates differences in the proximity of the *aldo* carbonyl to the $Ca$ enamine carbon depending on substrate chain length. b) Non-phosphorylated aldol acceptor substrates. Acetaldehyde (grey), Glycolaldehyde (magenta), and Ribose (yellow) are also oriented to present the $Re$ face of their *aldo* carbonyls to give the expected $S$-enantiomer products. Non-phosphorylated substrate binding is dominated by hydrogen-bonding to H26, H261, and D469 (except acetaldehyde and ribose which hydrogen-bonds to R520 even without the phosphate).
2.3.5 Comparison of docked and crystallized DR5P conformations

Asztalos and co-workers previously obtained a structure of E. coli TK with a mixture of both the cyclic and acyclic forms of DR5P bound in the active site [21]. However, the apparently ambiguous electron density due to low occupancy of the acyclic form, positioned it such that it would clash sterically with our modelled enamine intermediate, indicating that this might not be the exact productive binding conformation. While modelling of the yeast TK enamine structure coordinates by Asztalos et al found the acyclic DR5P to approach the enamine Cα at a distance of 1.6 Å with a favourable angle of 112° for the C1 aldo carbon relative to the C2-Cα of the enamine, inspection of their structure indicates that the enamine would be poised to attack the incorrect Se face of the C1 aldo carbonyl. Our docked acyclic DR5P gives a more plausible (or earlier) conformation for binding prior to a productive reaction in which the reactive enamine is 3.13 Å from the Re face of the aldehyde carbon atom of DR5P. The aldehyde O-atom is coordinated by His 26 (3.0 Å) and His 261 (2.8 Å), the C2 hydroxyl hydrogen bonds with Ser 385 (3.0 Å), the C3 (2.6 Å) and C4 (2.8 Å) hydroxyl groups both hydrogen bond with Asp 469, and the phosphate is coordinated to Arg 358, Arg 520, Ser 385 and His 461 as usual. The contacts observed in the binding of the cyclic form by Asztalos et al are different to those in our modelled acyclic form with the C1 to C3 hydroxyl groups being rotated around the protein side-chains by one position (Figure 2.5 b-c). Comparison of the two suggests that TK binds the cyclic DR5P, catalyses the ring opening, and then re-arranges the position of the aldehyde closer to the enamine prior to reaction.
His 261 is highly polarised by the phosphate in ThDP (2.7 Å) making it a good base for catalysis of the ring opening via deprotonation of the C1-hydroxyl group in DR5P which is 3.33 Å away. Ser 385 was also found in two occupancies with the first coordinating the DR5P phosphate [21], and the second in position to stabilise the protonation of the furanose O-atom of DR5P during ring opening. As the aldehyde forms at C1 upon ring opening, the hydrogen bond to His 261 is retained via the O-atom which moves by 3.2 Å into the position previously occupied by the C2 hydroxyl group, and forms an additional hydrogen bond with His 26. This movement is facilitated by a 180° rotation of the C2-C3 bond, as suggested by Asztalos et al, but forms a new hydrogen bond between the C2 hydroxyl group and Ser 385. Meanwhile the hydrogen bond between Asp 469 and the C3 hydroxyl group becomes shorter by 0.15 Å and the C3-C4 bond rotates by 180° to form a new hydrogen bond between the C4 hydroxyl group and Asp 469. Collectively these hydrogen bond exchanges serve to unwind the now acyclic DR5P backbone and free up the space near His 26 and the enamine-ThDP intermediate for the newly formed C1 aldehyde.
Figure 2.5 a) Cyclic and acyclic forms of DR5P. Cyclic DR5P (yellow) from a TK crystal structure [21] and the modeled acyclic DR5P (orange). b) and c) Hydrogen bonding network changes upon ring opening of DR5P. Cyclic (left) and acyclic (right) forms of DR5P with hydrogen bond interactions with transketolase shown schematically as dashed lines.

2.3.6 Glycolaldehyde docking

Glycolaldehyde (GA) is one of the smallest aldehydes accepted by TK and the reaction of GA with TK has been studied extensively [1, 26, 41, 74, 78]. As for DE4P, the docking of GA was performed both in the presence and absence of the modelled enamine. Comparison of the two docks indicated that they differed by an RMSD of only 0.53 Å and that their hydrogen bonding interactions were identical.
As with the other aldehydes, GA binding presented the Re face of the aldehyde to the nucleophilic enamine such that it would give the expected L-erythulose product (Figure 2.6 a). Overall the very small differences, for both GA and DE4P docked with and without the enamine, indicate that the productive positioning of the aldehyde acceptors relative to the enamine is sufficiently optimised through specific interactions other than those to the enamine in the active site. This observation is analogous to that by Asztalos et al where the active site is catalytically poised to maximally translate substrate binding interactions directly into the strained tetrahedral intermediate with the ketol donor. It is possible then that the reaction between aldol acceptor and the enamine may introduce a similarly strained conformation, driven by optimised substrate interactions without the need for induced fit or the unfavourable flexibility this introduces.
2.3.7 Comparison of docking for the ketol donors

Hydroxypyruvate (HPA) and D-xylulose-5-phosphate (DX5P) are both donor substrates for the TK catalysed reaction for which no structures of the non-covalent complexes have been obtained experimentally. Fluoropyruvate (FPA) is an analogue of HPA for which no measurable reaction is observed with TK. FPA and
HPA are both known competitive inhibitors of the pyruvate decarboxylase from *Zymomonas mobilis* [79] and preliminary data from our lab suggest that FPA can also inhibit the TK reaction when using HPA as the ketol donor (unpublished). Both of the ketol donor substrates and also FPA were found to dock in conformations that position the carbonyl within a reactive distance from the C2 carbon of the ThDP thiazolium ring (Figure 2.6 b). However, the DX5P docked conformation and position was found to be unusually different to all other substrates and the carbonyl was at an unfavourable orientation for reaction with the ThDP thiazolium ring (not shown). This dock was deemed to be suspect upon visual inspection, and indeed removal of DX5P from the plot of calculated and experimental log(Km) values in Figure 2.3 gave a slightly improved Pearson R2 correlation of 0.84 (not shown).

To examine the mechanism of binding for ketol donors, and their reaction with ThDP, the docking of HPA was compared to the covalently bound DX5P intermediate from an available crystal structure with *E. coli* TK [21], and also to the modelled ThDP-enamine intermediate (Figure 2.6 b). The most striking observation is that the reaction mechanism for HPA binding and subsequent covalent bond formation with ThDP (as observed by analogy in the DX5P complex), followed by the final enamine intermediate formation, involves a stepwise shortening of the emergent covalent bond between the C2 carbanion of ThDP and the substrate carbonyl, accompanied by an almost 90° rotation of the molecule. Throughout this reaction, the carbonyl O-atom and α-hydroxyl group in HPA remain in the same position, both maintaining hydrogen-bond interactions to His 473 and the amino
group on the pyrimidine ring of ThDP, and additionally between the carbonyl and His 100.

Hydrogen bonds from His 26 and His 261 to just one O-atom of the carboxyl group in HPA mirror those observed previously by crystallography in the covalently associated DX5P complex [21], where the two histidines interact with the C3 hydroxyl group (Figure 1.6). Asp 469 hydrogen bonds to the other carboxyl group O-atom in HPA which also mirrors an interaction with the C4 hydroxyl group in DX5P. However these interactions are all lost upon enamine formation by cleavage of the two-carbon ketol unit from HPA and DX5P.

Surprisingly, the Bergi-Dunitz angle for nucleophilic attack of the carbonyl of HPA by the ThDP carbanion is only 68°, which is a long way from the traditionally preferred 107° [80]. Such an unusual acute angle of attack is potentially made possible through a number of factors. First, the substrate is tightly anchored by interactions to the amino group on the pyrimidine ring of ThDP, His 100 and His 473, whereas Bergi-Dunitz angles of 107° are typically preferred in reactions between two unconstrained molecules. Second, the ThDP may attack in the carbene form, especially in the proximity of the negatively charged carboxyl group in HPA, which can take place from above a pi-system such as in a carbonyl. Finally, the conserved Asp 469 residue appears to be hydrogen bonded to one of the carboxyl group O-atoms in HPA (equivalent to the C1 hydroxyl group of DX5P). This may serve to lengthen the carbonyl bond in HPA giving it less double bond character, along with protonation of the carbonyl by the amino group on the pyrimidine ring of ThDP to
remove electron density from the O-atom, thus making an acute angle of attack more favourable.

FPA was found to dock in an identical manner to HPA with an RMSD of 0.75 Å suggesting that the fluorine atom can hydrogen bond with equivalent contacts to those found with HPA. FPA is less reactive than HPA as the high electronegativity of the fluorine atom decreases the polarity of the neighbouring carbonyl making it less reactive to the ThDP carbanion in TK. The identical binding of docked FPA and HPA indicates that FPA should provide a great opportunity for crystallographers to obtain a non-covalent complex of TK with an analogue of the ketol donor HPA.

The initial formation of a non-covalent ketol donor complex at an unusual angle of attack, through maximised substrate interactions, is consistent with the previous observation of strain in the subsequently formed tetrahedral intermediate [21]. This builds up a picture of the enzyme active site being poised catalytically such that the binding energetics allow the unusually constrained attack angle. The strain in the complex is released partially to form the out-of-plane covalent intermediate, and then released further upon formation of the enamine.
2.4 **Conclusions**

Although computational docking is not completely accurate, the most obvious errors can be eliminated by visual inspection as was the case for DX5P in this work. Furthermore, the lack of significant side-chain movements in the TK active site upon substrate binding eliminates errors that might otherwise have arisen by not modelling induced fit mechanisms. Computational docking in TK provided supportive evidence for understanding the enzyme mechanism where experimental structural studies have been difficult, and more importantly it has generated interesting hypotheses that can be tested in future experimental studies. The AutoDock algorithm used was able to determine binding conformations of several ketol donor and aldehyde acceptor substrates in the active site of *E. coli* TK with a good correlation between experimental and computationally derived $K_m$ values. We were also able to model aldol acceptors into TK with the ThDP-enamine intermediate present as would be formed after reaction with the ketol donor substrate. The binding of aldol acceptors was found to be identical both in the presence and absence of the enamine intermediate indicating that the interactions made independently of the enamine are sufficient for binding in a reactive conformation. This has an important consequence for protein engineering attempts to alter the substrate specificity of a two-substrate reaction in which the interactions of the second substrate with the enzyme are partially overlapping with those made by the first.

An interesting potential mechanism for the ring opening of cyclic DR5P is suggested by comparison of a previous crystal structure to our docked acyclic DR5P. We have
identified potential general acid and general base residues for the ring opening, and also suggest how the molecule moved around a network of hydrogen bonding interactions to bring the linearised DR5P into place for reaction with the enamine intermediate.

Finally, a comparison of HPA docking to the modelled enamine and to a structurally determined covalent DX5P intermediate indicates a nucleophilic attack by the deprotonated ThDP cofactor with an acute Bergi-Dunitz angle of just 68°, rather than the typically preferred obtuse angle of 107°. A repeat of the HPA docking conformation but with that of the non-reactive analogue FPA suggests a possible method for testing this unusual hypothesis by crystallography.
3 Statistical Coupling Analysis of *Escherichia coli* transketolase

3.1 Introduction

Elucidation of the structure and mechanism of transketolase has led to the identification of several key residues involved in substrate recognition [18, 21], protonation of cofactor [23], transition state stabilisation [7], and deprotonation of substrate. Efforts to engineer the substrate specificity and enantioselectivity of TK have both utilised and built upon this knowledge of structure and mechanism.

Despite the detailed understanding of TK, work to date has focussed on the function of individual residues in isolation. Early experimental mechanistic work on yeast TK involved the site directed mutagenesis of individual residues [18, 19, 24], and engineering of *E. coli* TK has focussed on saturation mutagenesis targeted to individual residues [9-11, 47]. In addition to identifying individual residues with key functions, it is also desirable to identify key synergistic networks of residues. Identification of such networks will lead to a better understanding of function and aid in the engineering of TK.

Statistical coupling analysis (SCA) is a powerful tool for identifying co-evolved residues in protein multiple sequence alignments (MSA) [65, 66]. The co-evolution of residues indicates potential synergy between them that is driven by overall protein fitness or function. Various properties can impact on protein function; these include expression, folding, solubility, stability and allostery; synergistic relationships could be linked to any one of these properties. Here we have used the
SCA method to identify potentially synergistic networks of residues in *E. coli* TK. Other methods to measure co-evolution of residues include Explicit Likelihood of Subset Variation (ELSC), mutual information and correlation-based methods.

Statistical coupling analysis is based on two hypotheses, firstly that without evolutionary constraint, all amino acids at a specific position in a multiple sequence alignment will approach their mean distribution in all proteins. Secondly, that the epistatic coupling of two positions in a protein should mutually constrain their evolution [65]. In order to identify such coupled positions in a multiple sequence alignment, the distribution of amino acids at one site is altered (by removing a subset of sequences from the MSA) and other sites are monitored for concurrent changes in distribution of amino acids. Notably, identification of such relationships does not require that the level of conservation change at the second site, just that the distribution of amino acids be altered.

TK exists as a homodimer of two 680-residue chains, each consisting of three domains, the PP (pyrophosphatase binding), the Pyr (pyrimidine binding), and the C-terminal domain. All TPP-dependent enzymes contain catalytic PP and Pyr domains which bind the ThDP cofactor, although different enzyme types have different domain architecture [81, 82]. Although the TPP-binding sites of these enzymes are very similar, each enzyme type has different substrate specificity. As the different enzyme types have diverged from a common ancestor, any synergistic networks of residues could be expected to create evolutionary constraints on the residues involved. Identification of such co-evolved networks could provide important
insights into substrate specificity and aid in the engineering of substrate specificity of *E. coli* TK.

The wild type *E. coli* TK enzyme has been engineered for improved substrate specificity towards the non-hydroxylated aldehyde acceptor substrate, propionaldehyde (PA) [10]. The D469T mutant was identified as giving the greatest improvement in activity on PA with a 5-fold increase in specific activity relative to wild type. A further mutant at site 469, D469Y, displayed the greatest substrate specificity with a 64-fold higher activity on PA relative to GA. Finally the D469E mutant displayed the greatest improvement in enantioselectivity with PA [10]. The repeated identification of Asp 469 variants among seventeen active site residues probed for three different selection criteria illustrates the importance of this residue. Identification of networks of synergistic residues involving Asp 469 could lead to further improvements in the engineering of *E. coli* TK.

Using SCA, we have identified residues that are statistically coupled to Asp 469. These include both a proximal connected network and a more distal and network distributed throughout the protein fold. Some of the residues had been previously identified as improving activity on non-natural substrates [9, 10] but the vast majority have not been investigated by mutagenesis before and represent new potential targets for rational design of expanded variant libraries.

In addition to measuring coupling between individual residues in an alignment, SCA allows global coupling analysis across all positions in the alignment that meet certain criteria. Two-dimensional hierarchical clustering, originally developed for microarray data analysis, can be used to identify networks of similarly coupled
residues throughout the protein sequence. For robust and conserved networks of residues, perturbations at network positions are expected to redundantly identify each other, clustering the coupling matrix allows the identification of such networks. We applied this method to identify inter- and intra-domain networks of coupled residues throughout the PP and Pyr domains of *E. coli* TK. Many of these networks were located in the interface between the two domains but we also identified interesting connected networks of residues in and around the active site.
3.2 Materials and Methods

3.2.1 Multiple sequence alignments

Costelloe et al carried out a phylogenetic analysis of seventeen different TPP dependant enzymes using sequence alignments of the conserved PP and PYR domains [82]. Transketolase (TK), D-xylulose-5-phosphate synthase (DXPS), dihydroxyacetone synthase (DHAS), phosphoketolase (PKL), 2-oxoisovalerate dehydrogenase (2OXO), pyruvate ferredoxin reductase (PFRD), pyruvate decarboxylase (PDC), indolepyruvate decarboxylase (IPDC), phenylpyruvate decarboxylase (PhPDC), pyruvate oxidase (PO), acetolactone synthase (ALS), glyoxylate carboligase (GXC), benzoylformate decarboxylase (BFDC), benzaldehyde lyase (BAL), oxalyl CoA decarboxylase (OCADC), sulfopyruvate decarboxylase (SPDC), and phosphopyruvate decarboxylase (PPDC) enzymes were included.

