

# **Combining Large Scale *In Silico* Analysis with Fragment Screening to Identify Novel, Ligandable Secondary Sites in Cancer- Associated Proteins**

**Catherine Fletcher**

Hit Discovery and Structural Design

Cancer Research UK Cancer Therapeutics Unit

The Institute of Cancer Research

University of London

This thesis is submitted as partial fulfilment of the  
requirements for the degree Doctor of Philosophy

## Acknowledgements

I am grateful to many people for their help throughout the course of my PhD. Firstly, I would like to thank my supervisors for the opportunity to work and study at the ICR. To Rob, thank you for devoting so much time to this project and for your support throughout, and to Bissan, thank you for your unwavering faith in my research. To Rosemary, thank you for patiently sharing your expertise and providing honest feedback, which was much appreciated, and to Yann-Vai, for patiently and enthusiastically training me in crystallography. And finally, to Costas, for teaching me how to code, for your kind support and always having the time to answer my questions, no matter how busy you were.

I would also like to thank the past and present members of the HDSD Team and Data Science – thank you for being so patient and welcoming, for offering your time, help and experience. There are too many people to individually name, but I would especially like to thank Patrizio, for helping me with canSAR3D, and Craig for answering many wide, varied questions about all things protein. In addition, I would like to thank Matt for his help with XChem, Jemima, Kathy and Caroline for help with setting up biophysical and biochemical assays. To Ellen and Suzanne and Nicola, thank you for answering my pretty stupid chemistry questions, very patiently. I am also grateful to Rosemary and Sandra, who designed and synthesised a series of analogues to push the project forwards, and Alice for her help at XChem.

The ICR has been a great place to work, and many people have been instrumental in this. The Baking Club, who have turned whole weeks around with their amazing creations, and those who shared lunches, and ice-creams, and drinks at the pH bar; my time here would have been much less enjoyable without you all.

A massive thank you goes to my family and friends for supporting me, especially those who are no longer with us. Thank you for your love, and your support, and for being proud of me. And to Nik – thank you for putting up with me. I couldn't have done it without you.



## **Declaration**

To the best of my knowledge, the research presented in this thesis is original work unless otherwise indicated. This thesis has not been previously submitted for a degree at this or any other university or institution, in part or in whole.

A handwritten signature in black ink, appearing to read 'C Fletcher', written in a cursive style.

Catherine Fletcher

## Abstract

Proteins often have multiple binding sites involved in interactions with other molecules. The majority of currently approved drugs bind a protein's primary site, the major functional site in the protein. Targeting the primary site can be challenging. Secondary site binders can allow for efficient inhibition of difficult-to-drug protein targets and there are now multiple examples of secondary site inhibitors in the clinic. However, the majority of known secondary sites were discovered through serendipity and the systematic identification of ligandable secondary sites remains challenging. This thesis integrates high throughput *in silico* analysis of publicly available protein structures based on canSAR3D with fragment screening to identify novel, ligandable and functionally relevant secondary sites in clinically relevant protein targets. Following the analysis, triaging of identified sites for functional relevance, clinical impact and technical feasibility identified a short-list of four targets – p53, ESR1, PIK3CA and IDH1. The novel secondary site in isocitrate dehydrogenase 1 (IDH1) was selected for validation.

The tumour-promoting IDH1-R132X mutation is found in up to 80% of glioma patients and 15% of acute myeloid leukaemia patients. Fragment screening identified 19 fragments binding specifically to the novel secondary site in IDH1-R132H. Following up these fragments in biochemical assays confirmed that binding to this pocket inhibits enzyme activity. My work shows that the newly discovered secondary pocket of IDH1-R132H is both ligandable and functionally relevant, and that my *in silico* analysis can be used to identify novel secondary sites in therapeutically relevant proteins.

## Abbreviations

<b>2HG</b>	D-2-hydroxyglutarate
<b>Å</b>	Angstrom
<b>AGI-5198</b>	$\alpha$ KG competitive, mutant selective IDH1 inhibitor
<b><math>\alpha</math>KG</b>	$\alpha$ -ketoglutarate
<b>BDC</b>	Background density correlation
<b>BisTris</b>	2-[Bis(2-hydroxyethyl)amino]-2-(hydroxymethyl)propane-1,3-diol
<b>BSA</b>	Bovine serum albumin
<b>CC<sub>1/2</sub></b>	Correlation coefficient of two halves of the data
<b>CGC</b>	Cancer Gene Census
<b>D2HGDH</b>	D-2-hydroxyglutarate dehydrogenase
<b>Da</b>	Dalton
<b>DMSO</b>	Dimethyl sulphoxide
<b>DTT</b>	Dithiothreitol
<b>GSK-864</b>	Mutant selective IDH1 inhibitor
<b>H-bond</b>	Hydrogen bond
<b>HEPES</b>	(4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid )
<b>IC<sub>50</sub></b>	Half maximal inhibitory concentration
<b>IDH1-R132H</b>	Cancer associated mutant IDH1 with an arginine to histidine substitution at position 132
<b>IDH1-WT</b>	Wild type isocitrate dehydrogenase
<b>IPTG</b>	Isopropyl $\beta$ -D-1-thiogalactopyranoside
<b>k<sub>cat</sub></b>	1st order rate constant
<b>K<sub>m</sub></b>	Michaelis constant
<b>KS</b>	Kolmogorov-Smirnov
<b>M</b>	Molar
<b>mol</b>	Moles
<b>MR</b>	Molecular Replacement
<b>MW</b>	Molecular weight
<b>NADP<sup>+</sup>/H</b>	Nicotinamide adenine dinucleotide phosphate /reduced form
<b>OCC</b>	Occupancy
<b>OD</b>	Optical density
<b>PanDDA</b>	Pan-Dataset Density Analysis
<b>PDB</b>	Protein Data Bank
<b>PEG</b>	Polyethylene glycol
<b>PEG5000MME</b>	Polyethylene glycol monomethyl ether with average molecular weight of 5000
<b>RFU</b>	Relative fluorescence units
<b>RMSD</b>	Root mean square deviation
<b>ROS</b>	Reactive oxygen species
<b>rpm</b>	Revolutions per minute
<b>RSCC</b>	Real space correlation coefficient
<b>RSZO</b>	Real space observed Z-score
<b>SDS</b>	Sodium dodecyl sulphate
<b>SOC</b>	Super optimal broth with catabolite repression
<b>SMGs</b>	Significantly Mutated Genes
<b>TB</b>	Terrific Broth
<b>TCGA</b>	The Cancer Genome Atlas
<b>TEV</b>	Tobacco Etch Virus
<b>T<sub>m</sub></b>	Melting temperate

<b>Tris</b>	Tris(hydroxymethyl)aminomethane
<b>Triton X-100</b>	Polyethylene glycol p-(1,1,3,3-tetramethylbutyl)-phenyl ether
<b>Tween-20</b>	Polysorbate 20
<b>UV</b>	Ultra violet
<b>V<sub>0</sub></b>	Initial rate
<b>V<sub>max</sub></b>	Maximum rate
<b>Z-value</b>	PanDDA specific statistic, the number of standard deviations from the ground state.
<b>Z'</b>	Z-prime
<b>ΔT<sub>m</sub></b>	Change in melting temperature compared to a control.

# Table of Contents

<b>Chapter 1: Introduction.....</b>	<b>20</b>
1.1 The need for new targeted therapeutics.....	20
1.2 Secondary sites as therapeutic targets.....	22
1.3 Target Evaluation.....	24
1.4 Computational predictions of druggability.....	26
1.5 Fragment-based lead discovery .....	28
1.6 Aims.....	30
1.7 Strategy .....	31
1.8 Isocitrate dehydrogenase .....	33
1.8.1 IDH1 structure .....	33
1.8.2 IDH1 as a cancer therapeutic target .....	35
<b>Chapter 2: <i>In silico</i> identification of ligandable sites.....</b>	<b>39</b>
2.1 Introduction .....	39
2.1.1 Pocket identification methods .....	39
2.1.2 Training predictors .....	40
2.1.3 canSAR3D structure-based ligandability predictor.....	41
2.2 Results.....	46
2.2.1 Defining the training sets.....	46
2.2.2 Statistical considerations for developing the secondary site ligandability predictor .....	48
2.2.3 Identifying ligandable secondary sites .....	54
2.2.4 Triaging pockets for target selection .....	54
2.2.5 Target selection.....	69

2.3	Conclusions .....	70
-----	-------------------	----

### **Chapter 3: Establishing enabling technologies for fragment**

<b>screening .....</b>	<b>71</b>
------------------------	-----------

3.1	Introduction .....	71
-----	--------------------	----

3.1.1	Thermal shift assays .....	72
-------	----------------------------	----

3.1.2	Crystallographic fragment screening.....	74
-------	--	----

3.2	Results.....	80
-----	--------------	----

3.2.1	Protein production and purification.....	80
-------	--	----

3.2.2	Variant characterisation .....	83
-------	--------------------------------	----

3.2.3	Establishing IDH1-R132H TSA for fragment screening .....	85
-------	--	----

3.2.4	Establishing IDH1-R132H crystallographic conditions for fragment screening .....	94
-------	--	----

3.2.5	Establishing IDH1-WT crystallographic conditions.....	99
-------	---	----

3.2.6	IDH1 structures .....	100
-------	-----------------------	-----

3.2.7	Confirming pocket ligandability .....	101
-------	---------------------------------------	-----

3.3	Conclusions .....	103
-----	-------------------	-----

<b>Chapter 4: Fragment screening to investigate the ligandability of the novel secondary site in IDH1-R132H.....</b>	<b>104</b>
--	------------

4.1	Introduction .....	104
-----	--------------------	-----

4.2	Fragment screen by TSA.....	105
-----	-----------------------------	-----

4.2.1	Binding Site elucidation.....	109
-------	-------------------------------	-----

4.3	Crystallographic fragment screening .....	110
-----	---	-----

4.3.1	Fragment soaking and data collection .....	110
-------	--	-----

4.3.2	Ground state map optimisation .....	112
-------	-------------------------------------	-----

4.3.3	Summary of results .....	114
-------	--------------------------	-----

4.3.4 Thermal shift assays to support the presence of fragments identified by crystallographic fragment screening .....	115
4.4 Overview of fragment hits from both fragment screens .....	117
4.4.1 Non-conserved binding modes: CCT371095 and CCT370874 .....	120
4.4.2 Benzoimidazole series .....	121
4.4.3 Fragments binding to Trp205 .....	122
4.4.4 Fragments binding with re-organisation of the pocket-forming loop .....	125
4.5 Conclusions .....	132

## **Chapter 5: Investigating the functional relevance of the novel**

<b>pocket .....</b>	<b>134</b>
5.1 Introduction .....	134
5.2 IDH1 NADPH fluorescence assay .....	135
5.2.1 Establishing an IDH1-R132H inhibition assay .....	135
5.2.2 Establishing an IDH1-WT biochemical assay to investigate compound selectivity .....	146
5.3 Inhibition of IDH1-R132H .....	151
5.3.1 Inhibition of IDH1-R132H by hit fragments .....	151
5.3.2 Investigation of fragment analogues against IDH1-R132H .....	154
5.4 Inhibition of IDH1-WT .....	159
5.4.1 Inhibition of IDH1-WT by fragments and analogues .....	159
5.5 Structural rationale for IDH1-R132H inhibition .....	160
5.6 Investigation of mutations in the novel pocket .....	162
5.6.1 Characterisation of IDH1-R132H double mutants .....	163
5.6.2 Investigating the effects of secondary site mutations on IDH1-R132H activity .....	163

5.7	Conclusions .....	166
-----	-------------------	-----

## **Chapter 6: Conclusions, lessons learnt and remaining**

<b>questions</b>	.....	<b>167</b>
------------------	-------	------------

6.1	Conclusions .....	167
-----	-------------------	-----

6.2	Lessons learnt .....	168
-----	----------------------	-----

6.2.1	Limitations of the computational predictor.....	168
-------	---	-----

6.2.2	Training set bias.....	172
-------	------------------------	-----

6.2.3	Crystal form limitations.....	172
-------	-------------------------------	-----

6.2.4	Non-isomorphous crystals in PanDDA.....	174
-------	---	-----

6.3	Questions remaining.....	177
-----	--------------------------	-----

6.4	Outlook .....	178
-----	---------------	-----

<b>Chapter 7: Materials and Methods</b>	.....	<b>180</b>
---	-------	------------

7.1	Computational Methods.....	180
-----	----------------------------	-----

7.1.1	Building the ligandability predictor.....	180
-------	---	-----

7.1.2	Triaging and target selection.....	182
-------	------------------------------------	-----

7.2	Experimental investigation.....	185
-----	---------------------------------	-----

7.2.1	General .....	185
-------	---------------	-----

7.2.2	Generation of IDH1 constructs.....	187
-------	------------------------------------	-----

7.2.3	Protein expression and purification.....	189
-------	--	-----

7.2.4	Protein characterisation by label-free thermal shift.....	193
-------	---	-----

7.2.5	SYPRO Orange thermal shift assays.....	193
-------	--	-----

7.2.6	Crystallisation.....	195
-------	----------------------	-----

7.2.7	Fragment soaking experiments with TSA fragment hits.....	197
-------	--	-----

7.2.8	Data collection, processing and structure solution.....	197
-------	---	-----



7.2.9	XChem Crystallography based fragment screen .....	198
7.2.10	NADPH fluorescence assay .....	200
7.2.11	Kinetic characterisation of IDH1-R132H double mutants.....	207
7.2.12	Compound Mass spectrometry .....	207
<b>Chapter 8: Appendix .....</b>		<b>208</b>
8.1	<i>in silico</i> analysis .....	208
8.1.1	Roc curves and randomised Roc curves for properties identified as being significant .....	208
8.1.2	Summary statistics for pocket properties .....	210
8.1.3	Pockets identified with ligandable secondary sites .....	211
8.1.4	Roc curves for sequence conservation .....	213
8.1.5	Publically available IDH1 structures by variant and conformation ..	213
8.1.6	Sequence conservation of mammalian IDH enzymes .....	214
8.2	<i>in vitro</i> investigation .....	216
8.2.1	Coding sequences and primers .....	216
8.2.2	Expression tests.....	217
8.2.3	Mass spectrometry of IDH1 variants.....	218
8.2.4	Label-free TSA using Prometheus for IDH1 variant characterisation .....	221
8.2.5	SYPRO Orange TSA.....	223
8.2.6	Compound mass spectrometry .....	226
8.2.7	Crystallographic refinement statistics tables .....	227
8.2.8	XChem fragment screening .....	233
8.3	Introduction to Crystallography .....	235
8.3.1	Crystallisation.....	235

8.3.2	X-ray diffraction .....	237
8.3.3	Data collection strategies .....	238
8.3.4	Data processing and assessing data quality.....	239
8.3.5	Molecular Replacement to solve crystal structures.....	241
8.3.6	Model correction and refinement.....	242
8.3.7	Validation of model.....	244
<b>Chapter 9: Bibliography.....</b>		<b>246</b>

## List of Figures

Figure 1.1: Relationship between fragment screening hit rate and subsequent development of high affinity ligand. ....	28
Figure 1.2: Affinity range of various screening techniques.....	29
Figure 1.3: Strategy used to identify a novel ligandable secondary site in a cancer-associated protein. ....	32
Figure 1.4: Structure of IDH1. ....	34
Figure 1.5: IDH1 mutations common in cancer.....	36
Figure 2.1: Primary and secondary sites are refined differently by the PickPocket algorithm as part of the canSAR3D pipeline.....	43
Figure 2.2: Flowchart out-lining the approach to identify novel, ligandable and functionally relevant secondary sites in cancer-associated proteins.....	45
Figure 2.3: Examples of pockets included in the training set. ....	47
Figure 2.4: Overview of the triaging process used to shortlist the four potential targets for experimental investigation.....	60
Figure 2.5: Overview of the novel secondary site in p53.....	63
Figure 2.6: Overview of the novel site in ESR1.....	64
Figure 2.7: Overview of the novel site in PIK3CA. ....	65
Figure 2.8: Overall structure of the IDH1-R132H dimer showing the location of the three pockets predicted to be ligandable.....	66
Figure 2.9: IDH1 pockets change with the conformation.....	68
Figure 3.1: Thermal denaturation curve of a given protein with SYPRO Orange. ....	72

Figure 3.2: Overview of PanDDA processing to identify minor conformations and states.....	75
Figure 3.3: Schematic overview of the XChem crystallographic fragment screening.....	79
Figure 3.4: Purification schema for IDH1 variants.....	80
Figure 3.5: Representative chromatograms and SDS-PAGE gels from IDH1 purifications. ....	82
Figure 3.6: Label-free thermal shift results.....	84
Figure 3.7: Typical melting curves for IDH1-R132H with NADPH.....	86
Figure 3.8: Bar chart showing different fluorescent measurements obtained for different SYPRO Orange and IDH1-R132H concentrations.....	87
Figure 3.9: Bar charts showing the change in melting temperature with increasing concentrations of NADPH (A) and $\alpha$ KG (B) .....	89
Figure 3.10: 2D structures of IDH1-R132H tool compounds.....	90
Figure 3.11: Bar chart showing $\Delta T_m$ values for IDH1-R132H with tool compounds.....	91
Figure 3.12: Bar chart showing $\Delta T_m$ values for IDH1-WT with NADP <sup>+</sup> , isocitrate or with both NADP <sup>+</sup> and isocitrate .....	92
Figure 3.13: Bar chart showing the increase in IDH1-WT melting temperature with increasing concentrations of tool compound GSK-864 in the presence and absence of NADP <sup>+</sup> .....	93
Figure 3.14: Hanging and sitting drop vapour diffusion setups. ....	95
Figure 3.15: Representative crystals from optimisation of IDH1-R132H (A-C) and IDH1-WT (D-F) crystal systems. ....	96

Figure 3.16: Bar chart showing the $\Delta T_m$ of IDH1-R132H with NADPH induced by hits from the Hampton Research Solubility and Stability screen .....	98
Figure 3.17: Structures of IDH1-R132H and IDH1-WT. ....	101
Figure 3.18: Confirmation of ligandability in the novel secondary site in in-house IDH1-R132H structures. ....	102
Figure 4.1: Overview of the fragment-screening cascade to identify hit matter binding to the novel secondary site. ....	104
Figure 4.2: Overview of the thermal shift fragment screening cascade .....	105
Figure 4.3: Scatter plot showing $\Delta T_m$ of all 2595 fragments screened against IDH1-R132H.....	106
Figure 4.4: Overlay of IDH1-R132H chain A from two different structures.....	113
Figure 4.5: Summary of sites identified by the PanDDA analysis .....	115
Figure 4.6: Overlay of the 19 fragments identified binding into the novel secondary site by thermal shift and crystallographic fragment screening. ....	117
Figure 4.7: Comparison of PanDDA and normal maps for fragments CCT370974 and CCT371095.....	120
Figure 4.8: Comparison of PanDDA and $2mF_o - DF_c$ maps for the two structurally similar fragments identified through XChem crystallographic screening.....	121
Figure 4.9: Thermal shift hit CCT242817 was identified binding to the novel pocket through an edge-face pi-stack on Trp205.....	123
Figure 4.10: Comparison of PanDDA and $2mF_o - DF_c$ maps for Trp205-stacking fragments identified through XChem crystallographic fragment screening....	124
Figure 4.11: Fragments can bind in the space occupied by Ile112 and induce remodelling of the pocket-forming loop. ....	126

Figure 4.12: Fragments binding with remodelling of the pocket-forming loop that is too low occupancy to be seen in normal $2mF_o - DF_c$ maps. ....	127
Figure 4.13: Comparison of PanDDa event and Z-maps, and $2mF_o - DF_c$ maps for loop-remodelling hits. ....	129
Figure 4.14: Four fragments binding to the novel pocket with reorganisation of the pocket-forming loop showed clear electron density in normal $2mF_o - DF_c$ maps.....	131
Figure 5.1: Overview of the IDH1-R132H biochemical assay. ....	136
Figure 5.2: Change in NADPH fluorescence over time in the presence or absence of Tween20 and BSA.....	137
Figure 5.3: Change in fluorescence signal over time with increasing concentrations of IDH1-R132H. ....	138
Figure 5.4: Initial rates and $K_m$ curves for IDH1-R132H with.....	140
Figure 5.5: Plot of initial rate of reaction of IDH1-R132H with increasing DMSO concentrations. ....	142
Figure 5.6: Comparison of signal window and percentage conversion .....	143
Figure 5.7: $IC_{50}$ curves for two tool compounds AGI-5198 and GSK 864. ....	144
Figure 5.8: Representative curves from the fluorescence interference assay. ....	145
Figure 5.9: Overview of IDH1-WT biochemical assay.....	146
Figure 5.10: Plot of fluorescence against time, monitoring production of NADPH .....	147
Figure 5.11: Michaelis-Menten curves for IDH1-WT cofactor $NADP^+$ and substrate isocitrate. ....	148

Figure 5.12: Comparison of signal windows and percentage NADP <sup>+</sup> conversion .....	149
Figure 5.13: Comparison of raw fluorescent signals from IDH1-WT and IDH1-R132H assays.....	150
Figure 5.14: The novel $\alpha$ -helical conformation may clash with the regulatory segment in the active conformation and offer structural rationale for inhibition of IDH1 activity. ....	161
Figure 5.15: Initial rates of reaction for different IDH1-R132H variants at different enzyme concentrations .....	164
Figure 6.1: Definition of the known allosteric site in IDH1 is dependent on both the protein conformation and completeness of the regulatory segment. ....	169
Figure 6.2: Structures showing the novel secondary site in PIK3CA, which since been validated by crystallographic fragment screening. ....	171
Figure 6.3: Key stabilising interactions in the IDH1-R132H novel secondary site.....	173
Figure 6.4: Comparison $2mF_o - DF_c$ (A) and Z- and event (B) maps from the same dataset from an IDH1-R132H crystal soaked with CCT242635. ....	175
Figure 6.5: Example of class averages for ground state generation.....	176
Figure 6.6: Diffraction pattern from an IDH1-R132H crystal soaked with high concentrations of CCT239686.....	177

## List of Tables

Table 1.1: Secondary site inhibitors under development or with FDA approval .....	24
Table 2.1: Pocket properties identified as being statistically significant and the thresholds used. ....	52
Table 2.2: Summary of statistics of properties .....	53
Table 2.3: Examples of known, ligandable secondary sites initially predicted to be ligandable by the computational predictor.....	55
Table 2.4: Overview of shortlisted targets .....	61
Table 3.1: PanDDA specific statistics used to identify and evaluate low occupancy events.....	77
Table 3.2: assay conditions for thermal shift fragment screening. ....	93
Table 4.1: Thermal shift dose response for the 20 fragments identified as hits. ....	108
Table 4.2: fragments that showed stabilisation of IDH1-WT with NADP <sup>+</sup> . ....	109
Table 4.3: Thermal shift values for fragments identified as XChem hits. ....	116
Table 4.4: Overview of fragment screening hits identified binding to the novel secondary site .....	119
Table 5.1 Kinetic parameters for IDH1-R132H.....	141
Table 5.2: Kinetic parameters for IDH1-WT. ....	148
Table 5.3: IC <sub>50</sub> values or maximum inhibition for hit fragments tested in the IDH1-R132H biochemical assay. ....	153
Table 5.4: Summary of biochemical assay data for analogues based on fragment hit CCT242817.....	156



Table 5.5: Summary of biochemical assay data for analogues based on fragment hit CCT242635.....	158
Table 5.6: Summary of biochemical assay data for analogue based on fragment hit CCT239544.....	159
Table 5.7: Melting temperatures from label-free thermal shift data for IDH1- R132H double mutants.....	163
Table 5.8: Kinetic parameters for IDH1-R132H variants with respect to co- factor $\alpha$ KG. ....	165
Table 7.1: Search models used for molecular replacement.....	198

## Chapter 1: Introduction

---

### 1.1 The need for new targeted therapeutics

Cancer is the second leading cause of death globally, accounting for one in six deaths<sup>6</sup>. In the UK, 28% of patients will undergo chemotherapy as part of their primary treatment; these cytotoxic agents are effective against any rapidly dividing cell type, leading to serious and severe side effects.

In contrast, targeted therapies interfere specifically with a molecule, usually a protein, shown to be critical in tumour cell survival or cancer progression<sup>7, 8</sup>. These drugs can elicit strong response rates by exploiting vulnerabilities in cancer cells, leading to selective killing of tumours over healthy tissues. For example, the BCR-ABL inhibitor imatinib shows response rates of up to 80% in chronic-phase CML patients<sup>9</sup>. Despite advances in targeted treatment of many cancer types, there remains an unmet need for treatment of less common and refractory cancers, and to overcome the emergence of resistance to current therapeutics.

The identification of appropriate targets is based on extensive research into the complex biology behind malignant transformation and identification of key drivers of these processes. The underlying causative processes are, however, very complicated, and vary between patients as well as tumour types. This adds a layer of complexity to the identification of biologically compelling targets. Multiple large-scale projects such as The Cancer Genome Atlas<sup>10</sup> (TCGA) and the International Cancer Genome Consortium<sup>11</sup> (ICGC) have been established

to collate patient-derived mutational data in order to identify recurrent mutations, with the aim of elucidating new tumour-driving mechanisms and potential targets for therapeutics. Multiple analyses of such datasets have identified gene sets that are recurrently mutated and implicated in driving malignant transformation. The expertly curated Cancer Gene Census (CGC)<sup>12</sup> contains 574 genes with genomic alterations that promote oncogenic transformation, and a further 145 newly identified genes with strong evidence for their involvement in cancer. A list of 127 Significantly Mutated Genes (SGMs) was published by Kandoth et al.<sup>13</sup> in 2013 following analysis of 3,281 tumour samples from the TCGA. Identifying genomic and transcriptional alterations within patient cohorts can aid in selection of targets likely to show clinical impact.

Despite the wealth of potential therapeutic targets and the positive clinical impact of targeted therapeutics, the development of successful oncology drugs remains very challenging. Over 90% of drug discovery projects fail before reaching the market, costing billions of dollars and many years of research<sup>14</sup>. Therefore, identification of targets more likely to be chemically tractable is important for risk mitigation. However, many targets that are biologically compelling may not be considered tractable by standard medicinal chemistry approaches, or represent a family not yet exploited, and are therefore high-risk targets. Recent examples of drugging novel, challenging protein targets such as Bcl family members show the potential reward of targeting these more challenging proteins and expanding the target space<sup>15</sup>.

## 1.2 Secondary sites as therapeutic targets

The majority of currently approved targeted therapeutics bind to the major functional, or primary, site in a protein. While targeting these primary sites has led to development of highly successful therapeutics, there are also challenges associated with this approach. For example, the physico-chemical properties associated with a given pocket may be unfavourable, as is the case of the phosphate binding site in protein phosphatases<sup>16</sup>. Some phosphatases are clinically relevant targets, such as PHLPP which de-phosphorylates AKT resulting in promotion of tumour growth in squamous cell carcinoma cell lines<sup>17</sup>. However, the primary site in PHLPP is small and highly polar to facilitate binding to phosphate groups. Due to this, it is highly challenging to develop primary inhibitors with acceptable bioavailability, and this class has historically been considered undruggable, though several inhibitors targeting secondary sites are now being developed<sup>18</sup>.

Primary sites may also have a high affinity natural ligand, such as the nucleotide binding site of Hsp70<sup>19</sup> and Ras<sup>20</sup>. Inhibitors targeting these sites require exceptionally high affinity to compete with the natural ligands. Inhibitors targeting the Hsp70 primary site show a large drop off in potency when characterised in cellular studies due to the high concentration of ATP in cells. Both Hsp70<sup>21, 22</sup> and Ras<sup>23</sup> have recently been successfully inhibited by targeting secondary sites.

Protein primary sites can share high sequence or structural homology within families, such as in kinases, which can hinder efforts to develop selectivity<sup>24, 25</sup>. While poly-pharmacology plays an important role in efficacy for some

inhibitors<sup>26</sup>, including Sorafenib, which was launched as a pan-kinase inhibitor targeting VEGFR2, VEGFR3, KIT, FLT3 and PDGFR  $\beta$ <sup>27</sup>, off-target inhibition can also result in toxicity and reduce the treatment window. Secondary sites, in contrast, tend to show lower sequence and structural conservation than primary sites, which can aid in development of selective inhibitors and reduce off-target toxicity<sup>22, 28</sup>.

Finally, exposure of cancer cells to targeted therapeutics inevitably results in emergence of resistance, often through mutations that abrogate the ability of drugs to bind their target<sup>2, 29-31</sup>. This can result in highly efficacious therapeutics losing potency and their impact on patients. For example, the EGFR T790M mutation confers resistance to front line therapeutics gefinitib and afatinib<sup>31</sup>, while the ABL-kinase T315I mutation confers resistance to all therapeutics developed prior ponatinib<sup>32</sup>. New drugs are currently continuously required to overcome emerging resistance.

Alternative approaches to modulating challenging but functionally relevant protein targets to allow translation into clinic include targeting functionally relevant secondary sites, including allosteric sites. Allosteric modulators of GPCRs (such as benzodiazapines) have been widely used for many decades in the treatment of psychological, neurological and CNS disorders<sup>33-35</sup>. In cancer therapeutics specifically, several secondary site inhibitors against diverse and challenging targets are showing efficacy in clinical trials, with some now approved and showing impact in patients (Table 1.1). In addition, targeting inhibitors to a secondary site presents another opportunity to overcome

resistance mutations, as they can also slow the development of resistance mutations when used synergistically with inhibitors targeting the primary site<sup>36</sup>.

Despite their relevance and potential for clinical impact, identifying functionally relevant, chemically tractable secondary sites remains challenging. The majority of currently known sites were discovered through serendipity. Relating inhibition at these novel sites to the functional modulation of a target within a cellular context presents a further level of complexity.

<b>Drug</b>	<b>Clinical Stage</b>	<b>Target</b>
Trametinib <sup>37</sup>	Approved for Braf V600E melanoma	Mek1; adjacent to primary site
Ivosidenib <sup>38, 39</sup>	Approved for relapsed/refractory AML with IDH1-R132X mutation	Induced allosteric pocket above active site
Ispinesib <sup>40</sup>	Stage II	Eg5; locks conformation
Asciminib <sup>41</sup>	Phase II for CML and Ph+ AML	BCR-Abl1 myristoyl pocket

Table 1.1: Secondary site inhibitors under development or with FDA approval

### 1.3 Target Evaluation

Given the cost of drug discovery project failure, selecting projects with reduced risk is an important aspect of drug discovery. Target evaluation assesses the biological, technical and competitive risks associated with a given target<sup>42, 43</sup>. Biological risk assesses the likelihood of a potent inhibitor having clinical impact, through identification of a suitable patient population, evidence for anti-cancer effect and knowledge of potential resistance mechanisms, amongst other considerations. The competitive risk assesses the competitive landscape,

as well as the unique selling point of a drug or target, and the potential for collaboration.

Technical risk assesses the likelihood of developing a potent molecule against the given target. This not only includes the availability of chemical tools and *in vivo* animal models, but also the presence of enabling technologies such as biophysical and biochemical assays, and structural biology. A significant aspect of technical risk is the druggability of the given target. This assesses whether there is a site in the protein that is considered to be ligandable, and if binding of a small molecule to that site will affect the protein function and lead to therapeutic benefit. In well-established drug targets, where second and third generation inhibitors are under development to mitigate resistance, the druggability of a given site is already well established. In addition, proteins from privileged families, such as kinases, are generally considered to be ligandable and present less of a risk.

When considering novel targets and secondary sites, the ligandability may be unknown. In these instances, computational predictors of ligandability and druggability can be used to assess the likelihood of both developing a potent molecule against the target, as well as the potent molecule having an impact on cellular viability.

## 1.4 Computational predictions of druggability

Prioritisation of druggable targets can reduce attrition during the drug discovery process<sup>42, 44</sup>. Druggability refers to the likelihood of finding bioavailable small molecules that bind to the given target and subsequently impact both the target function and the disease state. This can be split into two aspects: the ligandability of the target - the likelihood of identifying a small, drug-like molecule that binds with high affinity; and the functional relevance – the likelihood of small molecule binding resulting in modulation of both protein function and the disease state. Ligandability of a protein target is necessary but not sufficient for druggability using small molecule approaches. The majority of currently available ligandability predictors can be split into three groups: precedence; chemical; and structural predictors.

Precedence is the most straightforward predictor and is based solely on knowledge of previously drugged protein targets. For example, protein kinases are one of the most extensively pursued class of cancer therapeutic targets<sup>45, 46</sup>, and their primary, ATP-binding sites are generally considered to be ligandable. Further, kinases are involved in many cancer-driving pathways, and inhibiting these enzymes will often have impact on cellular viability<sup>47</sup>. Pursuing targets from families with high precedence may lower the risk associated with the target, but also limits the proteomic space that can be explored.

Ligand-based or chemical druggability assesses compounds tested against the target and its homologues for their bioactivity, molecular weight, tractability for medicinal chemistry elaboration and ligand efficiency<sup>48, 49</sup>. Proteins that have, or have close homologues with, ligands with good physicochemical properties



and high efficacy are considered ligandable. This approach requires chemical matter to have been tested against either the target or a close homologue. This information is not available for many proteins.

Structure-based ligandability predictors utilise the 3D structures of proteins to identify pockets that may be chemically tractable<sup>50</sup>. There are many tools available to do this, including canSAR3D<sup>49</sup>, which uses a variety of physical and chemical properties associated with each identified pocket to predict its ligandability. Its use is limited to proteins with experimentally determined structures, or with close homologues whose structures have been solved.

Both ligand- and structure-based predictors focus only on the chemical tractability of a given target, without considering the impact of small molecule binding on either the function of the protein target or on the cell. While a small molecule competing with the endogenous ligand for binding to the primary site will have clear impact on protein function, this may not translate to into a change in phenotype. Network druggability describes the likelihood that modulation of a protein will cause modulation of the disease state given its position in the interactome<sup>51</sup>. Genuine anti-cancer targets show a higher degree of connectivity within the network, with more first neighbours, and are part of larger communities than either non-drug targets or targets of drugs in different therapeutic areas. In contrast to the other three, network druggability does not directly predict the likelihood of developing an inhibitor: a protein target can have high network connectivity but lack a ligandable site.

## 1.5 Fragment-based lead discovery

A fragment-screening hit rate can also predict ligandability (Figure 1.1)<sup>52</sup>, and can be used in combination with computational predictors to evaluate targets. Fragments are low molecular weight molecules, typically less than 300 Da with fewer than 12 heavy atoms. Fragment-based lead discovery (FBLD) involves screening a relatively small number of fragments, hundreds to thousands, to identify protein binders. Despite the complex geometry of protein surfaces, ligands bind selectively to very specific locations, termed hot spots<sup>53</sup>. Protein targets with such a hot spot often yield high affinity, non-covalent drugs, regardless of the affinity of the initial hits. Fragment screening approaches have been successfully used to identify hit matter in known and putative secondary sites<sup>28, 53, 54</sup>

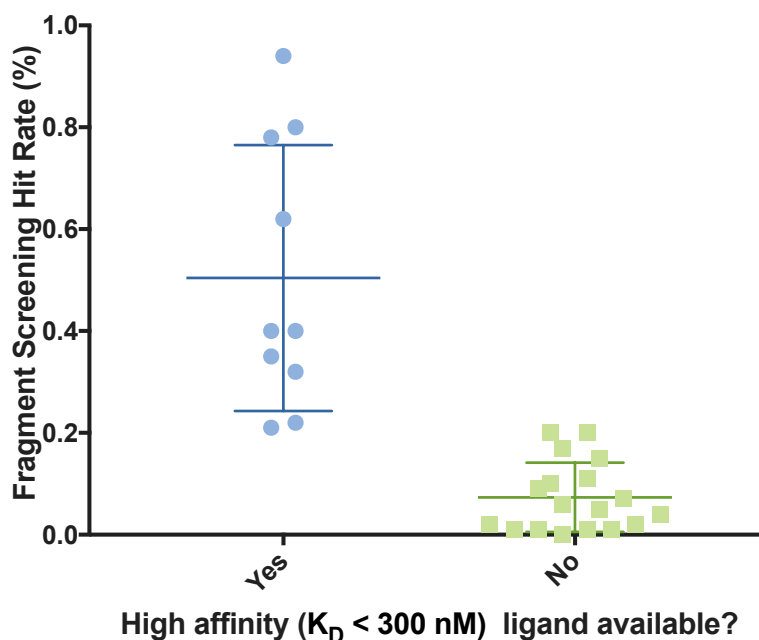


Figure 1.1: Relationship between fragment screening hit rate and subsequent development of high affinity ligand. Data from Hajduk et al. 2005<sup>25, 31, 53</sup>, plot made in Graphpad Prism

Molecules that obey Lipinski's Rule of Five are considered drug-like. Estimation of drug-like chemical space suggests that it may consist of  $10^{30}$  to  $10^{60}$  compounds<sup>55-57</sup>. Based on public databases such as PubChem<sup>58</sup>, ChemSpider<sup>59, 60</sup> and ChEBML<sup>10</sup> amongst others, it is estimated that only  $10^8$  compounds have been synthesized, representing a tiny proportion of the available space. Even the largest screening libraries using molecules of this mass covers very little of the potential space. When considering fragments, the possible chemical space is estimated to consist of approximately  $10^8$  molecules. A library of a thousand molecules this size would cover a far greater proportion than covered by using larger molecules. The use of fragments also allows a more efficient sampling of chemical space than using more drug-like molecules, leading to an increased hit rate.

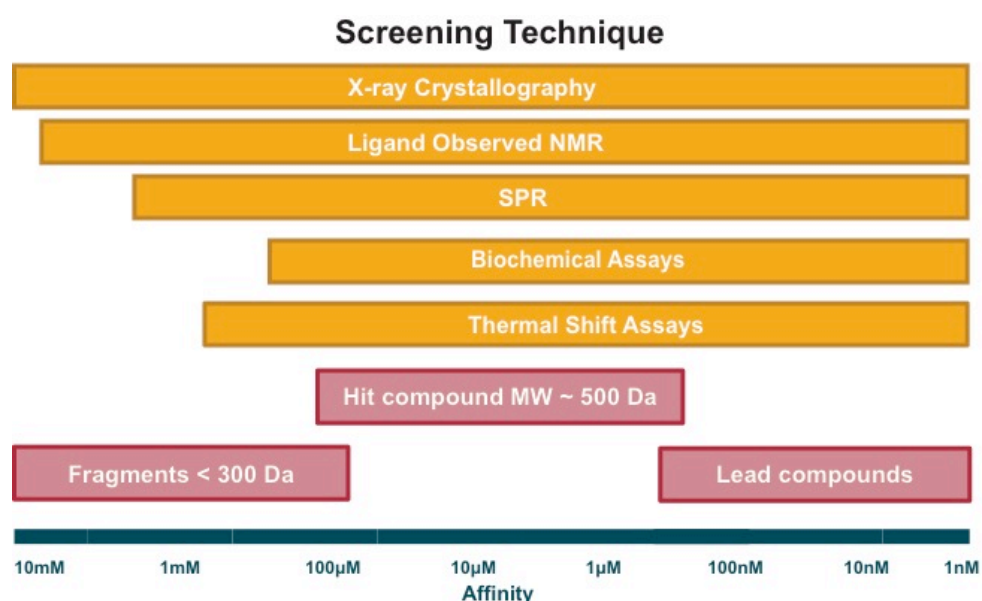


Figure 1.2: Affinity range of various screening techniques. Adapted from Hubbard et al. 2011<sup>25, 61</sup>.