Costelloe et al identified TPP dependent enzyme sequences using BLASTP searches of the Swissprot and nr databases (Blosum62 matrix and default settings). Query sequences for each enzyme were selected from those with known structure or well defined biochemistry and homologous sequences were identified with >30% sequence identity. In total 382 sequences representing the 17 different enzymes were identified. Following removal of putative sequences these hits were aligned using ClustalW.

Alignments of the PP and PYR domains were generated separately. The PP domain was defined as residues 1-350 in E. coli TK and the PYR domain was defined as residues 323-528 in EcoTK. Crystal structures of EcoTK (1QGD.pdb), P. putida 2OXO
(2BP7.pdb), *D. africanus* PFRD (1B0P.pdb), *S. cerevisiae* PDC (1PVD.pdb), and *L. plantarum* PO (1POX.pdb) were used to refine the alignment of functionally important residues and secondary structural elements. Finally, alignments were degapped to leave only residues found in *E. coli* TK.

### 3.2.2 Statistical coupling energy calculation

SCA Version 1.5 was used to calculate statistical coupling energies between sites in multiple sequence alignments (MSA’s) of the PP and PYR domains. SCA Version 1.5 is a MATLAB implementation of the calculation described by Lockless et al., 1999 [65]. The calculation gives a quantitative measure of the change in amino acid distribution at site *j* given a perturbation at another position *i*. This is calculated as a statistical coupling energy ($\Delta G_{\text{stat}}^{i,j}$).

Briefly, each site in the MSA is described by a 20 element vector of binomial probabilities of individual amino acid frequencies given their frequencies in all proteins ($P_j = [P_{\text{ala}}^j, P_{\text{cys}}^j, P_{\text{asp}}^j, ..., P_{\text{tyr}}^j]$). $P_j^x$ gives the probability of the observed number of *x* amino acids at position *j*, given its mean frequency in all proteins. The 20 element vector of $P_j^x$ can then be converted into a vector of statistical energies that represents the evolutionary constraint at site *j* ($\Delta G_{\text{stat}}^j = [\Delta G_{\text{ala}}^j, \Delta G_{\text{cys}}^j, \Delta G_{\text{asp}}^j, ..., \Delta G_{\text{tyr}}^j]$). Each term in the vector is the value for amino acid *x* at site *j* and is given by $\Delta G_{j}^{x} = kT^* \ln(P_j^x/P_{\text{MSA}}^x)$, where $kT^*$ is an arbitrary energy unit and $P_{\text{MSA}}^x$ represents a hypothetical site where all amino acids are observed at their mean frequencies in the MSA as a whole. This hypothetical site serves as a reference state for all sites and $\Delta G_{j}^{x}$ represents the statistical free energy separating site *j* from the
hypothetical site for amino acid x by the Boltzmann distribution. The magnitude of
the $\Delta G_{\text{stat}}^j$ vector represents an evolutionary conservation parameter for site j.

To measure functional coupling between two positions j and i in the MSA, two
statistical energy vectors are calculated, one from the full MSA ($\Delta G_{\text{stat}}^j$) and one
from a subalignment representing a perturbation of the amino acid frequencies at a
second site i ($\Delta G_{\text{stat}}^i|_j$). The magnitude of the difference in these two statistical
energy vectors represents a quantitative measure of the degree to which the
probability of individual amino acids at site j is dependent on the perturbation at i,
$\Delta \Delta G_{\text{stat}}^{j,I} = \Delta G_{\text{stat}}^j - \Delta G_{\text{stat}}^i|_j$. $\Delta \Delta G_{\text{stat}}^{j,I}$ is calculated for all sites j given a perturbation at
position i.

3.2.3 Matrix assembly and cluster analysis

All acceptable perturbations, and their associated coupling energies to other
positions in the sequence, are displayed as a matrix of statistical coupling energies.
Each row in the matrix represents a position in the alignment and each column
represents a specific perturbation.

Iterative clustering methods originally developed for microarray analysis are then
used to identify co-evolving networks of positions. The process of iterative
clustering involves sequential rounds of 2-dimensional clustering, after each round,
sub-matrices representing areas of low signal are eliminated. The next round of
clustering further refines the clusters focussing around positions and perturbations
with significant statistical coupling. This process is repeated until the clusters
converge and no further refinement is possible.
3.3 Results and Discussion

3.3.1 TPP-dependent multiple sequence alignment validation and perturbation of position 469

A 382 protein multiple sequence alignment (MSA) of the PP and PYR domains of TPP-dependent enzymes was converted into a matrix of residue frequencies and positions using SCA version 1.5. Calculation of evolutionary constraint at each site in the MSA reveals a diverse and well evolved collection of proteins. Figure 3.1 a illustrates graphically the evolutionary constraint at each position as measured by ΔG\text{stat}. High values of ΔG\text{stat} indicate a high degree of conservation or evolutionary constraint. Among the residues which display a very high degree of evolutionary constraint are Asp 155, Glu 411, and His 473. These three residues have important functions in the active site of TK. Asp 155 acts as a ligand for the metal ion CA\text{2+} or Mg\text{2+} [14], Glu 411 protonates the N_1’ Nitrogen of the pyrimidine ring [23], and His 473 has a potential involvement in transition state stabilisation [7].

The MSA must be validated before carrying out SCA to ensure that it meets certain criteria. In order for functional constraints in the MSA to be exposed, the alignment should have diversified to the extent that frequencies of amino acids at un-conserved sites have relaxed near to their mean values in all natural proteins [65]. This is evident from the frequencies of residues at the un-conserved site 38 which are approaching the mean frequencies of residues in all proteins (Figure 3.1 b). This is in contrast to a conserved site such as 469 where the amino acid frequencies differ considerably from the mean values in all proteins (Figure 3.2 a).
Secondly, the MSA should be large enough that random elimination of sequences does not change the amino acid frequencies at un-conserved sites [65]. To validate the MSA for this criteria, the five least conserved positions that still represent at least 85% occupancy were identified (1QGD numbered positions: 38, 211, 363, 508, and 509) and average $\Delta G^{\text{stat}}$ values were determined across these five sites following elimination of different proportions of the MSA. Random elimination of large proportions of the MSA was possible without increasing the average $\Delta G^{\text{stat}}$ at these un-conserved positions (Figure 3.1 c). The alignment was therefore judged to have reached statistical equilibrium in sequence space, a necessary condition for applying Boltzmann statistics.
Figure 3.1 a Overall conservation across the MSA as measured by $\Delta G_{\text{stat}}$, an expression of divergence from the frequency of residues expected in all proteins. b Frequency of residues at position 38, the least conserved position in the MSA. c The 5 least conserved sites that retain at least 85% occupancy were selected and average $\Delta G_{\text{stat}}$ at these 5 sites was evaluated following random elimination of increasing proportions of the MSA.
3.3.2 Identification of residues statistically coupled to Asp 469

Residue Asp 469 is known to be a functionally important residue in *E. coli* TK. Early studies identified the equivalent residue in yeast TK as having an important function in enantioselectivity [24] and recent work to engineer *E. coli* TK to accept non-natural substrates has identified several variants at this position that confer improved and desirable properties to the enzyme [9, 10]. Given the important function of this residue we used SCA to identify residues which display statistical coupling to Asp 469. Identification of such a network has important implications for the understanding of TK evolution and function but could also be applied to aid in the design variant libraries for engineering TK.

The residue Asp 469 is highly conserved in TK and in some of the other TPP-dependent enzymes (Figure 3.2 a). Using the SCA Toolbox we created a perturbation of the MSA to remove all sequences that do not contain an Asp at position 469, this perturbation resulted in a subalignment of 173 sequences. Statistical coupling to position 469 was then determined by calculation of the $\Delta \Delta G^{\text{stat}}$ values at all other positions in the alignment (Figure 3.2 b).

The statistical coupling energies between the 469D perturbation and other positions in the alignment are generally insignificant, however a relatively small number of positions display significant coupling to the 469D perturbation. Focusing on the seven sites displaying the highest levels of coupling to position 469 identifies an interesting cluster of residues surrounding the active site.

Mapping these sites onto the structure of *E. coli* TK reveals a network of connected residues within the protein core that display coupling with position 469.
Interestingly these are not restricted to residues in close proximity to position 469, coupling is also seen in distant residues and across the PP-PYR domain boundary (Figure 3.2 c).

Focusing on the active site of *E. coli* TK and the Asp 469 residue itself, four of the seven coupled residues form a connected cluster in and around the active site tunnel which is centred on Asp 469 (Figure 3.2 c). This cluster spans from residue Phe 437 to Thr 472 and then to Asp 469 of the PP domain. Asp 469 then forms a connecting bridge across the inter-domain barrier to His 26 which in turn interacts with His 66. Arg 520 is not physically in contact with this contiguous cluster but is relatively close to Asp 469 and reaches into the active site forming part of the entrance to the active site tunnel.

The two remaining residues, Tyr 72 and Pro 486, are more distant from the active site. Tyr 72 is located on the α-helix which leads towards His 66 of the active site and may therefore be responsible for positioning the loop containing His 66 to maintain the integrity of the active site. Pro 486 forms a hairpin loop at the opposite end of an α-helix which leads to residues Thr 472 and Asp 469. This residue may therefore have a similar function to Tyr 72 in maintaining tertiary structure and active site integrity.

Of the seven residues identified as coupling with Asp 469, two have been identified previously for their propensity to improve activity on non-natural substrates when mutated [9, 10]. Arg 520 was identified in screens for improved activity on both GA and PA and His 26 was identified in screens for improved activity on PA [10]. As Arg 520 is known to interact with the phosphate group of natural substrates, a rational
explanation for the increase in activity for non-phosphorylated substrates was a removal of the steric bulk around the entrance to the active site. The coupling between Arg 520 and Asp 469 suggests there may be a subtler synergistic explanation for the increased activity. Mutation of His 26 in E. coli TK led to variants with reversed enantioselectivity for PA [11]. This residue forms a pocket with Asp 469 which interacts with the α-hydroxyl group of natural substrates, coupling is not surprising for a residue with such a close functional and structural relationship with Asp 469.

Apart from Asp 469, His 26, and Arg 520, none of the remaining three active site residues in this network have been targeted for mutagenesis. The high hit-rate for obtaining improved mutants among these three sites indicates that similar success may be obtained at the other sites in the cluster. Identification of this network provides a good starting opportunity for creating targeted libraries of single and multiple mutants with potentially improved function on non-natural substrates.
Figure 3.2 a Frequency of residues at position 469 in the MSA versus the expected frequency of residues in all proteins. b $\Delta\Delta G^{\text{stat}}$ values between the 469D perturbation and different positions in the MSA. The seven positions representing the highest coupling energies to position 469 are coloured red. c The seven positions coupled to 469 mapped onto the structure of E. coli TK (Asp 469 is also shown). Residues in Chain A are coloured red and in chain B are coloured blue. A structurally contiguous group of coupled residues spans the PYR domain of chain A and the PP domain of chain B around the active site tunnel.
3.3.3 Global SCA analysis of TPP-dependent enzymes

The results above demonstrate the ability of SCA to identify residues displaying evolutionary constraint to perturbations at residue Asp 469. Both proximal connected and distant unconnected residues were identified that express an evolutionary dependence on the identity of the residue at position 469. The residues identified as coupled to Asp 469 may be part of a larger interconnected network of coupled residues, not necessarily directly coupled to Asp 469. To identify such an extended network of interactions and to identify other networks that don’t include Asp 469 we need to carry out a global SCA across the whole alignment.

To conduct a global SCA, each site in the MSA is subjected to perturbations to create subalignments. Perturbations are allowed that produce sub-alignments large enough to not represent global changes in conservation relative to the parent alignment [65]. To identify the cut off for sub-alignment size, the statistical coupling energy ($\Delta \Delta G^{\text{stat}}$) at un-conserved sites is monitored following random eliminations of different fractions of the MSA. The fraction of sequence elimination that begins to show coupling at un-conserved sites represents the limit of sequence elimination tolerated for the specific alignment. For the MSA of TPP dependent enzymes, un-conserved residues begin to display coupling where sub-alignments of less than 80 sequences are selected (Figure 3.3). Therefore for global SCA, perturbations were allowed that resulted in sub-alignments greater than 0.21 as a fraction of the total alignment size (382 sequences).
3 – Statistical coupling analysis

Figure 3.3 The 5 least conserved sites that retain at least 85% occupancy were selected and average $\Delta \Delta G^{\text{stat}}$ values were calculated at the 5 sites following random elimination of different proportions of the MSA.

Global SCA was carried out on the MSA in line with the criteria determined above and $\Delta \Delta G^{\text{stat}}$ values were determined for all positions in the alignment. Figure 3.4 shows the initial matrix with perturbations represented by columns and positions represented by rows. There is not a perturbation for each position in the alignment as some positions could not meet the cut-off for subalignment size. Therefore the matrix is naturally taller than it is wide. $\Delta \Delta G^{\text{stat}}$ values between residues are displayed as a linear colour scale from blue (0) to deep red (4). The highest level of coupling to position 469, identified following the Asp 469 perturbation, was Tyr72 with a $\Delta \Delta G^{\text{stat}}$ of 1.72. In contrast, many of the coupling interactions identified following global SCA are associated with coupling energies ($\Delta \Delta G^{\text{stat}}$) greater than 3 (Figure 3.4).
Figure 3.4 Matrix of $\Delta\Delta G^{\text{stat}}$ values following global statistical coupling analysis of TPP dependent enzymes. Columns represent specific perturbations of the MSA. Rows represent positions in the PP and PYR domains.

We carried out 2-dimensional hierarchical clustering to identify networks of coupled residues. Figure 3.5 shows the initial round of clustering. Large areas of the matrix are made up of low level coupling. Following clustering, large groups of perturbations or positions with low coupling energy were sequentially removed and the matrix was re-clustered to iteratively focus on the networks representing the highest statistical coupling.
Following five rounds of iterative clustering we identified a highly coupled network of 30 residues and 45 specific perturbations (Figure 3.6). Based on the clustering dendrogram, the 30 residues can be clustered into six groups of different sizes representing different coupling profiles. Six of the positions identified are known to have important functions in *E. coli* TK (Table 3.1). These include His 26 which forms a hydrogen bond with the α-hydroxyl group of natural acceptor substrates, His 66 which forms a hydrogen bond with the diphosphate group of ThDP and interacts
with the C1 hydroxyl group of donor substrates, Tyr 440 and Phe 434 which interact with the pyrimidine ring of ThDP, His 461 which binds the phosphate group of both donor and acceptor substrates, and Ile 187 which interacts with the divalent metal ion through its main chain oxygen. The identification of functionally defined residues demonstrates the robustness of SCA and indicates a potential application of the method where structure and functional information is lacking. Perhaps more interesting than the known residues are the residues for which we are unaware a function.

Figure 3.6 Final global SCA matrix following iterative focusing and reclustering around areas of high signal. As in previous figures, perturbations are represented by columns and positions by rows. The dendrogram of positions is coloured according to clusters.

The thirty residues identified with strong coupling interactions are not the most conserved residues in the alignment. This can be expected as highly conserved positions in an alignment are less likely to display changes in amino acid distribution
upon perturbation. Comparison of the thirty residues identified by SCA and the thirty most conserved residues in the alignment, as measured by evolutionary constraint (ΔG^{stat}), identifies only three residues present in both lists. These are His 26, His 66, and Gly 117. ΔG^{stat}, a measure of conservation, for the thirty highly coupled residues ranges from 0.27 to 1.35 with a mean of 0.51; for the thirty most conserved residues, ΔG^{stat} ranges from 0.88 to 2.86 with a mean of 1.50. Conservation is often used as the first tool to indicate potentially functional residues where structural and functional information is lacking, here we demonstrate an alternative approach to the identification of potentially functional residues which can complement the simple test of conservation.
When mapped onto the surface of the *E. coli* TK enzyme, the thirty residues identified display an interesting pattern. Relatively few are exposed on the surface of the holoenzyme apart from in and around the active site (Figure 3.7 a). However, when the two chains of TK are separated (Figure 3.7 b) we see several of the coupled residues exposed on the subunit surface in a connected network which spans out from the active site itself. Rotating the two chains apart further displays the surface residues identified that make up the interface between the two

<table>
<thead>
<tr>
<th>Group</th>
<th>EcoTK Residue</th>
<th>Position</th>
<th>Known function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>His 26</td>
<td>Active site surface</td>
<td>Catalysis/stereospecificity</td>
</tr>
<tr>
<td></td>
<td>His 66</td>
<td>Active site surface</td>
<td>Substrate recognition</td>
</tr>
<tr>
<td></td>
<td>Tyr 72</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Tyr 80</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Val 109</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ala 444</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asp 462</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asp 183</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gly 117</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>His 406</td>
<td>Interdomain surface</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Val 409</td>
<td>Interdomain surface</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arg 442</td>
<td>Interdomain surface</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Trp 41</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tyr 440</td>
<td>Interdomain surface</td>
<td>Forms hydrophobic pocket</td>
</tr>
<tr>
<td></td>
<td>Ile 517</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asn 443</td>
<td>Interdomain surface</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gln 478</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ser 168</td>
<td>Interdomain surface</td>
<td></td>
</tr>
<tr>
<td></td>
<td>His 461</td>
<td>Interdomain surface</td>
<td>Phosphate binding</td>
</tr>
<tr>
<td></td>
<td>Ile 187</td>
<td>Internal</td>
<td>Metal binding</td>
</tr>
<tr>
<td>4</td>
<td>Gly 67</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Glu 439</td>
<td>Interdomain surface</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Met 242</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Leu 74</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Phe 434</td>
<td>Interdomain surface</td>
<td>Forms hydrophobic pocket</td>
</tr>
<tr>
<td></td>
<td>Pro 115</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gly 413</td>
<td>Internal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Met 447</td>
<td>Interdomain surface</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leu 382</td>
<td>Interdomain surface</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Leu 62</td>
<td>Internal</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1 Hierarchical clusters of coupled positions identified (Figure 3.6) together with their identities in *E. coli* TK and their function where known.
subunits (Figure 3.7 c). In total, fourteen of the thirty residues identified are located on the subunit surface in and around the active sites.

Some of the interface residues form the binding pocket for the ThDP cofactor, these include Asp 183, Ile 187, Val 409, Phe 434, and Tyr 440. Four residues that make up the active site entrance tunnel are also identified, Lue 382 and His 461 form one side of the active site tunnel and His 26 and His 66 make up the other side. The other five surface residues are located on the interface between the two subunits and are buried in the holoenzyme. These include His 406, Arg 442, Asn 443, Glu 439 and Met 447. Presence of coupled networks of residues on the interface between the two subunits of transketolase could be expected, as a change in the residue identity on one surface of the inter-subunit interface may require a corresponding change on the other face to avoid deleterious effects on subunit aggregation. This also holds true for residues on opposite sides of the active site or on opposite sides of the cofactor binding site.
Figure 3.7 Surface rendering of TK showing two chains in different shades of gray and the coupled residues identified in red. a E. coli TK showing the active site tunnel with the TPP cofactor in green. b Separation of the two chains to expose the coupled residues that form the interaction face. c Rotation of the A chain by -45 degrees and the B chain by +45 degrees to expose the fourteen coupled residues in the interaction face and the active site.

In order to see the other sixteen coupled residues identified through SCA we need to look into the centre of the protein structure. In Figure 3.8 we see the same view
points but the individual clusters of coupled residues have been surface rendered and they are visualised with the cartoon secondary structure of the PP and PYR domains of chain B and coloured by hierarchical clustering. Although connected networks are apparent among the thirty residues, they are relatively widely distributed throughout the subunit. Examination of the residues in the structure indicates the presence of connected networks, but despite one or two exceptions, grouping of the residues by hierarchical clustering does not resolve all of these networks (Table 3.1). Several connected networks include residues from more than one clustering group. Despite the difficulty in resolving the individual networks, the majority of the residues identified are located on the subunit interface or one layer back from this interface, suggesting that the overriding pressure driving the coevolution of coupled networks in TK and other TPP-dependent enzymes is to support the interactions between the PP and Pyr domains which together form the active site.
Figure 3.8 All residues identified following global SCA of PP and PYR domains. Residues are coloured by the cluster groupings in Figure 3.6 and mapped onto one subunit of E. coli TK.

Although Asp 469 was not identified by global SCA (Asp 469 displays relatively low coupling compared to the highly coupled residues identified), three of the seven residues that are coupled with Asp 469 were identified and these residues were clustered into an independent group by hierarchical clustering. These residues were His 26, His 66 and Tyr 72, the three PP residues identified for their coupling to Asp 469.

This group of residues has a distinct coupling pattern involving good coupling signals with the perturbations 416T, 452Q, 103R, and 105E (Figure 3.6). In contrast
to the vast majority of the coupled residues identified, this group did not co-select their perturbations in the final clustered matrix. This may be due to a distinct group of perturbations specific to the His 26 cluster being removed during the iterative clustering process. This further demonstrates the independence of this cluster from the other residues identified. The location of the three residues, in the active site (His 26 and His 66) and deep within the protein fold (Tyr 72) further suggests a co-evolutionary constraint on catalytic function as opposed to structural dimerization pressures which may explain the coupling of other residues.

The highly coupled but independent residue 62 is also located close to His 26 group. Residue 62 is not highly conserved but it exhibits the highest level of coupling of all the residues in the sequence and falls in a separate, independent cluster. We are unaware of the function of this residue but it displays particularly high coupling energies with other functionally defined residues and its function warrants further investigation.

### 3.3.4 Single PYR domain global SCA analysis

Position 469 did not display strong enough coupling to be selected in the network of highly coupled residues identified above, where the strongest coupling was seen in the interface between the two subunits of TK. The homodimer of holo-TK is formed through interaction of the PP domain of one subunit interacting with the Pyr domain of the other subunit. In order to identify subtler networks of coupled residues within the Pyr domain of TK we carried out global SCA on just the Pyr domain alignment. By focusing on the Pyr domain we hope to identify networks
containing the 469 position which could aid in the design of multiple mutant libraries with enhanced properties over the individual 469 variant libraries.

The MSA was divided to leave just the Pyr domains and global SCA was carried out as described above. Following initial clustering, a similar set of residues were identified with high coupling energies indicating that the coupling networks even if driven by inter-domain interactions can be identified within one domain in isolation. In addition to the high signal clusters, an interesting cluster of coupled residues was identified in a separate part of the matrix (red sections of dendrograms in Figure 3.9 a below). Upon closer examination this cluster was found to contain Asp 469.
Figure 3.9 a Clustered matrix of $\Delta\Delta G^\text{stat}$ values following global SCA analysis of PYR domains. Interesting signal cluster containing position 469 identified by red colouring of dendrogram on perturbations and positions. b Matrix following iterative focusing and reclustering around interesting cluster identified in initial clustering.

Iterative clustering to focus down to this cluster results in the final matrix seen in Figure 3.9 b. Nine positions are identified in this group including the 469 position. In contrast to the three position His 26 cluster identified following global SCA of the
PP and Pyr domains, this cluster shows self-consistency: the nine positions are also represented in the perturbations that identify them. Such self-consistency gives confidence in the robustness of the evolutionary coupled network.

The nine positions identified include Gly 467, Asp 469, Gly 470, Thr 472, Pro 486, Pro 493, Asp 495, Glu 498, and Arg 520. Three of these residues, Thr 472, Pro 486 and Arg 520, were previously identified by their coupling energy with Asp 469 but the remaining residues have not been identified before. When these positions are mapped onto the PYR domain, two connected networks are visualised, the first containing residues Asp 469, Gly 467, Gly 470, Thr 472, and Arg 520 and the second containing residues Pro 493, Asp 495, and Glu 498 (Figure 3.10 a). The ninth member of this conserved network, Pro 486, is located at a distant point relative to the other groups. However, in E. coli TK, Pro 486 forms a tight hairpin turn in the protein structure between an alpha helix and a beta sheet (Figure 3.10 b). This turn forms at a midway point between residues in the two groups described above. It seems likely that the coupling of Pro 486 to the other two groups is related to its function in maintaining the tertiary structure around these two groups of important residues.

By just selecting residues evolutionarily coupled to Asp 469 we identified part of the cluster above. By carrying out a global SCA of the Pyr domain we have identified other physically connected members of this network together with a second, distant group linked to the first via the Pro 486 hairpin turn.
Figure 3.10 a Group of residues coupled to 469 when focusing on PYR domain. PYR domain of chain B displayed. b chain B is rotated away from view by 90 degrees in the X-plane to display relationship between the residues in the cluster. Secondary structure is only displayed for the region spanning residues 465 to 500 to emphasise the relationship between the residues.
3.4 Conclusions

We initially investigated the evolutionary coupling between Asp 469 and other positions throughout the MSA of TPP-dependent enzymes. This approach identified seven residues which were either close in proximity to Asp 469 or more distant but with a potential to impact on the tertiary structure around Asp 469. In addition to residues in the same subunit of TK, Asp 469 appears to express evolutionary coupling to sites in the other subunit of the enzyme across the interface barrier.

Having identified residues coupled to Asp 469 we carried out global SCA on the whole MSA to identify further networks within the protein fold. Global SCA identified coupling energies far greater than those associated with Asp 469. Focusing on those high energy positions identified thirty residues dispersed across the two domains. The location of these residues in the interface between the two subunits of TK suggests that co-evolution of the networks identified is driven by structural requirements for dimerization to form the active site. Identification of such interaction surfaces indicates a potential new application for SCA in the identification of protein-protein interaction sites by combining the multiple sequence alignments of proteins known to interact.

Although the clustering of the highly coupled residues did not appear to resolve the detailed sub-networks between the thirty residues, one cluster of three residues did demonstrate structural connectivity and did not appear to be involved in interface interactions. The three residues in this group had previously been identified as the only PP domain residues coupled to Asp 469. The independent identification of this small network by global perturbation analysis suggests that
these residues form a strongly coupled sub-network within the network centred on Asp 469.

In order to identify networks of evolutionary coupling containing the Asp 469 residue we focused the global SCA on the Pyr domain of the MSA. As the subunit interactions are formed between the PP and the Pyr domains of opposite subunits it was hoped that this would filter out the strong coupling signals from the interface between the subunits. The networks of coupled interface residues appear to be so strong that they can be identified by perturbations of the Pyr domain alone. However, an interesting lower signal cluster was investigated and found to contain Asp 469.

The Asp 469 cluster identified some of the residues that had previously shown coupling to Asp 469, however, an additional five residues were newly identified. Eight of the nine residues in the cluster, are arranged in two distinct connected groups in the Pyr domain of TK. In between these two structures, the ninth residue in the group forms a hairpin loop that maintains the tertiary structure around the active site.

SCA is a powerful tool that has previously been applied to several different families of proteins to discover evolutionarily coupled networks of residues [66]. Here we have used the tool to investigate evolutionary coupling within the TPP-dependent enzymes, and specifically the networks of coupled residues in the *E. coli* TK enzyme. Focusing on a key active site residue that has been mutated to engineer improved TK activity on non-natural substrates, we have identified a further eleven residues that could be targeted either singly or in combination to engineer the properties of
E. coli TK. Nine of these residues have not been targeted for directed mutagenesis before.
4 Statistical coupling guided library design for enzyme engineering of Transketolase

4.1 Introduction

Through the reversible transfer of a two carbon ketol unit from D-xylulose-5-phosphate to either D-ribose-5-phosphate or D-erythrose-4-phosphate, transketolase bridges the non-oxidative pentose-phosphate pathway and glycolysis [1]. The substrate specificity of transketolase is relatively broad and the reaction can be made irreversible by replacing the ketol donor with β-hydroxypyruvate (HPA) [8, 27-29, 44, 67, 83]. These attributes make transketolase a very attractive biocatalyst for the stereoselective formation of carbon-carbon bonds in industrial synthesis. *Escherichia coli* transketolase has become the favoured enzyme for biocatalysis applications as it has a higher specific activity with HPA compared to yeast or spinach derived enzyme [1, 28, 30]. Bacterial transketolase is also easier and cheaper to produce in the high yields required for industrial applications [29].

Although transketolase has broad substrate specificity, activity is typically lower with non-phosphorylated aldehyde acceptors such as glycolaldehyde. Use of non-phosphorylated substrates is desirable as it removes the requirement of a de-phosphorylation step later on in synthesis. Various strategies have been applied to increase transketolase activity with non-phosphorylated substrates such as glycolaldehyde. In one example, saturation mutagenesis was targeted at residues in the active site identifying variants with up to 5-fold improvement in specific activity with non-phosphorylated substrate relative to wild-type [9]. Transketolase activity is
lower still with non-hydroxylated aliphatic aldehydes [84] (typically 5-35% of those for α-hydroxylated substrates), however saturation mutagenesis targeted at the active site residues identified variants with up to 5-fold improvements in specific activity towards propionaldehyde, an aliphatic non-hydroxylated aldehyde [10].

The examples above demonstrate the potential for transketolase to be engineered to accept non-natural substrates, producing variants with beneficial activities relative to the wild type in industrial applications. Although specific activity has been improved up to 5-fold relative to wild-type for two different non-natural substrates, as we move further away from the natural substrates the ability to enhance activity with single changes to residues in the active site is expected to become more limited, with large increases in specific activity requiring multiple simultaneous mutations. The best mutations identified by saturation mutagenesis of E. coli transketolase have been recombined with the hope of achieving further improvements in activity on non-natural substrates. Unfortunately such improvements have proven elusive owing to the negative impact of multiple mutations on aspects of protein structure and function (unpublished data). New strategies are therefore required to identify multiple variants with the potential for enhanced activity towards non-natural substrates.

Different methods have been applied to identify residues which interact synergistically upon combined mutation to produce improvements that are greater than the sum of the respective single mutations. Synergistically interacting double mutants have been identified by targeting pairs of residues close in sequence [50, 85]. An alternative approach is to analyse the results of the first rounds of random
mutagenesis and recombine those predicted to be mutually beneficial [86, 87]. However, it is difficult to predict networks of synergistically acting residues not necessarily adjacent in sequence without prior experimental data on a considerable number of variants.

Statistical coupling analysis (SCA) is a powerful bioinformatics method for identifying co-evolved residues in aligned protein sequences. This technique has been applied to the transketolase enzyme and has identified potentially coupled networks of residues in and around the active site. The co-evolution of these residues may represent synergistic interactions for retained catalytic function, such synergy can be driven by various protein properties such as expression, folding, solubility, stability, activity or allostery.

Asp 469 has been identified as having a key role in substrate recognition and enantioselectivity in *E. coli* TK. Variants of Asp 469 were repeatedly identified from a large number of residues around the active site following directed evolution for improved catalytic activity on non-phosphorylated and non-α-hydroxylated substrates. Statistical coupling analysis of the TK enzyme also identified this residue among a cluster of co-evolving, potentially synergistic, residues. Individual or multiple variant libraries focused on these SCA identified residues could lead to further improvements in the activity of TK on non-natural substrates such as glycolaldehyde and propionaldehyde.