In addition, using fragments also reduces the molecular complexity – the number of interactions, both favourable and unfavourable, a molecule can have

with its protein target<sup>62</sup>. With a lower molecular complexity, fragments are less likely to possess an interaction that would abrogate its ability to bind a protein target. Identified fragment binders tend to make few, but higher quality interactions that act as a starting point for medicinal chemistry. Due to the low molecular weight of fragments, the affinity of initial hits is often in the high micromolar to low millimolar range. Sensitive biophysical techniques are required to detect binding events (Figure 1.2)<sup>61</sup>. Fragments can then be elaborated into larger, more potent hit and lead compounds.

## **1.6 Aims**

The aim of this project is to investigate how computational analysis can be combined with fragment screening to identify novel, ligandable secondary sites. It can be split into two individual aims:

### **1. Adaptation of canSAR3D to identify novel, ligandable secondary sites**

Various structure-based ligandability predictors are available. I used the ICR's predictor, canSAR3D, as it had been used to analyse all structures currently available in the PDB. This provides a large quantity of data on which to train and then test the predictor. The pocket properties used to predict ligandability within the canSAR3D pipeline were systematically analysed to identify those which are important for ligandability without biasing for primary sites. This included a retrospective validation using known secondary sites as well as the identification of novel secondary sites. Following identification of novel sites, the pockets were triaged for functional relevance using literature

evidence as well as patient-derived mutations and sequence conservation.

## **2. Fragment screening to investigate the ligandability of the secondary site**

Following selection of a target, fragment-screening approaches were used to identify molecules binding to the novel secondary site. Using fragment screening rather than a high-throughput screen of larger molecules may increase the likelihood of finding molecules that target the secondary site, and gives an overall assessment of the ligandability. Identified hit matter was then used to investigate the functional relevance of the secondary site through inhibition studies, which would represent the first step to confirming the druggability of the site.

## **1.7 Strategy**

The strategy used to meet these aims, is shown in Figure 1.3. The structure-based ligandability predictor canSAR3D was adapted to identify ligandable secondary sites as described in Chapter 2. These novel sites were triaged for technical feasibility and functional relevance, and four targets were shortlisted: p53, ESR1, PIK3CA and IDH1-R132H. While all four were clinically relevant, technically feasible targets, I selected the novel secondary site in IDH1-R132H as an initial target as a functional hypothesis could be developed. Fragment screening was then used to investigate the ligandability of the novel secondary site. Chapter 3 describes the establishment of enabling technologies for two fragment screening approaches, and Chapter 4 discusses the results of these fragment screens.

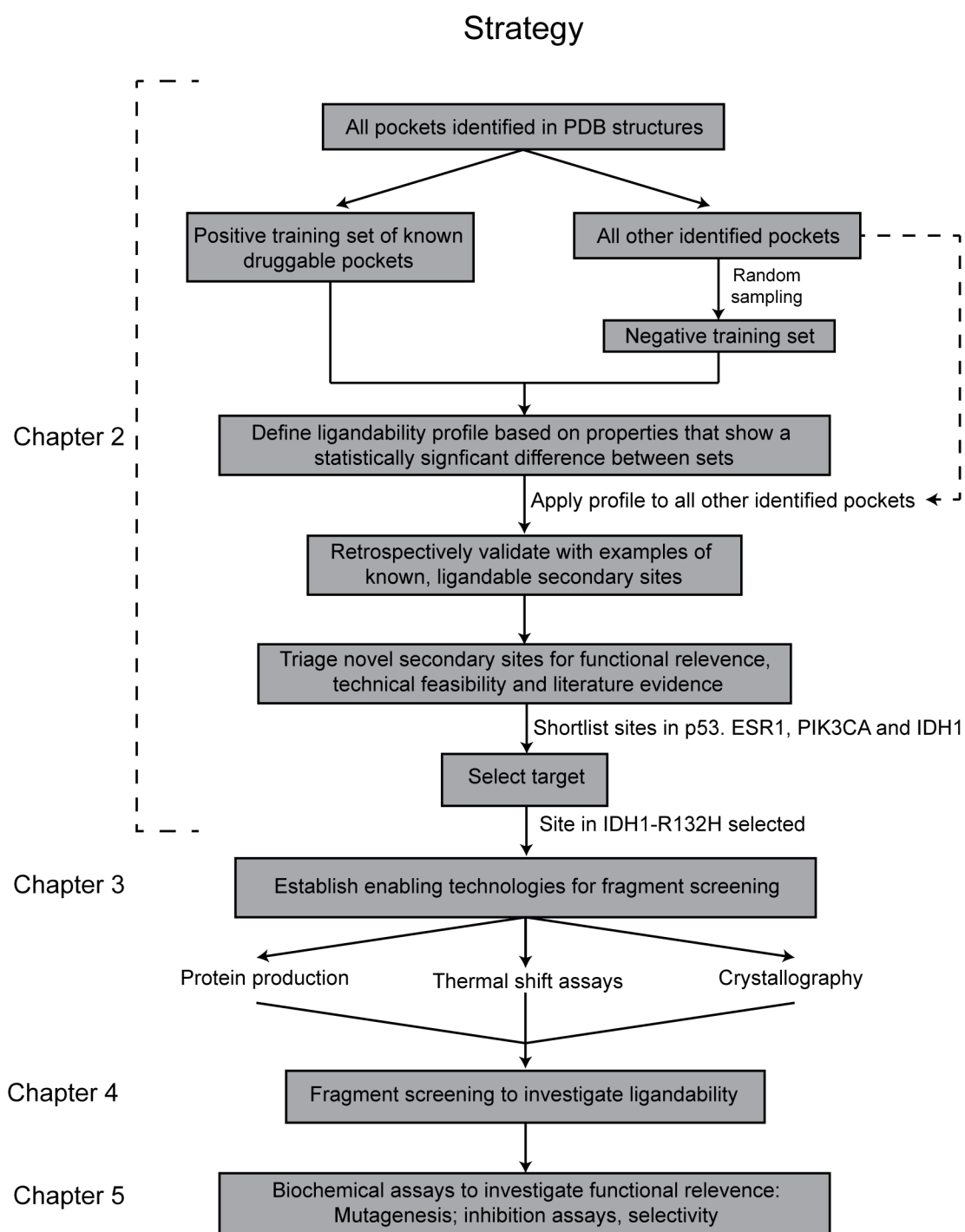


Figure 1.3: Strategy used to identify a novel ligandable secondary site in a cancer-associated protein.

After identification of hit matter targeting the novel secondary site, a biochemical assay was used to investigate the functional relevance of the secondary site, as described in Chapter 5.

## **1.8 Isocitrate dehydrogenase**

Isocitrate dehydrogenase (IDH) exists in three isoforms in humans. IDH3 is a heterotetramer that uses  $\text{NAD}^+$  to catalyse the conversion of isocitrate to  $\alpha$ -ketoglutarate ( $\alpha$ KG) as part of the citric acid cycle. IDH1 and IDH2 are homodimers that use  $\text{NADP}^+$  as a reducing agent to catalyse the same reaction. IDH2 is predominately localised to the mitochondria, while IDH1 localises to mostly to the cytoplasm. Wild type IDH1 (IDH1-WT) is the primary source of NADPH in most tissues, especially the brain, and is therefore involved in the mitigation of oxidation damage through the regeneration of glutathione<sup>63</sup>.

### **1.8.1 IDH1 structure**

IDH1 structures have been solved using X-ray crystallography. Each monomer is formed of a large, small and clasp domain (Figure 1.4). The large domain adopts a typical Rossmann fold, often associated with nucleotide binding<sup>64</sup>, while the small domain forms an  $\alpha/\beta$  sandwich. The two domains are connected by a  $\beta$ -sheet that forms the base of primary site. The dimer is held together by the clasp domains, consisting of two, two-stranded anti-parallel  $\beta$ -strands that interlock to form two, four-stranded anti-parallel  $\beta$ -sheets stacking on top of each other. In the inactive, co-factor bound conformation, the primary site is in an open conformation, with the regulatory segment, residues 271-286, forming a flexible loop stabilised by hydrogen bonding between Ser94 and

Asp279, and between His132 and Asp275. In all but one PDB structures (PDB 1T09), the electron density for this region is too weak to be modelled. With binding of the substrate and catalytic  $Mg^{2+}$ , the small and clasp domains move relative to the large domain. The previously unstructured regulatory segment adopts an  $\alpha$ -helix across the dimer interface, stabilised by magnesium. The primary site is formed of residues from both chains, allowing catalysis to occur.

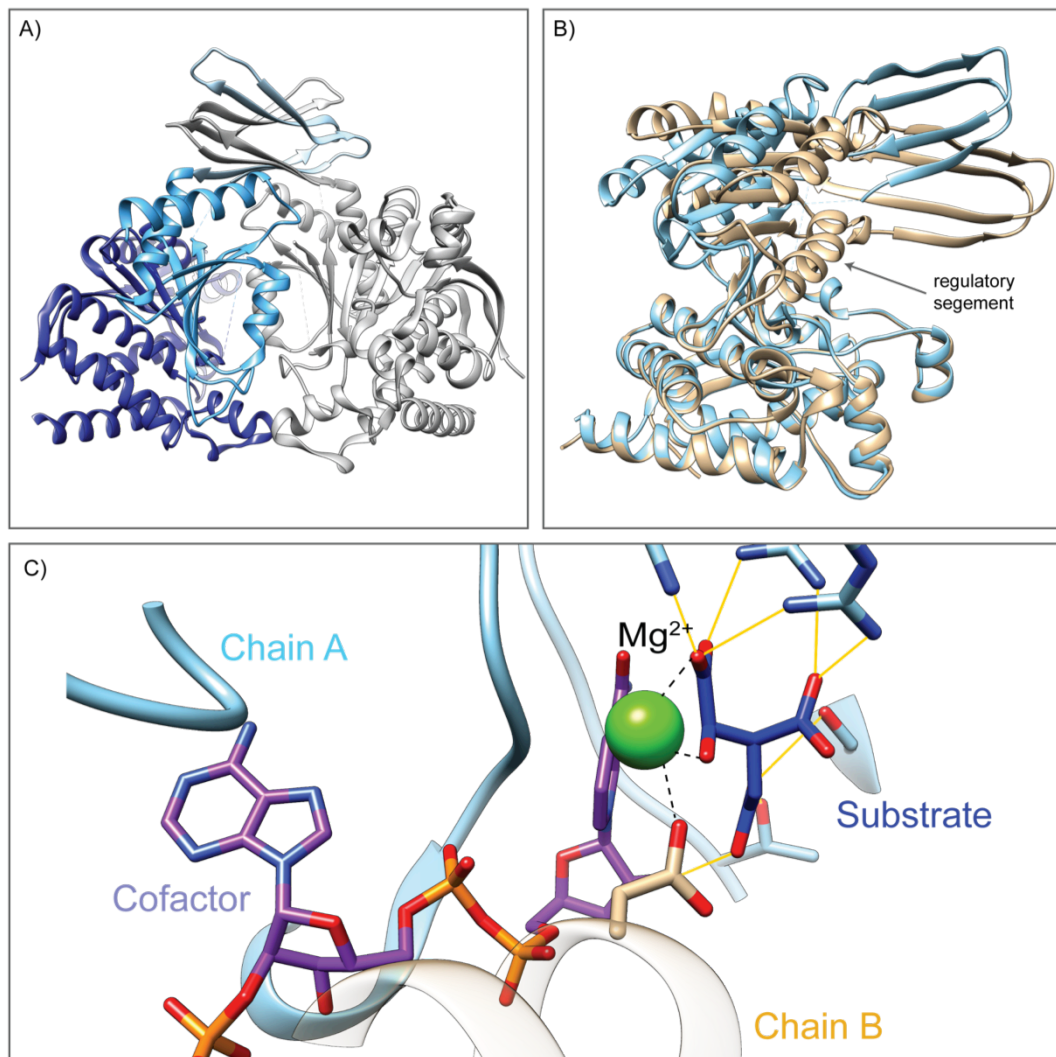


Figure 1.4: Structure of IDH1. A) IDH1 dimer, with the large domain of one monomer coloured in blue, the small domain in light blue, and the clasp domain in cyan. The second monomer is coloured grey for clarity. B) Binding of catalytic metal and substrate causes a large conformational change from the inactive (cyan) to active (tan) conformation. The small and clasp domains move relative to the large domain, with the regulatory segment adopting an  $\alpha$ -helical conformation. Monomer shown for clarity. C) The formation of the primary site involves residues from both chains. Figures made in Chimera<sup>65</sup>



### 1.8.2 IDH1 as a cancer therapeutic target

In IDH1-WT glioblastoma, IDH1-WT expression is up-regulated, resulting in increased production of NADPH, leading to the reduction of reactive oxygen species (ROS)<sup>66</sup>. Knock-out of IDH1-WT in glioblastoma cells results in increased sensitivity to radiation-induced senescence, and increased responses to fractionated radiotherapy in murine xenograft models<sup>67</sup>.

Heterozygous missense mutations in IDH1 are identified in up to 80% of glioma patients and 15% of AML patients, with an arginine to histidine substitution at residue 132 (IDH1-R132H) the most frequently observed<sup>68</sup>. Other substitutions at this position are observed with greater frequencies in other solid tumours, such as cholangiocarcinoma, prostate cancer and colorectal cancer (Figure 1.5). In addition, somatic mosaic mutations in IDH1 are known to cause both Ollier's Disease and Maffucci Syndrome, both characterised by multiple cartilaginous tumours<sup>69</sup>. IDH1-R132H mutations can sensitise cells to oxidative stress and PARP inhibition, which can be reversed upon treatment with IDH1-R132H inhibitors<sup>70, 71</sup>. The presence of IDH1-R132X mutations is favourable for patients with glioblastoma<sup>72</sup>, but is associated with decreased survival for patients with AML<sup>73</sup>.

Mutations at position 132 results in the loss of wild type IDH1 activity, and causes neomorphic conversion of  $\alpha$ KG to D-2-hydroxyglutarate (2HG), using NADPH as a reducing agent<sup>74</sup>. Under normal conditions, 2HG levels in cells are maintained by endogenous D-2-hydroxyglutarate dehydrogenase (D2HGDH), which catalyses the reverse reaction, converting 2HG back to  $\alpha$ KG<sup>75</sup>. The neomorphic activity of IDH1 mutants causes the accumulation of

2HG; elevated 2HG levels can be detected in the serum of patients with IDH1 mutant AML and glioma<sup>76</sup>.

2HG is structurally similar to  $\alpha$ KG and can competitively inhibit many  $\alpha$ KG-dependent dioxygenases, including JmjC histone demethylases<sup>77</sup>, prolyl hydroxylases<sup>78</sup> and 5-methylcytosine hydroxylases<sup>79</sup>, as well as alkylated DNA repair proteins<sup>80</sup>. This leads to global changes in both histone and DNA methylation patterns<sup>77, 81</sup> and the accumulation of DNA damage<sup>71</sup>. Production of 2HG by IDH1-R132H is sufficient to promote leukemogenesis, which can be reversed with inhibition of the IDH1-R132H<sup>82</sup>.

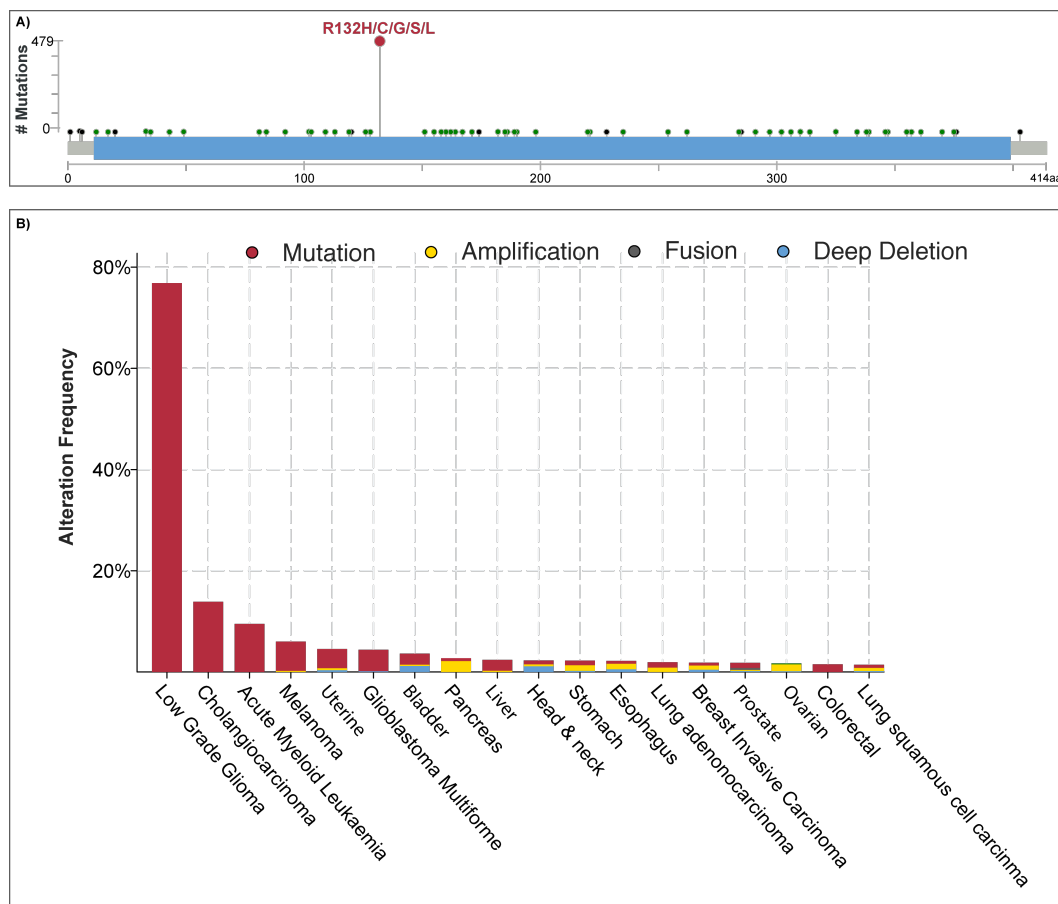


Figure 1.5: IDH1 mutations common in cancer. A) Distribution of mutations across IDH1. The vast majority of mutations are localised to arginine 132. B) Missense mutations are most frequently identified in Low Grade Glioma, but are also found in other tumour types at lower frequencies. Mutational data from TCGA: <https://www.cancer.gov/tcga>, visualisation adapted from cBioPortal<sup>8, 31, 83</sup>.

Although IDH1-R132H mutations are almost always heterozygous, which is common for oncogenic mutations<sup>84</sup>, patients with homozygous IDH1-R132H mutations have also been reported<sup>85</sup>. This indicates that loss of IDH1-WT is not lethal to the cell, which may be due to the presence of IDH2-WT. Although they have different subcellular localisations, IDH1-WT and IDH2-WT catalyse the same reaction. The IDH2-R172X and IDH2-R140X mutations are analogous to IDH1-R132X, resulting in the production of 2HG<sup>86</sup> and tumorigenesis. IDH1-WT and IDH2-WT may have functional redundancy that allows cells to survive complete loss of IDH1-WT activity. This is supported by the mutual exclusivity of IDH1 and IDH2 mutations<sup>87</sup>.

In patients with heterozygous IDH1-R132X mutations, approximately 50% of the IDH1 population *in vivo* will be as part of an IDH1-WT/IDH1-R132H heterodimer, assuming equivalent expression levels and non-discriminatory dimer formation. As IDH1-WT converts isocitrate to  $\alpha$ KG in the heterodimer, the local concentration of  $\alpha$ KG by the IDH1-R132H primary site is increased. *In vitro* studies of the IDH1-WT/R132H heterodimers show an increased production of 2HG in comparison to both the WT/WT and R132H/R132H homodimers<sup>88</sup>. Further, glioblastoma patients with homozygous IDH1-R132H mutations have approximately 14-fold lower mean 2HG in comparison to the patients bearing heterozygous IDH1-R132H mutations<sup>85</sup>. The increased local concentration of  $\alpha$ KG in close proximity to the IDH1-R132H primary site within WT/R132H heterodimers may decrease the efficacy of  $\alpha$ KG-competitive inhibitors.

Both IDH1-R132H and IDH1-WT are clinically relevant therapeutic targets. While current inhibitors show selectivity for IDH1-R132H over IDH1-WT, the impact of targeting both IDH1-WT and IDH1-R132H is unknown. The IDH1-R132H inhibitor ivosidenib<sup>39</sup> from Agios was granted FDA approval for treatment of refractory AML in July 2018. Although this compound is showing impact in patients, the first resistance mutation, S280F was also reported in July 2018<sup>89</sup>, showing the need for new targeted therapeutics. As the mutated residue is not located in the novel pocket, targeting the novel site may overcome resistance mutations and offer an alternative approach to inhibiting this clinically important protein.

## Chapter 2: *In silico* identification of ligandable sites

---

### 2.1 Introduction

Target ligandability refers to the likelihood of binding a small molecule to the site. Structure-based ligandability predictors are widely used for prioritisation of pockets during early stage evaluation of clinically relevant protein targets<sup>90</sup>. They generally have three components – a pocket identification method, a training set of pockets with known outcomes, and a discriminating function.

#### 2.1.1 Pocket identification methods

Multiple computational tools exist that identify pockets in protein structures. These can be grouped into energy-based and geometry-based methods. Energy-based methods, such as Q-SiteFinder<sup>91</sup>, calculate the interaction energy between the protein and a probe. Ligand-binding sites are predicted as regions of protein with clusters of favourable interaction sites<sup>92</sup>. Geometry-based methods include sphere-based methods such as SURFNET<sup>93</sup>. SURFNET places a sphere between two atoms, touching each one. If atoms from any neighbouring residue intrude into the sphere, then the sphere size is reduced until no atoms intrude. If the resulting sphere is less than a certain volume, it is rejected. This is repeated for all pairs of atoms. Pockets are defined as clusters of spheres and are reported as a surface contour in three dimensions. After pockets have been identified, a range of properties can be calculated. The properties that are calculated vary between different predictors, but tend to include geometric, physical and chemical properties such as volume, enclosure, and hydrophobicity of identified pocket.

### 2.1.2 Training predictors

Predictors are used to classify datasets according to a desired feature. Development of a predictor requires a positive training set formed of examples that have the desired feature and a negative training set formed of examples that do not have the desired feature. These sets are used to develop a decision rule that can then be applied to a new dataset – the test set.

In this context, the desired feature is ligandability. The positive training set is formed of pockets experimentally shown to be ligandable, although the pockets included varies between different predictors. The positive training set may be redundant, including multiple examples of the same pocket, or non-redundant and include only one example of the ligandable pocket. The negative training set may be formed of pockets considered to be 'less druggable'<sup>94</sup>, or simply pockets not in the positive training set. The two training sets of pockets and their associated properties are then used to build ligandability predictors, often through different machine learning approaches. For example PockDrugs<sup>95</sup> uses linear discriminant analysis while canSAR3D<sup>49</sup> uses a decision tree. The success of these predictors is benchmarked against a pre-defined set of pockets which were not used for training<sup>96</sup>.

Pocket volume is invariably found to be a primary determinant of ligandability. The majority of examples of druggable pockets are primary sites, which tend to be the largest and most geometrically complex pocket in the protein<sup>97</sup>. This leads to an implicit bias for the largest site being predicted as the most ligandable.

### **2.1.3 canSAR3D structure-based ligandability predictor**

The canSAR3D pipeline analyses all publically available protein structures. Up to ten pockets are identified in each separated chain in each PDB structure by SURFNET<sup>93</sup>, and 25 properties are calculated for each. These properties cover geometric features such as pocket volume and enclosure, as well as chemical properties such as the ratio of hydrophobic and polar residues in the pocket.

The pocket definition is subsequently refined based on sequence conservation by the PickPocket algorithm. A multiple sequence alignment using the target protein sequence is calculated by ClustalW<sup>98</sup>. Each residue in the sequence is given a score, which is calculated as the sum of the pair-wise residue similarity scores between the sequences and weighted for evolutionary distance between the sequences<sup>99, 100</sup>. Each sphere in the initial Surfnets cavity is subsequently given a score by summing the conservation scores of all residues in the protein with a weighting function that drops off rapidly with distance, such that close residues intruding on the sphere are given a higher weight. The origin of the refined pocket is calculated by selection of the sphere in the original pocket with the highest score. The scores of the surrounding spheres are also weighted by their distance from the origin sphere and compared to a minimum allowed score of 0.4, with spheres scoring higher than this included in the pocket. The overall pocket conservation score is likewise reported as a sum of the individual sphere scores weighted for distance from the origin. Weighting of the score by distance helps to maintain the 'pocket-like' nature of each identified pocket instead of allowing connection of multiple pockets through narrow channels.

Following calculation of the refined pocket, the properties are recalculated and reported alongside the conservation score. A decision tree machine-learning algorithm from ChEMBL Strudel<sup>10, 101, 102</sup> is then used to predict ligandability based on these properties, trained on a large and diverse set of known druggable sites. As with other predictors, pocket volume is a strong predictor of ligandability due to the inherent bias in the training set, but canSAR3D can overcome this bias in some instances.

Comparison of the pocket properties before and after refinement by PickPocket shows that primary sites show a large change in the pocket definition (Figure 2.1). In contrast, secondary sites tend to show much less refinement as the PickPocket algorithm cannot find a sufficient number of conserved residues in the secondary site to refine the pocket. This may be due to lower levels of sequence conservation in secondary site, or due to the smaller initial size of the secondary sites resulting in too few conserved residues being identified.

Predicting the presence of ligandable secondary sites is challenging due to the lack of examples of druggable secondary sites. Pocket identification methods are capable of identifying secondary sites in protein structures, but they are often considered less ligandable and subsequently deprioritised due to their smaller size in comparison to the primary sites. Consequently, structure-based ligandability predictors often perform very well when benchmarking with datasets formed of primary sites. The tendency to prioritise the largest pocket in the protein as the most ligandable leads to limited success detecting secondary sites.



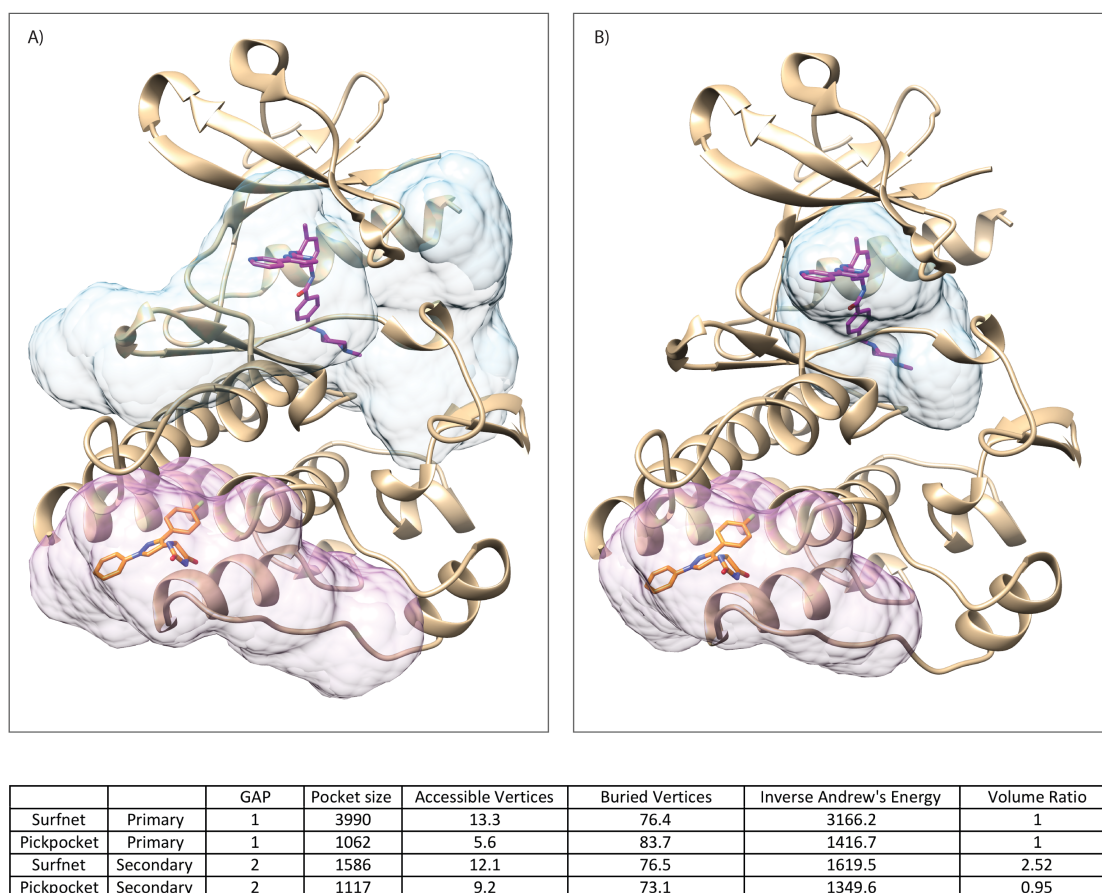


Figure 2.1: Primary and secondary sites are refined differently by the PickPocket algorithm as part of the canSAR3D pipeline. A) ABL (PDB 3PYY) and associated primary (blue) and secondary (pink) sites as defined by Surfnets. B) Pockets following refinement by PickPocket. C) Primary sites show much greater refinement by PickPocket than secondary sites. Using the PickPocket associated properties may therefore increase the bias for primary site-like pockets. Figures made in Chimera<sup>65</sup>

A large-scale analysis of crystallographic fragment screening data reported by Ludlow *et al.*<sup>28</sup> compared the distribution of physical and chemical properties between fragment-bound primary and secondary sites. They found that while the primary sites were in general larger than the secondary sites, the distributions of other pocket properties, such as number of polar contacts and atom mobility, were similar. This supports the assumption that properties important for ligandability, excluding volume, are conserved regardless of whether the pocket is a primary or a secondary site. Therefore, structure-based ligandability predictors based on primary sites may accurately predict

secondary site ligandability, if the tendency to prioritise the largest pocket as the most ligandable can be overcome.

This chapter describes the adaptation of canSAR3D in order to identify novel ligandable secondary sites in cancer-associated proteins. I used canSAR3D rather than other available predictors as it has precedence for identifying secondary sites. Furthermore, all structures in the PDB were analysed, with up to ten pockets identified and analysed per chain, yielding a wealth of data upon which to train the predictor. The new predictor was then used to identify ligandable secondary sites, which were triaged for biological deregulation, experimental feasibility and clinical relevance, leading to the selection of a novel site in IDH1-R132H as an initial target for experimental investigation (Figure 2.2).

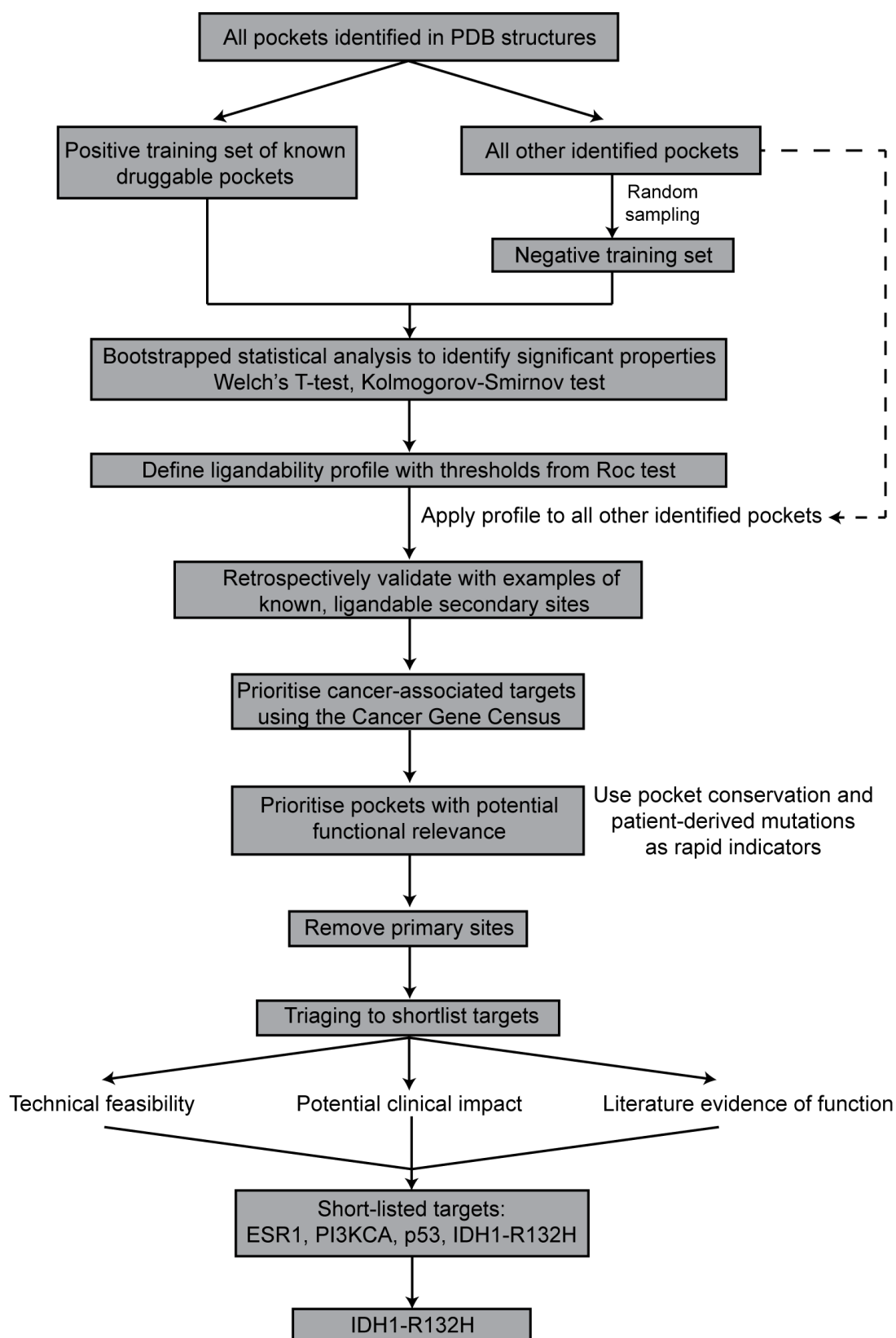


Figure 2.2: Flowchart out-lining the approach to identify novel, ligandable and functionally relevant secondary sites in cancer-associated proteins.

## 2.2 Results

### 2.2.1 Defining the training sets

At the time of analysis, more than 30,000 crystal structures of 5795 human proteins had been deposited in the PDB. This covers approximately 30% of the proteome as defined by Swiss-Prot<sup>103</sup>. As I was interested in targeting human proteins implicated in human cancers, I only considered human proteins. From the available structures, a total of 528,441 pockets were defined and had their properties calculated by the canSAR3D pipeline. I decided to use the SURFNET pocket definitions and associated properties to prevent any additional biasing towards primary sites (Section 2.1.3).

Positive and negative training sets were defined in order to identify which of the underlying pocket properties are important for secondary site ligandability. Ideally, the positive training set would have been formed of validated, druggable secondary sites, but there remain too few examples of these to build a robust training set. Based on the assumption that properties important for ligandability will be consistent whether a primary or secondary site is targeted, I formed the positive training set of 2,025 ligandable pockets. This included the catalytic sites of kinases and the ligand-binding sites of nuclear receptors, both of which are considered to be highly ligandable despite challenges in achieving target selectivity. I also included the binding sites of FDA-approved drugs from all species (Figure 2.3). In many cases, there were multiple examples of the same pocket with different ligands bound. For example, a kinase structure may be solved with ATP or approved drug bound to the primary site. Using different crystal structures for the same target can show variations for the calculated

ligandability, which can be attributed to flexibility in the protein<sup>104</sup>. Including multiple structures with different ligands and in different ligandable conformations gives a more robust representation of the ligandable pocket. These structures formed the redundant positive training set.

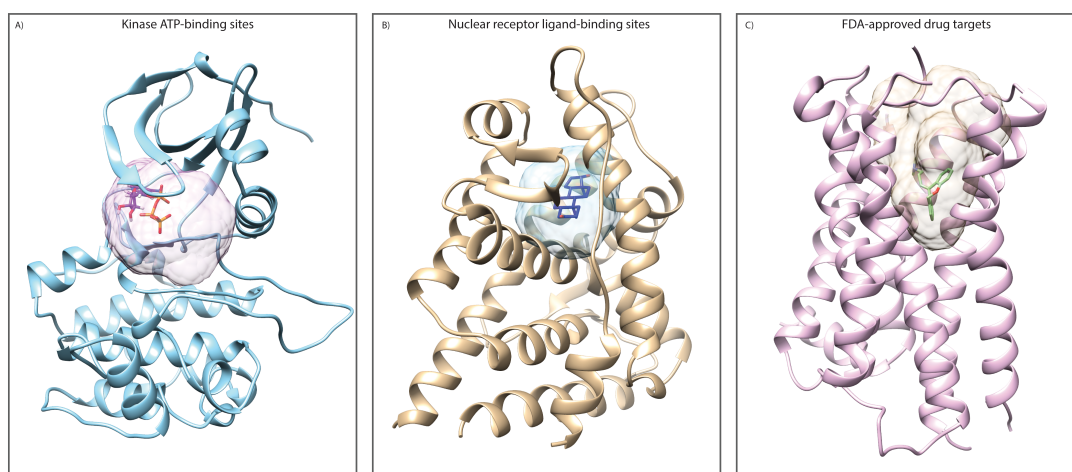


Figure 2.3: Examples of pockets included in the training set. A) ATP-bound kinase sites such as PDK1, PDB 4XX9; B) Ligand binding sites in nuclear receptors such as the testosterone binding site in the androgen receptor, PDB 2AMA; C) binding sites of FDA-approved drugs, such as Doxepin binding site in the histamine receptor  $\alpha$ H1, PDB 3RZE. Figures made in Chimera<sup>65</sup>

In addition, I chose not to enrich the positive training set with the few known examples of ligandable secondary sites. If there were a true difference in pocket properties between primary and secondary sites, the number of secondary sites included in the training set would be too small for the statistical analysis to identify these as a separate population. Importantly, retaining the secondary sites in the test set allows them to be treated as internal controls for validation of the predictor.

After defining the positive training set, 526,416 pockets with associated properties remained which formed the background set. The majority of these have unknown ligandability. Due to the size discrepancy between the positive

and background sets, I formed the negative training set by randomly sampling from the background set.

### **2.2.2 Statistical considerations for developing the secondary site ligandability predictor**

Using druggable primary sites to form the positive training set requires the assumption that the properties important for ligandability are the same regardless of which site on the protein is being targeted. However, training the predictor on primary sites leads to an inherent bias for the largest and most geometrically complex pocket in the protein. Properties that show a statistically significant difference because they are important for ligandability must be deconvoluted from those showing significant differences due to the training set bias.