Following pre-incubation with HPA as a ketol donor, we have challenged both single and multiple SCA libraries for enhanced activity with the non-natural, non-hydroxylated substrate propionaldehyde. We used Asp 469 mutations known to
improve activity towards PA as a required initial perturbation and explored double and triple mutants using only the most frequently observed natural variant of six evolutionary coupled residues. This mutagenesis strategy successfully identified a triple mutant cluster which behaves synergistically, whereby the combination of all three mutations leads to greater activity than would be expected from the additive effects of each mutation in isolation. Further evaluation of the triple mutant cycle by kinetic analysis identified a double mutant with significantly improved $k_{cat}$. For the production of (3S) 1,3-dihydroxypentan-2-one (DHP), this is the highest $k_{cat}$ obtained for any TK mutant to date. In industrial biocatalytic transformations the turnover constant $k_{cat}$ is a very valuable property to improve. The identified variant therefore has considerable potential in the production of DHP.
4.2 Materials and methods

4.2.1 Materials

All chemicals were obtained from Sigma and used as supplied except \( \beta \)-hydroxypyruvate (HPA) which was prepared as the lithium salt as described in section 4.2.1.5.

4.2.1.1 Luria Bertani medium

Luria Bertani (LB) medium was prepared with 10 g L\(^{-1}\) tryptone, 10 g L\(^{-1}\) NaCl, and 5 g L\(^{-1}\) yeast extract in pure water. pH was adjusted to 7 using 4M sodium hydroxide solution and media was sterilised by autoclave (15 minutes, 2 bar, 124 °C).

4.2.1.2 LB agar plates

LB medium was prepared as described in section 4.2.1.1 above, with the addition of 20 g L\(^{-1}\) select agar. LB agar was sterilised by autoclave (15 minutes, 2 bar, 124 °C). After cooling, ampicillin was added to a final concentration of 150 mg L\(^{-1}\) before the solution was transferred to petri dishes to solidify.

4.2.1.3 NZYM broth

NZYM broth was prepared with 22 g L\(^{-1}\) NZYM broth in pure water. pH was adjusted to 7 using 4 M sodium hydroxide solution and broth was sterilised by autoclave (15 minutes, 2 bar, 124 °C).
4.2.1.4 Ampicillin

Ampicillin was prepared in pure water to a concentration of 150 g L\(^{-1}\). 1ml stocks were sterilised by filtration and stored at -20 °C in 1.5 ml Eppendorf tubes.

4.2.1.5 β-Hydroxypyruvate

The lithium salt of β-hydroxypyruvate was prepared by modification of a previously described procedure [27]. A 0.06 M solution of bromopyruvic acid was prepared. Using a Metrohm autotitrator, 1 M LiOH was added to 100 ml bromopyruvic acid solution at 3 ml.min\(^{-1}\) to bring the pH to 7.0 (approximately 70 ml LiOH was required). LiOH flow rate was then reduced to 1 ml.min\(^{-1}\) and the pH was maintained at 9.0 until a total of 110 ml LiOH had been added. At the end of the reaction, the pH was immediately adjusted to pH 5 using glacial acetic acid and the solution was frozen at -20 °C (unless concentrated on the same day). The mixture was concentrated under low pressure (using a vacuum pump) to approximately 20 ml final volume. Concentrate was stored at 4 °C overnight before the crude product was filtered off and washed with ethanol. Crude product was suspended in 50 ml ethanol and 40 °C for 30 minutes using a rotavap with no vacuum. The white solid was then washed with further ethanol and dried under vacuum to give a white powder. This final product (approximately 2 g) was stored at 4 °C.

4.2.1.6 Stock co-factor solution

A 12x stock solution of co-factors was prepared with 28.8 mM TPP and 108 mM MgCl\(_2\). 12x co-factors were aliquoted into 1.5 ml Eppendorfs and frozen immediately at -20 °C until required.
4.2.2 Standard procedures

4.2.2.1 Streaking agar plates

Cultures were streaked out on LB agar ampicillin plates with a sterilised wire loop. Plates were incubated at 37 °C for 16 hours and then stored at 4 °C.

4.2.2.2 Overnight cultures

Single colonies were picked with a sterilised wire loop from an agar plate into 10 ml LB medium with ampicillin (to a final concentration of 150 mg L⁻¹) in 50 ml Falcon tubes. Tubes were incubated for 16 hours at 37 °C with shaking at 200 rpm. 500 µl glycerol stocks were prepared (4.2.2.4) and the remaining culture was spun down at 4500 rpm for 10 minutes. Supernatent was discarded and the pellets were used immediately or stored at -20 °C.

4.2.2.3 100 ml fermentations

10 ml overnight culture was added to 90 ml LBamp and incubated at 37 °C with shaking at 200 rpm. 1 ml samples were taken each hour to determine biomass by spectrophotometry at OD₆₀₀. Fermentation was grown until stationary phase which was determined by biomass measurements (approximately 8 hours). At stationary phase 1 ml glycerol stocks were prepared (4.2.2.4) and the remaining fermentation was distributed to two 50 ml Falcon tubes and centrifuged at 4000 rpm for 10 minutes. Supernatent was discarded and pellets were frozen at -20 °C.
4.2.2.4 Glycerol stocks

Glycerol stocks of cultures were prepared by adding filter sterilised 40% glycerol to overnight culture in a 1:1 volume ratio to give a final glycerol concentration of 20%. Glycerol stocks were aliquoted into 1.5 ml Eppendorf tubes and stored at -80 °C.

4.2.2.5 Preparation of plasmid DNA

A Qiagen QIAprep Spin Miniprep Kit was used with a microcentrifuge to prepare plasmid DNA from pelleted overnight cultures. Plasmid DNA was eluted in 50 µl 10 mM Tris·Cl, pH 8.5. Concentration and purity were assessed by absorbance at OD$_{260}$ and OD$_{280}$ using a Thermo Scientific NanoDrop Spectrophotometer. Final plasmid preparations were stored at -20 °C.

4.2.2.6 Quickchange™ site-directed mutagenesis

Polymerase chain reaction (PCR) mixture was prepared with 5 µl 10x reaction buffer (supplied by Stratagene), 2 µl pQR791 plasmid DNA (diluted to 10 ng/µl), 1.25 µl +ve primer (100 ng/µl), 1.25 µl -ve primer (100 ng/µl), 1 µl dNTP mix, 1 µl DMSO, 39.5 µl ddH2O (to total 50 µl), and 1 µl Pfu Turbo DNA Polymerase (2.5 U/µl). PCR was run with a 95 °C 30 second initialisation step followed by 16 cycles of PCR. Each cycle of PCR included a 95 °C 30 sec denaturing step, a 55 °C 1 min annealing step and a 68 °C 22 min elongation step (4 min/Kb). Finally, the plasmid mixture was digested by adding 1 µl Dpn1 and incubating at 37 °C for 2 hours.
4.2.2.7 XL10 Gold transformation

25μl of XL10 Gold cells (Stratagene Ltd.) were thawed on ice and aliquoted into prechilled falcon tubes. 1μl 2-mercaptoethanol was added and the mixture was incubated on ice for 10 minutes with gentle mixing by swirling every 2 minutes. 1μl of the dpn1 digest reaction was added and incubated on ice for 30 minutes. The mixture was heat pulsed at 42°C for 30 seconds then returned to ice for 2 minutes. 225μl NZYM broth (4.2.1.3) pre-heated to 42°C was added prior to incubation at 37°C for 1 hour with shaking at 225 rpm. The NZYM broth mixture was then plated out on LB agar plates and incubated overnight at 37°C.

For Multi Site-Directed Mutagenesis 45 μl cells, 2 μl 2-mercaptoethanol, 1.5 μl Dpn1 treated multi-site DNA, and 0.5 ml NZYM broth were used.

4.2.2.8 Sequencing

LBamp plates were streaked out from glycerol stock and incubated for 16 hours at 37 °C. Single colonies were used to inoculate 10 ml overnight cultures. After 16 hour incubation at 37 °C, 250 μl was taken to create 20% glycerol stocks and the remaining culture was spun down at 4000 rpm for 10 minutes. Plasmid DNA was extracted from pellets using the standard Miniprep protocol (4.2.2.5). Plasmid DNA was quantified using a NanoDrop spectrophotometer. DNA was then diluted to 16 fmole/μl using ddH2O. Plasmid DNA (24 μl) was sent to Wolfson Institute Scientific Services for sequencing together with 12 μl sequencing primer (TKLibSeqS) per reaction (diluted 1:25).
4.2.2.9 **His-tag protein purification**

Binding buffer was prepared with 0.5 M NaCl, 20 mM Tris-HCl, and 40 mM Imidazole. Wash buffer was prepared with 0.5 M NaCl, 20 mM Tris-HCl, and 60 mM Imidazole. Elution buffer was prepared with 0.5 M NaCl, 20 mM Tris-HCl, and 0.5 M Imidazole.

Cell pellets from 100 ml fermentations were resuspended in 10 ml binding buffer on ice. Suspended cells were then lysed by 7 cycles of sonication (20 sec on, 20 sec off). The sonicate was centrifuged at 4000 rpm for 10 minutes at 4 °C. The supernatant was recovered and filtered using a Whatman Puradisc 25 AS 1.0 µm Polyethersulfone membrane followed by Minisart Sterile-EO, non-pyrogenic, Hydrophilic 0.20 µm filters. 0.5 ml samples were taken of sonicate, supernatant and filtrate for evaluation by SDS page gel electrophoresis.

Filtrate was loaded onto a Novagen His bind quick 900 cartridge which had been wetted and equilibrated with 6 ml binding buffer. The cartridge was then washed with 20 ml binding buffer followed by 10 ml wash buffer (provided by Novagen). Finally, protein was eluted with 4 ml elution buffer. Purified protein was dialysed overnight against 2 L 0.5 M NaCl 20 mM Tris (pH 7.0).

4.2.2.10 **Protein concentration determination**

Concentration of purified TK was determined using densitometry. The extinction coefficient of the TK dimer was calculated from the Escherichia coli TK1 sequence (P27302) [88]. The dimer of *E. coli* TK contains 1326 amino acids including 22 tryptophan residues, 46 tyrosine residues, and 10 cysteine residues. The molecular
weight of the TK dimer is 144405.4. The extinction coefficient ($\varepsilon_{280}$) calculated as described by Pace et al is 190790 M$^{-1}$cm$^{-1}$.

$$\text{Concentration (mg/ml)} = \frac{\text{Abs}_{280} \times \text{Molecular weight}}{\varepsilon_{280}}$$

Concentration of TK in total cell lysates was determined by densitometry of TK bands in coomassie stained SDS gels using the ImageJ program available from http://rsbweb.nih.gov/ij/. Concentrations were benchmarked against bands of purified TK for which concentration had been determined by spectrophotometry.

### 4.2.2.11 SDS PAGE gel electrophoresis

An 8% acrylamide: bisacrylamide resolving gel solution was prepared using 13.35 ml ProtoGel (supplied by National Diagnostics), 12.25 ml Resolving buffer, 23.6 ml ddH2O, 0.5 ml Ammonium persulfate (10% w/v), and 0.05 ml TEMED. The resolving gel was poured into a preassembled electrophoresis gel chamber and overlayed with ethanol. Stacking gel was prepared with 1.3 ml ProtoGel, 2.5 ml Stacking buffer, 6.1 ml ddH2O, 0.05 ml Ammonium persulfate (10% w/v) and 0.01 ml TEMED. Once the resolving gel had solidified, the ethanol was poured off and the stacking gel solution was poured into the chamber. Finally a Teflon comb was inserted to create the wells.

After polymerisation had completed, the gel was mounted in the SDS chamber. Tris-glycine electrophoresis buffer was used to fill the chamber and flush the wells prior to sample loading. Samples were prepared for SDS by mixing with 2x Laemmli Sample Buffer (Bio-Rad Laboratories) and heating at 100 °C for 2 minutes. 20 µl of
prepared samples were loaded per well. Gels were run at 80 V for 30 minutes followed by 120 V for an additional 50 minutes. Gels were stained with Coomasie Blue.

4.2.3 Mutant library construction

Single point mutations were introduced to the PQR791 plasmid template using Quickchange™ site directed mutagenesis as described in section 4.2.2.6. Primers for mutagenesis were designed according to the criteria defined by Stratagene and are listed in Table 4.1. Primers were sourced from Operon Biotechnologies. Following transformation of DPN1 digests, colonies were picked for overnight cultures. Pellets from overnight cultures were miniprepped and sequenced to verify mutation. Following sequence verification, a 96-well plate (Deep Square Well) containing 900 µl LBamp in each well was inoculated in triplicate with each variant.

Double mutant libraries were generated using pools of variant plasmids as templates and pools of primers to introduce the second mutation. A G467X/D495X library was created by pooling the four G467 variant plasmids (G467X plasmids) and the four D495 variant plasmids (D495X plasmids) separately. Primers were also pooled to give G467X(-), G467X(+), D495X(-), and D495X(+) primer pools. Two Quickchange™ site directed mutagenesis reactions were set up, one with G467X plasmid templates and D495X(-)/D495X(+) primers, and one with D495X plasmid templates and G467X(-)/G467X(+) primers. The dpn1 digest of both reactions were pooled to create a mixture of plasmids harbouring different combinations of G467X/D495X double mutations. The final pool of plasmid DNA was transformed into XL10 Gold cells and the transformation was plated out on an LBamp agar plate.
The mutagenesis strategy described above was repeated to create three further double mutant libraries, D469X/E498X, D469X/R520X and E498X/R520X. Following incubation at 37 °C overnight, a QPix robot was used to pick 90 colonies from each double mutant library into 96-well plates (Deep Square Well) containing 900 µl LBamp in each well.

A triple mutant library was created using double mutant libraries as templates. Entire library plates of D469X/E498X, D469X/R520X and E498X/R520X double mutants (each containing approximately 200 colonies) were used to inoculate three 10 ml overnight cultures. Pellets from overnight cultures were miniprepped to isolate plasmid template pools of the D469X/E498X, D469X/R520X and E498X/R520X libraries. A quickchange reaction was set up using D469X/E498X plasmid template together with R520X(-)/R520X(+) primers. Two further quickchange reactions were set up, one with D469X/R520X template and E498X(-)/E498X(+) primers, and one with E498X/R520X template and D469X(-)/D469X(+) primers. The dpn1 digestion products from all three reactions were pooled to create a D469X/E498X/R520X triple library. The plasmid library was transformed into XL10 Gold cells and plated out on six LBamp agar plates. Following incubation at 37 °C overnight, a QPix robot was used to pick 540 colonies from the six triple mutant library plates into six 96-well plates (Deep Square Well) containing 900 µl LBamp in each well.

The 96-well plates of single, double and triple libraries were incubated for 16 hours at 37 °C with shaking at 1000 rpm. Following incubation, each 96-well plate was split into six reaction plates (100 µl/well), four glycerol plates (50 µl culture per
well, 50 µl 40% glycerol per well), and one OD read plate (20 µl culture, 180 µl LB per well) using a Tecan high-throughput robot. Reaction and glycerol plates were frozen at -80 °C. Absorbance at 600 nm was measured from the OD read plate using the Tecan spectrophotometer.

<table>
<thead>
<tr>
<th>Variant</th>
<th>(+) Primer sequence</th>
<th>(-) Primer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>G467D</td>
<td>ctccatggtctgtgatgataagacggcccgac</td>
<td>gtcgccgctcccccagaatcgatggag</td>
</tr>
<tr>
<td>G467V</td>
<td>ctccatggtctggtgagacggcccgac</td>
<td>gtcgccgctcccccagaatcgatggag</td>
</tr>
<tr>
<td>G467S</td>
<td>ctccatggtctgtgctagacggcccgac</td>
<td>gtcgccgctcccccagaatcgatggag</td>
</tr>
<tr>
<td>G467A</td>
<td>tcccatggtctggaagacggcccgac</td>
<td>gtcgccgctcccccagaatcgatggag</td>
</tr>
<tr>
<td>D469A</td>
<td>gtcgctggcgaagacgggccgactaccc</td>
<td>ggtgagctgccccggctcccccagacc</td>
</tr>
<tr>
<td>D469L</td>
<td>tcccatggtctggtcgcgtgatggag</td>
<td>gtgactgccctgcccagacagcatggg</td>
</tr>
<tr>
<td>D469T</td>
<td>tcccatggtctggtgagacggcccgaacttgcagggctcagacc</td>
<td></td>
</tr>
<tr>
<td>D469S</td>
<td>tcccatggtctggtcgcgtgatggag</td>
<td>gtgactgccctgcccagacagcatggg</td>
</tr>
<tr>
<td>D469K</td>
<td>ggtctgggcgaacccggtcagggctcagacc</td>
<td></td>
</tr>
<tr>
<td>D469N</td>
<td>ggtctgggcgaacccggtcagggctcagacc</td>
<td></td>
</tr>
<tr>
<td>D469L</td>
<td>ggtctgggcgaacccggtcagggctcagacc</td>
<td></td>
</tr>
<tr>
<td>D469T</td>
<td>ggtctgggcgaacccggtcagggctcagacc</td>
<td></td>
</tr>
<tr>
<td>T472S</td>
<td>gagaacggcccclaagacccgactacccgacttgc</td>
<td>caaccgccgctgctgggcccttcttc</td>
</tr>
<tr>
<td>T472D</td>
<td>gagaacggcccclaagacccgactacccgacttgc</td>
<td>caaccgccgctgctgggcccttcttc</td>
</tr>
<tr>
<td>T472A</td>
<td>gagaacggcccclaagacccgactacccgacttgc</td>
<td>caaccgccgctgctgggcccttcttc</td>
</tr>
<tr>
<td>D495E</td>
<td>tggcgtctcgttgtaaaaggttcttcccacgc</td>
<td>ggggcaaccgcgcaatcggtgatcagtg</td>
</tr>
<tr>
<td>D495N</td>
<td>tggcgtctcgttgtaaaaggttcttcccacgc</td>
<td>ggggcaaccgcgcaatcggtgatcagtg</td>
</tr>
<tr>
<td>D495Q</td>
<td>tggcgtctcgttgtaaaaggttcttcccacgc</td>
<td>ggggcaaccgcgcaatcggtgatcagtg</td>
</tr>
<tr>
<td>D495L</td>
<td>tggcgtctcgttgtaaaaggttcttcccacgc</td>
<td>ggggcaaccgcgcaatcggtgatcagtg</td>
</tr>
<tr>
<td>T472S</td>
<td>gagaacggcccclaagacccgactacccgacttgc</td>
<td>caaccgccgctgctgggcccttcttc</td>
</tr>
<tr>
<td>T472D</td>
<td>gagaacggcccclaagacccgactacccgacttgc</td>
<td>caaccgccgctgctgggcccttcttc</td>
</tr>
<tr>
<td>T472A</td>
<td>gagaacggcccclaagacccgactacccgacttgc</td>
<td>caaccgccgctgctgggcccttcttc</td>
</tr>
<tr>
<td>D495E</td>
<td>tggcgtctcgttgtaaaaggttcttcccacgc</td>
<td>ggggcaaccgcgcaatcggtgatcagtg</td>
</tr>
<tr>
<td>D495N</td>
<td>tggcgtctcgttgtaaaaggttcttcccacgc</td>
<td>ggggcaaccgcgcaatcggtgatcagtg</td>
</tr>
<tr>
<td>D495Q</td>
<td>tggcgtctcgttgtaaaaggttcttcccacgc</td>
<td>ggggcaaccgcgcaatcggtgatcagtg</td>
</tr>
<tr>
<td>D495L</td>
<td>tggcgtctcgttgtaaaaggttcttcccacgc</td>
<td>ggggcaaccgcgcaatcggtgatcagtg</td>
</tr>
</tbody>
</table>

Table 4.1 Primers used to create site directed mutations in E. coli TK.

### 4.2.4 Colorimetric screening of libraries for activity

Biotransformations were prepared using reaction plates containing 100 µl of cell culture per well, lysed by a double freeze-thaw from -80 °C. Cell lysate was pre-incubated with co-factors (2.4 mM TPP, 9 mM MgCl₂ final concentration) for 20 minutes. After pre-incubation, Lithium Hydroxypyrurate (50 mM final concentration) followed by Propionaldehyde (50 mM final concentration) were
Transketolase enzyme engineering

added to a total volume of 300 µl. All reagents were prepared in 50 mM Tris buffer at pH 7.0. Propionaldehyde was prepared fresh on the day of the biotransformation. The biotransformations were incubated for 48 hours at 21 ºC.

Colorimetric screening was carried out on biotransformation reaction plates as described previously [89]. Fresh 96-well plates (Shallow Round Well) were prepared with 50 µl Tris buffer (50 mM, pH 7.0) and 20 mg MP-Carbonate Scavenger Resin (supplied by Biotage) in each well. Individual biotransformation reactions (50 µl) were then added to each well and incubated for 3 hours at 21 ºC to quench any excess hydroxypyruvate. Following quenching, the reaction-resin mixture was diluted with 100 µl Tris buffer (50 mM, pH 7.0) and 100 µl of the reaction mixture was transferred to a new plate containing 20 µl 2,3,5-triphenyltetrazolium chloride solution (0.2% in methanol), leaving the resin beads in the first plate. A plate reader (Fluostar, BMG-labtech) fitted with an autoinjector was used to add 10 µl NaOH (3 M) to each well of the new plate. Following addition of NaOH, the plate was shaken for 10 seconds and left to stand for 1 minute before an OD measurement was taken for each of the wells at 485 nm (50 flashes per well).

4.2.4.1 Validation of colorimetric screen

The colorimetric screen was validated for its ability to identify positive variants relative to wild type level activity. Variants D469A, D469E, and D469T, known to demonstrate increased activity relative to wild type, were screened against wild type TK. In total, 93 positive controls and 93 wild type controls were screened (Figure 4.1). Results from the assay indicated twelve potential false positives in the
wild type group and five potential false negatives in the positive group. The colorimetric assay therefore displays good predictability of activity.

![Validation of the colorimetric assay to distinguish between wild-type transketolase and improved variants on the PA reaction.](image)

**Figure 4.1 Validation of the colorimetric assay to distinguish between wild-type transketolase and improved variants on the PA reaction.**

### 4.2.5 Solubilisation and refolding of insoluble variants

#### 4.2.5.1 Solubilisation of transketolase variants

Pellets from 100 ml fermentations were resuspended in 40 ml binding buffer, sonicated (20 sec on, 20 sec off, 20 cycles), and spun down for 15 min at 5000 rpm. This wash step was repeated once more with 20 ml binding buffer. The final pellet was resuspended in 5 ml binding buffer with 3 M Guanidine HCl (concentration required to remove TK from insoluble fractions), sonicated and incubated overnight at 4 °C to solubilise the TK enzyme. The solubilised TK was spun down at 5000 rpm for 30 minutes and the supernatant was purified on an NiNTA column as described previously (4.2.2.9) but with 3 M Guanidine HCl in each buffer. Imidazole
concentration in the wash buffer was reduced to 20 mM to account for the Guanidine HCl.

4.2.5.2 Refolding by dialysis

Eluted TK was dialysed against decreasing concentrations of Guanidine HCl in 0.5 M NaCl, 20 mM Tris over a course of 20 hours. Dialysed samples in 0.5 M NaCl, 20 mM Tris were spun down at 5000 rpm for 30 minutes at 4 °C and analysed for protein concentration at OD_{280} nm.

4.2.5.3 Refolding by drip feeding

5 ml of eluted TK was drip fed into 45 ml 50 mM Tris HCl, 5 mM DTT over the course of 8 hours. Dilute solution of refolded protein was concentrated to 1 ml sequentially using 10,000 micron followed by 3,000 micron spin columns. Concentrated protein solution was dialysed against 20 mM Tris, 0.5 M NaCl.

4.2.6 Activity of purified enzymes

Wild-type and mutant transketolases were over-expressed and purified using His-tag affinity chromatography, and enzyme kinetics were determined using HPLC. 300 µl reactions were prepared in sealed glass vials using 150 µl purified protein. Purified protein was pre-incubated with co-factors (2.4 mM TPP, 9 mM MgCl₂ final concentration) for 20 minutes prior to the addition of Lithium Hydroxypyruvate (50 mM final concentration). Propionaldehyde was then added to varying final concentrations. Samples (20 µl) of reaction mixture were taken at twelve separate time points (at least one hour apart) and quenched with 180 µl 0.1% Trifluoroacetic acid (TFA). Quenching was carried out in a microplate which was stored at -20 °C in-
between time points. For each variant, reactions were prepared with seven different final concentrations of propionaldehyde, from 8 mM to 60 mM.

The reaction product was measured by HPLC using a 15 cm C18 column and a 15 min isocratic protocol of 0.1% TFA in 5% acetonitrile with a flow rate of 0.6 ml min⁻¹. Product was detected by UV at 200 nm. Standard curves were produced with 10-50 mM 1,3-dihydroxypentan-2-one and these were used to determine the product concentrations. New standard curves were produced for each microplate analysed.

4.2.7 Enantioselectivity of purified enzymes

The stereoselectivity of purified TK variants was established by gas chromatography as described previously [11]. Reactions were carried out to completion using purified TK with 50 mM LiHPA, 50 mM PA, 50 mM Tris-HCl, 2.4 mM TPP, 9 mM MgCl₂, 250 mM NaCl, pH 7.0 at 300 µl scale in sealed glass vials. Reaction mixture (100 µl) was transferred to vials containing 300 µl EtOAc. Vials were shaken and allowed to partition. The organic phase (100 µl) was transferred to fresh vials and pyridine (containing DMAP (10 mg/ml), 20 µl) was added to each vial. Following conversion of DHP to the diacetate, enantiomeric purity was assessed by gas chromatography on a Perkin-Elmer Autosystem XL Gas chromatograph with a β-Dex 225 chiral column (Supelco, 30 m x 0.25 mm). GC conditions: injection volume, 1 µl; carrier gas, He; carrier gas pressure, 15 psi; injector temperature, 250 °C; oven temperature, 60 °C then increased at 3 °C/min; detector temperature, 300 °C; detection, flame ionised detector (FID). Retention times: (3R)-pentan-2-one diacetate, 29.9 min; (3S)-pentan-2-one diacetate, 30.3 min.
4.3 Results and Discussion

4.3.1 Library design strategy

We described the use of SCA to identify co-evolving networks of residues in TPP-dependent enzymes. The identification of such networks enables the selection of residues for mutagenesis that are functionally linked to the active site and the rational design of multiple-mutant libraries with synergistic potential. We have identified a nine residue network within the Pyr domain of TPP-dependent enzymes that includes the functional Asp 469 residue in *E. coli* TK (Figure 4.2). This network was comprised of two structurally contiguous groups of residues with a proline residue forming a hairpin turn at a midway point between the two groups. Here we describe the creation of variant libraries by site directed mutagenesis focused on this nine residue network.

The nine positions identified in the Pyr cluster included Gly 467, Asp 469, Gly 470, Thr 472, Pro 486, Pro 493, Asp 495, Glu 498, and Arg 520. Of these nine residues, we selected seven to target for mutagenesis. The two proline residues, Pro 486 and Pro 493, were excluded as mutagenesis of these residues is likely to disrupt the tertiary structure of the TK enzyme limiting the potential to identify active variants. A second strategy adopted to avoid the introduction of destabilising mutants was to limit mutations to the most frequently observed natural variants in the TPP-dependent enzyme MSA (Table 4.2). The exceptions to this strategy included Asp 469 for which mutations were selected that were known to improve activity towards PA, and Gly 467 for which one of the common residues is a proline which was avoided for the stability issues discussed above.
Figure 4.2 Cluster of co-evolving residues in E. coli TK identified by SCA of the PYR domain of TPP dependent enzymes. Two structurally contiguous groups of residues are linked by a proline residue (Pro 486) forming a hairpin turn in the tertiary structure. Arg 520, Gly 467, Asp 469, Gly 470, and Thr 472 make up the contiguous group proximal to the active site.

Table 4.2 Frequency (%) of residues at selected positions within the TPP-dependent enzyme multiple sequence alignment. Chosen variants (bold), including the wild type residues (italic) were the most frequently occurring natural variants from an alignment of 382 homologous TPP-dependent enzymes. Exceptions (underlined) are G467P and D469E/N/R.
4.3.2 Construction and screening of SCA directed single mutant library

The twenty-eight single point mutations were introduced into *E. coli* TK by site directed mutagenesis and verified by sequencing. 96-well microplates were prepared as described in section 4.2.3 containing the single-point mutants of TK in triplicate. Measurement of cell density across microplates indicated successful growth with no positional effects on growth across the plate. Eight wells displayed unusually low cell density (Figure 4.3) but comparison with equivalent variants displaying high cell density demonstrated no correlation between activity of TK and cell density. Cell density does not therefore appear to correlate closely with the active enzyme concentration and cannot be used as a surrogate measure for protein concentration for the determination of specific activities.

![Figure 4.3 Distribution of cell growth across the microplate as measured by cell density (OD$_{600}$).](image-url)
Activity of variants was determined for the TK catalysed conversion of propionaldehyde and HPA into DHP. The absolute activity of each variant was measured in triplicate using the colorimetric screen developed by Smith et al [89], relative activities were calculated using internal wild type standards (heptuplet). Good consistency was observed across the replicates giving confidence in the colorimetric screen to determine relative activities (Figure 4.4).

Variants with improved activities relative to wild type TK were identified for five of the seven residues in the SCA directed single mutant library, these sites included Gly 467, Asp 469, Gly 470, Tyr 472 and Arg 520 (Figure 4.4). Notably, these residues make up the structurally contiguous group of residues proximal to the active site (Figure 4.2). In contrast the two residues belonging to the more distant structural group failed to produce any variants with activity increases, all mutations at these positions resulted in seriously compromised TK enzyme with approximately 30% wild type activity (Figure 4.4).

As expected, all the Asp 469 variants displayed increased activity relative to the wild-type transketolase standards contained in the same plate (Figure 4.4). Asp 469 interacts with the α-hydroxyl group of natural aldehyde acceptors providing a possible explanation for the improved activity of variants at this position with non-α-hydroxylated substrates. Mutation of Arg 520 also produced a high proportion of variants with increased activity on PA relative to wild type TK. The Arg 520 variant R520V has previously been identified as improving TK activity on GA [9] and PA [10]. Arg 520 interacts with the phosphate group of natural TK substrates, it has been hypothesised that the reduction in steric bulk associated with variation of this
residue could explain the enhanced activity seen on non-phosphorylated substrates. Although R520G has been previously identified by directed evolution for enhanced activity on GA, R520K, R520Q and R520A have not.

The other three residues for which variants displayed improved activity- Gly 467, Gly 470, and Tyr 472- have not been previously targeted for saturation mutagenesis. The identification of variants at these positions with enhanced activity on PA demonstrates the potential of SCA to identify new positions which can be targeted for site directed mutagenesis. In this study only four variants were introduced at each site, the high frequency of improved variants despite the small sample size suggests that further improvements may be identified by full saturation mutagenesis of these SCA identified sites.

Figure 4.4 Relative lysate activities of single mutants relative to wild type on the reaction between HPA and propionaldehyde. Total activity determined using the high-throughput colorimetric assay with 50 mM LHPA, 50 mM PA, and 50 mM Tris-HCl, 2.4 mM TPP, 9 mM MgCl₂, 250 mM NaCl, pH 7.0 to produce 1,3-dihydroxypentan-2-one (DHP).
4.3.3 Construction of SCA directed double and triple mutant libraries

Analysis of the single variant SCA directed library identified variants with both improved and impaired activity on PA relative to wild type TK. If the residues of this co-evolving cluster are synergistically coupled we could expect to see non-additive effects when the different mutations are combined in multiple mutants. Even with such a focussed and restricted library of twenty eight variants, the theoretical total number of mutant combinations at these seven sites would be 78,124. Libraries of this size are not out of the range of screening by high-throughput techniques but this is beyond the scope of the current study.