The positive training set is formed of 2,025 pockets. The negative training was randomly sampled from the background set. The population distributions for these sets are unknown: they cannot be guaranteed to be normal (Gaussian). I therefore used two statistical tests, Welch's t-test and Kolmogorov-Smirnov (KS) test, to identify the pocket properties that showed a statistically significant difference between the positive and negative training sets.

#### ***2.2.2.1 Bootstrapped Welch's t-test***

Welch's t-test is a two-sample location test under the null hypothesis that the two population means are equal but the population variances are different. It is a variation of the more commonly used Student's t-test that is more reliable

when populations have unequal sample sizes and/or unequal variances, as is likely the case with these datasets.

T-tests assume the normality of the underlying distribution, which is unknown for these populations. Bootstrapping can help to overcome a non-parametric distribution based on the central limit theorem. Samples are taken from both the positive and negative training sets and a Welch's t-test is performed; both the sample mean and the p-value are recorded, and the sample is replaced. A total of 100,000 samples were taken and replaced, with the aggregate p-value reported as the proportion of the time the null hypothesis could not be rejected based on a significance cut-off of  $p \leq 0.05$ .

#### **2.2.2.2 *Kolmogorov-Smirnov test***

A KS statistic quantifies the distance between the empirical distribution of two samples, in this case the positive and negative training sets, under the null hypothesis that the two samples are drawn from the same empirical distribution. A KS test indicates whether the two samples are likely to be different. Unlike the Welch's T-test, it does not assume that the data is derived from a parametric distribution, and therefore can be used even where the data doesn't tend towards a normal distribution. The test was bootstrapped 100,000 times, with sample replacement, and the p-value and sample means recorded. The aggregate p-value was reported as the proportion of the time the null hypothesis could not be rejected based on a significance cut-off of  $p \leq 0.05$ .

Good correlation was observed between the two statistical tests, with eight properties identified as showing a statistically significant difference between the

positive and negative training sets (Table 2.1, Appendix 8.1.2). These are: Inverse Andrew's Energy, Pocket Volume, Buried Vertices Ratio, GAP, Volume Ratio, Accessible Vertices Ratio, PCA X and PCA Y. A description of these properties can be found in Table 2.1.

Both Welch's t-test and the KS test identify which properties show a statistically significant difference, but do not define thresholds to separate the populations. I used a Roc test to define these thresholds.

### **2.2.2.3 Roc tests**

Receiver operator characteristic (Roc) curves are non-parametric tests that calculate how well a given diagnostic can predict the binary classification as the threshold is varied. In this case, the diagnostic is the pocket property, and the classification is the ligandability. The area under the curve when the true positive fraction is plotted against the false positive fraction (AUC) is used to measure the accuracy of the prediction. This test was bootstrapped 10,000 times with sample replacement. Fewer bootstraps were used than previously due to the increase in processing power demand. An AUC of 80% was selected as a cut-off for significance. There was good correlation between the properties identified as showing a statistically significant difference between positive and negative training sets between the different statistical test (Table 2.2, Appendix 8.1.2).

Roc tests automatically report the best threshold, dependent on the pre-defined method. The R package used, pRoc<sup>105</sup>, has two alternative methods for identifying the 'best' threshold. Youdon's statistic maximises the distance from



the line of no discrimination, while the closest top left considers the optimal threshold to be the closest to perfect specificity and sensitivity - the point on the curve closest to the top-left of the plot. Both of these can be weighted dependent on the relative cost of false negative to false positive classification. For the purpose of triaging and experimental follow-up, a false positive is far more detrimental than a false negative, so I selected the closest top left method to favour specificity. Table 2.2 shows the thresholds calculated for the eight properties.

#### **2.2.2.4 Defining the ligandability profile**

The statistical analyses identified eight SURFNET properties that were statistically significant across at least two of the three statistical tests. While the majority of these properties are well-accepted descriptors of ligandability, such as pocket volume and enclosure<sup>106</sup>, the inverse Andrew's energy is a new descriptor. Andrew's energy is a theoretical maximum binding energy of a given small molecule if it's shape is complementary to the binding site<sup>107</sup>. It is based on the average binding energy of common functional groups as calculated by Andrews's *et al*<sup>108</sup>. The inverse Andrew's energy therefore is the theoretical binding energy achievable for a pocket if all of its side chains are involved in productive interactions with a small molecule partner, and is unique to the canSAR3D predictor. Whilst both PCA Y and PCA X, the lengths of the X- and Y-axis of the pocket, show statistically significant differences between positive and negative training sets, they are directly related to the pocket volume, and so were not used to define the ligandability profile.

Property	Definition	Threshold
Inverse Andrew's Energy	Andrew's energy is a theoretical maximum binding energy of a given small molecule if each chemical moiety is involved in productive binding with its target <sup>107</sup> . The inverse Andrew's energy therefore is the theoretical binding energy of a pocket if all of its side chains are involved in productive interactions with a small molecule partner.	$\geq 910$
Pocket Volume	The pocket volume is the calculated volume of the pocket based on the SURFNET sphere-rolling model in $\text{\AA}^3$ .	$\geq 750$
Buried Vertices Ratio	The property refers to how enclosed the pocket is; it is calculated using a series of points within the pocket from which projecting rays of fixed length are projected. The reported ratio is the proportion of points that have a large majority of rays contacting protein. It is significantly impacted by the size and shape of the pocket; for example, a small pocket would be calculated to have a greater enclosure than a large pocket of the same depth, as there would be fewer points in the centre of the pocket contacting only solvent.	$\geq 70$
GAP	Combined score of size and geometric complexity of a pocket. The GAP=1 pocket is the largest and most complex.	$\leq 3$
Volume ratio	Comparison of the volume of the pocket of interest in comparison to the largest and most geometrically complex pocket.	$\leq 3$
Accessible Vertices Ratio	This property refers to how open the pocket is; it is calculated in a similar way to Buried Vertices Ratio but considers rays that contact solvent.	$\leq 13.8$
PCA Y	This refers to the length of the principal Y-axis of the pocket. It is therefore linked to the pocket volume. It is not derived from a principal component analysis.	-
PCA Z	It refers to the length of the principal Z-axis of the pocket. It is therefore linked to the pocket volume.	-

Table 2.1: Pocket properties identified as being statistically significant and the thresholds used. PCA Y and PCA Z were not used for the profile due their link with the pocket volume.

To ensure that the remaining pocket properties had not been identified through chance, I repeated the statistical tests with randomisation of positive and negative training sets. None of the identified properties showed a statistically significant difference between the randomised sets (Table 2.2)

The ligandability profile was defined using the statistically significant properties and the thresholds provided by the Roc test as shown in Table 2.1.

Property	Welch's T-test P-value		KS test P-value		ROC test AUC %	
	Real	Randomised	Real	Randomised	Real	Randomised
Accessible Vertices	$9 \times 10^{-5}$	0.95	$9.1 \times 10^{-3}$	0.97	80.4	49.6
Andrew's Energy	$1.4 \times 10^{-2}$	0.96	0	0.97	88.9	51.0
Buried Vertices	0	0.95	0	0.97	91.7	50.5
GAP	$1.8 \times 10^{-4}$	0.95	$8.9 \times 10^{-3}$	0.99	81.4	50.5
Pocket Volume	$5.9 \times 10^{-2}$	0.96	$1.2 \times 10^{-3}$	0.97	81.6	49.5
Volume Ratio	$2.8 \times 10^{-2}$	0.95	0	0.97	81.0	49.8

Table 2.2: Summary of statistics of properties identified as showing statistically significant differences between the training and background set, and their associated randomised trial.

#### 2.2.2.5 *Use of p-values for significance testing*

The p-value is the probability that upon repetition of the experiment, the same or a more extreme result would be reported - how likely is it that this result would be seen if the Null Hypothesis were true. The standard significance cut off of  $p \leq 0.05$  means that if the Null Hypothesis were true, then these results or a more extreme result would be expected less than 5% of the time.

The p-value has become controversial over recent years<sup>109-112</sup>. In this case, p-values are used in combination with AUC to identify properties that may be important in ligandability. Both the p-values and the AUCs are supported by randomisation of the test and training sets. Further, the prediction can be

validated both by the use of known, ligandable secondary sites contained in the test set, and through experimental validation of a novel target.

### **2.2.3 Identifying ligandable secondary sites**

The ligandability profile defined through the statistical analysis was then used to identify ligandable secondary sites from within the background set of 526,416 pockets. The predictor identified 6,712 ligandable pockets in 1,391 proteins. In some instances, the same pocket was identified in multiple structures of the protein, while in others multiple pockets were identified in the same protein. Of these, 16 were known and validated ligandable secondary sites (Table 2.3). Overall, the inclusion of multiple known examples gives confidence in the predictor.

### **2.2.4 Triaging pockets for target selection**

#### ***2.2.4.1 Prioritisation of cancer-associated proteins***

In order to restrict the targets to cancer-associated protein, I used the 564 targets from Cancer Gene Census (CGC)<sup>113</sup> from COSMIC<sup>12</sup> and the 127 Significantly Mutated Genes (SMGs) from Kandoth et al.<sup>13</sup>. The CGC is an expert-curated database of proteins with mutations that are known to drive carcinogenesis, while the SMGs were identified through a large-scale analysis of TCGA. Across the two sets, 621 cancer-associated proteins were identified, with 80 found in both the CGC and as an SMG. Pockets identified in these cancer-associated proteins were retained. This step excluded most of the identified pockets, leaving only 696 pockets from 103 proteins, 10% and 7% of the ligandable pockets and proteins respectively.

Protein	PDB code	Ligand ID	Pocket description
<b>Caspase-7</b> <sup>114</sup>	1SHJ	NXN	Conserved allosteric site, 14A from active site. Binding of small molecule inactivates protease
<b>Fructose-1,6-bisphosphatase</b> <sup>115</sup>	3IFC	AMP	AMP binding site, involved in negative feedback loop
<b>Fructose-1,6-bisphosphatase</b> <sup>116</sup>	2WBB	RO3	As above
<b>Integrin <math>\alpha</math>-L</b> <sup>117</sup>	3M6F	BJZ	I-domain allosteric site; binding prevents conformational changes required for activity
<b>*Abl-1</b> <sup>118</sup>	3PYY	3YY	Myrisotyl binding site
<b>E2-R1</b> <sup>119</sup>	4MDK	U94	Pocket at interface of E2 and ubiquitin, stabilises interaction.
<b>Eg5</b> <sup>120</sup>	4BXN	6LX	Ispinesib binding site; prevents conformational changes required for activity,
<b>Farnesyl pyrophosphate synthase</b> <sup>121</sup>	5DIQ	5B9	Adjacent to IPP (substrate) binding site, at C-terminus
<b>Farnesyl pyrophosphate synthas</b> <sup>122</sup>	3N3L	MS0	As above
<b>Phosphodiesterase 4d</b> <sup>123</sup>	3G4I	D71	Stabilises binding of regulatory domain across active site
<b>Phosphodiesterase 4d</b> <sup>123</sup>	3IAD	15X	As above
<b>*Inducible T-cell Kinase (ITK)</b> <sup>124</sup>	4M14	QWS	Adjacent to, but not overlapping primary site in inactive conformation
<b>*Inducible T-cell Kinase (ITK)</b> <sup>124</sup>	4M15	QWS	As above
<b>Adenylate cyclase (solAC)</b> <sup>125</sup>	4OYA	1VE	Bicarbonate binding site
<b>Hexokinase-3</b> <sup>126</sup>	3HM8	BG6	Example from allosteric site database
<b>*K-Ras</b> <sup>23</sup>	4LUC	20G	Target G12C mutant; binding prevents communication between switch I and II
<b>*K-Ras</b> <sup>23</sup>	4LV6	20H	As above
<b>Erk5</b> <sup>127</sup>	5BYZ	4WE	Adjacent to primary site; binding displaces P-loop into primary site
<b>Erk5</b> <sup>127</sup>	4ZSJ	4R0	As above
<b>*Mek1</b> <sup>128</sup>	3MBL	LSG	Adjacent to primary site
<b>P53</b> <sup>129</sup>	5AOI	RZH	Bind to Y220C mutant specific pocket
<b>*IDH1</b> <sup>130</sup>	1T09	-	At dimer interface, competes with binding for catalytic cation

Table 2.3: Examples of known, ligandable secondary sites initially predicted to be ligandable by the computational predictor. \*Indicates a secondary site that was retained throughout the triaging process.

#### **2.2.4.2 *Triaging for potential functional relevance.***

For a pocket to be druggable by small molecule approaches, binding of the small molecule must have an impact on protein function and eventually on the disease state. Secondary sites may be associated with protein or cofactor binding, allosteric control of the primary site, or may be non-functional. There is currently no consensus approach for predicting functional relevance in secondary sites. Analyses aiming to investigate this tend to be based on either putative sites or on small datasets. With still a large number of potential targets, the pockets were triaged based on sequence conservation and mutation mapping from the over 10,000 patient cohort in TCGA, that were used together as rapid indicators of functional relevance.

The work previously discussed by Ludlow *et al.*<sup>28</sup> showed that the secondary sites identified by crystallographic fragments screening showed greater sequence conservation in comparison to the global sequence conservation between species orthologs, but to a lesser extent than observed in primary sites. The extent of secondary site sequence conservation is still under debate in the field. A pocket conservation score is calculated for each pocket automatically during the refinement of the SURFNET-identified pockets within the canSAR3D pipeline. I used sequence conservation as a rapid indicator of functional relevance, prioritising pockets with a sequence conservation of 70% or greater (Appendix 8.1.4).

The Allosteric Site Database (ASD)<sup>131</sup> collates sites in proteins involved in the allosteric control of protein function. Some of the entries are single residues implicated in the allosteric control of proteins, while others are enclosed

pockets, such as the myristoyl-binding site in Abl<sup>132</sup>. The ASD does not curate for ligandability. Shen *et al.*<sup>133</sup> reported an enrichment of deleterious mutations around primary and allosteric sites in comparison to tolerated mutations. This enrichment was not observed when considering non-functional sites. Further, they report an enrichment of patient-derived mutations around known allosteric sites in cancer-associated proteins in comparison to the overall mutation rate. Work by Dr. Al-Lazikani also shows a small but statistically significant enrichment of patient-derived mutations around allosteric sites in comparison to primary sites (personal communication). Based on these finding, I mapped normal tissue-matched mutations from the TCGA onto the identified pockets as a further indicator of functional relevance.

Of the 696 pockets in 103 proteins identified in cancer-associated proteins, only 273 pockets from 61 genes had associated mutations. For these, a mutation enrichment score was calculated (Equation 2.1). A score of one shows equal rate of mutation in comparison to the background protein, a score of less than one indicates a depletion of mutations and a score of greater than one indicates enrichment in comparison to the global.

$$Score = \frac{cavity\ mutation\ rate}{protein\ mutation\ rate} \quad \text{Equation 2.1}$$

Where

$$cavity\ mutation\ rate = \frac{\#mutations\ in\ cavity}{\#residues\ in\ cavity}$$

and

$$protein\ mutation\ rate = \frac{total\ mutation\ frequency}{length\ of\ protein}$$

Combining the conservation score and the mutation analysis leaves 247 pockets from 56 genes, which have a sequence conservation of greater than 70% and at least one associated mutation. In some cases, identified pockets are associated with high mutation rates due to the presence of a recurrent variant but do not show high levels of sequence conservation. For example, the recurrent Y220C mutation in p53 results in formation of a novel secondary site and subsequent destabilisation of the protein structure<sup>134</sup>. Small molecules targeting this site stabilise the protein fold<sup>135</sup> – it is both ligandable and functionally relevant. While my computational analysis predicts the pocket to be ligandable (Table 2.3), and shows mutation enrichment due to the recurrence of the Y220C mutation, it has low levels of sequence conservation and was therefore not taken forward as a pocket likely to show functional relevance.

#### **2.2.4.3 Target shortlisting**

The predictor identified multiple primary sites. These were manually removed, leaving 150 pockets in 40 proteins (Appendix 8.1.3). Of these, five were known and validated ligandable secondary sites in cancer-associated proteins (Table 2.3, proteins marked with an \*).

With 2.8% of the original 6855 pockets remaining, it then became feasible to look at each pocket in greater detail to prioritise targets further. Prioritisation was based on the evidence of functional relevance, reliability of the prediction, potential clinical impact and disease association, as well as the technical feasibility of the target.

Reliability of the prediction considers the quality of the structures predicted to be ligandable, which was assessed using the statistics reported in the PDB as



calculated by MolProbity<sup>136</sup>. Structures where the rate of Ramachandran outliers is greater than 1% in the chain of interest, or with poor fit to electron density around the pocket of interest as shown by RSRZ value greater than 2, as well as those with resolutions lower than 3.5 Å were excluded. The proportion of comparable structures, where the protein adopts the same conformation, in which the pocket is predicted to be ligandable was also considered. Recurrent predictions in high quality crystal structures may be more reliable.

Potential for clinical impact had already been considered through application of the CGC and SGMs, and conservation scores and mutation enrichment had been used as a rapid indicator of potential functional relevance. However, the mutation profile of the target is also an important consideration for drug discovery. High mutation rates associated with a single recurrent alternation may be indicative of biological relevance. They are also lower risk for drug discovery as the cancer-associated, recurrent variant can be targeted and offer potential tumour specificity. High mutation rates not associated with high recurrence can be more challenging as there is no clear variant to target, and resistance may emerge more rapidly. Given the lack of consensus around predicting functional relevance around secondary sites, literature evidence of function was used to prioritise those with more evidence of functional relevance and disease association.

The technical feasibility of the target considers how likely reliable experimental systems can be established in house. As structural studies will be required to identify binding sites following fragment screening, targets with straightforward

expression systems such *E. coli* and multiple high resolution crystal structures from different groups deposited in the PDB are indicative of a more amenable system. Further, the availability of tool compounds and published assays aids in the development of assays.

Four novel pockets in ESR1, p53, PIK3CA and IDH1 were identified as potential initial targets for validation. Table 2.4 shows a summary of these sites and some of the associated considerations.

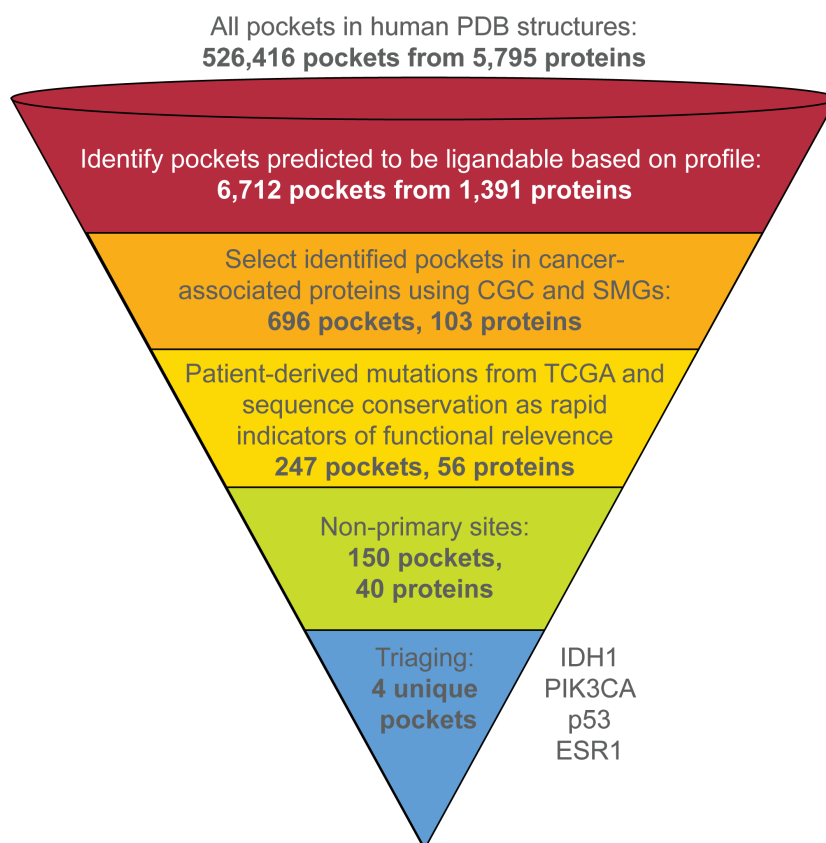


Figure 2.4: Overview of the triaging process used to shortlist the four potential targets for experimental investigation.

Protein	ESR1	IDH1	TP53	PIK3CA	PIK3CG
Family	NUCLEAR RECEPTOR	DEHYDROGENASE	TRANSCRIPTION FACTOR	PI3/PI4 KINASE	
Average Resolution	2.2 Å	2 Å	2.5 Å	2.7 Å	
Expression System	<i>E. coli</i>	<i>E. coli</i>	<i>E. coli</i>	<i>Spodoptera frugiperda</i>	
Druggable snapshots	50	1	1	3	24
Total number of structures	193	18	162	21	89
Mutation enrichment in the pocket	1.4	0.7	0.9	0.19	0.87
Sequence conservation in the pocket	0.96	0.93	1	0.83	0.96
Literature evidence and comments	No literature evidence. Close to ligand binding site; not AF-2. Also predicted druggable in AR and PGR (lost during mutation mapping)	Literature evidence: potential regulatory site. Changes conformation between active and inactive	Involved in aggregation. Formed by steric zipper region required for aggregation	No literature evidence. Some structures of PIK3CA show pocket occluded by His-tag. Suggests peptide binding function. Conserved between alpha and gamma isoforms	

Table 2.4: Overview of shortlisted targets; number of structure was correct when the initial computational assessment was completed, but will have since increased as more crystal structures are deposited in the PDB. The mutation enrichment was calculated specifically considering residues forming the pocket as described in equation 2.1. The sequence conservation is calculated as part of the canSAR3D pipeline.

#### **2.2.4.4 Shortlisted targets**

The nuclear transcription factor p53 regulates cell cycle arrest and apoptosis in response to DNA damage, and is tightly regulated by a network of post-translational modifications. It is the most frequently mutated protein in human cancer, with over 50% of tumours harbouring an inactivating mutation<sup>137</sup>. Contact mutations either abrogate the ability of p53 to bind to DNA or change the target-binding site, while stability mutations destabilise the protein core causing aggregation. Further, negatively regulating proteins such as MDM2 and MDMX are overexpressed in many tumour types, leading to increased p53 ubiquitylation and inactivation, which is an important step in tumourogenesis as it allows cancer cells to evade apoptosis. Reactivation of p53 is associated with cell cycle arrest and the induction of apoptosis. Small molecule re-activators of p53-Y220C have been reported to bind to the Y220C-specific cleft and induce apoptosis<sup>129, 138</sup>.

Two pockets in p53 were predicted to be ligandable. The first is the aforementioned Y220C-specific pocket that has previously been targeted by small molecular activators<sup>135</sup>. The second is formed by the  $\beta$ -sheet associated with amyloid fibre formation<sup>139</sup>, and the N-terminal loop (Figure 2.5). The pocket was only predicted to ligandable in one of the available 160 comparable structures. It does not appear to depend on the presence of specific mutant. If binding of a small molecule here could stabilise the structure, then it may be applicable to a broad range of p53-mutant tumours.

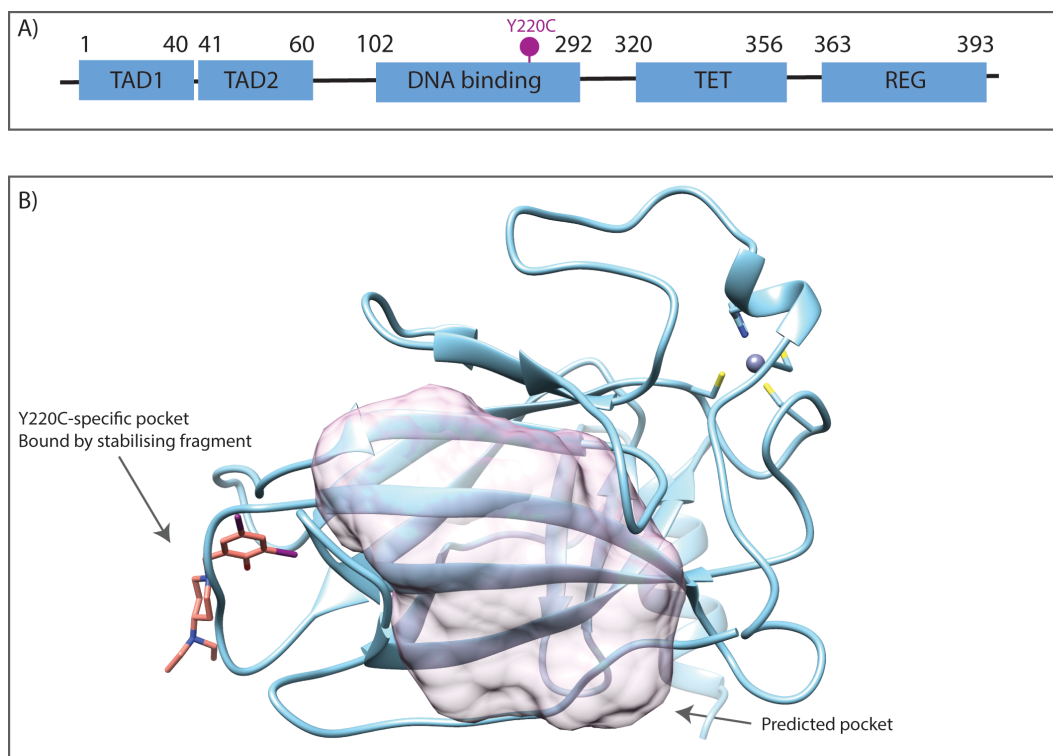


Figure 2.5: Overview of the novel secondary site in p53. A) Domain organisation of tumour suppressor p53. B) Structure of p53 DNA-binding domain (PDB 1KZY) with location of predicted pocket shown as pink transparent surface. Location of the Y220C-specific pocket bound by fragment also shown. Figure made in Chimera<sup>65</sup>

The Estrogen Receptor (ESR1) is a homodimeric steroid-hormone binding nuclear receptor that is activated upon estrogen binding. Approximately 70% of breast cancers are ER positive, with the receptor frequently over-expressed, leading to uncontrolled cellular proliferation<sup>140</sup>. Each monomer contains an N-terminal activation function (AF)-1 domain, a DNA binding domain, and a ligand-binding domain (LBD), which harbours both the ligand-binding site and the activation function (AF)-2 helix<sup>141, 142</sup>. The ESR1-LBD has been structurally characterized (Figure 2.6), and the novel secondary site was identified in this domain.

The predicted pocket is distinct from the ligand binding and AF-2 sites. It was predicted to be ligandable in 50 of the available ESR1-LBD chains from the

PDB. In addition, it was also predicted to be druggable in other NR3-family members for which crystal structures were available. Despite this structural conservation, this pocket does not seem to be referenced to in literature. Although it is a highly novel target, no therapeutic hypothesis can be developed.

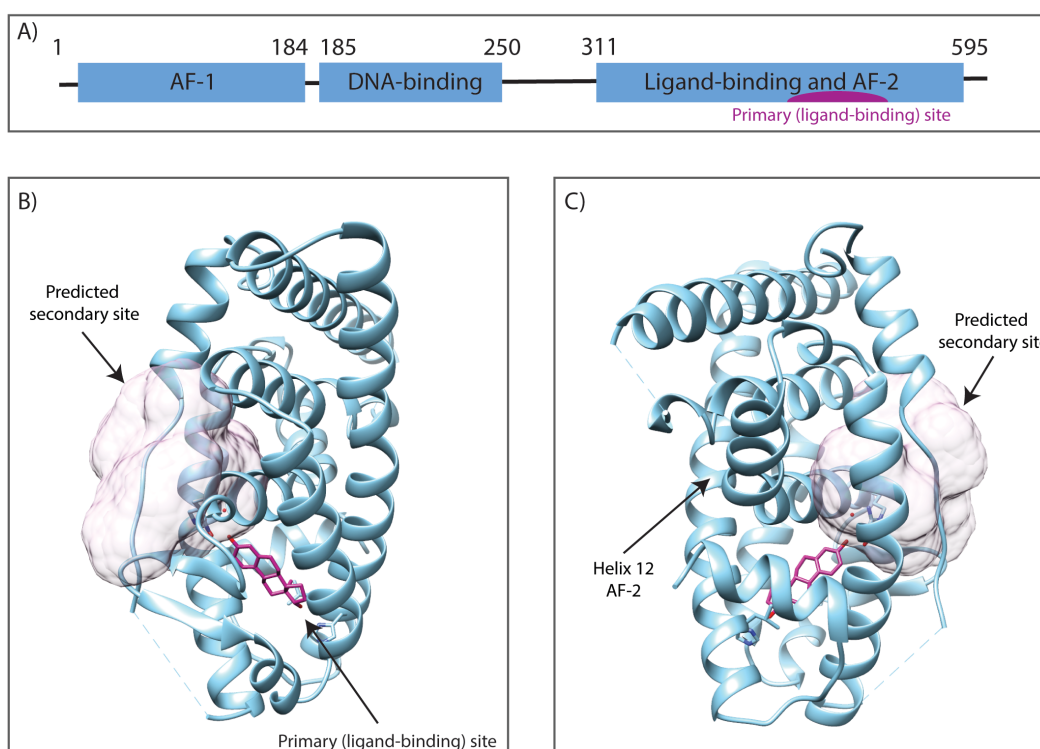


Figure 2.6: Overview of the novel site in ESR1. Domain organisation of ESR1; B) structure of ESR1 (PDB 1ERE) showing the location of the primary (ligand-binding) site and the predicted secondary site; C) Rotated view of ESR1 showing the AF-2 helix. Figure made in Chimera<sup>65</sup>

PIK3CA and PIK3CG are Class I PI3-Kinases that phosphorylate  $\text{PIP}_2$  to produce  $\text{PIP}_3$ .  $\text{PIP}_3$  is involved in the recruitment of various downstream proteins to the plasma membrane for activation, including Akt. Over-activation of PI3K signalling is one of the most common events in human cancers, and can occur through mutation of PI3Ks resulting in constitutive activation<sup>143, 144</sup>.

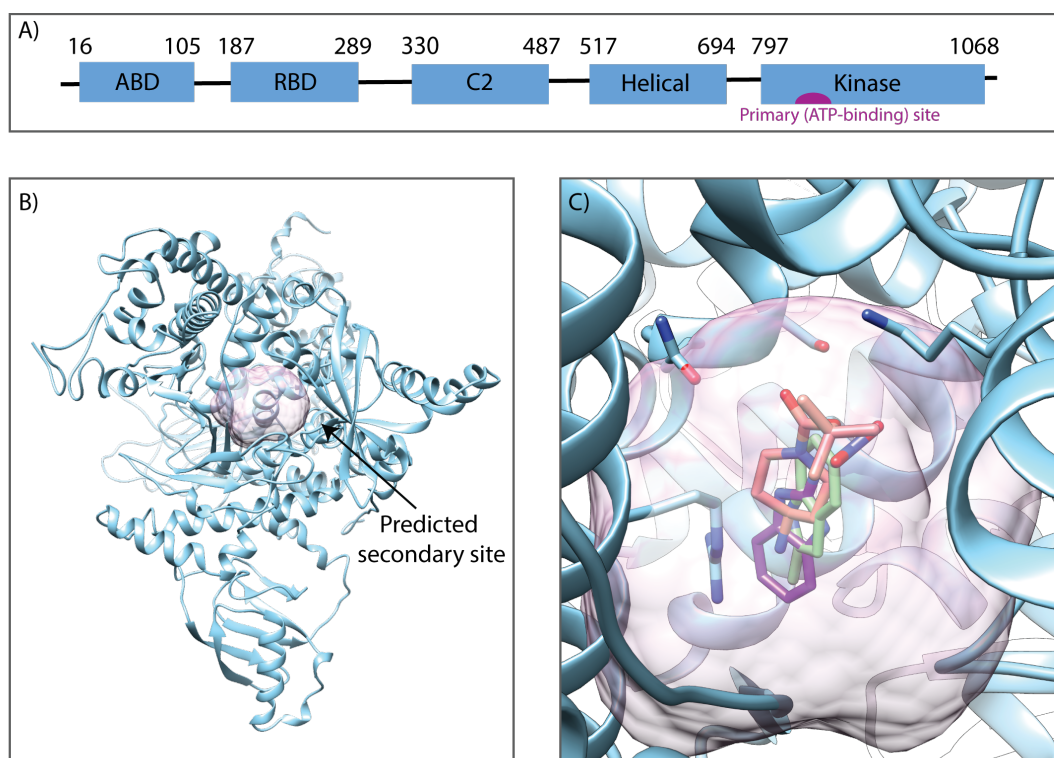


Figure 2.7: Overview of the novel site in PIK3CA. A) Domain organisation of PIK3CA showing location of primary (ATP-binding) site. B) Structure of PI3Kalpha showing location of predicted site. C) Fragments identified binding into the predicted site by Miller et al. 2017<sup>145</sup>. Figures made in Chimera<sup>65</sup>

The novel secondary (Figure 2.7) site is discrete from both the primary site, which is targeted by Copanlisib<sup>81</sup>, and the phospho-peptide binding site which has recently been shown to be ligandable by crystallographic fragment screening<sup>145</sup>. It is a deep pocket at the interface of the helical, kinase and RBD domains. It was predicted four times in PIK3CA during the initial analysis, and was also predicted to be druggable in the PIK3CG isoform. This may support potential functional relevance for this site in Class I PI3Ks, but there was no literature to support this at the time of the analysis. During my research, a crystallographic fragment screen against PI3Kα has identified multiple fragments binding into this novel predicted pocket<sup>145</sup>, validating the ligandability prediction.

Isocitrate Dehydrogenase 1 (IDH1) is a homodimeric, metabolic enzyme that catalyses the oxidative decarboxylation of isocitrate to  $\alpha$ -ketoglutarate ( $\alpha$ KG) with the concomitant reduction of  $\text{NADP}^+$  to NADPH. Substrate and catalytic metal binding causes a large conformational change from the inactive, co-factor bound conformation to the catalytically active form, with the formation of an  $\alpha$ -helix across the dimer interface that completes the primary site<sup>130</sup>.

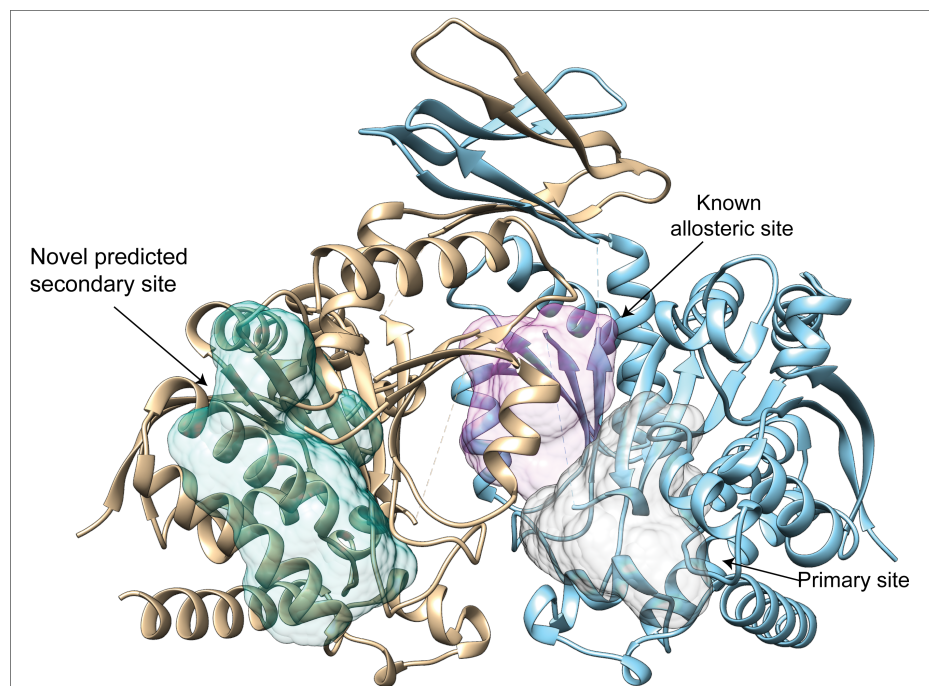


Figure 2.8: Overall structure of the IDH1-R132H dimer showing the location of the three pockets predicted to be ligandable. Figure made in Chimera<sup>65</sup>

Heterozygous missense mutations in IDH1 are identified in up to 80% of glioma patients and 15% of AML patients, with an arginine to histidine substitution at residue 132 the most commonly observed<sup>68</sup>. This causes the neomorphic conversion of  $\alpha$ KG to 2-hydroxyglutarate (2HG), using NADPH as a reducing agent<sup>74</sup>. 2HG is structurally highly similar to  $\alpha$ KG and can inhibit many  $\alpha$ KG-dependent enzymes, including those involved in epigenetic marking. This leads



to widespread epigenetic deregulation, and is sufficient to promote leukemogenesis, which can be reversed with inhibition of the mutant protein<sup>82</sup>.

Both the primary<sup>146</sup> and known allosteric<sup>130</sup> sites have been targeted by inhibitors, and both are predicted to be ligandable through my computational analysis. In this work, a third site was also predicted to be ligandable (Figure 2.8) in one of the four chains of IDH1-R132H in the inactive conformation, but none of the structures of IDH1-R132H in the active conformation. The conformational change from the inactive to the active form results in significant broadening of the secondary site, leading to a decrease in the overall enclosure (Figure 2.9). The broadening of this secondary site was described by Xu et al. in 2004<sup>130</sup>, who suggested that the pocket may have a role in regulating the conformational change required for adoption of the active conformation, but this has not been investigated further.

Targeting this pocket may prevent the conformational change required for catalytic activity and inhibit the production of 2HG. This site is also distant from the patient-derived S280F resistance mutation impacting the efficacy of ivosidenib in clinic<sup>89</sup>. Targeting this novel site may provide a potential way to overcome this resistance.