In order to test the hypothesis of synergy between the seven residues identified by SCA we have selected two groups of residues from which to produce smaller, more manageable multiple mutant libraries. Each of these groups includes residues from both the proximal contiguous cluster and the distal group of residues, for which no single variants with improved were identified.

The first group selected for combined mutagenesis includes residues Asp 469, Glu 498 and Arg 520. Variation at two of these residues has demonstrated improvements in activity on both GA and PA. In contrast, all four variants of Glu 498 resulted in severely impaired catalytic activity on PA. If synergistic relationships exist between these residues we could hope to see non-additive effects when the mutations are combined.
The second group selected for combined mutagenesis includes residues Gly 467 and Asp 495. Again this group includes positions from both contiguous structural clusters, one of which has produced improved variants (Gly 467), and one from which no improvements have been identified (Asp 495).

Following creation of the 540 variant triple mutant library (D469X/E498X/R520X) and the 90 variant double library (G467X/D495X), sample sequencing of ten wells from each was carried out to verify coverage of sequence space (Table 4.3). Full representation was established for the G467X/D495X library with all variants represented at each position. Although we did not attain full representation of variants across all the three sites of the D469X/E498X/R520X library, site D469 and E498 displayed three of the four possible variants, and at site R520 all four possible variants were observed. The diversity of the libraries therefore appeared very good, giving confidence that full coverage of the sequence space was attained. Notably, no wild type residues were identified in either library. In order to ensure that double as well as triple variants were captured for the D469X/E498X/R520X library, three further double mutant libraries were created for D469X/E498X, D469X/R520X, and E498X/R520X, each with 90 variants.
4.3.4 Double and triple mutant library colorimetric screen of activity

Libraries were screened in duplicate using the colorimetric screen developed by Smith et al [89]. For each plate, relative activity on the propionaldehyde HPA reaction was determined using internal wild type standards. The well-defined nature of the libraries produced (as determined by sample sequencing) presents an opportunity for analysis of variant libraries as a whole, comparing the average relative activities and the ranges of relative activities across libraries. The D469X/R520X library produced the highest mean relative activity. Average activities across the other four libraries were lower than wild type and the G467X/D495X library demonstrated the lowest average relative activity (Table 4.4).

<table>
<thead>
<tr>
<th>Library</th>
<th>Experimental Mean</th>
<th>Range</th>
<th>Expected Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>G467X/D495X</td>
<td>0.73</td>
<td>0.42 – 1.04</td>
<td>0.33</td>
<td>0.15 – 0.49</td>
</tr>
<tr>
<td>D469X/E498X</td>
<td>0.68</td>
<td>0.36 – 0.95</td>
<td>0.39</td>
<td>0.33 – 0.47</td>
</tr>
<tr>
<td>D469X/R520X</td>
<td>1.07</td>
<td>0.64 – 1.49</td>
<td>1.23</td>
<td>0.40 – 1.62</td>
</tr>
<tr>
<td>E498X/R520X</td>
<td>0.92</td>
<td>0.65 – 1.16</td>
<td>0.29</td>
<td>0.09 – 0.42</td>
</tr>
<tr>
<td>D469X/E498X/R520X</td>
<td>0.90</td>
<td>0.25 – 1.91</td>
<td>0.37</td>
<td>0.11 – 0.57</td>
</tr>
</tbody>
</table>

Table 4.4 Mean relative activities and ranges of activities across libraries compared with that expected in unbiased libraries based on single variant relative activities. Total activity determined with 50 mM LiHPA, 50 mM PA, and 50 mM Tris-HCl, 2.4 mM TPP, 9 mM MgCl₂, 250 mM NaCl, pH 7.0 to produce 1,3-dihydroxypentan-2-one (DHP).
If we assume no bias in the libraries, we can calculate the expected mean activities based on the activities of single mutants. Comparing such expected mean relative activities with the experimental mean relative activities indicates generally higher than expected activity across the libraries. The exception to this is the D469X/R520X library which produced a lower experimental mean relative activity than that expected (Table 4.4).

Comparison of the ranges of expected and experimental activities alleviates the requirement for a lack of bias, assuming full coverage of variants has been attained. The experimental range of activities across the G467X/D495X library was 2.1- to 2.8-fold greater than expected. Relative activities at the low end of the range were as expected for the D469X/E498X library but activities at the high end were 2.0-fold greater than expected. The range of activities for the D469X/R520X library was as expected although the experimental range was slightly tighter than the expected range. The experimental ranges for both the E498X/R520X and D469X/E498X/R520X were markedly higher than expected. For the E498X/R520X library this was most apparent at the low end with 7.2-fold greater activity than expected, and for the D469X/E498X/R520X library the biggest increase was at the high end with a 3.4-fold increase over that expected.

These data suggest potential synergy between the mutations in the multiple mutant libraries. General synergy is apparent between Gly 467 and Asp 495 variants with higher than expected activities. Variants of residue Glu 498 display potential synergy with both the Asp 469 (D469X/E498X library) and the Arg 520 variants (E498X/R520X library). However, no double variants have been identified in either
library with superior activity to the single variants of Asp 469 or Arg 520. This suggests that the synergistic relationships in both combinations are acting to alleviate the negative impact of variants at Glu 498. In the triple library, D469X/E498X/R520X, variants have been identified that display a synergistic increase in activity over and above the activity seen with any single variant. Although no synergy is apparent between Asp 469 and Arg 520 variants in the D469X/R520X library, these residues may have a more complex synergistic relationship incorporating the third variant, Glu 498.

The most active variant following the primary colorimetric screen was found in the triple mutant library D469X/E498X/R520X which displayed a relative activity of 1.91. Sequencing identified this variant as D469S/E498D/R520Q. Such a triple variant is of particular interest because the E498D mutation results in an inactive enzyme in isolation. In order to explore the synergistic relationships between these residues further, a mutant cycle of this triple variant was created and purified for further detailed analysis.

4.3.5 Total lysate activities of the triple mutant cycle D469S/E498D/R520Q

In order to produce a full triple mutant cycle, constructs were created by site directed mutagenesis for the double variants D469S/E498D, D469S/R520Q, and E498D/R520Q. Single variants had been produced previously by site directed mutagenesis. Following production and verification of the triple cycle, total lysate relative activities were determined for the triple mutant cycle using the colorimetric assay (Figure 4.5).
The D469S mutation and the R520Q mutation independently alleviate the negative impact of E498D, producing higher relative activities than expected with the D469S/E498D and the E498D/R520Q combinations. When all three mutations are combined, a triple variant is produced with considerably higher relative activity than could be expected taking each of the variants in isolation. These results suggest significant synergy between the three variants. However, in the absence of the E498D mutation, no synergy is apparent between D469S and R520Q based on this analysis.

Figure 4.5 Experimental and expected total lysate activities of Transketolase variants in the triple mutant cycle of D469S/E498D/R520Q. Total activity determined with 50 mM LiHPA, 50 mM PA, and 50 mM Tris-HCl, 2.4 mM TPP, 9 mM MgCl₂, 250 mM NaCl, pH 7.0 to produce 1,3-dihydroxypentan-2-one (DHP).
4.3.6 Purification of the triple mutant cycle D469S/E498D/R520Q

Lab scale fermentations of the selected triple mutant cycle variants were carried out and variants were purified on Ni-NTA columns as described in section 4.2.2.9. Expression and solubility profiles of the different variants provide information on the potential origin of synergy between these three positions.

Expression of the single variant D469S was marginally decreased relative to wild type TK, in contrast the expression of the R520Q variant was increased by 40%. However, introduction of the E498D variant reduced expression to less than 40% that of wild type TK (Figure 4.6). All of the codons introduced by site directed mutagenesis were the optimal for expression in *E. coli* TK so the decrease in expression with E498D is likely to be due to a loss of protein stability or solubility.

The combination of the two variants in the double mutant E498D/R520Q recovers the total protein expression of TK but the protein remains insoluble as determined from the concentration of TK in the soluble fraction. The double mutant D469S/E498D fails to improve total protein expression but does marginally improve the solubility relative to the E498D mutation in isolation. The triple mutant D469S/E498D/R520Q increases the total protein expression relative to E498D but no increase in the soluble fraction of TK is observed (Figure 4.6).

Based on the results above the enhanced activity of the triple mutant cannot be explained by an increase in expression or stability relative to wild type TK. Instead, the improvement appears to be due to enhanced specific activity or better folding of the soluble fraction to the active native state.
For all variants, a similar percentage of soluble protein was His-tag purified using His affinity resin. In all variants containing the E498D variant, the yield was too low to carry out further analysis on purified enzyme.

![Figure 4.6 Total, soluble and pure concentrations of TK variants.](image)

### 4.3.7 Solubilisation and refolding of E498D containing variants

In order to obtain kinetic data on the full triple mutant cycle of D469S/E498D/R520Q, efforts were made to obtain purified variants in suitable quantities for the three members of the cycle containing the E498D variant. New 100 ml fermentations were produced for each variant and prior to purification, TK was solubilised with Guanidine HCl as described in section 4.2.5.1. Following purification, refolding was attempted by both dialysis and drip-feeding. Unfortunately neither technique resulted in active protein as determined by HPLC
kinetic analysis. Further analysis of purified variants has therefore been limited to those variants that did not contain the E498D mutation.

4.3.8 Enzyme kinetics of purified variants

The enzyme kinetics of His-tag purified wild type TK and variants D469S, R520Q, and D469S/R520Q were established by HPLC with 50 mM HPA over a range of propionaldehyde concentrations (8 mM-60 mM) (Table 4.6 and Figure 4.7). At 50 mM, the concentration of HPA used here is approximately 10 times the $K_m$ of wild type TK.

The single mutation D469S produced a variant of TK with a 60% increase in $k_{cat}$ (57 min$^{-1}$ vs 35 min$^{-1}$) and a 40% decrease in $K_m$ (71 mM vs 181 mM) relative to wild type. The combined effect of these changes was a 4-fold increase in the corresponding $k_{cat}/K_m$ (0.81 min$^{-1}$ mM$^{-1}$ vs 0.19 min$^{-1}$ mM$^{-1}$) relative to wild type. The R520Q variant resulted in a similar $k_{cat}$ to wild type TK (31 min$^{-1}$ vs 35 min$^{-1}$). However, the $K_m$ for this variant almost doubled relative to wild type (329 mM vs 181 mM) resulting in a $k_{cat}/K_m$ approximately half that of the wild type enzyme (0.10 min$^{-1}$ mM$^{-1}$ vs 0.19 min$^{-1}$ mM$^{-1}$).

The most dramatic changes in kinetic parameters were observed with the double mutant D469S/R520Q. This variant displayed a 20-fold increase in $k_{cat}$ (700 min$^{-1}$ vs 35 min$^{-1}$) which more than compensated for a 3.5-fold increase in $K_m$ (628 mM vs 181 mM). The resulting $k_{cat}/K_m$ was 6-fold higher than that of wild type (1.11 min$^{-1}$ mM$^{-1}$ vs 0.19 min$^{-1}$ mM$^{-1}$), representing the highest turnover constant determined for the production of DHP with any TK variant to date (Figure 4.7).
Although the high $K_m$ of D469S/R520Q indicates poor substrate affinity for PA, the concentration of PA used in screens (50 mM) is at the lower end of the spectrum of what is used in cost-effective industrial biocatalytic transformations. Such low substrate affinity would be acceptable in industrial biocatalysis as long as substrate concentration could be increased sufficiently to achieve reasonable saturation of the active site. It is also important that enantioselectivity is not negatively impacted by a reduction in the number, or strength, of enzyme substrate interactions.

Previous kinetic analysis of purified wild type TK with propionaldehyde determined the $K_m$ to be 140±50 mM [10]. The $K_m$ determined here is higher than that previously determined but is within the calculated error. The source of this discrepancy could also be the reaction conditions. In the current study, reactions were carried out in a salt concentration of 250 mM NaCl which was included to aid in the purification of the enzyme. Hibbert et al carried out their kinetic analysis in the absence of NaCl. The salt concentration used here was the same for each variant so we do not expect this discrepancy to adversely impact on our analysis of relative kinetic parameters.

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Initial Velocities (mM min$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[PA] 8mM</td>
</tr>
<tr>
<td>WT</td>
<td>0.005</td>
</tr>
<tr>
<td>D469S</td>
<td>0.006</td>
</tr>
<tr>
<td>R520Q</td>
<td>0.003</td>
</tr>
<tr>
<td>D469S/R520Q</td>
<td>0.027</td>
</tr>
</tbody>
</table>

Table 4.5 Initial velocities of a triple mutant cycle of TK mutants at varying propionaldehyde (PA) concentrations. Initial velocities were determined from the concentration of DHP using HPLC and purified protein. Activity determined with 50 mM LiHPA, and 50 mM Tris-HCl, 2.4 mM TPP, 9 mM MgCl$_2$, 250 mM NaCl, pH 7.0 to produce 1,3-dihydroxypentan-2-one (DHP). *Not available.
Table 4.6 Kinetic parameters and enantioselectivities of a triple mutant cycle of TK mutants for the propionaldehyde (PA) substrate. Lysate activities were determined using a colorimetric assay; kinetic parameters were determined using HPLC and purified protein. Total activity determined with 50 mM LiHPA, 50 mM PA, and 50 mM Tris-HCl, 2.4 mM TPP, 9 mM MgCl₂, 250 mM NaCl, pH 7.0 to produce 1,3-dihydroxypentan-2-one (DHP). Wild-type specific activity is 0.029 ± 0.001 μmol/mg/min [10]. Obtained after purification. Enantiomeric excess of the S-isomer. d Not available.

Figure 4.7 Comparison of $k_{cat}$, $K_m$ and $k_{cat}/K_m$ for the double mutant cycle of D469S and R520Q.

### 4.3.9 Enantioselectivity of variants compared to wild type

In addition to potential synergistic effects on turnover rate and substrate affinity, the synergy between co-evolving residues may also be associated with effects on enantioselectivity. Mutants of TK have been identified in previous work that have both improved and reversed enantioselectivity [11], evaluation of such changes in
the SCA directed variants may indicate further synergistic associations between the residues. It is also important to ascertain whether the reduced substrate affinity of the double mutant D469S/R520Q has negatively impacted on the enantioselectivity of the enzyme.

Gas chromatography was used to determine the stereospecificity of wild type TK and the three purified TK variants (Table 4.6). Wild type TK catalysed the production of S-DHP in 57% enantiomeric excess, this correlates closely with previous results for wild type TK [11]. The D469S variant demonstrated improved enantioselectivity relative to wild type, producing S-DHP in 65% enantiomeric excess, this is similar to the increase observed with another variant, D469T [11]. In contrast, the R520Q variant appeared to have no effect on enantioselectivity (S-DHP 58% ee). When the two variations were combined in the double mutant D469S/R520Q, enantioselectivity was increased to the same extent observed with D469S in isolation (S-DHP 67% ee).

These results suggest that the co-evolutionary pressures on Asp 469 and Arg 520 are more likely to be associated with turnover rate and substrate affinity rather than enantioselectivity. However, these particular variants were selected based on increased activity on the PA substrate. Full evaluation of the enantioselective pressures on co-evolution would require the enantiomeric evaluation of many more variants. Importantly for the use of the D469S/R520Q variant in industrial biocatalytic transformations, we have demonstrated that the potential decrease in substrate affinity has not negatively impacted on enantioselectivity and in fact the enantioselectivity has been improved relative to wild type TK.
4.4 Conclusions

We have produced libraries of single, double and triple TK variants based around a cluster of co-evolving residues identified by SCA. The cluster of residues comprised of two structurally contiguous groups of residues and although only a small number of variants were introduced at each position, single variants with improved activity on the non-natural substrate propionaldehyde were identified for each of the positions in the structural group proximal to the active site. This demonstrates the potential of the SCA technique to identify new residues to target for site directed mutagenesis in directed evolution experiments.