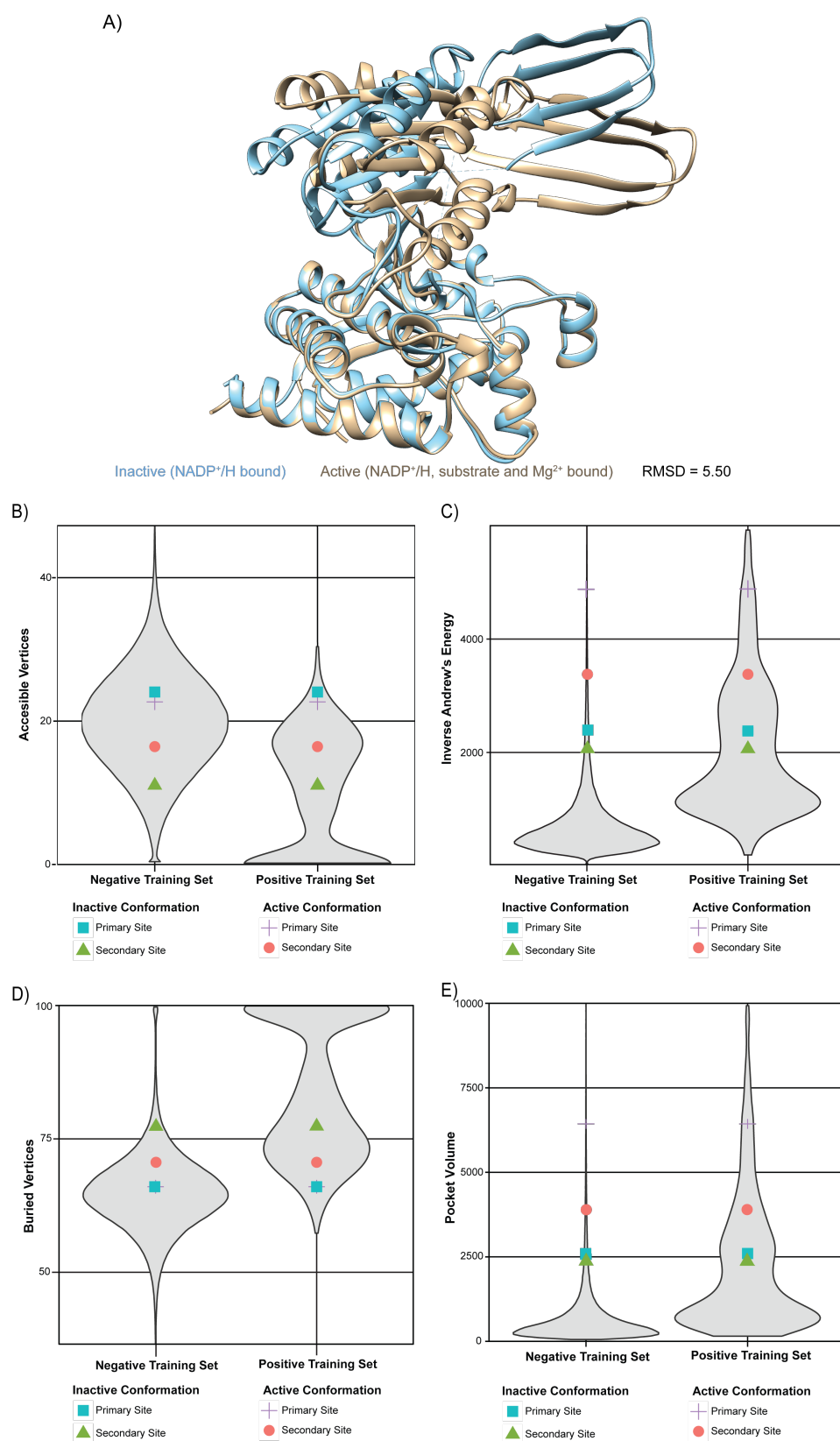


Figure 2.9: IDH1 pockets change with the conformation. A) Overlay of inactive (cyan) and active (tan) conformers of IDH1 showing the large conformational change that occurs upon substrate and catalytic metal binding; a single chain is shown for clarity. B-E) Comparison of primary and novel secondary site pocket properties for the active (4KZO) and inactive conformations (4UMY). Structure figure made in Chimera<sup>65</sup>, violin plots from ggplot2<sup>147</sup>

### 2.2.5 Target selection

ESR1, PIK3CA, p53 and IDH1 are all highly relevant therapeutic targets and have clear potential for clinical impact. Multiple crystal structures are available for all of the proteins, including structures of the protein in complex with a ligand. This indicates that all of the proteins can be readily crystallised and are DMSO tolerant. IDH1 and PIK3CA are both enzymes with *in vitro* biochemical activity assays reported, while ESR1 and p53 activity assays are cell based. Commercial tool compounds are available for all of the short-listed targets to use as experimental controls. All of the systems can be considered experimentally tractable.

The novel pocket in IDH1-R132H was the only pocket to which a potential function could be assigned, as it has been suggested that it may have a role in regulating the conformational change from the inactive to the active conformation<sup>130</sup>. As the pocket was only predicted to be ligandable in the inactive conformation, binding of a small molecule to this site may impact the protein's ability to change conformation, and subsequently modulate enzymatic activity. I therefore selected the novel secondary site in IDH1-R132H as an initial target for experimental investigation.

## 2.3 Conclusions

Computational structure-based ligandability predictors can help to prioritise targets for drug discovery. While they are very successful at predicting primary site ligandability, they are generally less successful at identifying ligandable secondary sites due to an inference bias for the largest and most geometrically complex pocket in the protein. Based on the observation that canSAR3D can occasionally predict ligandability in secondary sites, I adapted the predictor to overcome this limitation.

The novel sites were subsequently triaged in order to prioritise a target most likely to show functional relevance and clinical impact. This process identified 40 cancer-related proteins with novel secondary sites, including five known and experimentally validated sites. I short-listed four targets – p53, ESR1, PIK3CA and IDH1. Crystallographic fragment screening has since experimentally validated the short-listed pocket in PIK3CA<sup>145</sup>. The novel secondary site in IDH1-R132H was selected for experimental investigation by fragment screening.

## **Chapter 3:        Establishing enabling technologies for fragment screening**

---

### **3.1 Introduction**

In order to use fragment screening to identify hits binding to the novel secondary site in IDH1-R132H, several enabling technologies needed to be established. Pure IDH1 variants needed to be produced on a large enough scale to allow experimental investigation. The produced protein also needed to be characterised to ensure that it is correctly folded and suitable for use in assays. Further, the fragment screening assays required establishing. I chose two fragment screening approaches to investigate the ligandability of the novel secondary site – a thermal shift assay (TSA) and crystallographic fragment screening. TSA is a rapid, high throughput approach that has relatively low protein requirements. X-ray crystallography-based fragment screening is highly sensitive to weak fragment binders and has been successfully used to identify and investigate secondary sites<sup>28, 148</sup>. However, it requires significantly more protein, is more difficult to establish and data processing is more challenging. This chapter describes establishment of the key enabling techniques required for fragment screening to investigate the ligandability of the novel secondary site in IDH1-R132H.

### 3.1.1 Thermal shift assays

The thermal stability of a protein is affected by a range of factors, including type of buffer, pH, salt concentration, as well as binding of a small molecule<sup>149</sup>. There are two thermal shift assays that are commonly used – label-free TSA based on intrinsic tryptophan fluorescence, and TSA using a reporter dye such as SYPRO Orange. Both of these techniques are based on the change in fluorescent signal with protein unfolding.

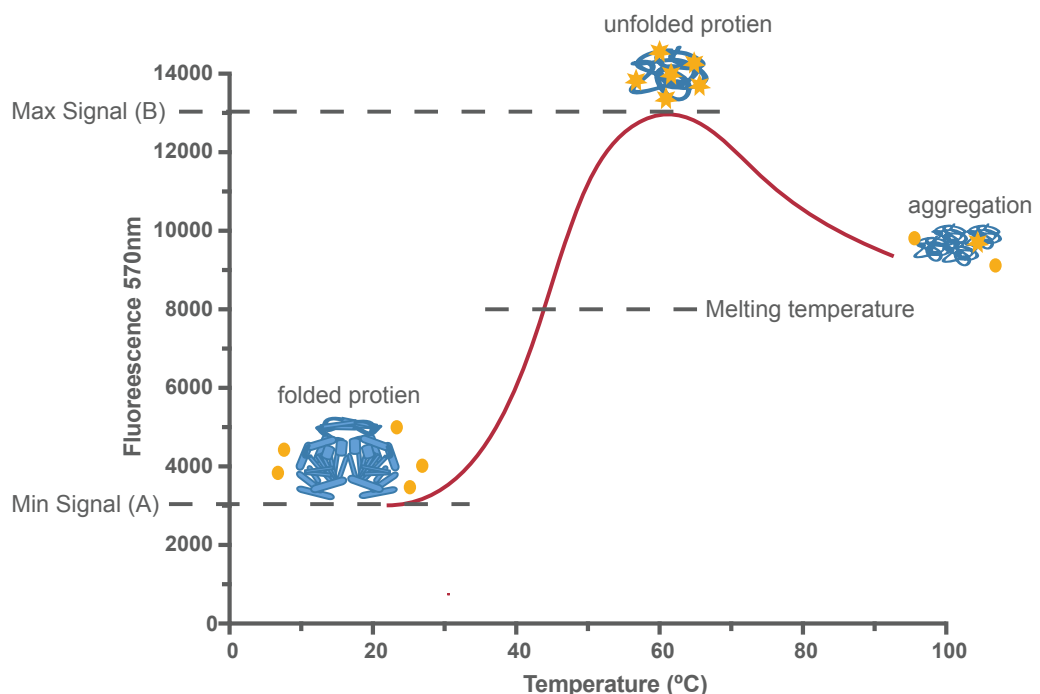


Figure 3.1: Thermal denaturation curve of a given protein with SYPRO Orange. Label-free TSA is based on the intrinsic fluorescence of tryptophan and other aromatic residues, but the overall outline is the same. At low temperature, the hydrophobic patches are buried within the protein, so the dye is in a polar environment in the bulk solvent. As the temperature increases, the protein unfolds, exposing the hydrophobic regions to which the dye can bind resulting in increased fluorescence. Figure based on Miyazaki et al. 2017<sup>28, 150</sup>.

In folded proteins, hydrophobic residues are mainly buried within the protein. As a protein is heated it unfolds, exposing the hydrophobic residues to the polar solvent. In a label-free TSA, the change in the wavelength of tryptophan fluorescence as it is exposed to the polar solvent is measured at 330 nm, 350

nm or as a ratio of fluorescence units at 330 and 350 nm<sup>151, 152</sup>. While label-free TSA is useful for characterisation of purified proteins, it is low throughput due to the use of individually filled capillaries. Dye-based thermal shift systems, such as those using SYPRO Orange, are much higher throughput due to their plate-based format and have lower protein requirements. This technique measures the increase in fluorescence of the dye that occurs when it binds to exposed hydrophobic patches upon protein unfolding (Figure 3.1). TSA with SYPRO Orange is routinely used to examine protein stability and investigate stabilisation by ligands<sup>150, 153-155</sup>, and has also been reported for investigation of interactions between IDH1-R132H and compounds<sup>156</sup>.

For fragment screening by TSA, compounds are mixed with the protein of interest and the fluorescent dye. The samples are subsequently heated and the fluorescence measured over time. The melting curve generated is sigmoidal, with the melting temperature ( $T_m$ ) defined as the midpoint of the unfolding transition. This can be calculated using either the Boltzmann equation (Equation 3.1) or by calculating the maximum of the first derivative of the melting curve<sup>157</sup>. The change in melting temperature ( $\Delta T_m$ ) is calculated based on comparison to a protein control.

$$y = A + \frac{B - A}{1 + e^{\frac{T_m - x}{\text{slope}}}} \quad \text{Equation 3.1}$$

High affinity, potent molecules such as tool compounds can give thermal shifts up to 10 °C or more depending on the system<sup>158</sup>. This technique can also identify fragment binders with  $\Delta T_m$  values as low as 0.5 °C making it suitable for fragment screening<sup>153</sup>.

### **3.1.2 Crystallographic fragment screening**

Crystallographic fragment screening has several advantages over other fragment screening approaches<sup>159, 160</sup>. It is capable of identifying fragments binding with a large range of affinities, including weak millimolar binders, due to both the sensitivity of the technique as well as the comparatively high fragment concentrations used for soaking. In addition, binding sites are immediately elucidated, allowing identification of hot spots and novel sites. As high fragment screening hit rates is associated with ligandability<sup>53</sup>, this allows immediate assessment of the overall ligandability of the site. It also allows evaluation of the potential for structure-based fragment linking, growing and merging, facilitating movement from weak initial fragment hits to more potent molecules.

#### **3.1.2.1 *PanDataset Density Analysis***

Electron density is calculated as an average of the signal derived from a large number of protein molecules in the crystal, where low occupancy states such as binding events or minor conformations may be obscured by the major conformation (Figure 3.2A). Low occupancy fragment binding can be due to many reasons, including weak, although potentially very specific, binding or poor ligand solubility. In these cases, the fragment density can be obscured due to crystallographic averaging.



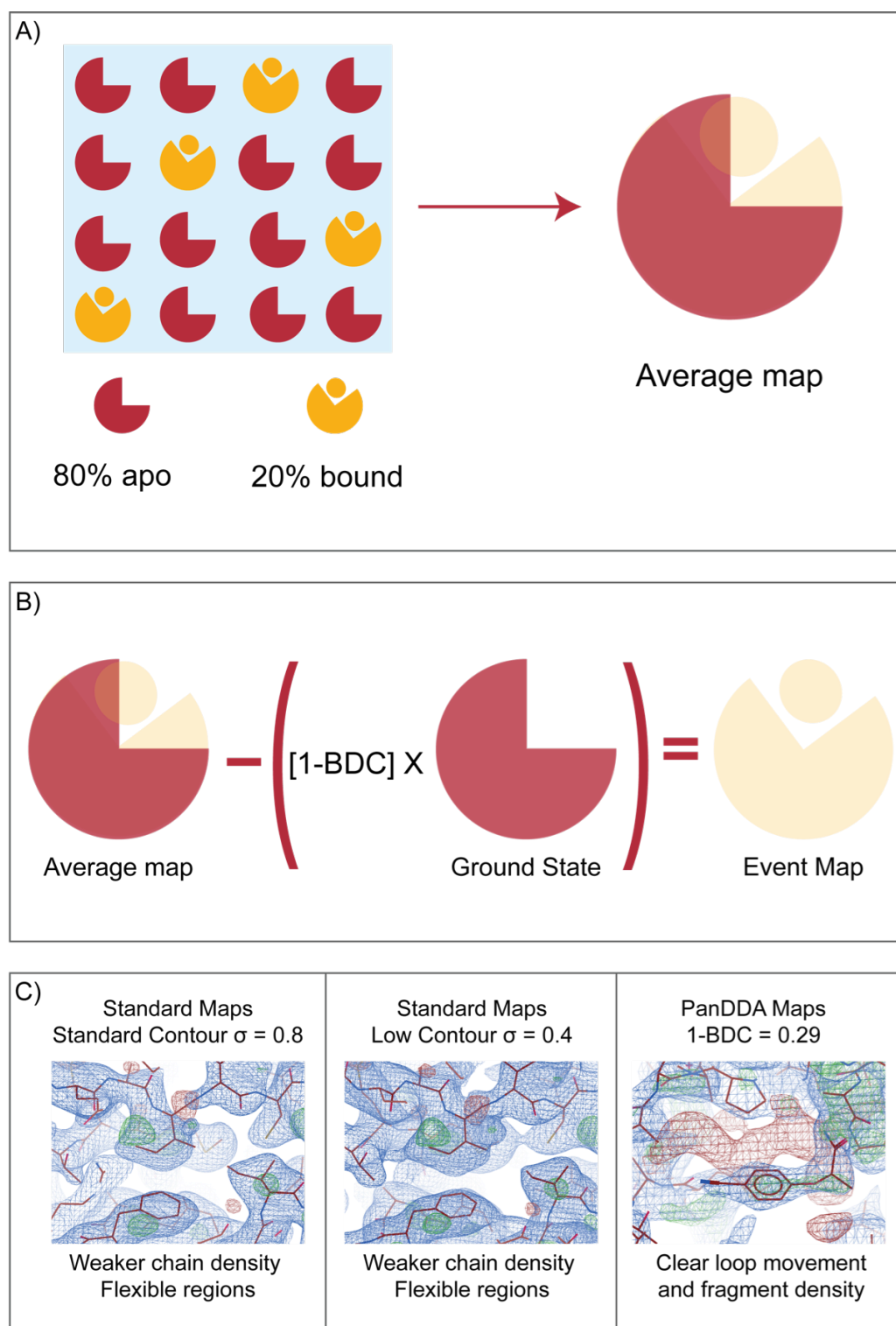


Figure 3.2: Overview of PanDDA processing to identify minor conformations and states. A) Electron density is an average of all states in the protein crystal. Low occupancy states are obscured by high occupancy states. B) PanDDA calculates an average ground state map that is subsequently subtracted from each dataset in real space, to leave only the event map where the density deviates. C) Using PanDDA, low occupancy states are revealed that cannot be visualised using normal approaches, even at low contour. The event map is shown as a blue mesh, whilst the Z-map is shown as green and red meshes. Data shown is from IDH1-R132H crystallographic fragment screening and visualised in COOT<sup>161</sup>.

PanDataset Density Analysis (PanDDA<sup>162</sup>) is crystallographic data analysis software that was developed to allow de-convolution of low occupancy states, including weakly binding fragments. The analysis aligns 50 or more sigma-weighted  $2mF_o - DF_c$  maps from isolated datasets in real space to generate an averaged **ground state map**. The ground state map shows a much lower level of noise in comparison to any individual map used to calculate it. This map is then subtracted from each of the individual  $2mF_o - DF_c$  maps in real space to calculate **Z-scores** at each voxel of the asymmetric unit, producing the **Z-map**.

Clusters of large **Z-scores**, greater than  $\pm 3$ , are representative of significant local deviation in the analysed map in comparison to the ground state and are reported as **events**. Positive density in the corresponding **Z-map** indicates where additional density is observed in comparison to the ground state, while negative density indicates where density has been lost in comparison to the ground state (Figure 3.2C). Z-scores are an objective and statistically meaningful measure of potentially interesting deviations from the ground state, allowing identification of low occupancy **events** that do not yield strong density in  $mF_o - DF_c$  maps.

Once **events** have been located using the Z-map, the program estimates the fraction of the data that contains the event, the occupancy of the fragment in a ligand-binding context. This estimated occupancy, reported as one minus Background Density Correction (1-BDC), is then used to calculate a de-convoluted **event map** through subtraction of the ground state scaled for the estimated occupancy (Figure 3.2B). The resulting event map shows density only for the ligand binding fraction or minor protein conformer.

In addition to the normal statistics used to validate structure quality, additional statistics are reported specifically to validate low-occupancy ligands (Table 3.1). Large B-ratios and low RSZO/OCC values do not necessarily indicate errors in the model, but highlight weak features. Low values of RSCC are indicative of poor fit between the model and the data, and will be lower in the early stages of refinement. All of these statistics will be affected by the quality of the data, and especially by the resolution of the system. They are considered in combination with event and Z-maps to assess the reliability of the binding event.

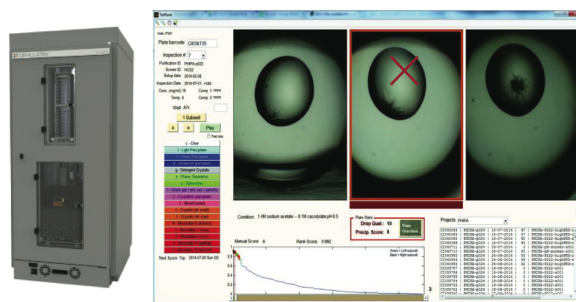
<b>1 – BDC</b>	1 – background density correction	Selected to maximise the contrast between ground state and event maps. It is the fraction of the ground state to be subtracted from the dataset to give the event maps, and is therefore related to the occupancy.
<b>Z-score</b>		Shows the extent of deviation from the ground state. Large Z-scores ( $\pm 3$ ) indicate significant deviations from the ground state and indicate potentially interesting features.
<b>RSCC</b>	Real space correlation coefficient	Correlation between model and observed density. RSCC > 0.7 required for a 'good' model.
<b>RMSD</b>	Movement of ligand after refinement	Measure of how far a ligand moves following refinement; for a good fit, only small movements would be expected (< 1)
<b>RSZO/OCC</b>	Real-Space Observed Z-Score divided by the estimated occupancy of the fragment	Signal-to-noise ratio, reports strength of density for the ligand in comparison to the noise, normalised for the occupancy. Dependent on resolution and structure quality post-MR. Should be as high as possible.
<b>B-ratio</b>	Ratio of ligand/protein B-factors	Ratio of B-factors for ligand and surrounding residues; whilst ligands are expected to be more mobile (hence ratio > 1), ratios >>2 are likely to indicate mis-modelled ligands.

Table 3.1: PanDDA specific statistics used to identify and evaluate low occupancy events

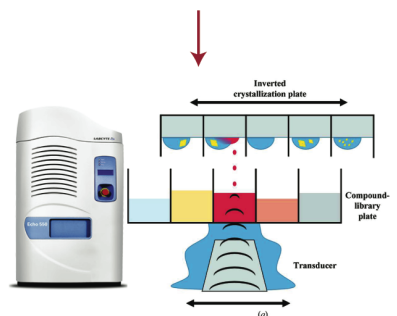
### **3.1.2.2 XChem crystallographic fragment screening platform**

The XChem facility at Diamond Light Source (DLS) is a dedicated facility for crystallographic fragment screening<sup>23, 163</sup> (Figure 3.3). Crystals are grown in plates that are compatible with ECHO acoustic dispensing, and imaged. Crystals to be targeted for by fragment soaking are selected, and the location in the drop the fragment will be dispensed to is selected. Fragments are then acoustically dispensed to the targeted locations in crystal drops using an Echo acoustic dispenser, which allows rapid and accurate addition of fragments without directly hitting and potentially damaging crystals. Harvesting is robot-assisted and data collection is unattended, allowing approximately 1000 crystals to be soaked, harvested and data collected within a week.

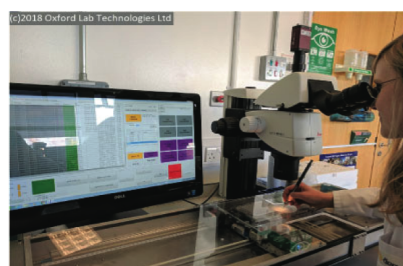
Auto-processing results are imported directly into XChemExplorer<sup>164</sup> (XCE) for rapid, parallel processing, including molecular replacement and generation of ligand restraint files. The PanDDA analysis is run directly from XCE and automatically reports datasets showing significant deviation from the ground state, which are then manually inspected. Fragments are modelled into the event map and local corrections are made, followed by a cycle of refinement in Refmac5<sup>165</sup>. PanDDA statistics are used alongside the normal refinement statistics to track the progress of correction and refinement.



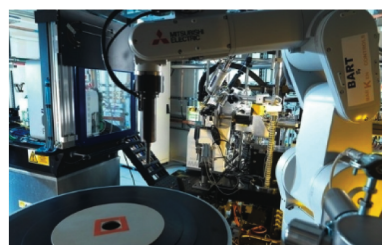
Imaging of crystal plates, crystal selection and in-drop targeting



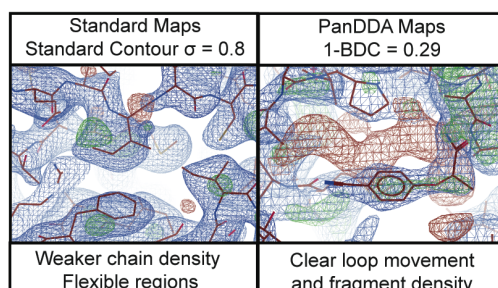
Fragments acoustically dispensed using an ECHO



Robot-assisted harvesting using a Shifter



Unattended data collection on IO4-1



PanDDA identifies low occupancy events including weakly binding fragments

Figure 3.3: Schematic overview of the XChem crystallographic fragment screening<sup>163</sup>. Crystals are imaged using a RockImager and ranked; figure from Diamond. Drops containing crystals are selected, and the location within the drop to which the dispense will be targeted is also selected. The fragments are then dispensed using an Echo acoustic dispenser; figure from Diamond. Crystal harvesting is semi-automated using a Shifter; figure from Diamond. Data collection is unattended on IO4-1; figure from Diamond. Data is imported into XCE for rapid parallel processing and PanDDA is used to identify events.

## 3.2 Results

### 3.2.1 Protein production and purification

Both IDH1-WT and IDH1-R132H variants have been crystallised using full-length constructs. The DNA sequence was codon-optimised for *E. coli* expression and cloned into a pET-28a vector, and mutants generated using site-directed mutagenesis against this construct (Chapter 7.2.2, Appendix 8.2.1). All IDH1 variants expressed well and no optimisation of expression conditions was required (Appendix 8.2.2). The general purification protocols were adapted from IDH1-R132H purifications reported in the literature and a general purification strategy for His-tagged proteins provided by Dr Yann-Vaï Le Bihan. All IDH1 constructs were purified using a four-step schema (Figure 3.4), except IDH1-R132H, which was purified using an optimised five-step schema.

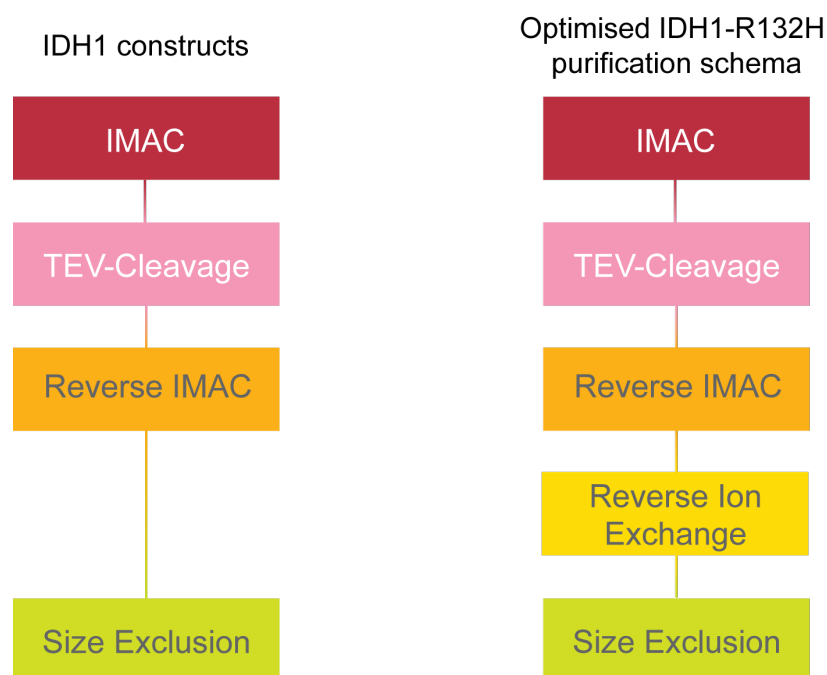


Figure 3.4: Purification schema for IDH1 variants. All variants of IDH1 were purified using the protocol shown on the left, except the main IDH1-R132H variant that was purified using the optimised protocol on the right. IMAC – immobilised metal affinity chromatography; TEV – tobacco etch virus

Full details of the optimised purification strategy can be found in Chapter 7.2.3. Representative chromatographs and SDS-PAGE gels are shown in Figure 3.5 with bands corresponding to IDH1 variants boxed in red. In brief, clarified cell lysates were applied to a HisTrap FF Immobilised Metal Affinity Column (IMAC). IDH1 variants bound to the column through the hexahistidine tag under low imidazole concentrations and were eluted using increasing concentrations of imidazole (Figure 3.5A). Fractions containing IDH1 were pooled and dialysed overnight in the presence of His-tagged TEV-protease to remove the hexahistidine tag. The protein was subsequently re-applied to the IMAC column. Uncleaved IDH1, His-TEV and contaminant proteins bound to the column, whilst the cleaved IDH1 was not bound and was collected in the flow through. Fractions were analysed using SDS-PAGE showing IDH1 at greater than 90% purity (Figure 3.5B).

Fractions containing IDH1 were pooled and concentrated, and subsequently loaded onto a size exclusion column. IDH1 elutes as a symmetrical peak with a maximum at 85 mL. Analysis by SDS-PAGE shows greater than 95% purity (Figure 3.5D). The  $A_{260}/A_{280}$  was measured between 0.60 and 0.62 for all purifications, close to the theoretical value for pure protein with no nucleic acid contaminants,  $A_{260}/A_{280} = 0.57^{166}$ .

IDH1 variants co-purify from reverse-IMAC with a contaminant that is not fully separated by SEC (Figure 3.5B, blue box), so an additional purification step was designed between the reverse-IMAC and SEC for IDH1-R132H. The theoretical isoelectric point of IDH1-R132H is 6.4 as calculated by EXPASY ProtParam<sup>167</sup>.

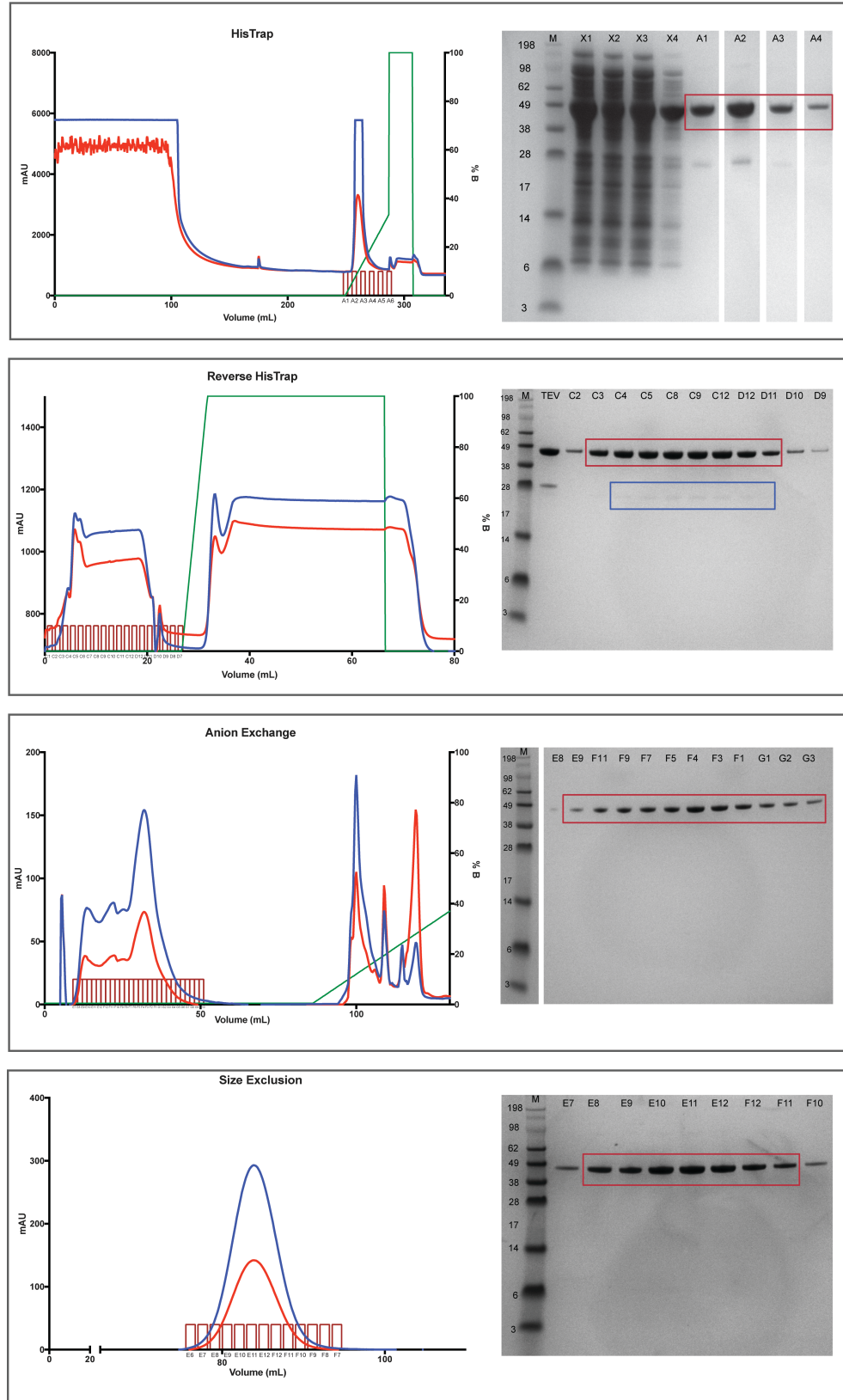


Figure 3.5: Representative chromatograms and SDS-PAGE gels from IDH1 purifications. The blue trace shows the UV 280 nm (mAu), the red trace shows the UV 260 nm (mAu) and the green trace shows the proportion of buffer B loaded in v/v. Bands corresponding to IDH1 variants are boxed in red, while the contaminant is boxed in blue. All variants are purified by affinity chromatography (HisTrap and Reverse HisTrap) and size exclusion chromatography. Optimisation of the IDH1-R132H purification adds an additional ion exchange step that successfully removes this contaminant.



At pH 8, IDH1-R132H would be expected to have a negative charge and should therefore bind to an anion exchange column under low salt conditions, to be eluted by increasing the salt concentration. A buffer screen against IDH1-R132H showed that IDH1-R132H is stable in low salt conditions (Appendix 8.2.5.1), indicating this approach may be suitable without causing IDH1-R132H precipitation. IDH1-R132H was dialysed into a 5 mM NaCl buffer and loaded onto a ResourceQ column equilibrated with the same low salt buffer. Unexpectedly, IDH1-R132H eluted directly in the flow-through, but as the contaminant(s) interacted with the column separation was still achieved (Figure 3.5C). Fractions were analysed by SDS-PAGE, showing removal of the main contaminant. Appropriate fractions were then pooled and loaded onto a SEC column as before. The correct mass of each IDH1 variant was confirmed by mass spectrometry (Appendix 8.2.3). Each variant was purified to greater than 95% purity and with high yields of 30 mg/L for IDH1-R132H and 40 mg/L for IDH1-WT, sufficient for experimental investigation.

### **3.2.2 Variant characterisation**

Following production of pure IDH1 variants, label-free TSA was used to investigate the thermal stability of IDH1-R132H and IDH1-WT and ensure they were correctly folded through their ability to bind their native co-factors and substrates – NADP<sup>+</sup> and isocitrate for IDH1-WT, and NADPH and  $\alpha$ KG for IDH1-R132H.

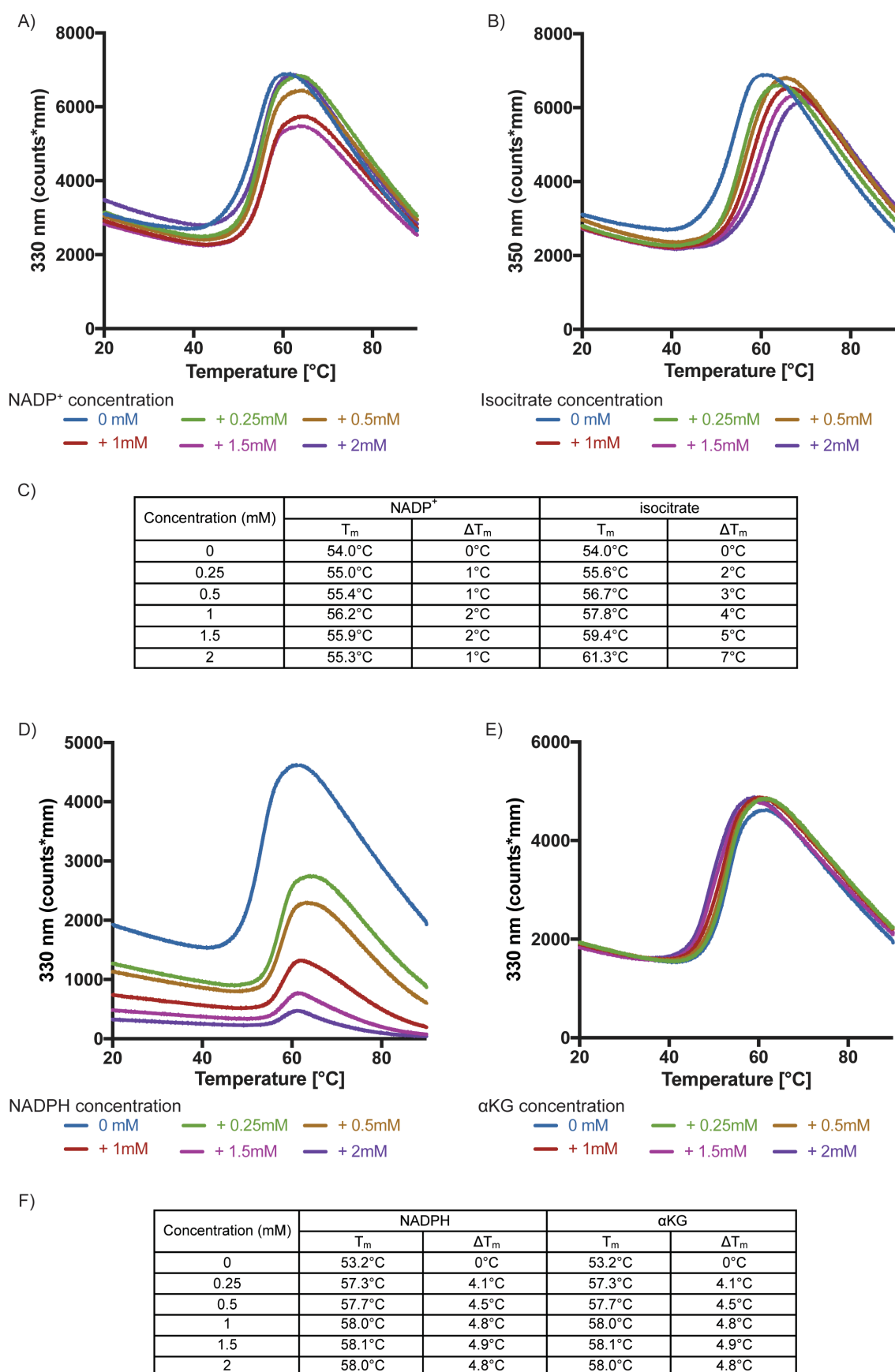


Figure 3.6: Label-free thermal shift results for IDH1-WT with NADP<sup>+</sup> (A,C) and isocitrate (B,C), and for IDH1-R132H with NADPH (D,F) and αKG (E,F). Plots made in GraphPad Prism

The melting temperature of apo IDH1-WT was 54 °C. It was stabilised by its natural co-factor NADP<sup>+</sup> with a maximum  $\Delta T_m$  of 2.2 °C (Figure 3A, C). Apo IDH1-R132H was found to be slightly less stable than IDH1-WT under the same buffer conditions, with a  $T_m$  of 53.2 °C. However, binding of the natural substrate NADPH induced a much larger thermal shift than the IDH1-WT system, with  $\Delta T_m$  values up to 5 °C (Figure 3D, F). NADPH has an absorbance peak at 350 nm, which overlaps with the emission spectra of aromatic residues, resulting in a decrease in fluorescent signal with increase NADPH concentration. Still, the ability of each protein to bind their natural co-factor indicates that they are properly folded.

IDH1-WT was also stabilised by its natural substrate isocitrate to a greater extent than induced by the co-factor, with a  $\Delta T_m$  of 7 °C (Figure 3B, C) In contrast, IDH1-R132H is slightly destabilised by its natural substrate  $\alpha$ KG, with a  $\Delta T_m$  of - 4 °C (Figure 3E, F). The reason for this difference is not clear. Full experimental details for label-free TSA can be found in Chapter 7.2.4.

### **3.2.3 Establishing IDH1-R132H TSA for fragment screening**

A SYPRO Orange TSA was established for fragment screening as it is higher throughput and requires less protein. High concentrations of NADPH are required to favour the inactive conformation for screening and to block the primary site. The excitation/emission spectra of NADPH and SYPRO Orange do not overlap, so NADPH should not interfere with the SYPRO Orange TSA. A SYPRO Orange thermal shift assay has been reported in the literature using 10 X SYPRO Orange and IDH1-R132H at 0.5 mg/mL (7.8  $\mu$ M)<sup>156</sup>. These

conditions were used as a starting point to establish an assay suitable for fragment screening.

### 3.2.3.1 Binding Curve analysis

IDH1-R132H melting curves show variations in shape, with high initial signals. Binding events can also further change the shape of the curve. These features make analysis of the binding curves more complex. The melting temperature from SYPRO Orange thermal shift assays can be calculated through either a Boltzmann analysis or the first derivative. The Boltzmann analysis identifies the highest signal in the curve and calculates the  $T_m$  based on the Boltzmann distribution below this temperature. In Figure 3.7A, the initial peak is lower than the true maximum and the algorithm is able to identify the correct peak and use it for the analysis. In Figure 3.7B, the initial peak is higher than the true maxima, and therefore is selected for the analysis, resulting in a large reported negative shift, greater than 10 °C, reported.

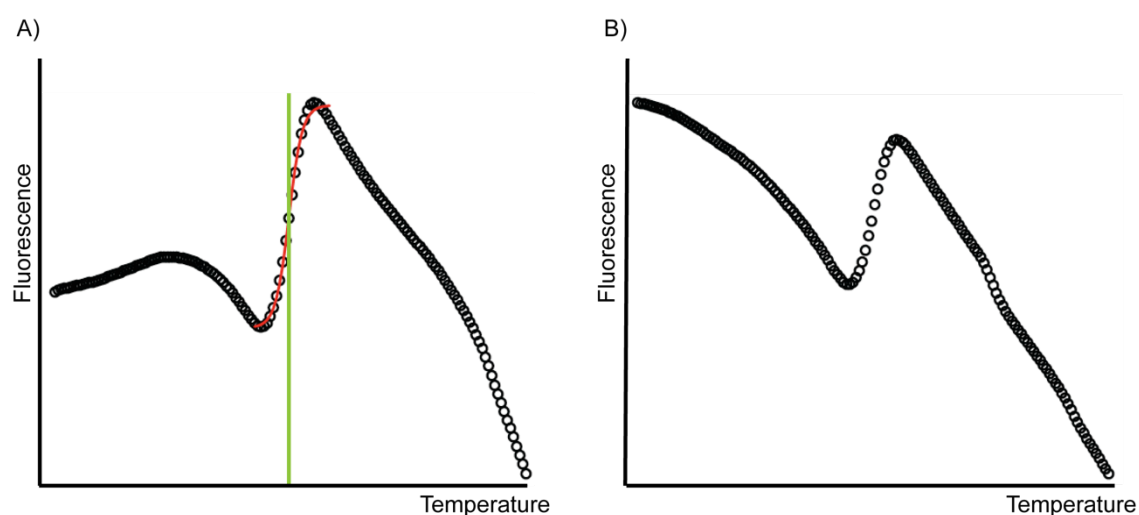


Figure 3.7: Typical melting curves for IDH1-R132H with NADPH, plotting the fluorescence against melting temperature. Panel A shows a plot where the correct peak has been identified as the maximum. The red curve indicates the curve that is analysed by Boltzmann distribution, with the green line showing the calculated melting temperature. Panel B shows a curve where the high initial fluorescence does not allow identification of the correct peak. This results in either no reported melting temperature, or a very large (> 10 °C) negative shift as the algorithm attempts to use the initial peak. Plots made in R<sup>168</sup>

To overcome this, I adapted the R script used for analysis to exclude data points below 40 °C, to exclude the initial peak. Even with this adaptation, the Boltzmann analysis was not always able to identify the correct peak. In these instances I used the  $T_m$  as calculated by the first derivative. The precision of the first derivative is limited by the thermocycler used, which only collects data at 0.5 °C temperature increments.

### 3.2.3.2 Investigation of IDH1-R132H and SYPRO Orange concentrations

I investigated four enzyme concentrations - 3  $\mu$ M, 5  $\mu$ M, 10  $\mu$ M and 15  $\mu$ M - and three SYPRO Orange concentrations – 5 X, 10 X and 15 X - to identify conditions that gave a sufficient signal (Figure 3.8). Fluorescence signal increased with protein concentration, with the largest increase between 5  $\mu$ M and 10  $\mu$ M; the increase was much smaller between 10  $\mu$ M and 15  $\mu$ M.

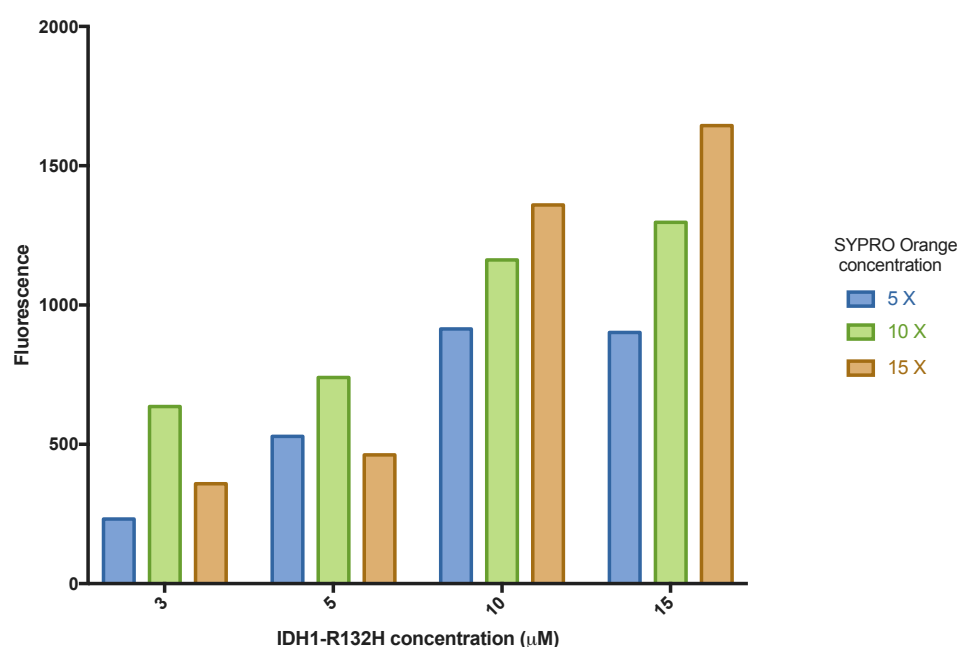


Figure 3.8: Bar chart showing different fluorescent measurements obtained for different SYPRO Orange and IDH1-R132H concentrations. Plot made in GraphPad Prism.

Similarly, the fluorescence increased with increasing SYPRO Orange concentration, with the largest increase between 5 X and 10 X, and a smaller increase between 10 X and 15 X. IDH1-R132H at 7  $\mu$ M with 10 X SYPRO Orange was therefore selected for further thermal shift experiments as it gave a sufficient signal whilst limiting protein consumption.

### ***3.2.3.3 Investigating of NADPH, $\alpha$ KG and tool compounds binding to IDH1-R132H by TSA***

The addition of NADPH was required during fragment screening to favour the conformation in which the novel secondary site was predicted ligandable. It also blocked the co-factor binding site from fragment binding, although the known allosteric and substrate-binding sites were still accessible. I therefore investigated the concentration of NADPH to use during fragment screening. To identify a positive control compound to use in the screen, I also investigated the effects of the natural substrate and known tool compounds on the thermal stability of IDH1-R132H in the presence of NADPH.

In the absence of its co-factor, IDH1-R132H has a melting temperature of 50.4 °C as measured by SYPRO Orange TSA. This is approximately 2.8 °C lower than measured using label-free TSA (Section 3.2.2), which is likely due to variations in how the melting curve is measured and the different fluorophore used. A dose-dependent increase in IDH1-R132H thermal stability was observed with increasing concentration of NADPH (Figure 3.9A). An NADPH concentration of 500  $\mu$ M was used for further experiments.

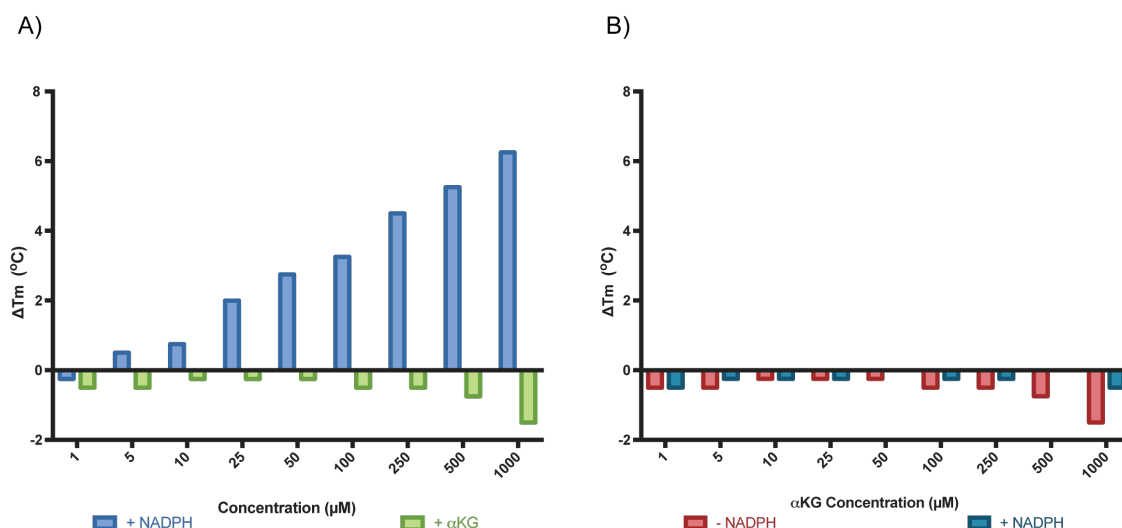


Figure 3.9: Bar charts showing the change in melting temperature with increasing concentrations of NADPH (A) and  $\alpha KG$  (B) as calculated by the first derivative. Graphs made in GraphPad Prism

In contrast, increasing concentrations of  $\alpha KG$  alone caused a small decrease in stability, with  $\Delta T_m = -1.5$   $^{\circ}C$  (Figure 3.9B). In the presence of NADPH,  $\alpha KG$  showed little to no impact on the thermal stability of IDH1-R132H, and was therefore unsuitable as a positive control for fragment screening. I therefore investigated commercially available IDH1-R132H tool compounds GSK-864<sup>158</sup> and AGI-5198<sup>169</sup> (Figure 3.10). GSK-864 (Figure 3.10A) selectively inhibits IDH1-R132H, with  $IC_{50}$  values of 15 nM and 466 nM against IDH1-R132H and IDH1-WT respectively, as determined using *in vitro* RapidFire Mass Spectrometry experiments measuring 2-HG production<sup>158</sup>. It is  $\alpha KG$ -competitive, which was determined using a diaphorase/resazurin-based *in vitro* biochemical assay measuring resorufin production<sup>170</sup>. The binding site for this compound is unknown, although a crystal structure showing binding of a structurally similar molecule, GSK-321, to the known allosteric site is available (Figure 3.10B, C). Similarly, AGI-5198 (Figure 3.10D) is a selective,  $\alpha KG$ -competitive<sup>170</sup> IDH1-R132H inhibitor with an  $IC_{50}$  value of 70 nM against IDH1-R132H as measured using a diaphorase/resazurin-based *in vitro* biochemical

assay, with no measurable activity against IDH1-WT when tested using the same assay<sup>169</sup>. The binding site and binding mode for this compound is unknown. The thermal stabilisation of IDH1-R132H by GSK-864 and AGI-5198 was investigated both in the presence and absence of NADPH (Figure 3.11).

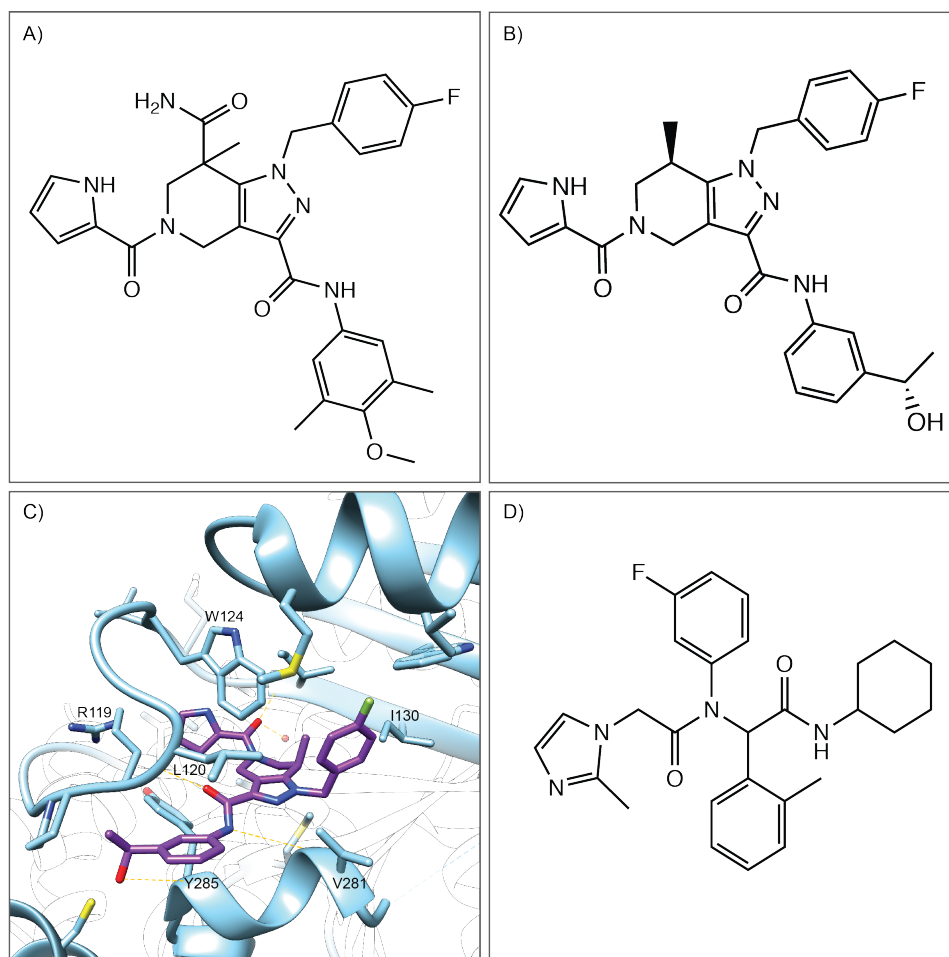


Figure 3.10: 2D structures of IDH1-R132H tool compounds GSK-864 (A) and structurally similar molecule GSK-321 (B). The binding mode of GSK-864 is unknown, but the GSK-321 has been shown to bind the known allosteric site, mainly through Hydrogen bonds with the protein backbone (C; PDB 5DE1). The binding mode for the second tool compound used, AGI-5198 (D), is also unknown.

AGI-5198 was able to stabilise IDH1-R132H, with a  $\Delta T_m$  value of 2.5 °C without NADPH, and a  $\Delta T_m$  value 4 °C with NADPH. In contrast, GSK-864 was able to stabilise IDH1-R132H to a greater extent than AGI-5198, with  $\Delta T_m$  values of 9



°C in both the presence and absence of NADPH. As both GSK-864 and AGI-5198 show consistent thermal stabilisation of IDH1-R132H in the presence of NADPH, these were selected for use as positive controls for the fragment screen. Conditions selected for fragment screening are shown in Table 3.2.

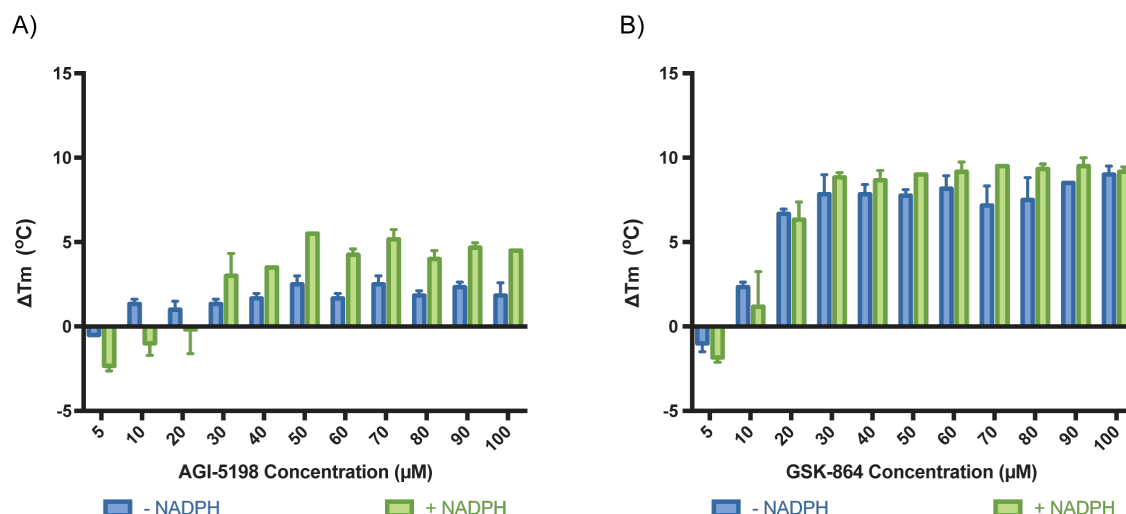


Figure 3.11: Bar chart showing  $\Delta T_m$  values for IDH1-R132H with tool compounds AGI-5198 (A) or GSK-864 (B) in the presence and absence of NADPH as calculated by the first derivative. Graphs made in GraphPad Prism.

#### 3.2.3.4 Investigating of $NADP^+$ , isocitrate and tool compounds binding to IDH1-WT by TSA

A SYPRO Orange thermal shift assay for IDH1-WT to investigate potential binding selectivity was also established. To enable this, I investigated substrate and co-factor binding for IDH1-WT. As buffer components can have a significant impact on thermal stability, I used the same assay conditions for the WT system as for the mutant system.

IDH1-WT was stabilised by both  $\text{NADP}^+$  and isocitrate, with maximum  $\Delta T_m$  values of  $2.75^\circ\text{C}$  and  $2^\circ\text{C}$  respectively (Figure 3.12). When isocitrate is incubated with IDH1-WT in the presence of  $\text{NADP}^+$ , the measured thermal shift is  $8.25^\circ\text{C}$  in comparison to IDH1-WT with  $\text{NADP}^+$  alone. Unlike IDH1-R132H, which is more stable with only co-factor present, IDH1-WT is more stable with both co-factor and substrate present.

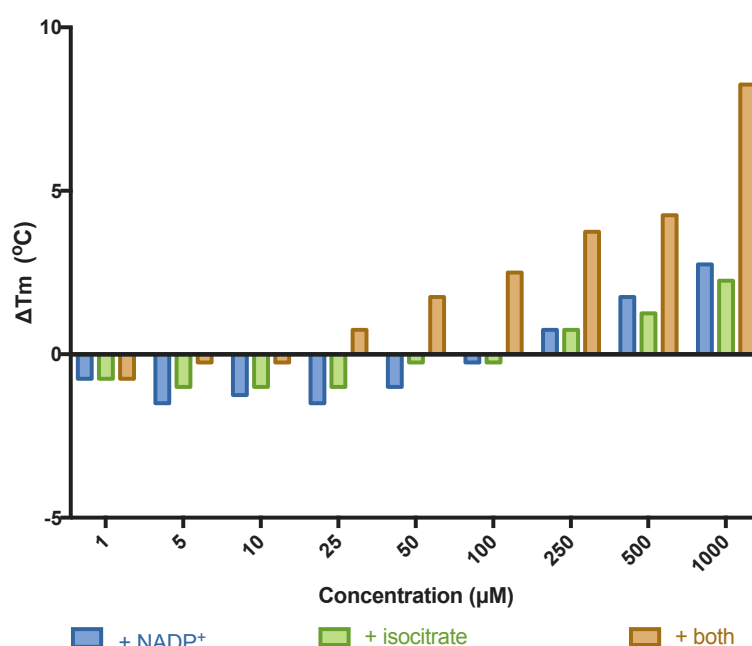


Figure 3.12: Bar chart showing  $\Delta T_m$  values for IDH1-WT with  $\text{NADP}^+$ , isocitrate or with both  $\text{NADP}^+$  and isocitrate. Graphs made in GraphPad Prism.

Whilst the two tool compounds, AGI-5198 and GSK-864, are selective for IDH1-R132H, GSK-864 is reported to show measurable activity against IDH1-WT. I therefore investigated the ability of GSK-864 to stabilise IDH1-WT to use as a control in the thermal shift assay. Like IDH1-R132H, IDH1-WT was equally stabilised by GSK-854 in both the presence and the absence of co-factor, although to slightly less extent, with a  $\Delta T_m$  of  $6^\circ\text{C}$  for WT in comparison to  $9^\circ\text{C}$  for R132H (Figure 3.13). GSK-864 was therefore suitable for use as a control in

for counter-screening against IDH1-WT. Conditions selected for investigation of fragment binding specificity are shown in Table 3.2.

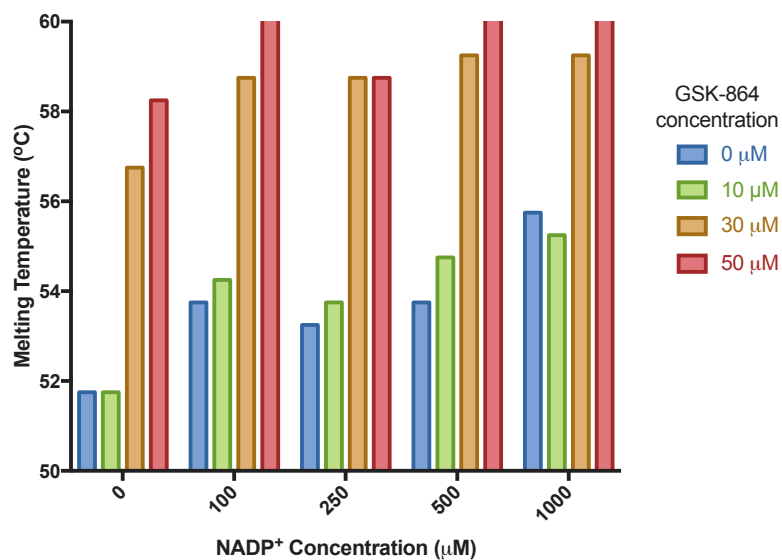


Figure 3.13: Bar chart showing the increase in IDH1-WT melting temperature with increasing concentrations of tool compound GSK-864 in the presence and absence of NADP<sup>+</sup>, as calculated by the first derivative. Graphs made in GraphPad Prism.

Protein	Variant	IDH1-R132H	IDH1-WT
	Concentration (μM)	7	10
Co-factor	Co-factor	NADPH	NADP <sup>+</sup>
	Concentration (μM)	500	500
AGI-5198 control	Concentration (μM)	10	-
GSK-864 control	Concentration (μM)	10	10

Table 3.2: assay conditions for thermal shift fragment screening. Final buffer was 75 mM HEPES pH 7.5, 100 mM NaCl, 2% DMSO with 10X SYPRO Orange. Fragments were screened at 300 μM.

### **3.2.4 Establishing IDH1-R132H crystallographic conditions for fragment screening**

To establish crystallography, I initially investigated the conditions used to produce the structure in which the novel pocket was predicted to be ligandable<sup>130</sup>. Different buffers, Tris and BisTris, were reported in the PDB and the published paper. Both buffers were initially investigated, with 220 mM ammonium sulphate and a range of polyethylene glycol monomethylether 5000 (PEG5000MME) as reported in the literature<sup>156</sup>.

Initial crystallisation trials were carried out in 15-well EasyXstal plates, which use hanging drop vapour diffusion for crystallisation (Figure 3.14). IDH1-R132H was pre-incubated with NADPH to favour the inactive conformation in which the novel pocket was identified. Addition of NADPH also increases the protein stability as shown by TSA, which can improve crystal quality<sup>171</sup>. Crystals formed readily but were irregular in shape and poorly diffracting (Figure 3.15A). To slow the rate of vapour diffusion, the incubation temperature was reduced from 18 °C to 12 °C. This improved crystal morphology, but diffraction remained poor (Figure 3.15B).

I then switched from the EasyXstal plates to 96-well SwissCi plates to test the effect of sitting drop vapour diffusion experiments (Figure 3.14) and reducing drop and reservoir volume. SwissCi plates are also compatible with the XChem crystallographic fragment-screening platform. The range of buffer conditions and drop ratios were the same as investigated in EasyXstal plates. The use of SwissCi plates resulted in improved crystal morphology, reproducibility and diffraction.

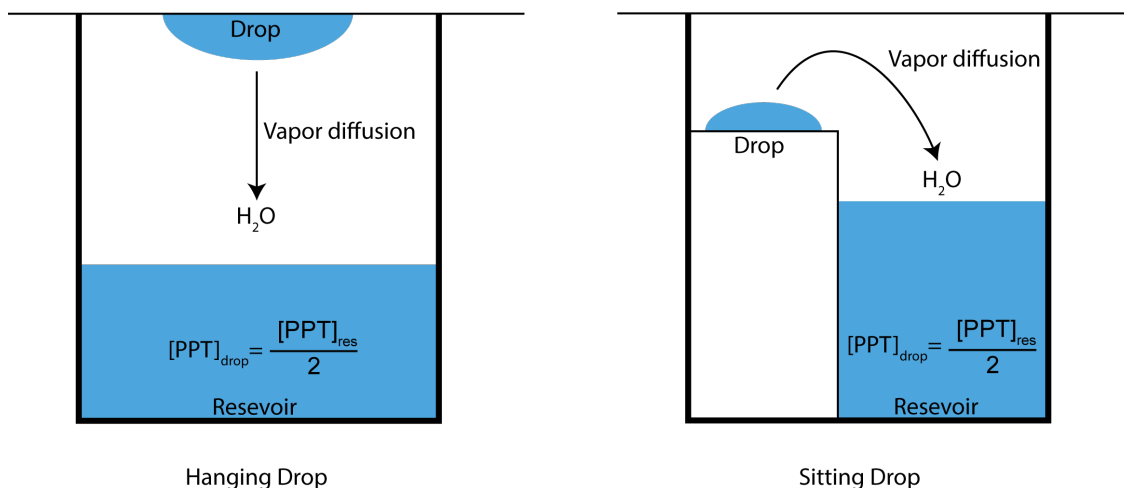


Figure 3.14: Hanging and sitting drop vapour diffusion setups. The difference in precipitant concentrations between the reservoir and the drop facilitates vapour diffusion.

The range of buffer conditions and drop ratios were the same as investigated in EasyXstal plates. The use of SwissCi plates resulted in improved crystal morphology, reproducibility and diffraction. Crystals grown under these conditions were large, approximately 150  $\mu\text{m}$  in length, cubic, but with an internal growth defect (Figure 3.15C). Despite this, the crystals were robust, and relatively easy to manipulate. They showed intolerance to glycerol cryo-protection, but withstood transfer to perfluoropolyether cryo-oil, which allowed me to solve the first high resolution, in-house structure of IDH1-R132H (1.89 Å, Appendix 8.2.7).

Although the IDH1-R132H crystallisation conditions yielded highly diffracting crystals, it showed limited reproducibility with approximately 50 crystals per plate. As crystallographic fragment screening requires approximately 1000 crystals, I further optimised the system to increase the reproducibility.

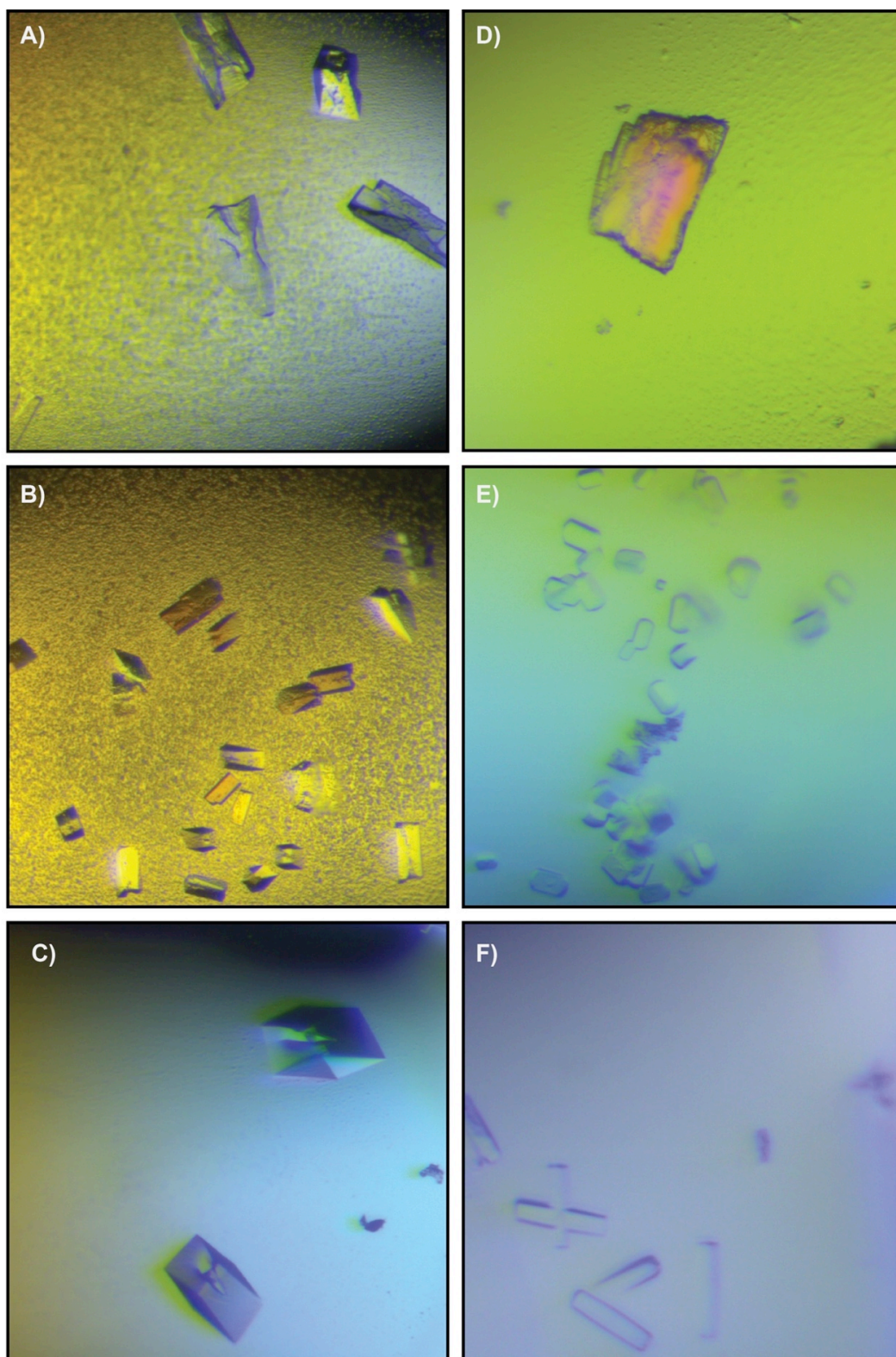


Figure 3.15: Representative crystals from optimisation of IDH1-R132H (A-C) and IDH1-WT (D-F) crystal systems. A) Initial IDH1-R132H crystal hits under published conditions at 18 °C. B) Reducing incubation temperature improved crystal morphology but not diffraction. C) Movement to SwissCi plates with reduction of well and drop volumes improved morphology and diffraction. D) Initial crystal hit for IDH1-WT under published conditions at 4 °C. E) IDH1-WT crystal hit with 1/10 dilution of IDH1-R132H seeds. F) IDH1-WT crystal hit with 1/100 dilution of IDH1-R132H seeds.

Across the tested conditions, 100 mM Tris at pH 6.9 or pH 7.1 with 24% PEG5000MME were the most reliable, with the crystals grown at pH 7.1 diffracting to slightly higher resolution than those grown at pH 6.9. Under these conditions, one plate yielded approximately 100 to 150 crystals showing the characteristic growth defect and diffracted to an average resolution of 2.5 Å. This condition varies slightly from that used for the first high-resolution structure and the average resolution is lower. However, for crystallographic fragment screening it was more important to have a reliable, reproducible system that consistently diffracts well, as opposed to a system that occasionally gives well diffracting crystals, but is less reproducible.

I investigated additional approaches to improve crystal reproducibility and resolution further. Additives are commonly used to improve crystal quality<sup>172</sup>. These small molecules bind to the protein in solution and stabilise it, such that it can form more regular crystals. I screened 96 additives from the Hampton Research Solubility and Stability using SYPRO Orange TSA. Five molecules were identified stabilising IDH1-R132H (Figure 3.16), with Trimethylamine N-oxide dehydrate (TMAO) inducing the largest  $\Delta T_m$ , of 3 °C.

These five additives were repurchased and investigated at a range of concentrations between 1 mM and 250 mM, with 250 mM being the concentration recommended by suppliers. The lowest concentration tested yielded crystals that were morphologically similar and grew in similar buffer conditions to the previous crystals, and diffracted to similar resolution. Higher concentrations of additives caused precipitation under the buffer condition used. When the structures were solved, no additional density for the additive

was observed. In this system, inclusion of additive yielded no improvement in either crystal quality or reproducibility.

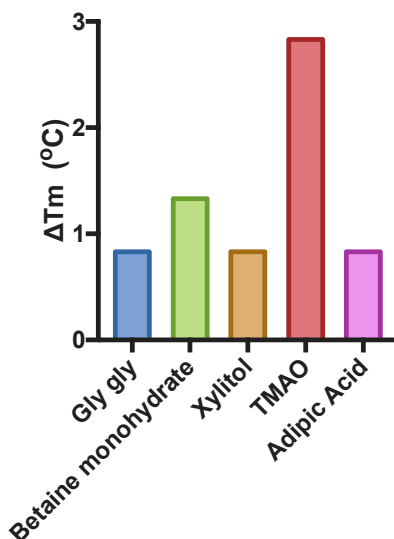


Figure 3.16: Bar chart showing the  $\Delta T_m$  of IDH1-R132H with NADPH induced by hits from the Hampton Research Solubility and Stability screen in comparison to IDH1-R132H with NADPH alone, as calculated by the first derivative. Graphs made in GraphPad Prism.

In addition, I investigated the effect of drop ratio on crystal quality. The drop ratio was increased from the one-to-one ratio of protein to crystallisation buffer, to a two-to-one ratio. These conditions yielded large, cubic crystals that lacked the characteristic growth defect, and also diffracted strongly. Due to the lack of defect they were more robust, easier to handle, and easier to cryo-protect. However, co-factor crystals of almost identical morphology were found growing in the same wells. The similarity in morphology made specific selection of the protein crystals impossible - of the ten crystals tested, five were found to be co-factor crystals. In order to prevent the growth of co-factor crystals under these conditions, I reduced the concentration of NADPH added to protein during pre-incubation, but this caused precipitation of the protein. The two-to-one drop



ratio was therefore considered to be unsuitable for crystallographic fragment screening.

Finally I investigated seeding with IDH1-R132H crystals, but found while this increased the number of crystals forming per well, it did not improve crystal reproducibility or diffracting resolution. Although none of the conditions investigated further improved the crystal system, I was able to routinely grow 100 to 150 crystals per plate, which would routinely diffract to 2.5 Å and were tolerant to 10% DMSO for up to one hour. The system was deemed to be sufficiently reproducible for fragment screening.

### **3.2.5 Establishing IDH1-WT crystallographic conditions**

The original computational analysis did not predict the pocket to be ligandable IDH1-WT. This was based on a single structure at 2.7 Å, with side chains observed out of density around the pocket and may therefore not be reliable. I therefore wished to re-assess the ligandability of the pocket in IDH1-WT in an in-house structure, and established IDH1-WT crystallography to achieve this.

Initial trials of IDH1-WT crystallisation also followed the published conditions, but this yielded irregular, poorly diffracting crystals (Figure 3.15D). Therefore, I investigated similar approaches that helped yield high-resolution crystals in the IDH1-R132H system, including using SwissCi plates and screening a wider range of buffers. However, this failed to improve both crystal morphology and diffraction.

Following this, I investigated seeding with IDH1-R132H in SwissCi plates. Both the 1/10 and 1/100 dilutions of IDH1-R132H seeds yielded crystals in the same

crystallisation buffer that yielded the high-resolution IDH1-R132H crystals (Figure 3.15E and F respectively). The 1/100 dilution produced slightly larger crystals that were easier to manipulate and withstood transfer to oil for cryoprotection. This system yielded the first in-house, high-resolution, in-house IDH1-WT structure (1.85 Å, Appendix 8.2.7). Full experimental details can be found in Chapter 7.2.6.

### 3.2.6 IDH1 structures

Both the IDH1-WT and IDH1-R132H formed crystals with one functional homodimer per asymmetric unit (Figure 3.17A). The unit cell and cell parameters for both IDH1-WT and IDH1-R132H are the same as reported in the literature. The overall structures are highly similar to those published and to each other. Clear density is observed for cofactor NADP<sup>+</sup>/H in both structures (Figure 3.17B), although the resolution is not sufficient to distinguish between the different ring puckers of the oxidised and reduced forms of NADPH.

In the IDH1-R132H structure, the density at position 132 indicates a histidine residue in two conformations (Figure 3.17C). The density at the same position in the IDH1-WT structure corresponds to part of an arginine residue, with density for the terminal amine groups missing (Figure 3.17D). In the inactive conformation, Arg132 makes no stabilising interactions and is therefore highly flexible, leading to weaker electron density for the end of the side chain. This confirms that the mutagenesis was successful and that the variants are as expected.