Although the distal residues did not confer enhanced activity in isolation, the co-evolution of such residues suggests potential synergistic relationships between these and the proximal residues. We sought to identify evidence of such synergistic relationships by producing targeted double and triple variant libraries. The positions selected for combination were chosen such that negatively impacting variants would be combined with positively impacting variants. Synergistic relationships are expected to be more easily identifiable for such combinations.

Across the double and triple libraries, one variant was identified as significantly increasing activity on PA. This variant was D469S/E498D/R520Q and a triple mutant cycle was created to further investigate the synergistic relationships between each mutation. Analysis of the total lysate activities of the members of this triple mutant cycle indicates a lack of synergy between D469S and R520Q, however, both of these mutations alleviate the negative impact of the E498D mutation and when all three mutations are present together, activity is increased considerably. Analysis of the
total concentration and soluble fraction for each member of the cycle suggested that the combined effect of these mutations on activity cannot be explained by synergistic effects on protein expression or solubility. Instead, synergy is likely to be related to specific activity or better folding of the soluble fraction to the active native state.

Although negative effects on expression and solubility meant that no E498D containing variants could be purified for further kinetic analysis, the kinetic analysis of the other members of the triple mutant cycle has revealed some interesting relationships between the residues. Both the D469S and the R520Q mutations have a relatively small impact on enzyme kinetic parameters in isolation. However, when these two mutations are combined in D469S/R520Q, the turnover constant is increased by 20-fold relative to wild type TK. The $K_m$ is also increased relative to wild type for the double mutant, suggesting a decrease in substrate affinity, but the combined effects of these changes still resulted in a 5.8-fold increase in $k_{cat}/K_m$.

In contrast with earlier results based on total lysate relative activity, enzyme kinetic analysis indicates significant synergy between the two mutations D469S and R520Q. However, total lysate relative activities were determined at the low PA concentration of 50 mM. This is 12.6-fold lower than the $K_m$ of D469S/R520Q and as such, any synergistic increase in turnover constant could be masked by the concurrent increase in $K_m$. The synergy between these residues may have been apparent in earlier screens had a higher concentration of propionaldehyde been adopted.
Without the enzyme kinetics for the triple mutant D469S/E498D/R520Q it is difficult to explain why combination of all three mutations had such a beneficial effect in earlier screens. However, we can hypothesis that by altering the conformation of the active site, the E498D mutation may have alleviated the increase in $K_m$, thereby unmasking the synergistic increase in turnover constant at the lower substrate concentration used in the screen.

Interestingly, the D469S/R520Q double mutant of TK obtained by this approach is not found in any of the natural 382 TPP-dependent enzyme sequences used to perform the SCA. For the naturally occurring R520Q mutation, the equivalent to Asp 469 in TK is found to be Lys in various PDC and PPDC enzymes, Tyr in one TK and Asp in several ALS enzymes, but not Ser. When D469S occurs, the equivalent to Arg 520 in TK is found to be Glu or Arg. However, as the natural variants have been selected during their evolution for a particular range of functions useful to the cell, it is not necessarily expected that precisely the same natural combinations would be found when selecting for a new property such as altered substrate specificity.

In D469S/R520Q we have identified a multiple variant of TK that has a higher turnover constant for the PA reaction than any other TK variant discovered to date. Although the $K_m$ was also increased, this does not pose a problem as long as substrate concentration can be increased sufficiently and the lower substrate affinity doesn’t negatively impact on enantioselectivity. We have shown that enantioselectivity is actually increased with PA relative to wild type. The D469S/R520Q mutant is therefore very promising for the synthesis of chiral aliphatic keto diols, particularly (3S)-1,3-dihydroxy pentan-2-one.
5 General Discussion

5.1 Overall summary of thesis

The aim of this thesis was to utilise computational and bioinformatics tools to further our understanding of the transketolase enzyme and to aid in our efforts to engineer transketolase for biocatalytic applications.

Initially we sought to increase our mechanistic understanding by modelling substrate binding in the active site of \textit{E. coli} transketolase and the results presented in chapter 2 add to the growing body of structural information on this enzyme. Computational docking allowed us to model transient, reactive structures that would be very difficult or impossible to determine experimentally. Of particular interest was the docked conformation of the ketol donor HPA, which indicates an interesting angle of nucleophilic attack, which although unusual, is supported by recently published structural data.

In chapter 3 we have used a powerful bioinformatics approach termed Statistical Coupling Analysis to identify networks of co-evolving residues across the TPP-dependent enzymes. Statistical Coupling Analysis or SCA, provides an opportunity to delve deeper than the structure of proteins, uncovering networks of residues within the protein fold that have acted together through evolution to maintain or improve fitness. The identification of such networks in transketolase has important implications for enzyme engineering, suggesting combinations of residues to target simultaneously and identifying distant residues that are functionally coupled to key residues within the active site.
In chapter 4 we used the results of statistical coupling analysis to guide the design of multiple transketolase mutants with improved activity on non-natural substrates. Small targeted mutagenesis libraries were produced based on one of the co-evolving networks and simultaneous mutation of multiple residues within this cluster displayed synergistic effects on enzyme activity that would not have been anticipated from the effects of each mutation in isolation. Notably, we have identified a double mutant with a 20-fold improvement in turnover-rate using the non-natural substrate propionaldehyde. This is greater than any improvement identified to date from single variant libraries.

In these three results chapters we have achieved the main aims of this thesis. However, many questions have been raised in the course of this work and considerable additional work is required to test and expand on the conclusions presented here. In the following discussion we will review the conclusions presented and propose further work which we hope will develop these conclusions further.

5.2 Computational automated docking in transketolase

Computational automated docking is a powerful technique that can help to rationalise experimental observations from enzyme kinetics, identify potential transient structures along an enzyme reaction pathway that are difficult to obtain by experimental structure determination, and generate hypotheses to test further by experiment. Although computational docking is not completely accurate, the most obvious errors can be eliminated by visual inspection, as was the case with DX5P in this work. Transketolase is a particularly good model for this approach as
the lack of induced-fit side chain movements within the enzyme eliminates the requirement to model side chain flexibility and the potential errors this can introduce.

We used the AutoDock algorithm to reproduce the structures of known substrate-protein complexes as an initial benchmark for using the docking technique in other transketolase complexes. AutoDock was able to accurately reproduce the complex of DE4P bound in yeast TK within 1.65 Å RMSD of the crystal structure. Although AutoDock appears to struggle with explicitly charged groups, resulting in a slight shift of DE4P towards the phosphate interacting residues, the hydrogen bond network of the docked substrate was predicted accurately. Docking DE4P in *E. coli* TK demonstrated a near identical binding conformation to that for yeast TK. This is an important finding as it allows us to confidently apply the knowledge gained from studies of yeast TK to *E. coli* TK and visa versa.

Having demonstrated the accuracy of AutoDock we docked a series of known ketol donor and aldehyde acceptor substrates into the active site of *E. coli* TK and calculated their $K_m$ values from AutoDock reported binding energies. Comparison with experimentally determined $K_m$ values demonstrated good correlation, further validating the approach. Although AutoDock appears to systemically underestimate $K_m$ values, this is not of concern for the use of the binding energy calculation to select minimal energy docked conformations.

Examination of the structures of TK-substrate complexes reveals a common binding conformation, similar to that of DE4P. Phosphorylated substrates all bound with the phosphate group interacting with a conserved group of residues near the entrance
to the active site. This strong interaction leads to low experimental and calculated $K_m$ values, and the phosphate groups anchoring effects result in an inversely proportional relationship between substrate chain length and the distance of docked substrate to the ThDP cofactor. Both phosphorylated and non-phosphorylated aldol acceptor substrates bound with the Re-face of the aldehyde carbonyl exposed to nucleophilic attack by the $\alpha$-carbon of the enamine to give $S$-enantiomer products.

Two of the aldehyde acceptors, glycolaldehyde and DE4P were docked both in the presence and absence of the reactive ThDP-enamine intermediate. Such structures would be difficult or impossible to determine experimentally given their reactive nature. Binding of the two substrates was found to be identical in the presence and absence of the enamine, indicating that formation of the enamine is not a prerequisite for binding in a reactive conformation.

Comparison of docked acyclic DR5P with the recently published structure of the bound cyclic form suggests a potential mechanism for the ring opening of the substrate. A series of C-C bond rotations and hydrogen bond exchanges has been proposed that collectively unwind the DR5P molecule and orientate the newly acyclic form in readiness for nucleophilic attack by the enamine.

We have also docked ketol donors HPA and DX5P in the active site of *E. coli* TK. Unfortunately the docked conformation of DX5P was orientated unfavourably relative to the ThDP thiazolium ring; however, examination of the docked conformation of HPA has produced some of the most interesting findings of the computational docking work. HPA was docked in such a conformation that the
angle of nucleophillic attack by the deprotonated ThDP cofactor is 68° rather than the favoured Bergi-Dunitz angle of 107°. This unusual result correlates well with the observed strain in the covalent ThDP-DX5P intermediate and supports the growing picture of a catalytically poised active site in which substrate interaction energies are converted into the strained high-energy intermediate. The identical docking conformation of unreactive FPA provides an opportunity to test this conclusion experimentally.

5.3 Statistical coupling analysis of transketolase

Computational modelling of substrate binding in the active site of TK has provided evidence supporting our previous assumptions that substrate binding is equivalent between *E. coli* and yeast TK, and that many of the natural substrate interactions are maintained with non-natural substrates. This work gives us confidence in using such assumptions and has also increased our mechanistic understanding of TK. However, we have not identified any new residue functions that could lead to the improved engineering of TK for biocatalytic applications. This is perhaps unsurprising given the extensive mutagenesis and enzyme engineering work which has focussed on the residues around the active site.

When it comes to engineering TK, and enzyme engineering as a whole, the main gaps in our experience relate to multiple mutation of residues and the identification of target residues distant to the active site. These types of interactions are governed by complex relationships that are difficult to tease apart with structural data alone. The relatively new technique of statistical coupling analysis (SCA) has the potential to guide our search for beneficial multiple and distant mutations and
the following results describe the application of this technique to the transketolase enzyme.

Using a multiple sequence alignment of the PP and Pyr domains of TPP-dependent enzymes we initially used SCA to identify residues evolutionarily coupled to the important functional residue Asp 469. Seven residues were identified, five of which made up a structurally contiguous group around the active site. The other two residues were more distant; however, both were positioned on elements of secondary structure leading directly to the active site and may play a role in fine adjustment of the tertiary structure. Two of the five identified active site residues, Arg 520 and His 26, had been previously identified in enzyme engineering for modified substrate specificity. The improved activity of Arg 520 variants on non-natural substrates has been rationalised by the role of Arg 520 in phosphate binding. Evolutionary coupling between Arg 520 and Asp 469 suggests there may be a subtler synergistic explanation for the increased activity of these variants.

In order to expand our analyses of evolutionary coupling we carried out global SCA on the whole PP and Pyr domain MSA. Focusing on areas of high coupling energy we identified thirty highly coupled residues, six of which have a known function in TK. Evolutionary conservation is often used to determine potentially functional residues; SCA offers an alternative approach to the identification of functional residues. Notably, only three of the thirty SCA identified residues would also have been identified in a cut of the thirty most conserved residues. Two of these residues, His 26 and His 66, are well characterised but the third, Gly 117, would
make an interesting target for saturation mutagenesis given its redundant identification by both approaches.

Residues identified by global SCA were dominated by those located at the subunit interface between the PP and Pyr domains; this is not unexpected as changes on one side are likely to require changes on the other side to maintain interactions. However, one small sub-cluster did stand out among the thirty residues, this sub-cluster included residues His 26, His 66 and Tyr 72. This group of residues displayed a distinct coupling pattern and was also identified among the residues coupled to Asp 469. Leu 62 was also of interest, displaying the highest levels of coupling in the whole alignment and clustered independently of the other twenty-nine residues.

In order to identify subtler networks of coupled residues that may have been masked by strong subunit interaction networks, SCA global analysis was repeated with just the Pyr domains. This strategy identified an interesting network of nine coupled residues which included Asp 469. When mapped onto the structure of TK, this cluster forms two structurally contiguous groups of residues linked midway by the ninth residue which forms a hairpin turn. Confidence in this cluster is enhanced by its self-consistency and its structurally contiguous nature.

The identification of co-evolving networks of residues in the TPP-dependent enzymes demonstrates the power of this technique. The co-dependence of variation through evolution is likely to mirror co-dependency between newly introduced mutations. As such, the networks identified provide a new recourse of residues to target both in isolation and in combination for the engineering of E. coli TK. The secondary finding of this work, that many coupling interactions are located
on the interface between interacting subunits, suggests a new application of the SCA technique. Determination of the sites of protein-protein interaction in multi-protein complexes is an experimentally intensive process. Multiple sequence alignments of two proteins known to interact could be concatenated prior to statistical coupling analysis, providing a new method for the prediction of protein-protein interaction sites.

5.4 SCA guided library design for engineering transketolase activity

Based on the networks of evolutionarily coupled residues identified by SCA, we sought to engineer the activity of transketolase by targeting multiple residues for combined mutagenesis. As a high-throughput colorimetric screen had been developed in the lab for activity on the propionaldehyde substrate, improved activity on this reaction was selected as an engineering goal. We chose the nine residue Pyr domain cluster of co-evolving residues to target for mutagenesis as it displays self-consistency, is structurally contiguous, and contains the Asp 469 residue (mutation of which has improved activity on PA).

Single point mutation of seven of the nine cluster residues to common variants in the multiple sequence alignment produced interesting results, with all residues proximal to the active site producing improvements in activity on PA. In contrast, mutation of the two residues making up the structural group distal to the active site did not confer any improvement in activity, in fact activity was severely impaired. The identification of improved activity in a relatively small library of single variants demonstrates the potential of SCA to select residues for mutagenesis. Full
saturation mutagenesis at all the sites identified by SCA could be a productive strategy in the future directed evolution of TK.

The potential for SCA to identify combinations of residues for multiple mutation was investigated by selecting two groups of residues from the cluster for combined mutagenesis. Average relative activities and ranges of relative activities across these libraries indicated synergy between the residues. This synergy was also displayed in a triple mutant cycle of a particularly active triple mutant, D469S/E498D/R520Q. Although these results support our assumption of synergistic relationships between co-evolving residues, we have not investigated whether any synergy is present between non-co-evolving residues. Such an experiment would represent an important negative control for further work.

Detailed kinetic analysis of a subset of the triple mutant cycle demonstrated that the double mutant D469S/R520Q increases turnover rate of TK on the PA reaction by 20-fold. Although $K_m$ was also increased for this variant, the combined effect of these changes was a 5.8-fold increase in $k_{cat}/K_m$. The apparent synergy between these mutations contradicts earlier results which suggested the requirement of a third mutation at Glu 498 to unlock the synergistic relationships at these sites. However, it is likely that the increase in $K_m$, combined with the relatively low PA concentration used in earlier analyses, masked the synergistic increase in turnover rate of this double mutant.

Using SCA to guide the selection of multiple variants we have identified a double variant in D469S/R520Q that has a higher turnover rate for the production of DHP than any other TK variant discovered to date. These residues were known
previously for their propensity to improve activity on the PA reaction following mutagenesis. Based on this prior knowledge it could be argued that combination of such variants would be expected to give further improvements in activity. However, the improvements demonstrated here are greater than could be expected based on the effects of each mutation in isolation and earlier work to combine beneficial mutations in *E. coli* TK has failed to produce any improvements in activity.