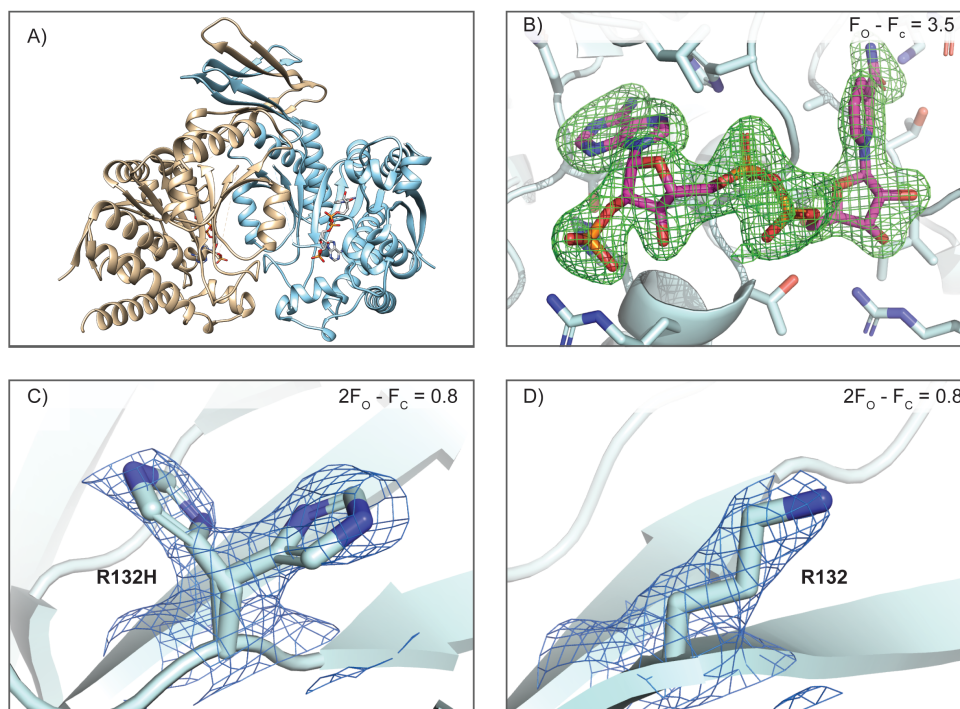


Figure 3.17: Structures of IDH1-R132H and IDH1-WT. A) Overall structure of IDH1-R132H homodimer. B)  $mF_o - DF_c$  density contoured at  $3.5 \sigma$ , showing density for  $NADP^+/H$ , representative of both IDH1-R132H and IDH1-WT. C)  $2mF_o - DF_c$  map contoured at  $0.8 \sigma$ , showing electron density at position 132 in IDH1-R132H. D)  $2mF_o - DF_c$  map contoured at  $0.8 \sigma$ , showing electron density at position 132 in IDH1-WT. Figures made in Pymol<sup>173</sup>

### 3.2.7 Confirming pocket ligandability

I assessed the ligandability of the high-resolution, in-house structure of IDH1-R132H in order to confirm that the novel secondary site is predicted to be ligandable. The presence of the secondary site was confirmed by the computational analysis, but was not initially predicted to be ligandable (Figure 3.18A). Despite being at higher resolution, no electron density was associated with the side chain of Lys345, and I was unable to model it. Density was observed for Lys345 in the published structure and the whole residue is modelled, forming the border of the pocket. Without Lys345, the pocket is not completely formed, leading to the loss of ligandability. Modelling of Lys345

rotamers in the in-house structure rescues pocket ligandability (Figure 3.18B), showing the importance of residue completeness on the ligandability assessment.

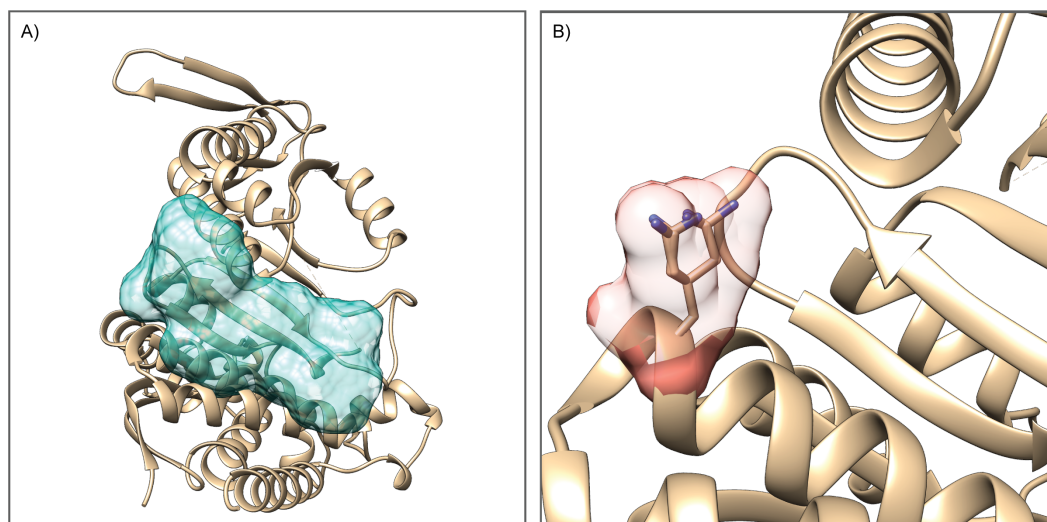


Figure 3.18: Confirmation of ligandability in the novel secondary site in in-house IDH1-R132H structures. A) In house structure IDH1-R132H showing the predicted secondary site. B) Lys345 rotamers modelled in to assess ligandability. Figures made in Chimera<sup>65</sup>.

Similarly, the ligandability of the IDH1-WT structure was re-assessed using the high-resolution in-house structure. In the published structure, the secondary site was identified but not predicted ligandable. In the new structure, the pocket was predicted to be ligandable. Small changes in the backbone and side chain orientation changed how the edges of pocket are defined and subsequently the calculated enclosure for this pocket. This may be due to the increased resolution of the in-house structure in comparison to the published structure, or the use of the different crystallisation conditions and IDH1-R132H seeds. The predictions show the sensitivity of the predictor to both protein mobility and model completeness, as discussed in Chapter 6.2.1. Following fragment screening against IDH1-R132H, I also investigated fragment hits for their effect on IDH1-WT to investigate whether this ligandability prediction is correct.

### 3.3 Conclusions

The chapter describes the enabling technologies established in order to carry out a fragment screening campaign against IDH1-R132H. IDH1-WT and IDH1-R132H constructs were cloned, and the variants expressed with high yields and were purified to greater than 95% purity. IDH1-WT and IDH1-R132H were characterised using label-free thermal shift assays, confirming their ability to bind co-factor and therefore their suitability for use in assays and crystallography. I established conditions for two fragment screening approaches - a SYPRO Orange thermal shift assay, and reproducible crystal systems for both IDH1-R132H and IDH1-WT with the aim of completing a crystallographic fragment screen against IDH1-R132H. I solved high-resolution structures of both IDH1-R132H and IDH1-WT, and computationally confirmed the ligandability of the novel pocket in these structures. Fragment screening against IDH1-R132H was therefore considered to be feasible.

## Chapter 4: Fragment screening to investigate the ligandability of the novel secondary site in IDH1-R132H

### 4.1 Introduction

I used two fragment-screening approaches to investigate the ligandability of the novel pocket - SYPRO Orange TSA and crystallographic fragment screening. TSA has previously been described for in the literature<sup>156</sup>, and is also rapid to establish and relatively high-throughput. Crystallographic fragment screening is highly sensitive to weak binders and immediately elucidates the binding mode, but requires significantly more protein, is more challenging to establish and analysis is more complex.

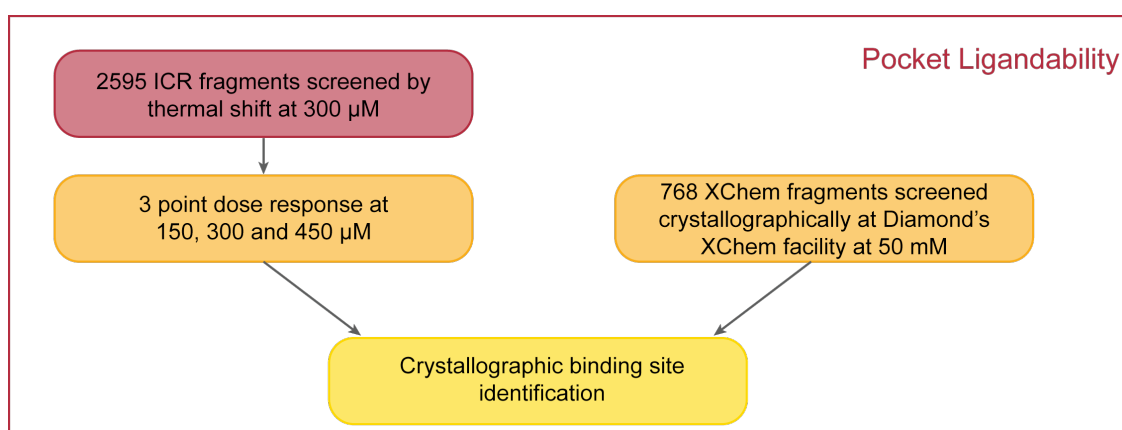


Figure 4.1: Overview of the fragment-screening cascade to identify hit matter binding to the novel secondary site.

This chapter describes the TSA and crystallographic fragment screens (Figure 4.1). In both systems, NADPH was added in order to favour the inactive conformation in which the novel pocket was predicted to be ligandable, and block the primary site. I discuss the results from the two screens in

combination, including the emergence of conserved binding modes for the identified fragments and remodelling of the novel pocket.

## 4.2 Fragment screen by TSA

The SYPRO Orange TSA fragment screen was carried out as shown in Figure 4.2. Both the ICR's and the 3D Fragment Consortium<sup>174</sup> libraries, 2595 fragment in total, were screened at 300  $\mu\text{M}$  against IDH1-R132H in the presence of 500  $\mu\text{M}$  NADPH (Figure 4.2). The conditions used for fragments screening are shown in Table 3.2.

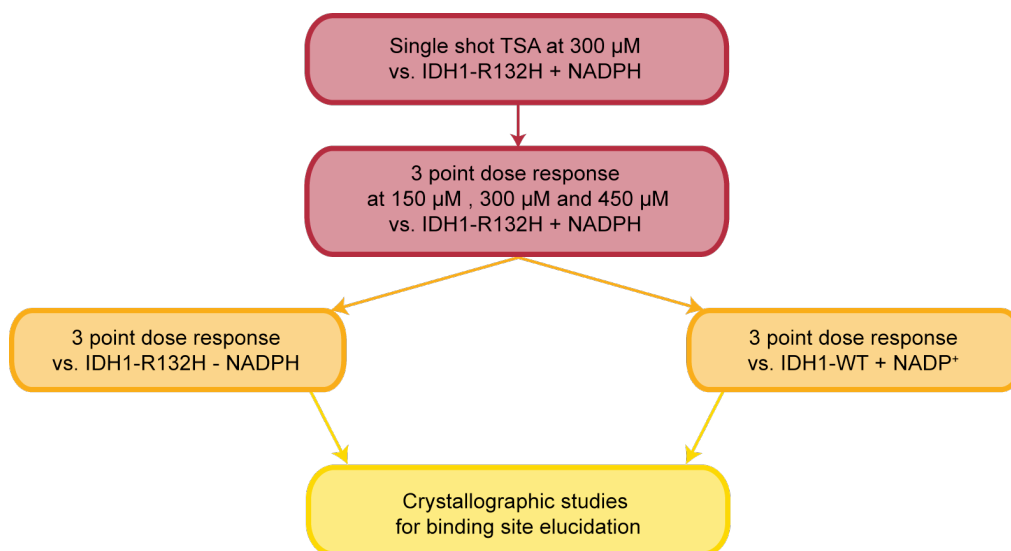


Figure 4.2: Overview of the thermal shift fragment screening cascade to identify hit matter targeting the novel secondary site and investigate selectivity between different conformations of IDH1-R132H.

In the screen, the mean melting temperature for IDH1-R132H was found to be  $56.75\text{ }^{\circ}\text{C} \pm 0.5\text{ }^{\circ}\text{C}$ . Fragments with a  $\Delta T_m$  values greater than 1.5 standard deviations above the mean ( $\Delta T_m \geq 0.86\text{ }^{\circ}\text{C}$ ) were selected as primary hits. This threshold was chosen to be inclusive, but above the variation seen in the baseline measurements. In addition, curves were manually inspected to identify

datasets where the Boltzmann analysis failed (see Chapter 3.2.3.1). For these datasets, fragments with  $\Delta T_m \geq 0.9^\circ\text{C}$  as calculated by the first derivative were included. This yielded 170 primary fragment hits (Figure 4.3).

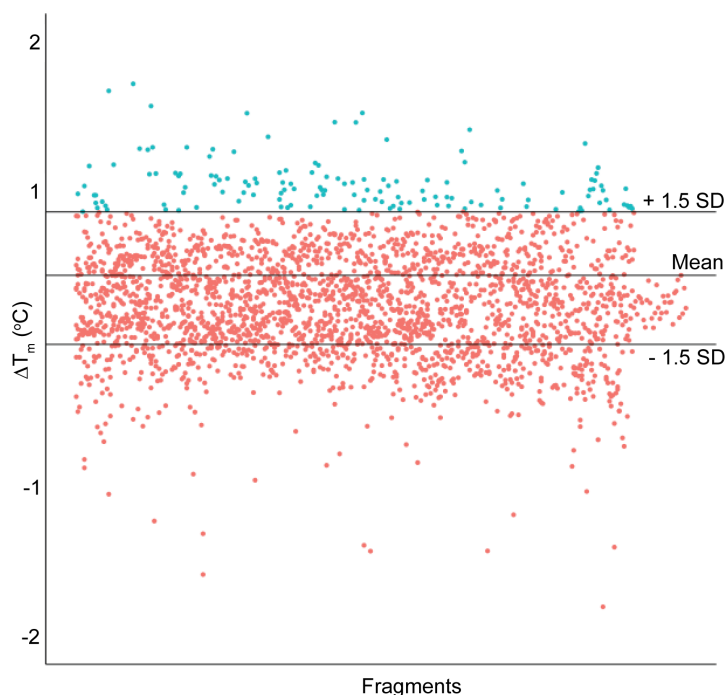
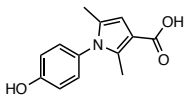
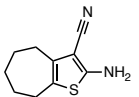
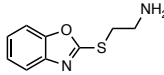
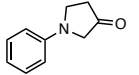
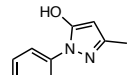
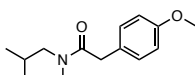
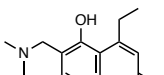
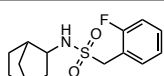
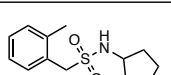
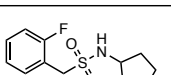
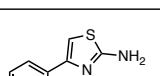
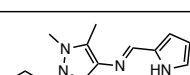


Figure 4.3: Scatter plot showing  $\Delta T_m$  of all 2595 fragments screened against IDH1-R132H, with the 170 fragments identified as hits with a  $\Delta T_m \geq 1.5$  standard deviations above the mean ( $\Delta T_m \geq 0.86^\circ\text{C}$ ; blue). These were taken forwards into a three-point dose response. Plot made in R<sup>168</sup> with ggplot2<sup>147</sup>

The primary hits were retested at three concentrations, 150  $\mu\text{M}$ , 300  $\mu\text{M}$  and 450  $\mu\text{M}$ , against IDH1-R132H with NADPH. Of the 170 primary hits, 20 showed either a concentration-dependent increase in  $\Delta T_m$  across the three concentration points measured, or maintained the same stabilisation across the three concentrations tested (Table 4.1).



Fragment ID	Structure	Concentration (μM)	IDH1-R132H + NADPH		IDH1-R132H - NADPH	
			Boltzmann ΔTm	1st Derivative ΔTm	Boltzmann ΔTm	1st Derivative ΔTm
CCT175011		150	0.18	0.08	-1.39	0.50
		300	1.05	0.58	0.87	1.00
		450	-	1.08	0.00	44*
CCT202357		150	0.19	0.08	0.00	1.00
		300	-0.39	0.08	0.00	1.50
		450	-0.01	0.58	4.07*	0.50
CCT239544		150	0.71	1.00	0.00	0.50
		300	1.36	1.50	-2.05	1.00
		450	1.40	1.50	-1.52	1.00
CCT239559		150	-	0.50	0.00	0.50
		300	-	1.00	-0.76	0.50
		450	-	1.00	0.00	0.50
CCT239686		150	-0.02	0	-0.78	0.00
		300	0.09	0.5	0.71	0.50
		450	-	0.5	2.04	0.50
CCT240016*		150	0.26	0.50	0.78	0.00
		300	-0.03	0.50	-2.05	0.00
		450	0.13	0.50	0.00	21.5*
CCT240509		150	-	37.5*	0.00	0.50
		300	0.68	1	0.00	0.50
		450	0.53	1	0.00	0.50
CCT240768		150	-0.08	0	-1.43	0.00
		300	0.23	0.5	-0.34	0.50
		450	0.30	0.5	-1.13	0.00
CCT240771		150	-0.08	0	0.00	0.00
		300	0.23	0.5	0.00	0.00
		450	0.30	0.5	-1.13	0.00
CCT240772		150	-0.30	0.5	0.00	-17.5*
		300	-	38.5*	0.00	44*
		450	-0.04	1	-0.75	0.00
CCT242635		150	1.04	1	1.21	0.50
		300	2.25	1.5	2.69	0.50
		450	3.19	1.5	6.54*	0.50
CCT242645*		150	0.84	1	0.00	0.50
		300	0.58	1	0.00	0.50
		450	0.96	1.5	0.00	0.50

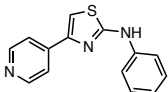
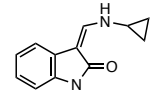
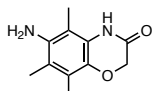
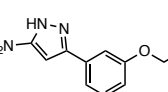
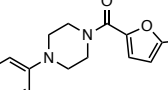
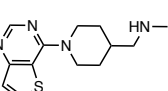
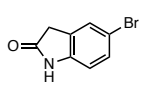
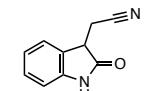
Fragment ID	Structure	Concentration ( $\mu\text{M}$ )	IDH1-R132H + NADPH		IDH1-R132H - NADPH	
			Boltzmann $\Delta T_m$	Derived $\Delta T_m$	Boltzmann $\Delta T_m$	Derived $\Delta T_m$
CCT242649		150	1.04	1	0.00	-0.50
		300	1.61	1	15.30*	-0.50
		450	1.65	1	0.10	0.00
CCT242650		150	0.55	0.5	0.00	1.00
		300	2.30	2.5	0.00	2.00
		450	0.57	0.5	0.00	0.50
CCT242722°		150	0.28	1	-0.98	0.00
		300	0.33	1	1.29	0.50
		450	0.50	1	0.00	0.50
CCT242759		150	0.20	0.5	-3.43	-1.00
		300	0.46	1	-1.43	0.00
		450	3.64	1.5	-1.18	0.00
CCT242817		150	-0.48	0.50	0.00	0.00
		300	0.08	0.50	-1.08	0.00
		450	0.58	0.50	-25.63*	-0.50
CCT243076°		150	0.57	1	-0.83	0.00
		300	0.66	1	-0.48	0.00
		450	0.66	1	-51.60*	-26.5*
CCT243079°		150	1.33	1.5	0.42	0.50
		300	1.72	1.5	0.94	0.50
		450	1.47	1.5	5.61*	1.00
CCT289127		150	0.13	0.5	-1.75	0.00
		300	0.19	0.5	-1.31	0.00
		450	0.46	0.5	-1.45	0.00

Table 4.1: Thermal shift dose response for the 20 fragments identified as hits. Most fragments induced a greater shift in the IDH1-R132H+NADPH system than in either the absence of NADPH, or in the IDH1-WT system. Thermal shifts as calculated by both the Boltzmann distribution and the first derivative are reported. Unreported values indicates failure of the Boltzmann analysis to report a melting temperature

\* indicates either selection of the incorrect peak by the Boltzmann, or the protein precipitating out of solution leading to flat curves, resulting in large ( $>10^\circ\text{C}$ ) shifts.

° indicates a fragment that could not be repurchased.

These 20 fragments were also screened against IDH1-R132H in the absence of NADPH. Under these conditions, only CCT175011, CCT240509 and CCT242650 stabilised IDH1-WT (Table 4.1); most fragments gave a little to no shift, indicating that they may bind in to a pocket only present with the co-factor.

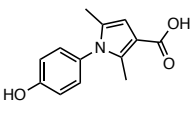
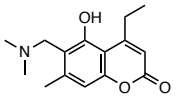
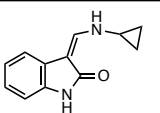
Fragment ID	Structure	Concentration ( $\mu\text{M}$ )	IDH1-WT + NADP <sup>+</sup>	
			Boltzmann $\Delta T_m$	1st Derivative $\Delta T_m$
CCT175011		150	0.33	0.17
		300	0.78	0.67
		450	1.13	0.67
CCT240509		150	0.31	0.67
		300	0.37	0.17
		450	0.34	0.67
CCT242650		150	0.34	0.17
		300	0.60	0.17
		450	0.80	0.67

Table 4.2: fragments that showed stabilisation of IDH1-WT with NADP<sup>+</sup>.

The 20 confirmed fragment hits were also tested against IDH1-WT with NADP<sup>+</sup>. Only three of the fragments caused a measurable  $\Delta T_m$  in this system (Table 4.2). This may indicate that the 17 fragments that did not affect IDH1-WT are selective for IDH1-R132H, or that the IDH1-WT system is less sensitive to weakly binding fragments. Of the 20 confirmed fragments, 15 were available for re-purchase as indicated in Table 4.1, and taken forwards for crystallographic studies.

#### 4.2.1 Binding Site elucidation

In order to elucidate their binding sites, fragments were directly soaked at into IDH1-R132H crystals. Fragment stocks were made at 500mM, and added to drops with a final concentration of 50 mM fragment with 10% DMSO, which is in the range of concentrations commonly used for fragment screening<sup>106, 175</sup>. Several crystals were soaked and fished for each fragment, and 44 datasets from individual crystal soaking experiments were collected at the Diamond Light Source (DLS). Structures were solved by molecular replacement using

my in-house IDH1-R132H structure as a search model, followed by a round of refinement with Buster<sup>176</sup>. Manual corrections were made to the structure before a second round of refinement, after which the electron density was inspected to identify fragment hits. The datasets were also analysed by PanDDA where possible, to identify fragments binding with low occupancy or high mobility. Of the 15 fragments soaked, five yielded fragment-bound crystal structures. Pleasingly, all of these fragments were identifying binding into the novel secondary site. The binding modes are described alongside the crystallographic fragment hits later in this chapter (Chapter 4.4).

## **4.3 Crystallographic fragment screening**

### **4.3.1 Fragment soaking and data collection**

The experiment was run across three visits to DLS, with the help of Dr Yann-Vaï Le Bihan (ICR), Dr Matthew Rodrigues (ICR) and Dr Alice Douangamath (DLS). The Diamond-SGC-iNEXT Poised (DSiP) library of 768 fragments at 500 mM was soaked into IDH1-R132H crystals. Final soaking conditions were 50 mM fragment with 10% DMSO, which is in line with the conditions from the manual soaking of TSA fragment hits. Plates were incubated for one hour at room temperature before harvesting. Although perfluoropolyether had been previously used as a cryo-protectant, this cryo-oil was too viscous to be acoustically dispensed by the ECHO instrument and was therefore not usable for these experiments. While other ECHO-compatible cryo-protectants such as Ethylene glycol (MEG) were investigated, I found that 10% DMSO alone was sufficient for robust cryo-protection of the crystals.

Of the 768 fragments investigated, ten caused crystals to melt. In total, 910 datasets were collected, including 39 datasets soaked with DMSO only, and 103 soaked with fragment hits from TSA.

Data was automatically processed on the DLS servers, and imported into XChem-Explorer (XCE). The 'best' auto-processing result was selected by XCE based on considerations such as the resolution, completeness,  $R_{merge}$  and  $Mn<I/sig(I)>$ . Although the space group,  $P4_32_12$ , was pre-specified during set up, the auto-processing often found the incorrect enantiomorphic space group and processed the data in  $P4_12_12$ . This does not cause the molecular replacement to fail, but gives lower quality results due to poorly assigned systematic absences causing low completeness. Where possible, the processing with the correct space group was manually selected, but was not always available. Following processing, the structures were then solved within XCE by molecular replacement using the DIMPLE software. The initial search model was based on multiple high-resolution, in-house IDH1-R132H structures. As well as the protein and co-factor, conserved water and buffer molecules were also retained to give the best initial electron density maps possible. Simultaneously, the ligand restraint files for the soaked fragments are generated using Grade<sup>176, 177</sup>, or aceDRG<sup>178</sup> where Grade failed. Finally, the output from molecular replacement was used for PanDDA analysis, which was run from within XCE.

### 4.3.2 Ground state map optimisation

The ground state map that PanDDA builds is an average of all of the  $2mF_o - DF_c$  electron density maps that are used to calculate it (Chapter 3.1.2.1). In regions where protein shows less mobility and a single conformer, the ground state map is very clear, but the density becomes highly blurred where differences between the datasets are observed. In order to calculate the cleanest ground state map, the individual datasets should show little structural variation. I therefore used an initial PanDDA after each visit to identify which datasets deviated from the ground state and should be excluded from future iterations of ground state map calculation. Datasets at low resolution ( $> 3 \text{ \AA}$ ) were also excluded as they could have a negative impact on the quality of the ground state map.

The majority of datasets showed IDH1-R132H in the same conformer as the model used for molecular replacement, with two recurrent deviations that were not mutually exclusive. The first deviation was the movement of a loop at the edge of the novel secondary site, residues 110-126, which is observed in two discrete conformations (Section 4.4.4). Approximately 3% of the datasets collected showed this loop in a different conformation from the molecular replacement model, with occupancies varying between 10 and 100%. These datasets were removed from ground state analysis as it represented a large deviation from the ground state.

The second recurrent deviation was the movement between the dimers, and between the domains within the dimer. Whilst the majority of the monomer aligns well, the clasp domain adopts a range of conformations between two

extremes (Figure 4.4). This results in significant deviation from the ground state at multiple sites, resulting in a large number of false positive events identified by PanDDA.

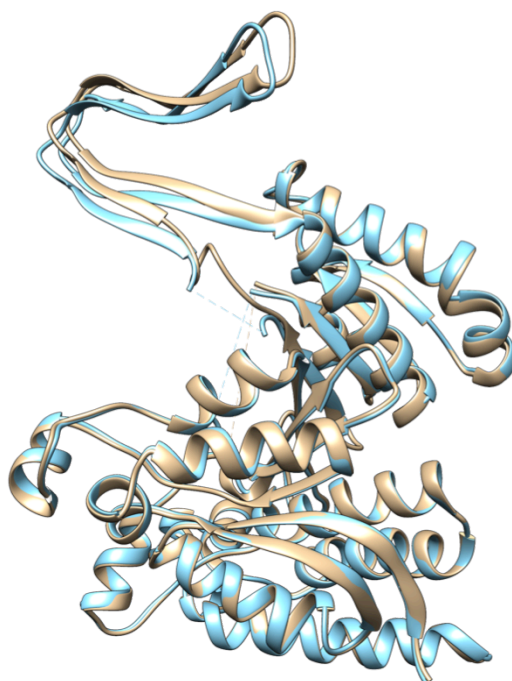


Figure 4.4: Overlay of IDH1-R132H chain A from two different structures (monomer shown for simplicity). Whilst the majority of the structure overlays well, the clasp domain shows variation between the structures, which is sufficient to be identified as an event during PanDDA analysis, and also acts to blur the ground state map in this region. Figure made in Chimera<sup>65</sup>.

Once a clean ground state map was generated by discarding datasets that showed significant deviation, I corrected the original molecular replacement model to improve the fit to the ground state map. Re-solving the structures using this improved model and re-running PanDDA can further improve the initial electron density and ground state maps, and therefore improves event detection and cleaner density in PanDDA maps. Whilst theoretically each round will continuously improve initial electron density maps and the subsequent Z-maps and PanDDA maps, like during normal crystallographic refinement, the initial model used for molecular replacement fitted the ground state map well,

and did not require extensive correction. Given this, the molecular replacement model was only corrected once against the optimised ground state.

The generation of this cleaner ground state map allowed identification of additional events that would have otherwise been missed. For example, no event was identified when dataset IDH1-x0056 was initially analysed, but was when the data was reanalysed with an optimised ground state (Figure 4.13A), allowing for identification of CCT370982 as an XChem fragment hit.

### 4.3.3 Summary of results

After optimisation of the ground state and the MR model as described above, a final PanDDA analysis identified 1536 events in the 801 datasets retained for characterisation (Figure 4.5). Whilst 910 datasets had been collected, some were excluded as they were known to be apo, or were rejected due to too low resolution ( $> 3.5 \text{ \AA}$ ), incorrect space group assignment or molecular replacement failure (initial  $R_{work} > 40\%$ ). For the latter two, attempts were made to change the auto-processing result so that they could be included. However, in the majority of cases, these two features were symptomatic of poor quality, low-resolution data.

Each of the identified events was manually inspected. The majority were due to the variations in clasp conformation or monomer movement as described previously, as well as changes in the co-factor occupancy that were also recognised as events. Events where the density matched the shape of the soaked fragment were modelled and refined once with Refmac5. Fragments



that remained clear following the refinement and were supported by PanDDA statistics (Chapter 3.1.2.1) were exported for further refinement with Buster.

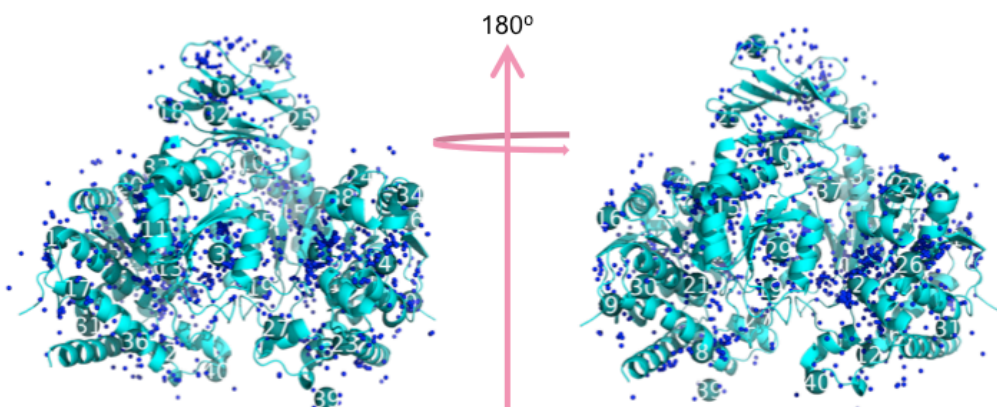


Figure 4.5: Summary of sites identified by the PanDDA analysis, as generated by the PanDDA program. Many of the identified sites are due to clasp movement as described earlier; in addition, differences in co-factor occupancy were also identified as events.

In total, 14 fragments from the DSiP library were identified binding specifically to the novel site, showing that this secondary site is ligandable. In addition, five fragments were identified binding into the known allosteric site (Appendix 8.2.8.2). This shows that the crystal system is amenable to fragments binding to both sites, but fragments were preferentially and specifically bound to the novel site.

#### **4.3.4 Thermal shift assays to support the presence of fragments identified by crystallographic fragment screening**

Fragments identified through crystallographic fragment screening were also investigated by SYPRO Orange TSA. Fragments identified through XChem were therefore tested at 500  $\mu$ M, 1mM and 2mM (Table 4.3) to reflect the higher fragment concentration used during screening.

XChem fragment hits - novel pocket binding					
Fragment	Binding mode	$\Delta T_m$ (1st derivative) °C			1-BDC
		0.5mM	1mM	2mM	
CCT370971	Benzoimidazole	0	0	0	0.37
CCT370970	Benzoimidazole	0	0.5	0.5	0.45
CCT370974	Loop move	0	0	0	0.38
CCT370982	Loop move	0	0	0	0.46
CCT370980	Loop move	0	0	0	0.4
CCT370978	Loop move	0.07	0.09	0.18	0.44
CCT370979	Loop move	0	-0.5	-0.5	0.42
CCT370977	Loop move	0	0	0	0.46
CCT370973	Singlet	0	0	0	0.31
CCT371095	Singlet	0.02	0.03	0.18	0.34
CCT371098	Trp205-stack	0	0	0	0.42
CCT154567	Trp205-stack	0	0	0.5	0.55
CCT373604	Trp205-stack	0	0	0	0.51
CCT372954	Trp205-stack	0	0	0	0.54

Table 4.3: Thermal shift values for fragments identified as XChem hits. There doesn't appear to be a correlation between the extent of the shift with either the binding mode or the estimated occupancy

Of the 14 tested fragments, only CCT370970 and CCT154567 showed significant stabilisation of IDH1-R132H, while the other fragments showed little to no affect on the thermal stability of IDH1-R132H. There did not appear to be a correlation between estimated occupancy and measured thermal shift. Some XChem fragments that showed clear electron density, such as CCT370974, did not give a measurable shift. This reflects the sensitivity of crystallographic fragment screening to weakly binding fragments, and yields more hit matter in comparison to other techniques such as TSA.

## 4.4 Overview of fragment hits from both fragment screens

Across the two fragment screens, 19 fragments were found to bind specifically to the novel secondary site (Table 4.3, Figure 4.6). Fragments were identified binding across the breadth of the pocket, with 11 showing clear electron density in normal  $2mF_o - DF_c$  maps. These structures were prioritised for refinement. The remaining eight fragments showed weak or ambiguous density and were de-prioritised for refinement.

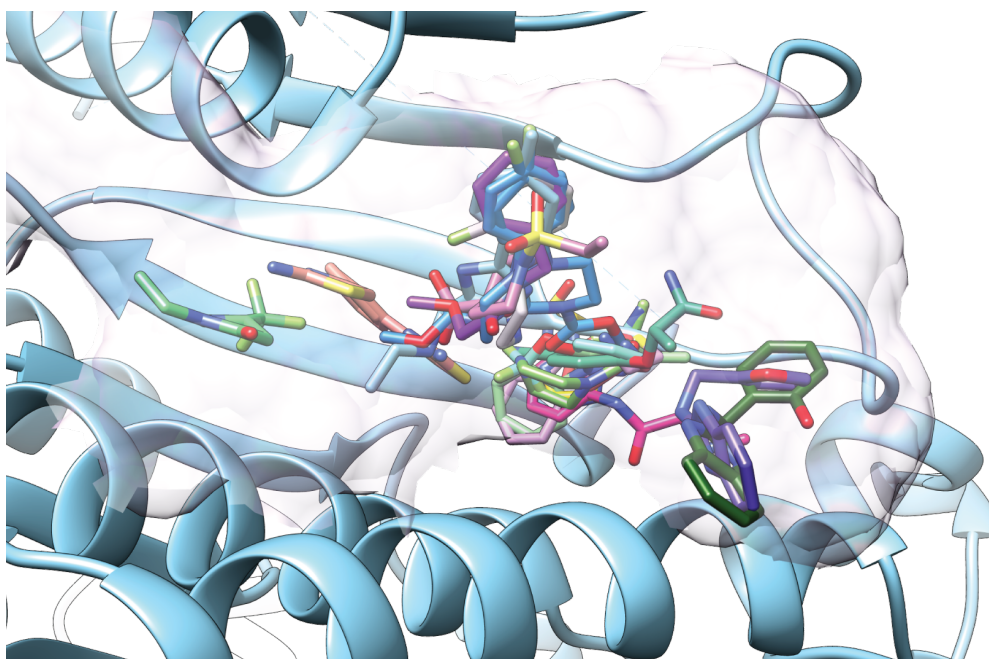
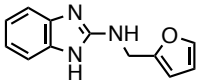
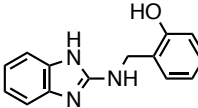
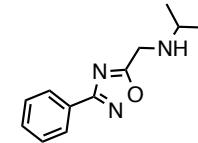
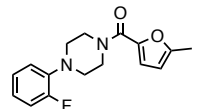
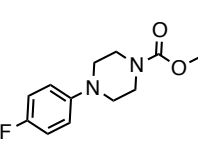
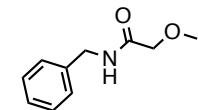
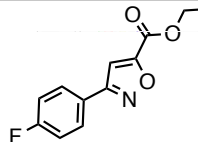
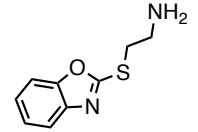
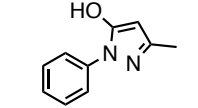
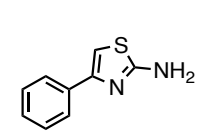
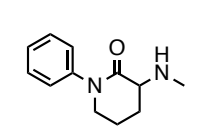


Figure 4.6: Overlay of the 19 fragments identified binding into the novel secondary site by thermal shift and crystallographic fragment screening. Fragments bind to the novel pocket with varying binding modes, across the width of the pocket. Figure made in Chimera<sup>65</sup>

Compound	Structure	RSCC	Occupancy	Binding Mode	Screen	Refined
CCT370970		0.85	1	Glu361 stacking	XChem	Yes
CCT370971		0.70	1	Glu361 stacking	XChem	Yes
CCT154567		0.85	1	Trp205 stacking	XChem	Yes
CCT242817		0.80	1	Trp205 stacking	TSA	Yes
CCT371098		0.70	1	Trp205 stacking	XChem	Yes
CCT372954		0.74	1	Trp205 stacking	XChem	Yes
CCT373604		0.70	1	Trp205 stacking	XChem	Yes
CCT239544		0.78	0.79	Loop remodelling	TSA	Yes
CCT239686		0.83	0.86	Loop remodelling	TSA	Yes
CCT242635		0.79	1	Loop remodelling	TSA	Yes
CCT370974		0.85	0.86	Loop remodelling	XChem	Yes

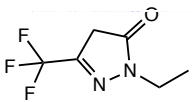
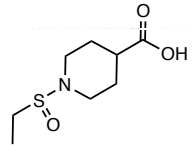
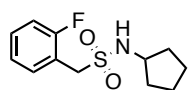
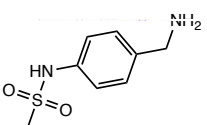
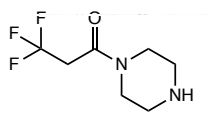
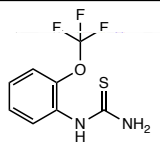
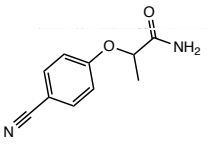
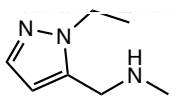
Compound	Structure	RSCC	Occupancy	Binding Mode	Screen	Refined
CCT370974		-	-	Singlet	XChem	No
CCT371095		-	-	Singlet	XChem	No
CCT240772		-	-	Loop remodelling	TSA	No
CCT370972		-	-	Loop remodelling	XChem	No
CCT370978		-	-	Loop remodelling	XChem	No
CCT370979		-	-	Loop remodelling	XChem	No
CCT370980		-	-	Loop remodelling	XChem	No
CCT370982		-	-	Loop remodelling	XChem	No

Table 4.4: Overview of fragment screening hits identified binding to the novel secondary site, their binding mode and screening technology through which they were identified. Refinement statistics for the eleven fully refined structures can be found in Appendix 9.X; the remaining eight structures showed low fragment occupancy and were deprioritised.

#### 4.4.1 Non-conserved binding modes: CCT371095 and CCT370874

PanDDA identified two maps showing deviation from the ground state next to Arg338 (Figure 4.7, left). The PanDDA maps allowed both fragments to be modelled, and the calculated RSCC following one round of refinement supports the presence of the fragment, although the RSZO/OCC is low (Appendix 8.2.8.1). Normal  $2mF_o - DF_c$  maps also show density next to Arg338 that overlaps with that observed in PanDDA, but does not allow clear modelling of these fragments (Figure 4.7, right). Due to this, these structures have not been fully refined.