Given the promising results obtained from targeting the nine residue Pyr domain cluster, further work should be carried out to investigate the potential of the other SCA clusters for mutagenesis. Of particular interest would be the small PP domain cluster of His 26, His 66 and Tyr 72, which demonstrated very robust clustering in our analysis. Many other residues were identified by SCA that have not been targeted for mutagenesis to date. The results presented here suggest that the creation of saturation mutagenesis libraries of these positions could be a very promising strategy for the future engineering of TK.
6 References


7 Appendix

7.1 Example Autodock input files

7.1.1 DPF example file format

```
outlev 1                             # diagnostic output level
intelec                              # calculate internal electrostatics
seed pid time                        # seeds for random generator
ligand_types C HD OA                 # atoms types in ligand
map 1QGD.C.map                       # atom-specific affinity map
map 1QGD.HD.map                      # atom-specific affinity map
map 1QGD.OA.map                      # atom-specific affinity map
elecmap 1QGD.e.map                   # electrostatics map
desolvmap 1QGD.d.map                 # desolvation map
move HPA.pdbqt                       # small molecule
about 909.9433 895.6123 902.4398      # small molecule center
tran0 random # initial coordinates/A or random
quat0 random                         # initial quaternion
ndihe 4                              # number of active torsions
dihe0 random                         # initial dihedrals (relative) or random
tstep 2.0                            # translation step/A
qstep 50.0                           # quaternion step/deg
dstep 50.0                           # torsion step/deg
torsdof 2 0.274000                   # torsional degrees of freedom and coefficient
rmstol 0.5                           # cluster_tolerance/A
extnrg 1000.0                        # external grid energy
e0max 0.0 10000                      # max initial energy; max number of retries
ga_pop_size 150                      # number of individuals in population
ga_num_generations 50000             # maximum number of energy evaluations
ga_elitism 1                         # maximum number of generations
ga_mutation_rate 0.02                # number of top individuals to survive to next generation
gacrossover_rate 0.8                 # rate of gene mutation
gawindow_size 10                     # rate of crossover
set_ga                               # # Alpha parameter of Cauchy distribution
sw_max_its 300                       # Beta parameter Cauchy distribution
sw_max_succ 4                        # set the above parameters for GA or LGA
sw_max_fail 4                        # iterations of Solis & Wets local search
sw_rho 1.0                           # consecutive successes before changing rho
sw_lb_rho 0.01                       # consecutive failures before changing rho
lg_search_freq 0.06                  # size of local search space to sample
set_sw1                              # lower bound on rho
compute_unbound_extended             # probability of performing local search on individual
ga_run 50                            # set the above Solis & Wets parameters
analysis                             # do this many hybrid GA-LS runs
```

7.1.2 GPF example file format

```
npts 64 64 64                        # num.grid points in xyz
gridfld 1QGD.maps.fld                # grid data file
spacing 0.20277777778                # spacing(A)
receptor_types A C HD N OA P SA      # receptor atom types
ligand_types C HD OA                 # ligand atom types
receptor 1QGD.pdbqt                  # macromolecule
gridcenter -11.806 24.875 37.708     # xz-coordinates or auto
smooth 0.5                           # store minimum energy w/in rad(A)
map 1QGD.C.map                       # atom-specific affinity map
map 1QGD.HD.map                      # atom-specific affinity map
map 1QGD.OA.map                      # atom-specific affinity map
desolvmap 1QGD.d.map                 # electrostatic potential map
dielectric -0.1465                   # <0, AD4 distance-dep.diel;>0, constant
```

177
7.1.3 PDBQT 1QGD file input
<table>
<thead>
<tr>
<th>ATOM</th>
<th>HE1 TRP A 54</th>
<th>2.539</th>
<th>9.094</th>
<th>9.868</th>
<th>1.00</th>
<th>0.00</th>
<th>0.165 H</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRP A 54</td>
<td>773</td>
<td>O</td>
<td>PRO A 52</td>
<td>-0.338</td>
<td>8.799</td>
<td>2.482</td>
<td>1.00</td>
</tr>
<tr>
<td>PRO A 52</td>
<td>769</td>
<td>N</td>
<td>PRO A 52</td>
<td>7.338</td>
<td>14.834</td>
<td>1.953</td>
<td>1.00</td>
</tr>
<tr>
<td>GLN A 50</td>
<td>744</td>
<td>CB</td>
<td>1.586</td>
<td>12.549</td>
<td>1.673</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>PRO A 49</td>
<td>727</td>
<td>C</td>
<td>PRO A 49</td>
<td>7.338</td>
<td>17.104</td>
<td>1.953</td>
<td>1.00</td>
</tr>
<tr>
<td>ASN A 48</td>
<td>723</td>
<td>OD1</td>
<td>-0.338</td>
<td>8.799</td>
<td>2.482</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ASN A 48</td>
<td>719</td>
<td>CG</td>
<td>1.586</td>
<td>12.549</td>
<td>1.673</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>HIS A 47</td>
<td>715</td>
<td>O</td>
<td>ASN A 48</td>
<td>-0.338</td>
<td>8.799</td>
<td>2.482</td>
<td>1.00</td>
</tr>
<tr>
<td>HIS A 47</td>
<td>706</td>
<td>CE1</td>
<td>1.586</td>
<td>12.549</td>
<td>1.673</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>HIS A 47</td>
<td>698</td>
<td>CB</td>
<td>1.586</td>
<td>12.549</td>
<td>1.673</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>HIS A 47</td>
<td>689</td>
<td>HZ1</td>
<td>-0.338</td>
<td>8.799</td>
<td>2.482</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>PHE A 44</td>
<td>647</td>
<td>CE2</td>
<td>6.462</td>
<td>12.325</td>
<td>2.057</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>PHE A 44</td>
<td>640</td>
<td>C</td>
<td>-0.338</td>
<td>8.799</td>
<td>2.482</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ASP A 43</td>
<td>632</td>
<td>CA</td>
<td>ASP A 43</td>
<td>-0.338</td>
<td>8.799</td>
<td>2.482</td>
<td>1.00</td>
</tr>
<tr>
<td>ARG A 42</td>
<td>615</td>
<td>2HH1</td>
<td>-0.338</td>
<td>8.799</td>
<td>2.482</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ARG A 42</td>
<td>612</td>
<td>CZ</td>
<td>ARG A 42</td>
<td>-0.338</td>
<td>8.799</td>
<td>2.482</td>
<td>1.00</td>
</tr>
<tr>
<td>ARG A 42</td>
<td>610</td>
<td>CD</td>
<td>ARG A 42</td>
<td>-0.338</td>
<td>8.799</td>
<td>2.482</td>
<td>1.00</td>
</tr>
<tr>
<td>ARG A 42</td>
<td>607</td>
<td>CA</td>
<td>ARG A 42</td>
<td>-0.338</td>
<td>8.799</td>
<td>2.482</td>
<td>1.00</td>
</tr>
<tr>
<td>TRP A 41</td>
<td>576</td>
<td>O</td>
<td>TRP A 41</td>
<td>-0.338</td>
<td>8.799</td>
<td>2.482</td>
<td>1.00</td>
</tr>
<tr>
<td>LEU A 40</td>
<td>554</td>
<td>CA</td>
<td>LEU A 40</td>
<td>-0.338</td>
<td>8.799</td>
<td>2.482</td>
<td>1.00</td>
</tr>
<tr>
<td>VAL A 39</td>
<td>540</td>
<td>C</td>
<td>VAL A 39</td>
<td>-0.338</td>
<td>8.799</td>
<td>2.482</td>
<td>1.00</td>
</tr>
<tr>
<td>VAL A 39</td>
<td>587</td>
<td>NE1</td>
<td>1.586</td>
<td>12.549</td>
<td>1.673</td>
<td>1.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

The table continues with similar entries for other atoms.
<table>
<thead>
<tr>
<th>ATOM</th>
<th>7038</th>
<th>CB</th>
<th>LEU A 466</th>
<th>19.660</th>
<th>14.508</th>
<th>55.577</th>
<th>1.00</th>
<th>0.00</th>
<th>0.163 HD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>7015</td>
<td>N</td>
<td>ILE A 464</td>
<td>16.671</td>
<td>13.865</td>
<td>57.535</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>7010</td>
<td>HG</td>
<td>SER A 463</td>
<td>17.398</td>
<td>10.369</td>
<td>56.752</td>
<td>1.00</td>
<td>0.00</td>
<td>0.209 HD</td>
</tr>
<tr>
<td>ATOM</td>
<td>7009</td>
<td>OG</td>
<td>SER A 463</td>
<td>17.322</td>
<td>11.223</td>
<td>56.343</td>
<td>1.00</td>
<td>13.38</td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>7000</td>
<td>N</td>
<td>SER A 463</td>
<td>16.261</td>
<td>12.917</td>
<td>54.042</td>
<td>1.00</td>
<td>7.05</td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>6990</td>
<td>CA</td>
<td>ASP A 462</td>
<td>15.725</td>
<td>12.254</td>
<td>51.730</td>
<td>1.00</td>
<td>8.68</td>
<td>0.186 C</td>
</tr>
<tr>
<td>ATOM</td>
<td>6975</td>
<td>O</td>
<td>HIS A 461</td>
<td>16.518</td>
<td>14.819</td>
<td>51.147</td>
<td>1.00</td>
<td>10.53</td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>6974</td>
<td>C</td>
<td>HIS A 461</td>
<td>17.321</td>
<td>13.891</td>
<td>50.935</td>
<td>1.00</td>
<td>9.96</td>
<td>0.241 C</td>
</tr>
<tr>
<td>ATOM</td>
<td>6957</td>
<td>HN</td>
<td>THR A 460</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>6955</td>
<td>HH</td>
<td>TYR A 459</td>
<td>9.307</td>
<td>15.008</td>
<td>41.203</td>
<td>1.00</td>
<td>0.00</td>
<td>0.217 HD</td>
</tr>
<tr>
<td>ATOM</td>
<td>6953</td>
<td>CZ</td>
<td>TYR A 459</td>
<td>10.116</td>
<td>13.833</td>
<td>42.621</td>
<td>1.00</td>
<td>10.89</td>
<td>0.065 A</td>
</tr>
<tr>
<td>ATOM</td>
<td>6937</td>
<td>CA</td>
<td>TYR A 459</td>
<td>14.614</td>
<td>12.481</td>
<td>44.670</td>
<td>1.00</td>
<td>5.44</td>
<td>0.180 C</td>
</tr>
<tr>
<td>ATOM</td>
<td>6927</td>
<td>CG1</td>
<td>VAL A 458</td>
<td>17.417</td>
<td>10.370</td>
<td>41.244</td>
<td>1.00</td>
<td>16.03</td>
<td>0.012 C</td>
</tr>
<tr>
<td>ATOM</td>
<td>6924</td>
<td>O</td>
<td>VAL A 458</td>
<td>15.419</td>
<td>12.891</td>
<td>42.130</td>
<td>1.00</td>
<td>10.83</td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>6908</td>
<td>CB</td>
<td>MET A 457</td>
<td>10.238</td>
<td>11.178</td>
<td>40.178</td>
<td>1.00</td>
<td>9.54</td>
<td>0.045 C</td>
</tr>
<tr>
<td>ATOM</td>
<td>6907</td>
<td>O</td>
<td>MET A 457</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>6874</td>
<td>O</td>
<td>GLN A 455</td>
<td>8.529</td>
<td>7.925</td>
<td>36.606</td>
<td>1.00</td>
<td>11.94</td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>6870</td>
<td>HN</td>
<td>GLN A 455</td>
<td>7.046</td>
<td>7.234</td>
<td>35.052</td>
<td>1.00</td>
<td>0.00</td>
<td>0.163 HD</td>
</tr>
<tr>
<td>ATOM</td>
<td>6865</td>
<td>2HH1</td>
<td>ARG A 454</td>
<td>9.552</td>
<td>1.626</td>
<td>30.677</td>
<td>1.00</td>
<td>0.00</td>
<td>0.174 HD</td>
</tr>
<tr>
<td>ATOM</td>
<td>6861</td>
<td>HE</td>
<td>ARG A 454</td>
<td>7.867</td>
<td>4.644</td>
<td>29.949</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>6860</td>
<td>NE</td>
<td>ARG A 454</td>
<td>8.328</td>
<td>3.790</td>
<td>30.263</td>
<td>1.00</td>
<td>17.17</td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>6851</td>
<td>CB</td>
<td>ARG A 454</td>
<td>6.465</td>
<td>5.235</td>
<td>31.922</td>
<td>1.00</td>
<td>11.59</td>
<td>0.036 C</td>
</tr>
<tr>
<td>ATOM</td>
<td>6841</td>
<td>NE2</td>
<td>GLN A 453</td>
<td>0.261</td>
<td>8.445</td>
<td>29.585</td>
<td>1.00</td>
<td>13.85</td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>6829</td>
<td>HN</td>
<td>GLN A 453</td>
<td>1.165</td>
<td>8.304</td>
<td>34.027</td>
<td>1.00</td>
<td>0.00</td>
<td>0.163 HD</td>
</tr>
<tr>
<td>ATOM</td>
<td>6812</td>
<td>CB</td>
<td>LYS A 452</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>6808</td>
<td>CA</td>
<td>LYS A 452</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atom</td>
<td>Type</td>
<td>Residue</td>
<td>Chain</td>
<td>Position</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>---------</td>
<td>-------</td>
<td>----------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8179</td>
<td>ATOM</td>
<td>GLY A</td>
<td>540</td>
<td>8.227</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8163</td>
<td>ATOM</td>
<td>ARG A</td>
<td>538</td>
<td>4.224</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8072</td>
<td>ATOM</td>
<td>LEU A</td>
<td>533</td>
<td>17.255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8062</td>
<td>ATOM</td>
<td>GLN A</td>
<td>532</td>
<td>16.091</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8047</td>
<td>ATOM</td>
<td>GLU A</td>
<td>531</td>
<td>15.210</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8013</td>
<td>ATOM</td>
<td>THR A</td>
<td>529</td>
<td>19.529</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8010</td>
<td>ATOM</td>
<td>THR A</td>
<td>529</td>
<td>19.241</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7997</td>
<td>ATOM</td>
<td>ARG A</td>
<td>528</td>
<td>18.847</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7987</td>
<td>ATOM</td>
<td>ARG A</td>
<td>528</td>
<td>20.697</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7974</td>
<td>ATOM</td>
<td>GLU A</td>
<td>527</td>
<td>22.068</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7966</td>
<td>ATOM</td>
<td>GLN A</td>
<td>526</td>
<td>19.675</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7959</td>
<td>ATOM</td>
<td>GLN A</td>
<td>525</td>
<td>25.536</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7941</td>
<td>ATOM</td>
<td>GLN A</td>
<td>525</td>
<td>9.821</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8481</td>
<td>ATOM</td>
<td>VAL A</td>
<td>561</td>
<td>18.367</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8450</td>
<td>ATOM</td>
<td>GLY A</td>
<td>558</td>
<td>17.601</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8438</td>
<td>ATOM</td>
<td>THR A</td>
<td>557</td>
<td>10.539</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8436</td>
<td>ATOM</td>
<td>THR A</td>
<td>557</td>
<td>11.947</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8424</td>
<td>ATOM</td>
<td>ALA A</td>
<td>556</td>
<td>12.350</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8393</td>
<td>ATOM</td>
<td>PHE A</td>
<td>554</td>
<td>13.185</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8366</td>
<td>ATOM</td>
<td>ILE A</td>
<td>553</td>
<td>11.195</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8354</td>
<td>ATOM</td>
<td>LEU A</td>
<td>552</td>
<td>8.353</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8319</td>
<td>ATOM</td>
<td>PRO A</td>
<td>550</td>
<td>8.322</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8295</td>
<td>ATOM</td>
<td>GLY A</td>
<td>548</td>
<td>8.501</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8260</td>
<td>ATOM</td>
<td>ASP A</td>
<td>545</td>
<td>14.159</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8232</td>
<td>ATOM</td>
<td>LEU A</td>
<td>543</td>
<td>17.490</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8217</td>
<td>ATOM</td>
<td>LEU A</td>
<td>543</td>
<td>13.985</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>