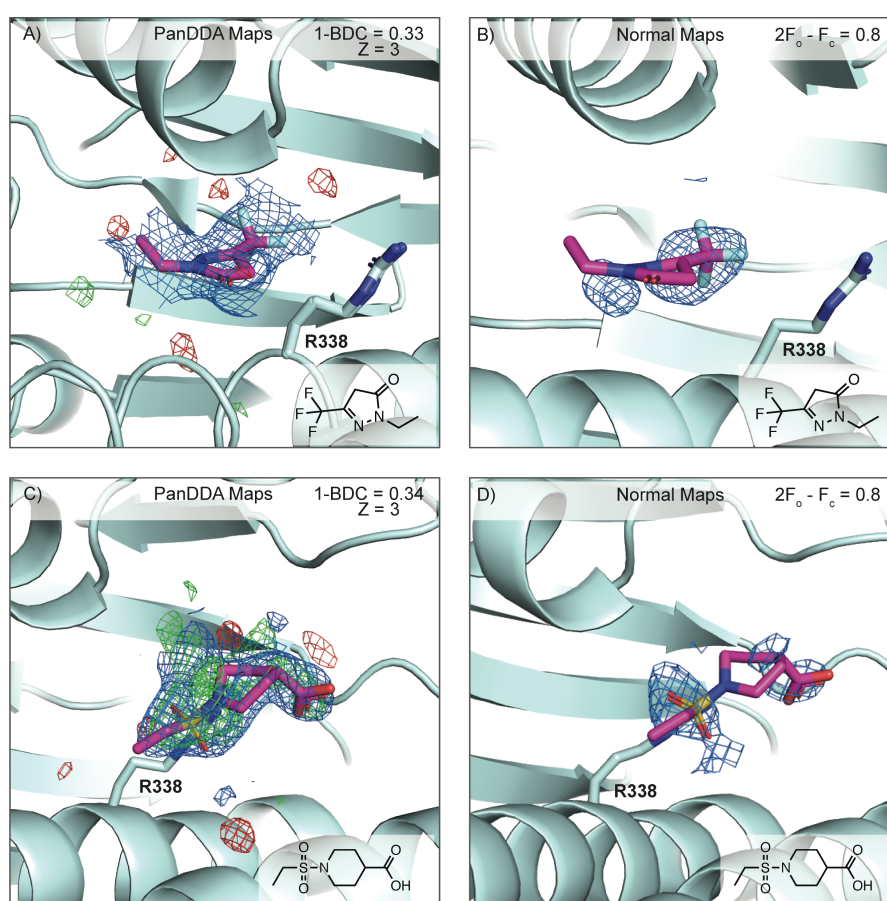


Figure 4.7: Comparison of PanDDA and normal maps for fragments CCT370974 and CCT371095. A) PanDDA event and Z maps for CCT370974; B)  $2mF_o - DF_c$  maps for CCT370974. C) PanDDA event and Z maps for CCT371095; D)  $2mF_o - DF_c$  maps for CCT371095. Figures made in Pymol<sup>173</sup>.

#### 4.4.2 Benzoimidazole series

Two structurally similar fragments from XChem, CCT370970 and CCT370971, were identified binding into the novel secondary site. The PanDDA maps show clear deviation from the ground state (Figure 4.8A, C), with the statistics reported from the initial round of refinement supporting the presence of the fragments (Appendix 8.2.8.1) The statistics in combination with the electron density maps supports the presence of these fragments identified by PanDDA.

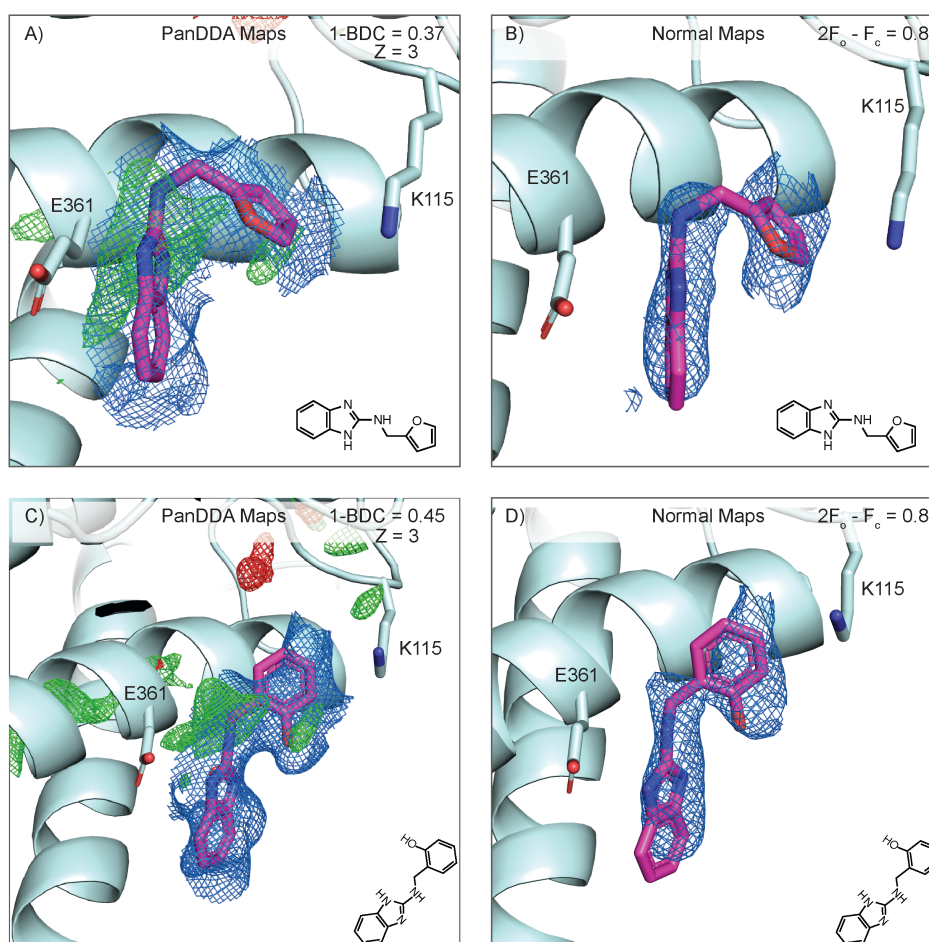


Figure 4.8: Comparison of PanDDA and  $2mF_o - DF_c$  maps for the two structurally similar fragments identified through XChem crystallographic screening. PanDDA maps for both CCT370970 (A) and CCT370971 (C) show clear deviation from the ground state. Normal  $2mF_o - DF_c$  maps show weaker density in the same location (B, D), which allows the fragments to be modelled with 100% occupancy but high mobility. Figures made in Pymol<sup>173</sup>.

Density matching the fragment shape and overlapping well with the PanDDA maps can also be seen in  $2mF_o - DF_c$  maps (Figure 4.8B, D). Occupancy refinement supports modelling of these fragments with 100% occupancies, but with twofold higher B-factor values than the global structure. This is likely due to the sub-optimal interactions these fragments make with the surrounding protein.

In both fragment-bound structures, the benzoimidazole group stack on top of the side chain of Glu361 of chain B, with the phenol and furan moieties forming an internal H-bond with the benzoimidazole. The fragments are not seen binding to the pocket in chain A as the side chain of Glu361 has rotated to stack on top of the backbone of the symmetry-related molecule, which would prevent the benzoimidazole forming this interaction. In addition, the fragments would clash extensively with the symmetry-related molecule. In solution, they could be expected to bind to both chains.

#### **4.4.3 Fragments binding to Trp205**

A total of five fragments from both screens were identified binding to the novel pocket through an edge-face  $\pi$ -stack between a phenyl group common to these fragments and the side chain of Trp205. The density for the groups involved in this interaction is strong in both PanDDA and normal  $2mF_o - DF_c$  maps with 100% occupancy. The other parts of these ligands, however, are more flexible. For CCT242817, which was identified by TSA, the density for the terminal furan is too weak to model this part of the fragment (Figure 4.9), although mass spectrometry shows that the fragment is intact (Appendix 8.2.6).



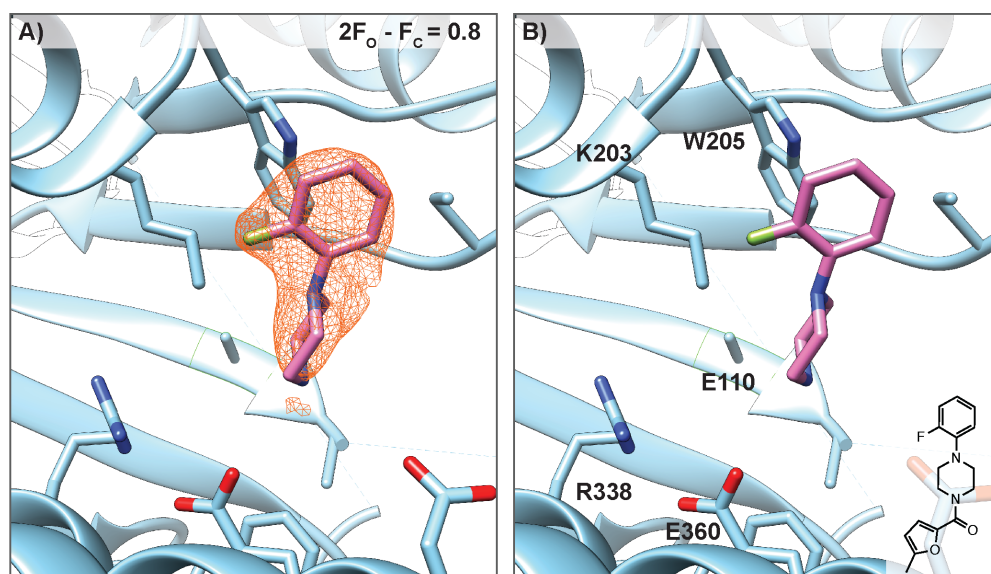


Figure 4.9: Thermal shift hit CCT242817 was identified binding to the novel pocket through an edge-face pi-stack on Trp205. A)  $2F_o - DF_c$  map contoured at  $\sigma = 0.8$ . C) Residues forming the CCT242817 binding site. The density for the fluorophenyl-piperazine group is strong. The methyl-furoyl group cannot be modelled, as the electron density for this region is weak. Figure made in Chimera<sup>65</sup>

Although CCT242817 does not make direct contacts with Arg338 or Glu110, the salt bridge between the two residues is broken. Arg338 rotates such that it faces the solvent and interacts solely with the side chain of Glu360. The side chain of Glu110 can no longer be modelled due to lack of electron density.

For the four fragments identified through crystallographic fragment screening, PanDDA maps showed clear deviation from the ground state and clear electron density in  $2mF_o - DF_c$  maps (Figure 4.10). These fragments share a phenyl group, which stacks on the side chain of Trp205 in the same way as CCT242817. The remaining groups are more flexible and adopt different conformations. For the structures of IDH1-R132H and CCT371098, and IDH1-R132H and CCT373604, the fragment can be modelled in two conformations, further indicating flexibility (Figure 4.10A, C).

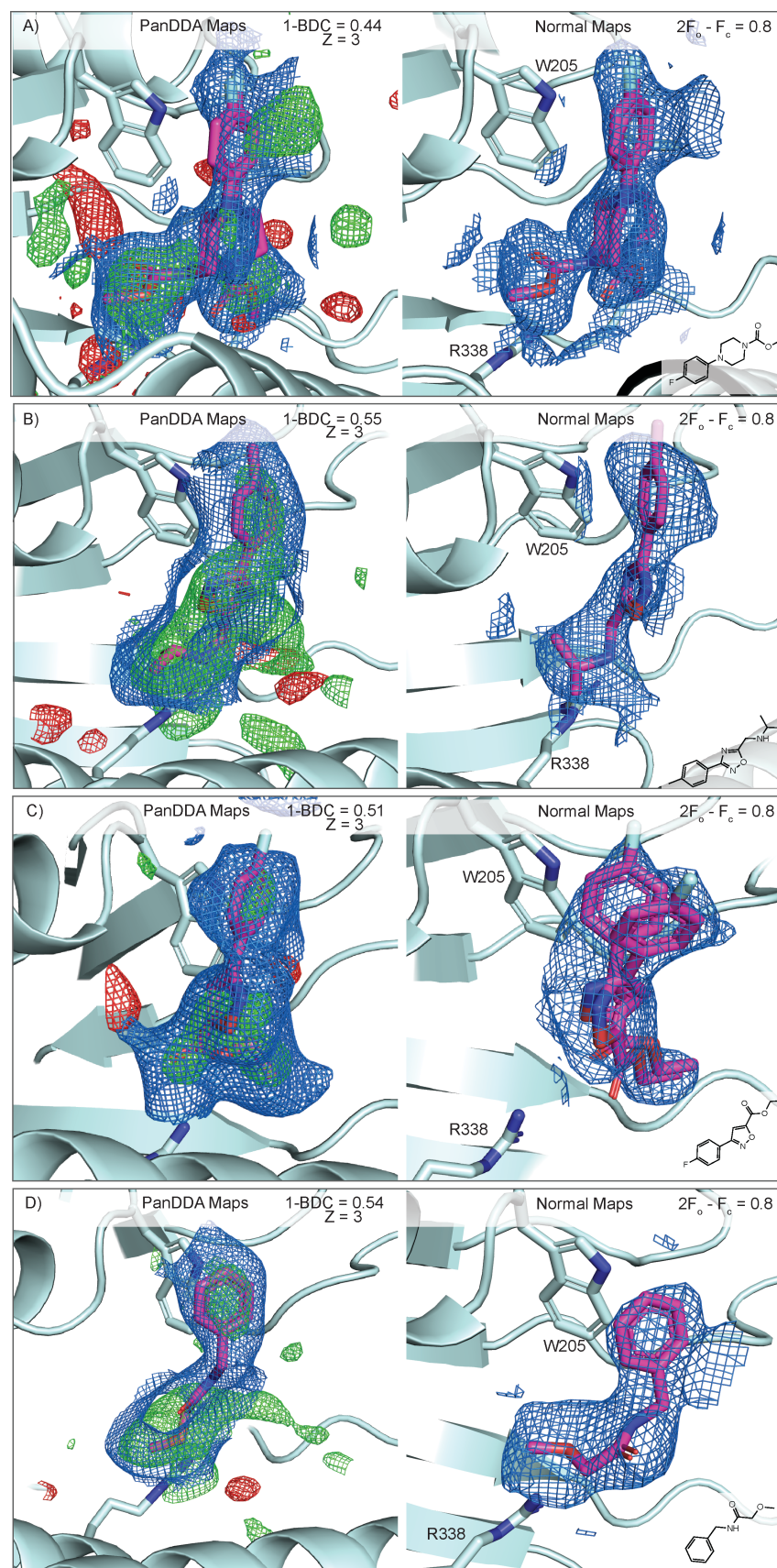


Figure 4.10: Comparison of PanDDA and  $2F_o - DF_c$  maps for Trp205-stacking fragments identified through XChem crystallographic fragment screening. A) CCT371098; B) CCT154567; C) CCT373604; D) CCT372954. PanDDA maps are contoured at  $\sigma = 1\text{-BDC}$  (absolute). Figures made in Pymol<sup>173</sup>.

In the structures with CCT242817, CCT371098 or CCT373604, fragment density is only observed in chain A. For the remaining two fragment-bound structures, with CCT154567 or CCT372954, density corresponding to the fragment can be observed in both chains. This indicates that fragments could be expected to bind to both chains in solution.

Fragment binding to the novel secondary site through an edge-face  $\pi$ -stack on Trp205 causes significant destabilisation of the pocket-forming loop, residues 110-126. This is seen as weaker density for this region. In the CCT242817-bound structure, residues 112, 113, 120-123 cannot be placed due to lack of density. It is also highlighted by an increase in the B-factor for this region in comparison to the structure as a whole. The average B-factor for this loop is 1.06 times the global average in the fragment-free IDH1-R132H structure, but 1.60 times the global average in the CCT242817-bound structure. The structural rationale behind this destabilisation is unclear, but is discussed further in Chapter 5.5.

#### **4.4.4 Fragments binding with re-organisation of the pocket-forming loop**

From the two fragment screening techniques, ten fragments were identified fragments binding to the novel pocket with significant reorganisation of the pocket-forming loop, residues 110-126 (Table 4.4). In the IDH1-R132H-NADPH structure, a salt bridge between Arg338 and Glu110 stabilise the pocket, with a hydrophobic interaction between Ile112 and Phe334 stabilising the pocket-forming loop. The side chain of Arg119 also forms an extensive hydrogen-bonding network to stabilise the  $\beta$ -turn in this loop (Figure 4.11A).

In the fragment-bound conformation, the interaction between Arg338 and Glu110 is broken, and the side chain of Arg338 rotates to face the solvent as was also observed in the structures of Trp205-stacking fragments. The interaction between Ile112 and Phe334 is also broken, allowing movement of the pocket-forming loop to adopt a novel  $\alpha$ -helical conformation that has not previously been reported for IDH1 (Figure 4.11B).

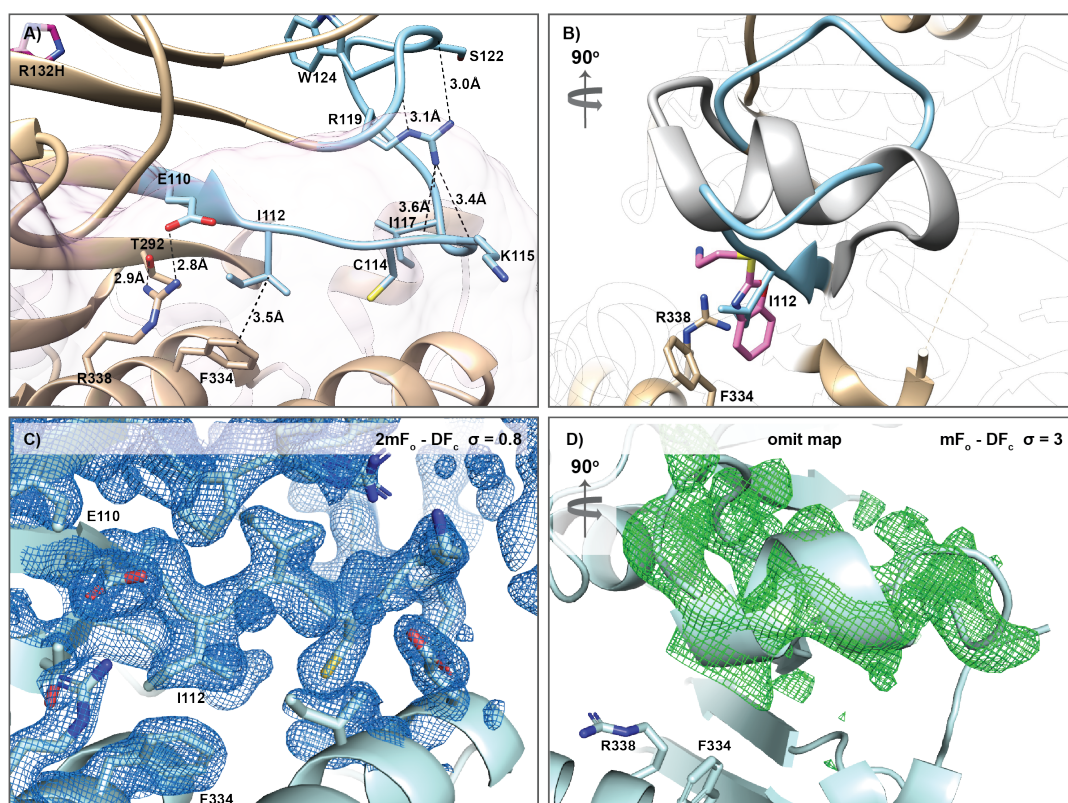


Figure 4.11: Fragments can bind in the space occupied by Ile112 and induce remodelling of the pocket-forming loop. A) The novel pocket predicted ligandable in IDH1-R132H, with the pocket forming-loop (residues 110-126) in cyan. Key stabilising interactions include a hydrophobic interaction between Phe334 and Ile112, as well as an extensive hydrogen-bonding network involving the side chain of Arg119 and the backbone carbonyls of multiple residues. B) Rotation of  $90^\circ$  relative to panel A; the normal loop conformer is shown in cyan. Fragment binding results in displacement of the loop to adopt a novel  $\alpha$ -helical conformation shown in grey. C)  $2mF_o - DF_c$  map ( $\sigma = 0.8$ ) of the pocket-forming loop in the normal conformation in the same orientation as panel A. D) Omit map contoured at  $\sigma = 3$  showing adoption of the novel  $\alpha$ -helix with the fragment binding, in this case with CCT240772 (Figure 4.13D). Figures made in Chimera<sup>65</sup> and Pymol<sup>173</sup>.



This loop movement was only observed in chain A. In chain B, adoption of the new helical conformation would lead to a steric clash between the side chain of Trp124 with the side chain of Pro149 and the backbone carbonyl of Glu174 in the symmetry-related molecule. The crystal packing therefore prevents adoption of the novel  $\alpha$ -helix in chain B. In solution, it could be expected that the fragments would bind to both chains. Within the remodelled pocket, all fragments bind through a hydrophobic interaction with Phe334 in the space previously occupied by Ile112 (Figure 4.12, Figure 4.13, Figure 4.14A-C), except for thermal shift hit CCT242635 (Figure 4.14D).

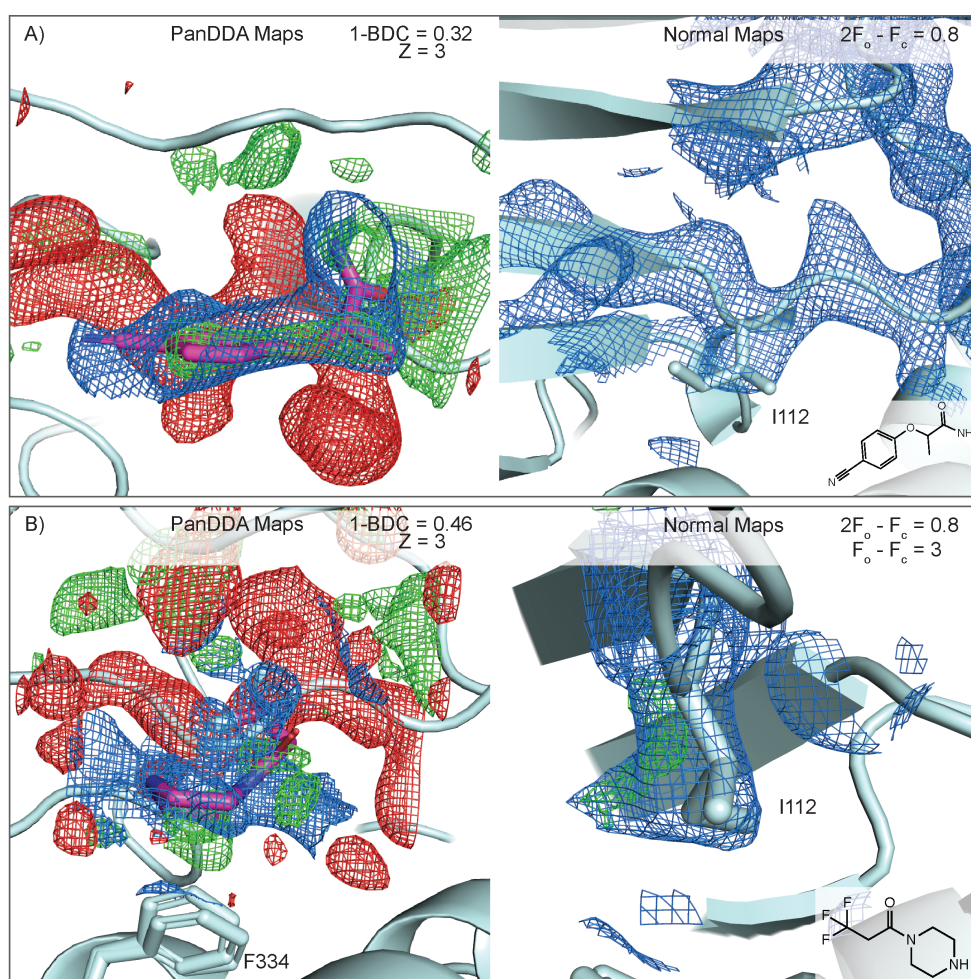


Figure 4.12: Fragments binding with remodelling of the pocket-forming loop that is too low occupancy to be seen in normal  $2mF_o - DF_c$  maps. Both CCT370980 (A) and CCT370978 (B) were identified as XChem hits. Figures made in Pymol<sup>173</sup>.

For two of the identified fragments binding with re-organisation of the pocket-forming loop, CCT370978 and CCT370980, the occupancy was too low to observe the loop movement or fragment binding in normal  $2mF_o - DF_c$  maps. The PanDDA maps show clear deviation from the ground state as negative density where the normal loop is no longer observed (red mesh, Figure 4.12, left) and positive density corresponding to the new helix and the fragment (green mesh, Figure 4.12, left). However, the normal  $2mF_o - DF_c$  maps shows the loop in the normal conformation (Figure 4.12, right). PanDDA statistics (Appendix 8.2.8.1) and maps supported the modelling of these fragments. For CCT370978, additional density that cannot be explained by the isoleucine side chain is observed in the normal maps, which further supports the presence of this fragment.

A further four fragments, CCT370982, CCT370979, CCT370972 and CCT240772, were identified with clear movement of the pocket-forming loop in PanDDA maps. CCT370982, CCT370979 and CCT370972 (Figure 4.13A-C) occupy the space normally taken by Ile112. The loss of the pocket-forming loop is identified as a very strong event, but the fragment overlaps and density therefore seems to be obscured (see Chapter 6.2.5 for a further discussion of this). In contrast, CCT240772 binds by Lys203 and Arg338, further from the remodelled loop. Clear deviation from the ground state can be observed for both the loop movement and the fragment-binding event (Figure 4.13D). The corresponding  $2mF_o - DF_c$  map show movement of the pocket forming loop, but very weak density for the fragment.

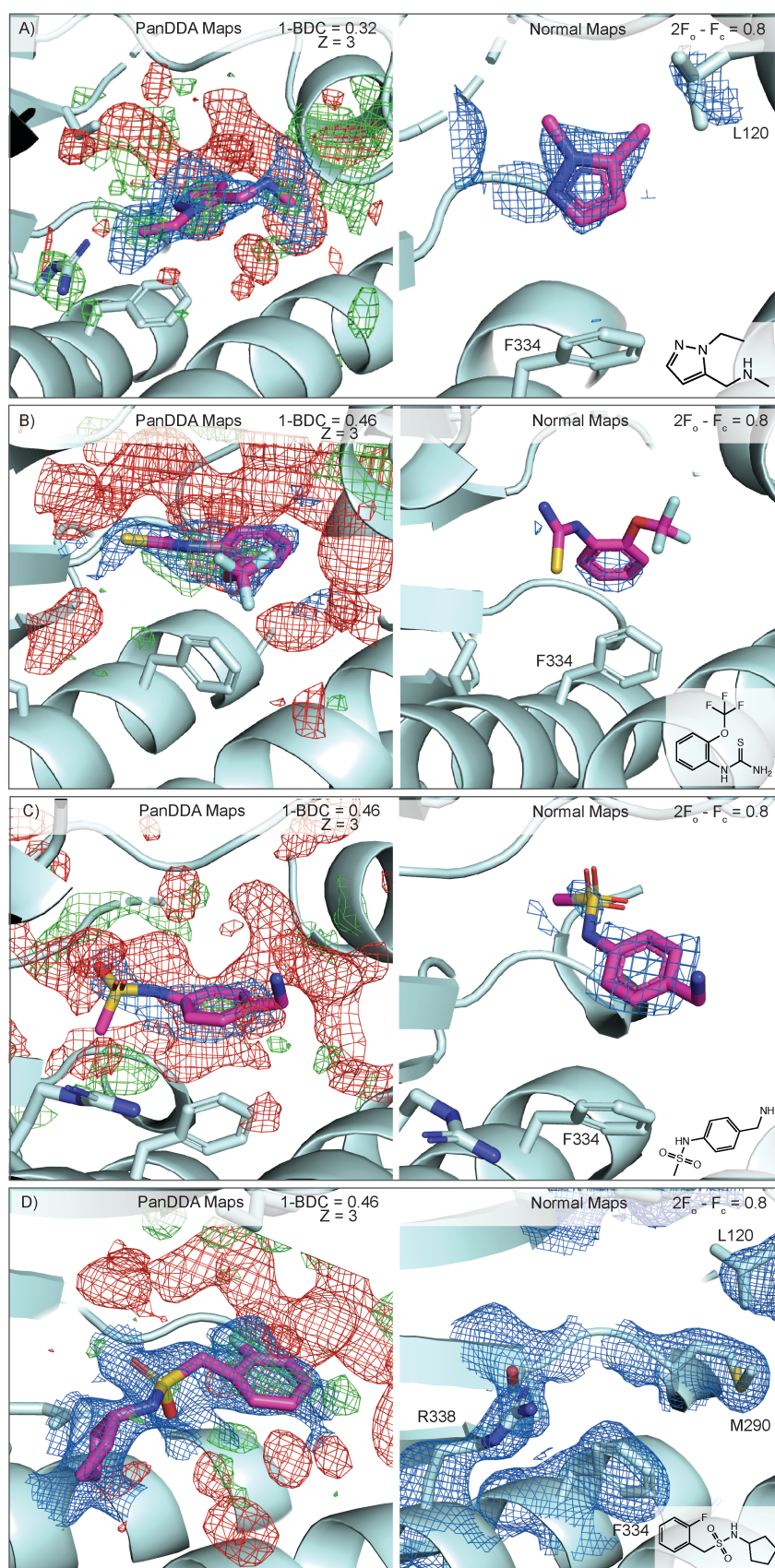


Figure 4.13: Comparison of PanDDa event and Z-maps, and  $2mF_o - DF_c$  maps for loop-remodelling hits. Fragment density in normal  $2mF_o - DF_c$  maps is weak A) CCT370982; B) CCT370979; C) CCT370972; D) CCT240772. Figures made in Pymol<sup>173</sup>.

PanDDA statistics supported the presence and inclusion of these four fragments in further analysis (Appendix 8.2.8.1). PanDDA therefore identifies multiple fragment hits that would have otherwise been overlooked.

Of the ten fragments inducing re-organisation of the pocket-forming loop, four - CCT239544, CCT239686, CCT370974 and CCT242635 - showed clear density in  $2mF_o - DF_c$  maps (Figure 4.14A-D). These fragments bind through a  $\pi$ - $\pi$  stack on the side chain of Phe334, except for CCT242635, which stacks on top of Arg338 in two distinct conformations (Figure 4.14D). In one conformation, the aminothiazole is stacked on top of the remodelled Arg338 side chain, with the primary amine substituent interacting with the side chain of Ser202. The phenyl group sits in a hydrophobic pocket formed by the protein backbone and the side chain of Val294, also stacking on top of the remodelled Arg338. In the second conformation, the placement of the phenyl group is conserved, but the aminothiazole rotates to point towards Glu360, with the amine forming a stabilising interaction with the side chain of this residue. The side chain of Phe334 rotates to adopt an edge-face  $\pi$ -stack with the thiazole. Breaking the hydrophobic interaction between Phe334 and Ile112 by fragment binding seems to be important for adoption of the novel  $\alpha$ -helical conformation of the pocket-forming loop.



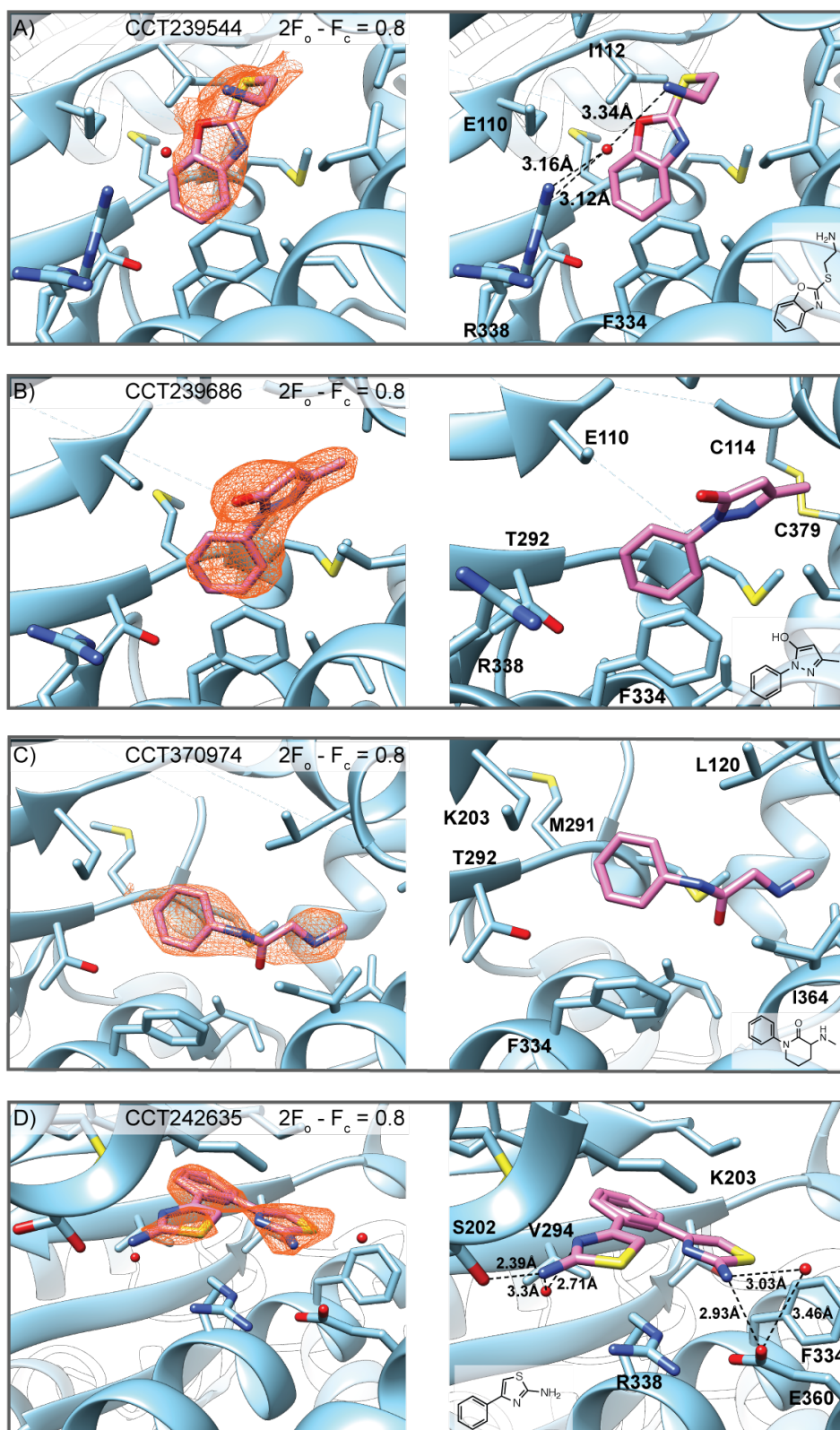


Figure 4.14: Four fragments binding to the novel pocket with reorganisation of the pocket-forming loop showed clear electron density in normal  $2mF_o - DF_c$  maps (contour at  $\sigma = 0.8$ ). These adopted a variety of binding modes, but all involved the side chain of Phe334. In addition, the salt bridge between R338, E110 and T292 was also broken in all, with R338 rotating to face the solvent. A) CCT239544; B) CCT239686; C) CCT370974; D) CCT242635; figure made in Chimera<sup>65</sup>

## 4.5 Conclusions

This chapter describes two fragment-screening approaches used to identify hit matter binding to the novel secondary site in IDH1-R132H. A total of 2595 fragments from the ICR's fragment library and the 3D Fragment Consortium library were screened using TSA, and a further 768 fragments were screened by crystallographic fragment screening at Diamond's XChem facility. In all, nineteen fragments were identified binding to the novel secondary site. The 14 fragments identified during crystallographic fragment screening were also investigated by TSA, but only two showed significant impact on IDH1-R132H stability. This likely due the higher fragment concentrations used for crystallographic fragment screening in comparison to TSA. Across the two fragment screens, clear electron density could be seen in normal  $2mF_o - DF_c$  maps for 11 fragments, which allowed full refinement of these fragment-bound structures. The remaining eight fragments only showed density in PanDDA maps, including the TSA hit CCT240772. PanDDA therefore identified hits that would have otherwise been overlooked.

Fragments were identified binding across the width of the pocket with various binding modes. Interestingly, ten fragments were identified binding to the novel pocket with extensive re-modelling of the pocket-forming loop to adopt a novel  $\alpha$ -helical conformation that has not previously been described in IDH1 structures.

From the TSA fragment screen, the binding sites of ten of the 15 hits remain unknown. Efforts to repeat soaks with fragment hits binding to IDH1-R132H with remodelling of the pocket forming loop indicates that only a small

proportion of crystals can allow this remodelling to occur, although the structural rationale for this limitation is unclear. Additional fragments binding with this remodelling may have been lost from both screens due to this limitation in the crystal system. In total, 19 fragments were identified binding to the novel secondary site, with a total hit rate of 0.56%, confirming that this novel secondary site is ligandable and the computational prediction was correct.

## **Chapter 5: Investigating the functional relevance of the novel pocket**

---

### **5.1 Introduction**

For a pocket to be considered druggable, binding of small molecules to the pocket needs to have an impact on protein function, and eventually on cellular phenotype. Chapter 4 described the fragment screening approaches used to identify chemical matter binding to the novel secondary site, confirming its ligandability. This chapter describes the investigation of the functional relevance of the novel secondary site using biochemical assays. Specifically, I analysed the effect of the fragment hits and selected mutants on IDH1 activity. To do this, I established a biochemical assay based on intrinsic NADPH fluorescence. Fragments hits were investigated for their ability to inhibit IDH1-R132H, and their selectivity over IDH1-WT. A series of analogues were designed and synthesised by Sandra Codony Gisbert and Dr Rosemary Huckvale based on fragment hits that inhibited IDH1-R132H activity, and were tested in the biochemical assays.

In addition, several IDH1-R132H variants with mutations in the novel pocket that mimicked recurrent fragment binding features were designed. Following cloning, expression and purification of these site-directed IDH1-R132H mutants, they were characterised using the biochemical assay to investigate the impact of secondary site mutations on IDH1-R132H activity.

## 5.2 IDH1 NADPH fluorescence assay

### 5.2.1 Establishing an IDH1-R132H inhibition assay

IDH1-R132H catalytic activity requires the co-factor NADPH, the substrate  $\alpha$ KG and the catalytic  $Mg^{2+}$  (Figure 5.1A, B). Biochemical assays for IDH1-R132H based on the intrinsic fluorescence of NADPH have been reported<sup>179</sup>. I established an NADPH-fluorescence assay to investigate IDH1-R132H inhibition by fragment hits. Although IDH1-R132H consumes NADPH as it converts  $\alpha$ KG to 2HG, leading to a decrease in fluorescence signal over time, I measured the change in fluorescence ( $\Delta$  fluorescence) over time by subtracting each fluorescent measurement from the control, the uninitiated reaction, to obtain reaction progression curves that increased with time (Figure 5.1C).

The hit fragments were identified binding to the novel secondary site in the presence of NADPH, and were therefore not expected to be competitive with respect to NADPH. The MOA of the fragment hits with respect to  $\alpha$ KG and  $Mg^{2+}$  is unknown. Fragments bind to IDH1-R132H when the protein adopts the inactive conformation. When substrate and  $Mg^{2+}$  are both bound IDH1-R132H undergoes a conformation change to adopt the catalytically active conformation. In this conformation, the novel secondary site was not predicted to be ligandable, and fragments were not expected to bind. To allow identification of both competitive and uncompetitive inhibitors, I aimed to maintain the  $\alpha$ KG concentration at its  $K_m$ . Therefore, the  $K_m$  for all components needed to be determined. The assay was established in 384-well plates with 10  $\mu$ L well volumes.

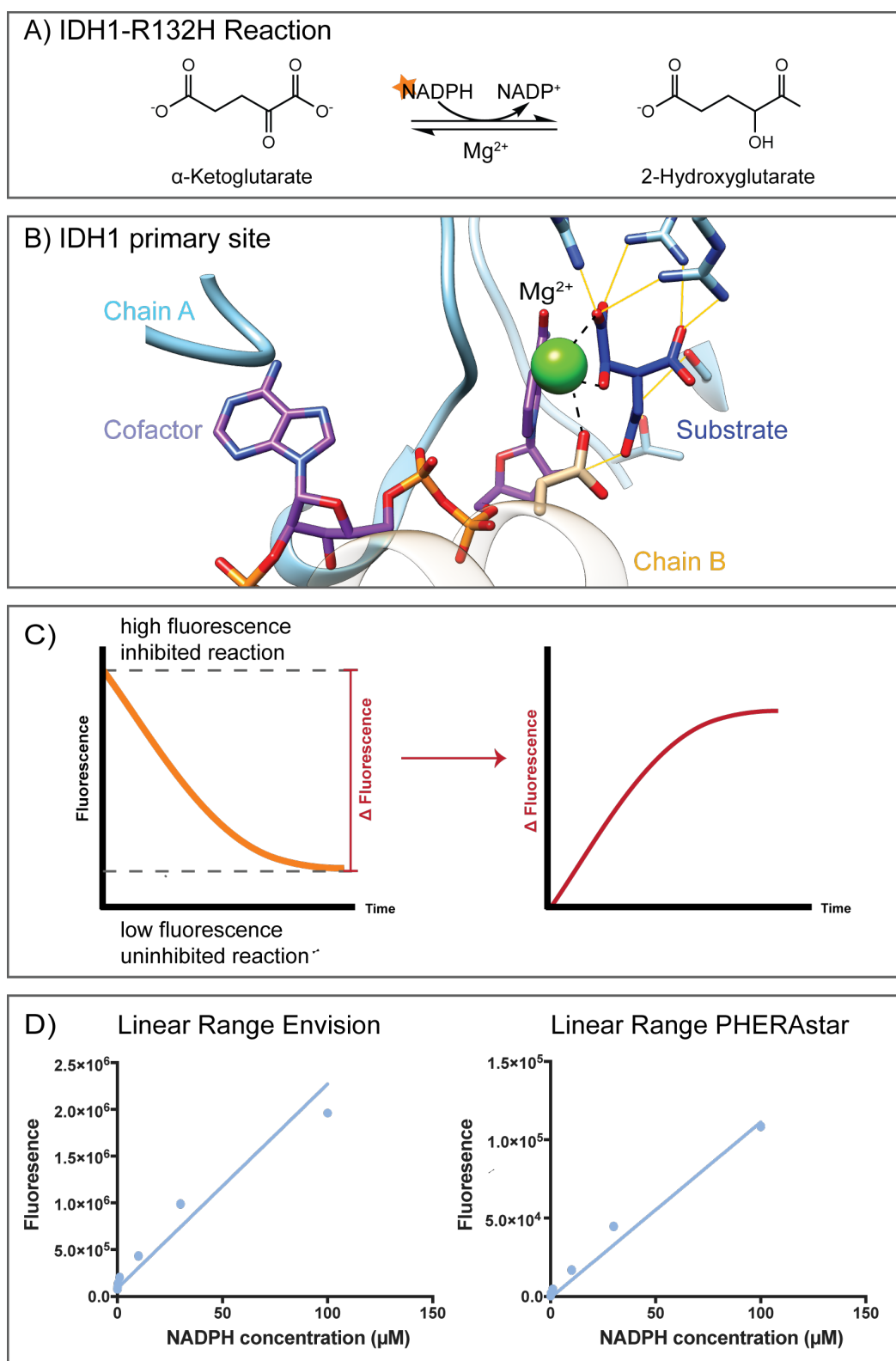


Figure 5.1: Overview of the IDH1-R132H biochemical assay. A) IDH1-R132H converts aKG to 2HG with the concomitant reduction of NADPH to NADP<sup>+</sup> in the presence of a catalytic Mg<sup>2+</sup>. B) Binding of cofactor, substrate and Mg<sup>2+</sup> is required for adoption of catalytically active conformation, which involves residues from both chains, shown in cyan and tan. PDB 1T0L C) NADPH has intrinsic fluorescence. As it is consumed by IDH1-R132H, the fluorescent signal decreases. Each measurement is subtracted from a control, and the change in fluorescent signal,  $\Delta$  fluorescence, is reported. D) Fluorescence signal against NADPH concentration showing the linear range. Figure made in Chimera<sup>65</sup> and plot made in GraphPad Prism.

### 5.2.1.1 Obtaining a NADPH signal

A NADPH calibration curve showed a linear relationship between NADPH concentration and fluorescence signal up to 100  $\mu\text{M}$  (Figure 5.1D). All experiments were therefore carried out with the concentration of NADPH at or below 100  $\mu\text{M}$ . For comparability, the initial activity assay was performed using the same buffer as for TSA experiments, with protein, co-factor substrate and metal concentrations of 20 nM IDH1-R132H, 25  $\mu\text{M}$  NADPH, 1 mM  $\alpha\text{KG}$  and 10 mM  $\text{Mg}^{2+}$ . However, no signal window could be obtained. After investigating the literature, I found that in most reported assays both Tween20 and Bovine Serum Albumin (BSA) were also added to the buffers. Addition of Tween20 alone was not sufficient for the reaction to progress, but addition of BSA was sufficient for the reaction to occur (Figure 5.2). When both were added, the rate was slightly slower than BSA alone, but gave smaller errors. In subsequent experiments,  $\text{Mg}^{2+}$ , Tween20 and BSA were added to the buffers.

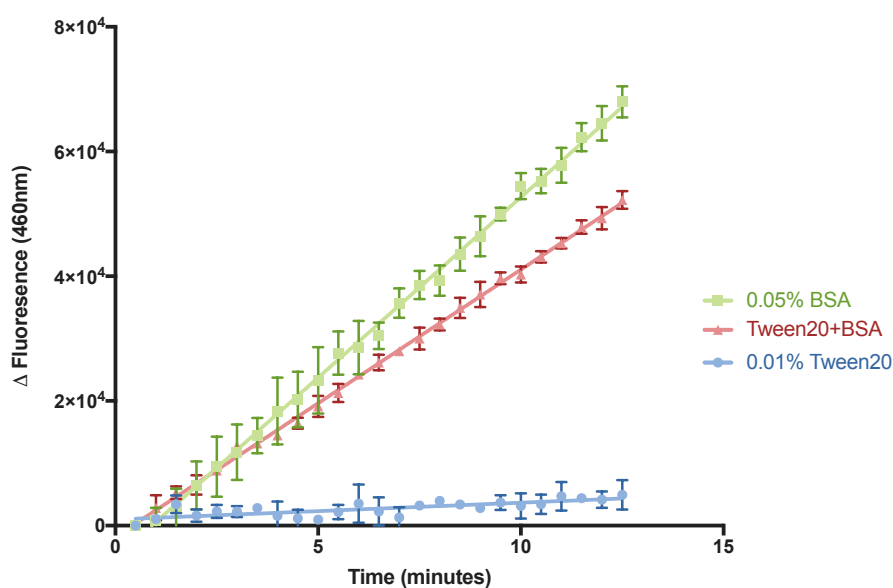


Figure 5.2: Change in NADPH fluorescence over time in the presence or absence of Tween20 and BSA. BSA is required for measurable NADPH fluorescence. Tween20 is not required, but reduces the errors when present with BSA. Data is from two technical repeats; error bars show the standard deviation. Plot made in GraphPad Prism.

### 5.2.1.2 Selection of optimal IDH1-R132H concentration

A range of IDH1-R132H concentrations between 20 nM and 100 nM were investigated in the presence of 100  $\mu$ M NADPH, 800  $\mu$ M  $\alpha$ KG and 10 mM  $\text{MgCl}_2$  (Figure 5.3). The  $\Delta$  fluorescence is calculated by subtracting each measurement from a control, which lacked IDH1-R132H. NADPH by itself undergoes some non-enzymatic oxidation, resulting in an initial drop in fluorescent signal even in the absence of IDH1-R132H. The presence of IDH1-R132H seems to partially protect NADPH from non-enzymatic oxidation, which leads to some initial  $\Delta$  fluorescence measurements being negative. Increasing concentrations of IDH1-R132H increased the rate of reaction. Concentrations above 60 nM reached a plateau within the 60 minutes reaction time. The lowest concentration tested, 20 nM, stayed linear across all time points and was therefore selected 20 nM IDH1-R132H for use in further assays.

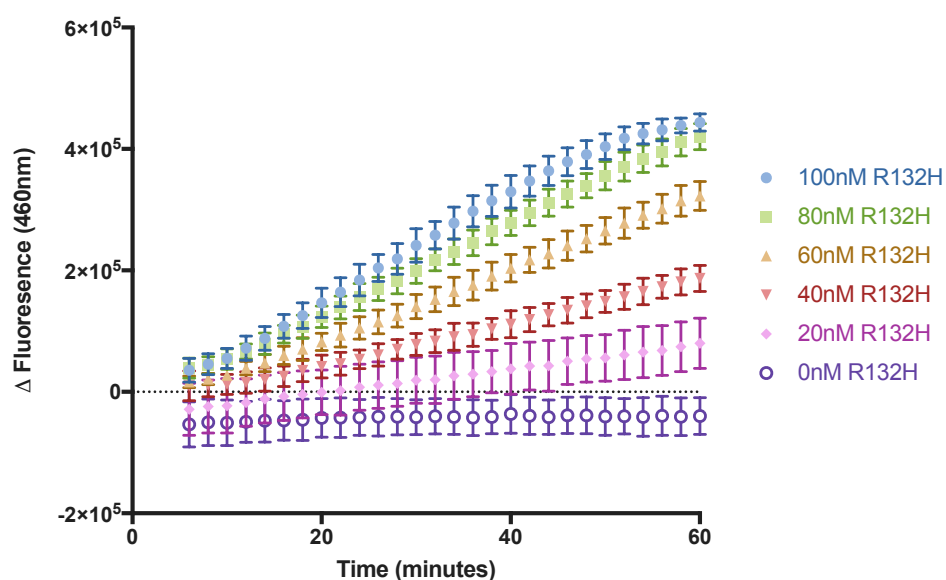


Figure 5.3: Change in fluorescence signal over time with increasing concentrations of IDH1-R132H. I selected 20 nM IDH1-R132H for use in inhibition assays. Data is from two technical repeats; error bars show the standard deviation. Plot made in GraphPad Prism



### 5.2.1.3 Kinetic characterisation of IDH1-R132H

The IDH1-R132H reaction requires co-factor, substrate and magnesium. I determined the kinetic parameters for each of these. To measure  $V_{max}$  and  $K_m$  values for NADPH, IDH1-R132H was incubated with varying concentrations of NADPH between 100 and 2.5  $\mu$ M, with the substrate  $\alpha$ KG in excess at 5 mM. Ideally, lower concentrations of NADPH would have also been tested, but at concentrations below 2.5  $\mu$ M, no reliable signal could be obtained. The  $\Delta$  fluorescence was determined over time and plotted in Graphpad Prism, using linear regression in order to calculate the initial rates of reaction at different NADPH concentrations (Figure 5.4A). Kinetic parameters were calculated by plotting the initial rate against NADPH concentration and fitting the curve with a non-linear regression (Equation 5.1, Figure 5.4B).

$$Y = V_{max} \times \frac{X}{K_m + X} \quad \text{Equation 5.1}$$

For kinetic characterisation of IDH1-R132H with respect to  $\alpha$ KG, a range of  $\alpha$ KG concentrations between 5 mM and 500  $\mu$ M were investigated, with the  $\Delta$  fluorescence measured, and the data processed in the same way as for the NADPH kinetic analysis (Figure 5.4C,D).

For kinetic characterisation of IDH1-R132H with respect to  $Mg^{2+}$ , a range of  $Mg^{2+}$  concentrations between 250 and 0.12 mM were investigated, with the  $\Delta$  fluorescence measured, and data processed in the same way as for the NADPH and  $\alpha$ KG analyses (Figure 5.4E,F).

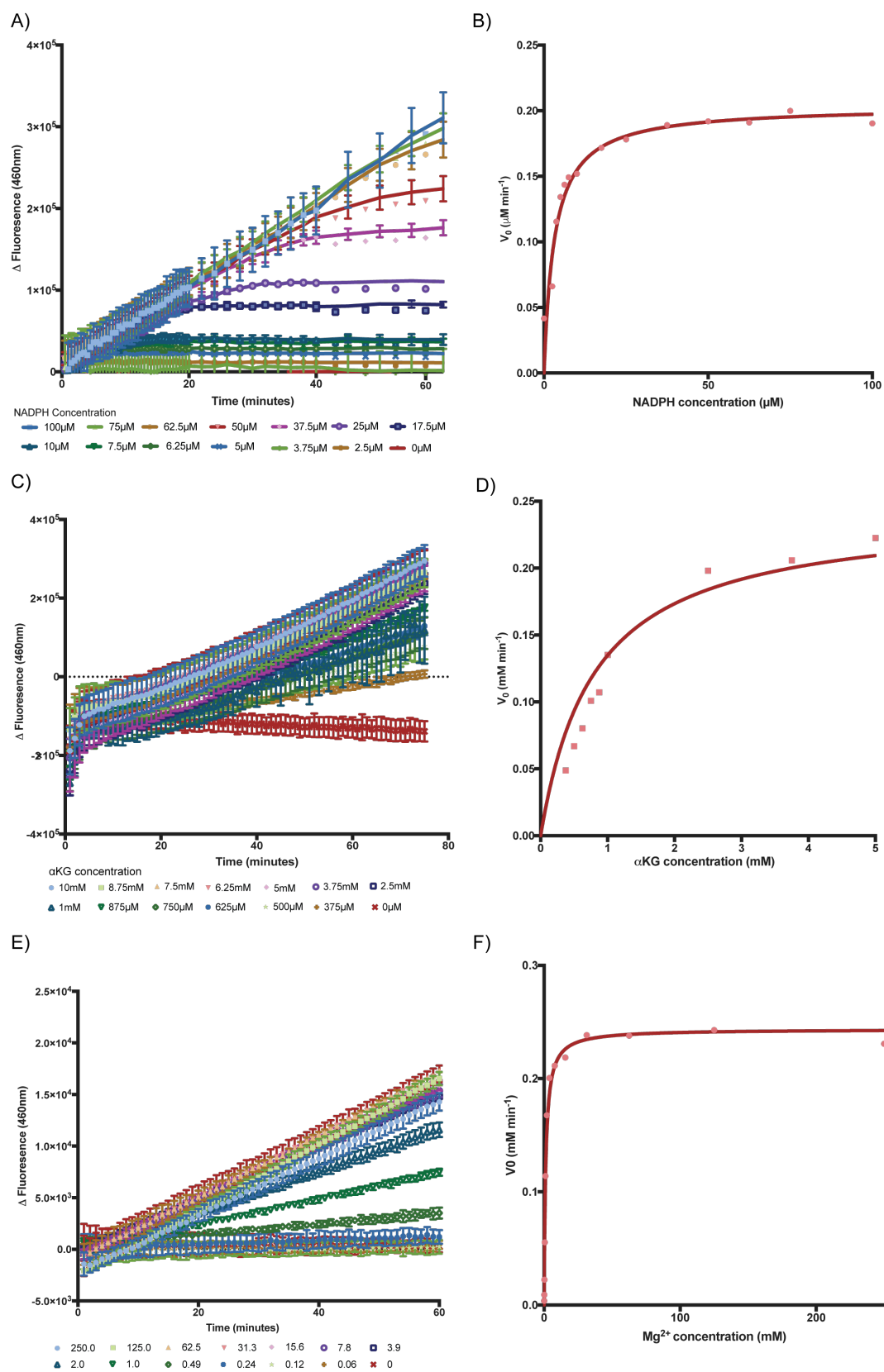


Figure 5.4: Initial rates and  $K_m$  curves for IDH1-R132H with NADPH (A, B),  $\alpha$ KG (C, D) and  $Mg^{2+}$  (E, F). Data shown is representative of three biological repeats; error bars show the standard deviation. Plot made in GraphPad Prism

The calculated  $K_m$  values for NADPH,  $\alpha$ KG and  $Mg^{2+}$  for IDH1-R132H were 3.17  $\mu$ M, 803  $\mu$ M and 1.16 mM respectively (Table 5.1). The  $K_m$  value for  $\alpha$ KG is in line with that reported in the literature, but the value for NADPH is approximately tenfold higher than reported, and the  $K_m$  for  $Mg^{2+}$  is approximately tenfold lower<sup>74</sup>. However, it should be noted that they were measured under different conditions.

	$V_{max}$ ( $\mu$ M min <sup>-1</sup> )	$K_m$ ( $\mu$ M)	$K_{cat}$ (min <sup>-1</sup> )
<b>NADPH</b>	0.204 $\pm$ 0.0087	3.17 $\pm$ 0.66	10.2
<b><math>\alpha</math>KG</b>	0.244 $\pm$ 0.012	803 $\pm$ 152	12.2
<b><math>Mg^{2+}</math></b>	0.253 $\pm$ 0.0053	1160 $\pm$ 130	12.7

Table 5.1 Kinetic parameters for IDH1-R132H. Values are an average from three biological repeats, with the standard deviation. Calculated by GraphPad Prism

#### **5.2.1.4 Investigating the impact of DMSO on IDH1-R132H activity**

As fragments are stored at fixed concentrations in 100% DMSO, I investigated the DMSO tolerance of the IDH1-R132H to find the maximum fragment concentration that could initially be tested from a 100 mM stock solution. I examined the initial rates of reaction for several different DMSO concentrations between 0 and 5% v/v (Figure 5.5). IDH1-R132H activity increased with increasing DMSO concentrations up to the maximum concentration tested, indicating that IDH1-R132H was tolerant to DMSO up to at least 5% v/v DMSO. I selected 3% v/v as the maximum DMSO concentration as it allowed investigation of fragments up to a concentration 3 mM.

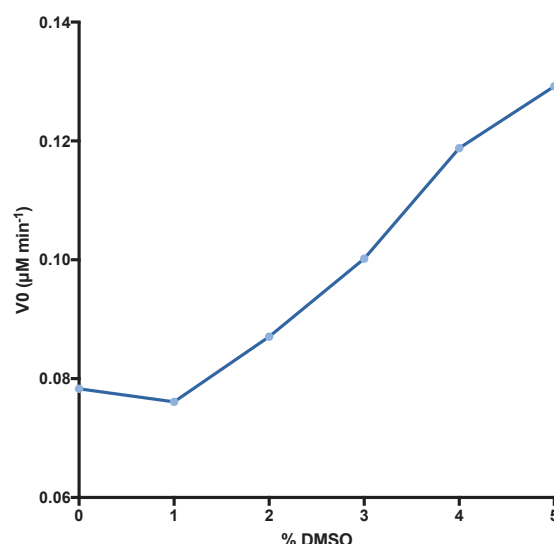


Figure 5.5: Plot of initial rate of reaction of IDH1-R132H with increasing DMSO concentrations. IDH1-R132H activity increases with increasing DMSO up to 5%. Plot made in GraphPad Prism

#### 5.2.1.5 Selecting NADPH concentration

The fluorescent signal is dependent on NADPH and is linear up to 100 μM. I chose to use 75 μM NADPH as it is within the linear detection range, and is 25-fold the measured  $K_m$  and is therefore in significant excess. To check whether a sufficient signal window (Equation 5.2) could be obtained whilst limiting the percentage conversion, I calculated both values at various time points (Figure 5.6).

$$SW = \frac{(\mu_{max} - 3\sigma_{max}) - (\mu_{min} - 3\sigma_{min})}{\sigma_{max}} \quad \text{Equation 5.2}$$

Where  $\mu$  is the mean and  $\sigma$  is the standard deviation

While the 60 minute incubation gave the largest signal window, the percentage conversion is greater than 10%, which is greater than would be desired<sup>180</sup>. A 45 minute incubation with 75 μM NADPH gave an acceptable signal window of 15, with 6% NADPH conversion. This leaves approximately 70 μM NADPH in the

system, a 24-fold excess over the  $K_m$ . Given that one molecule NADPH is required to convert one molecule of  $\alpha$ KG to 2HG, this percentage conversion maintains the  $\alpha$ KG concentration at approximately  $K_m$ .

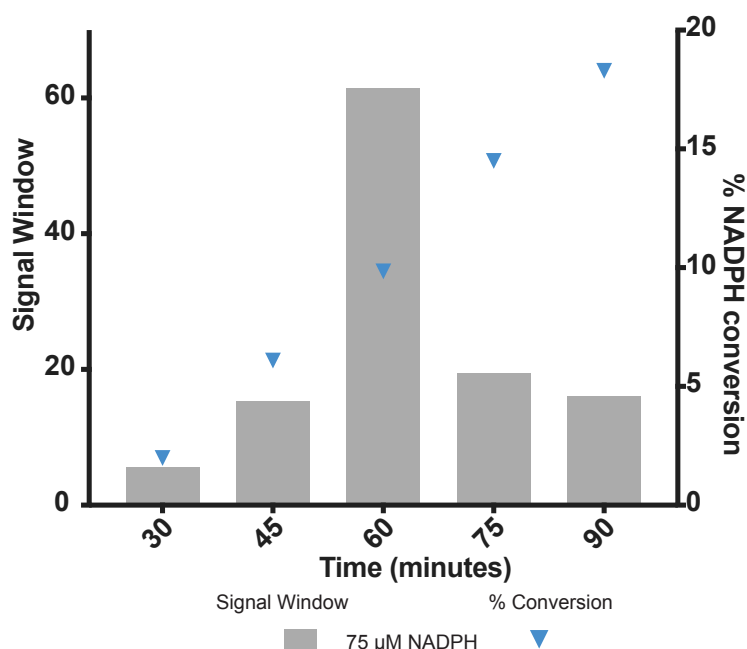


Figure 5.6: Comparison of signal window and percentage conversion for 20 nM IDH1-R132H and 75  $\mu$ M NADPH at different time points. Grey bars show the signal window as calculated by equation 5.2, while the blue points show the percentage NADPH conversion. Plot made in GraphPad Prism

Final assay conditions were 20 nM IDH1-R132H with 75  $\mu$ M NADPH, 800  $\mu$ M  $\alpha$ KG and 10 mM  $Mg^{2+}$ , with 3% DMSO. Fragments were tested in a 10-point, twofold dilution curve from 3 mM. Plates were incubated for 45 minutes before fluorescence was measured.

#### 5.2.1.6 Inhibition of IDH1-R132H by known inhibitors

The two tool compounds, AGI-5198 and GSK-864, were tested in the fluorescence assay (Figure 5.7). The measured  $IC_{50}$  value for AGI-5198 was 7 nM, which is approximately tenfold more potent than reported<sup>169</sup>. The measured  $IC_{50}$  for GSK 854 was 38 nM, which is in line with the reported

value<sup>181</sup>. The IC<sub>50</sub> values for both inhibitors were close to the tight binding limit of the assay, 10 nM, but were consistent across assay repeats, and were therefore used as positive controls for this assay.

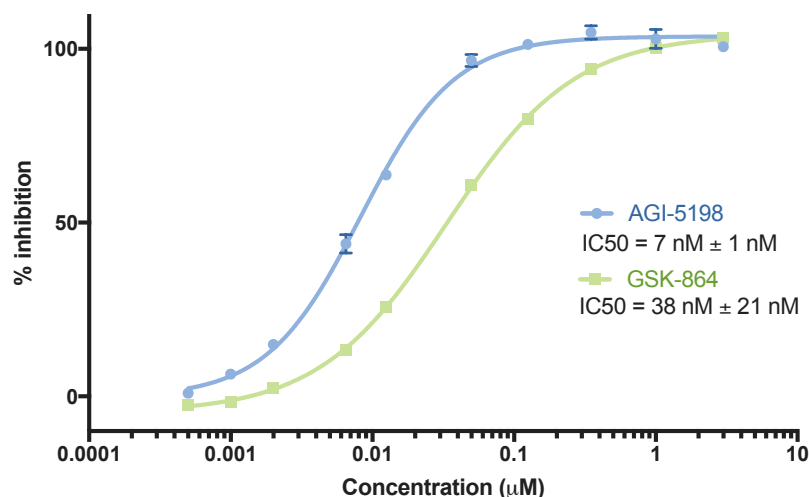


Figure 5.7: IC<sub>50</sub> curves for two tool compounds AGI-5198 and GSK 864. Data is representative of three biological repeats, and plotted in Graphpad Prism, error bars show the standard deviation.

#### 5.2.1.7 IDH1-R132H fluorescence interference assay

Structural features common to fluorophores can also be found in fragments, resulting in interference in fluorescence-based biochemical assays. Compounds can interfere with the fluorescence signal through either auto-fluorescence or through quenching<sup>182</sup>. NADPH fluorescence is relatively high energy, with maximum absorbance and emission at 340 nm and 460 nm respectively. Compound libraries tend to have greater interference in this blue-green spectral region. I therefore investigated the potential auto-fluorescence and quenching of the fragments through a fluorescence interference assay.

To investigate fragment fluorescence interference, the hit fragments were dispensed into plates in a 10-point twofold dilution curve starting at 3 mM. The low control was IDH1-R132H with 75  $\mu$ M NADPH to mimic the inhibited reaction, while the high control was IDH1-R123H with 40  $\mu$ M NADPH to mimic the uninhibited reaction.

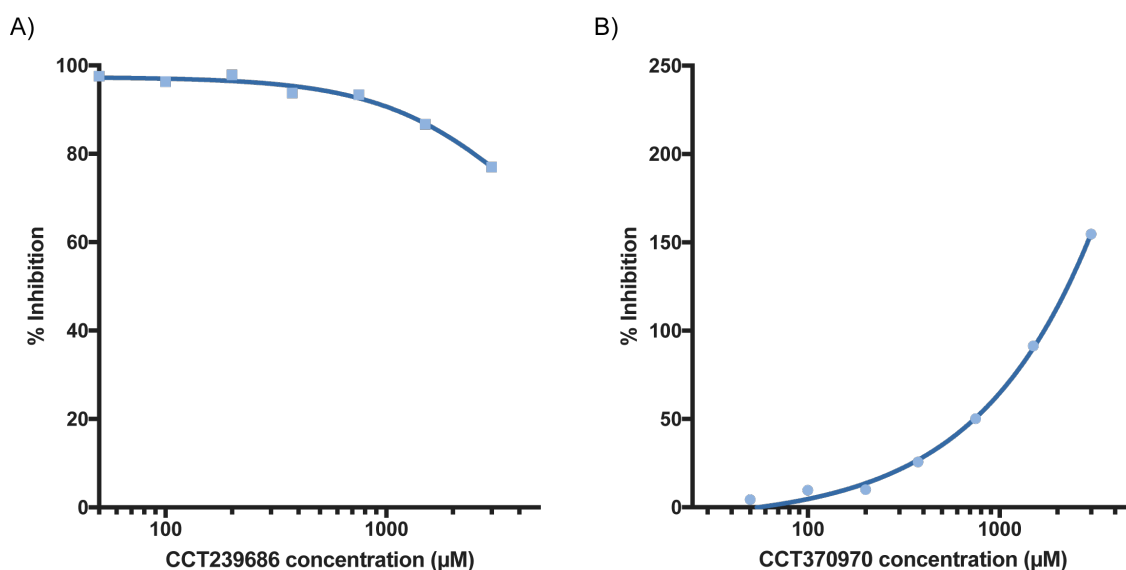


Figure 5.8: Representative curves from the fluorescence interference assay. A) CCT239686 shows less than 90% inhibition at the top two concentration points, indicating that this fragment quenches fluorescence at these concentrations. B) CCT370970 shows greater than 10% inhibition across multiple concentration points indicating that it is an auto-fluorescent fragment and strongly interferes with the biochemical assay. Data representative of two biological repeats. Plot made in GraphPad Prism

To investigate quenching, fragments were incubated with IDH1-R132H and 75  $\mu$ M NADPH. Fragments were considered to be quenching if the normalised fluorescence measurement was less than 90% of the minimum, such that it appears to show less than 90% inhibition activity in  $IC_{50}$  curves (Figure 5.8A). To investigate auto-fluorescence, fragments were incubated with IDH1-R132H and 40  $\mu$ M NADPH. Fragments were considered to be auto-fluorescent if the normalised fluorescence measurement was more than 110% of the minimum, such that it appears to show greater than 10% inhibition activity in  $IC_{50}$  curves (Figure 5.8B)

### 5.2.2 Establishing an IDH1-WT biochemical assay to investigate compound selectivity

The novel secondary site was predicted to be ligandable in the in house structures of both IDH1-R132H and IDH1-WT (Chapter 3). I therefore established IDH1-WT biochemical assays to investigate potential selectivity between the two variants when targeting the novel secondary site with inhibitors. IDH1-WT produces fluorescent NADPH as it converts isocitrate to  $\alpha$ KG (Figure 5.9), and thus the reaction progression can be followed by measuring the fluorescent signal corresponding to the production of NADPH over time. To directly compare compound activity against IDH1-WT with that against IDH1-R132H, the same assay conditions were required. The concentration of  $\text{Mg}^{2+}$  was maintained at 10 mM, the co-factor  $\text{NADP}^+$  was maintained at 25-fold  $K_m$ , and isocitrate concentration at its  $K_m$ .

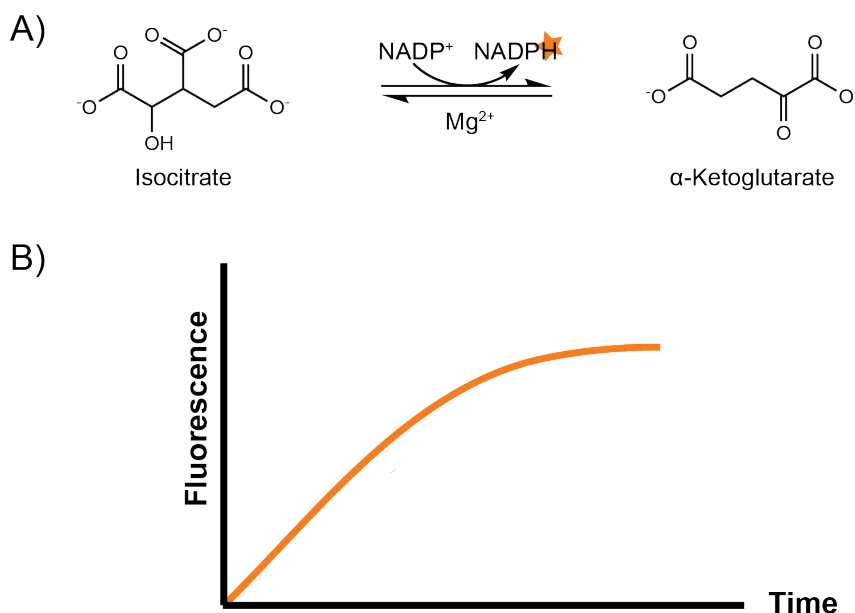


Figure 5.9: Overview of IDH1-WT biochemical assay. A) IDH1-WT produces fluorescent NADPH as it converts isocitrate to  $\alpha$ KG in the presence of  $\text{Mg}^{2+}$ . B) Intrinsic NADPH fluorescence allows the reaction progression to be measured by the increase in fluorescence over time.



### 5.2.2.1 Kinetic characterisation of IDH1-WT

Titration of different concentrations of IDH1-WT revealed it to be more active than IDH1-R132H, as a lower enzyme concentration was able to turnover the co-factor at a faster rate (Figure 5.10). Therefore, 0.5 nM IDH1-WT was used to enable subsequent  $K_m$  and kinetic characterisation. The measured  $K_m$  values were 17.1  $\mu\text{M}$  for  $\text{NADP}^+$  and 5.7  $\mu\text{M}$  for isocitrate under these conditions (Figure 5.11, Table 5.2). These values are approximately threefold and tenfold lower than the reported  $K_m$  values for  $\text{NADP}^+$  and isocitrate, 49  $\mu\text{M}$  and 65  $\mu\text{M}$  respectively<sup>74</sup>. However, these values were measured under different conditions. The  $k_{\text{cat}}$  of IDH1-WT was calculated to be  $\sim 550 \text{ min}^{-1}$ , which is significantly greater than IDH1-R132H. This increase is likely driven by the increased affinity for its natural substrate isocitrate, 5.7  $\mu\text{M}$ , in comparison to the lower affinity of IDH1-R132H for its substrate  $\alpha\text{KG}$ , with  $K_m$  803  $\mu\text{M}$ .

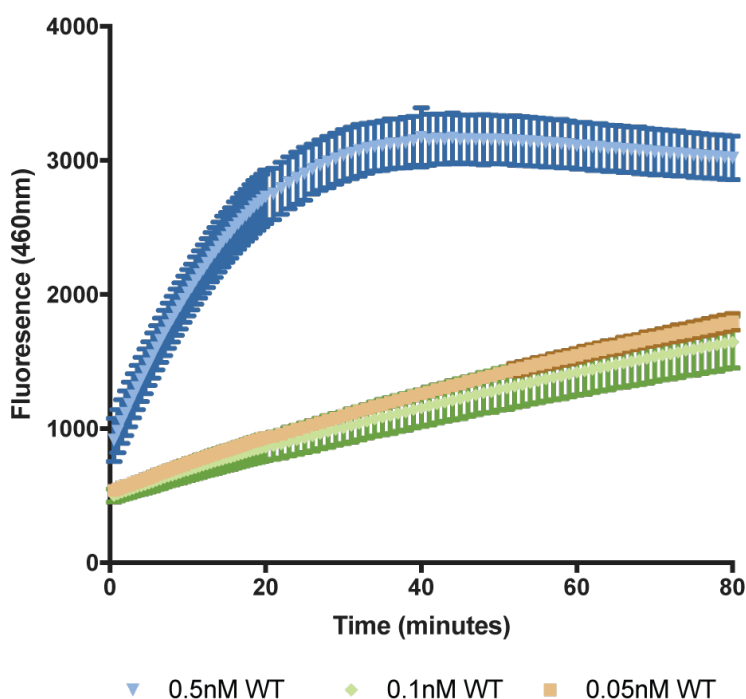


Figure 5.10: Plot of fluorescence against time, monitoring production of NADPH over time at three IDH1-WT concentrations. Data shown is from two technical repeats, error bars show the standard deviation. Plot made in GraphPad Prism

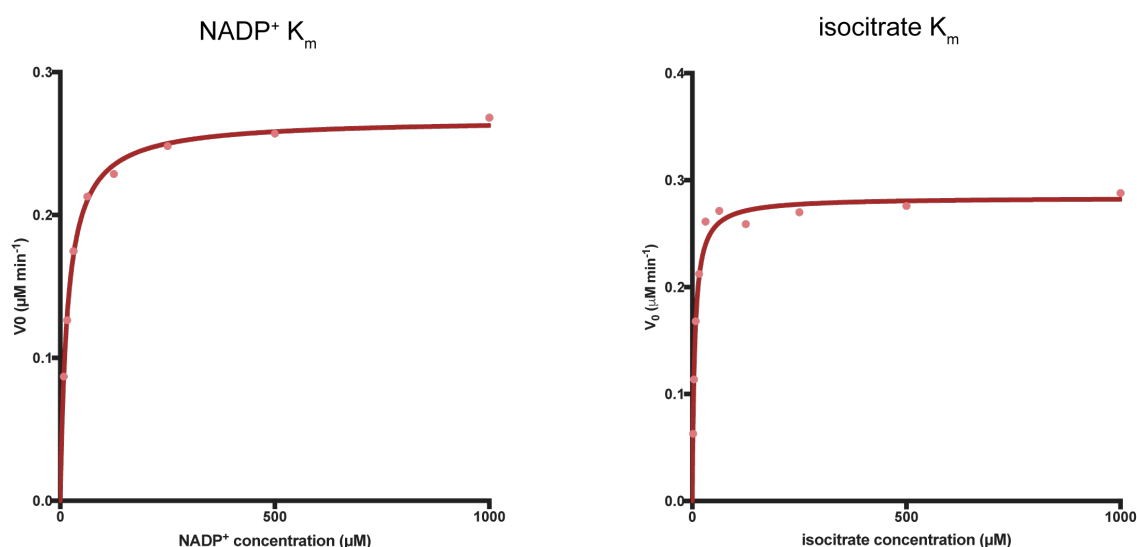


Figure 5.11: Michaelis-Menten curves for IDH1-WT cofactor NADP<sup>+</sup> and substrate isocitrate. Data shown is representative of 3 biological repeats. Plots made in GraphPad Prism

	$V_{\max}$ ( $\mu\text{M min}^{-1}$ )	$K_m$ ( $\mu\text{M}$ )	$K_{\text{cat}}$ ( $\text{min}^{-1}$ )
<b>NADP<sup>+</sup></b>	$0.267 \pm 0.0087$	$17.1 \pm 2.90$	534
<b>isocitrate</b>	$0.284 \pm 0.0064$	$5.70 \pm 0.69$	567

Table 5.2: Kinetic parameters for IDH1-WT. Data is an average of three biological repeats with the standard deviation.

### 5.2.2.2 Selection of optimal IDH1-WT concentration

As IDH1-WT is significantly more active than IDH1-R132H (Figure 5.10), I titrated IDH1-WT and calculated the signal window to identify an enzyme concentration suitable for inhibition assays. The  $\text{Mg}^{2+}$  concentration was maintained at 10 mM, the NADP<sup>+</sup> concentration at 25-fold  $K_m$  and the isocitrate concentration at  $K_m$  to mimic the IDH1-R132H assay. The maximum concentration investigated showed very rapid turnover of NADPH, reaching a plateau after 20 minutes, and was therefore unsuitable for inhibition assays.

Both 0.1 nM and 0.05 nM IDH1-WT showed similar rates of reaction, with linear ranges extending past one hour. At 45 minutes, the same incubation time as the IDH1-R132H reaction, 0.05 nM IDH1-WT gave a signal window (Equation 5.1) greater than 8 with less than 1% NADP<sup>+</sup> conversion, also in line with the IDH1-R132H reaction (Figure 5.12). I therefore selected 0.05 nM IDH1-WT to investigate fragment inhibition. Final reaction conditions were 0.05 nM IDH1-WT with 500  $\mu$ M NADP<sup>+</sup> and 5  $\mu$ M isocitrate. All other variables were kept the same as the IDH1-R132H fluorescence assay.

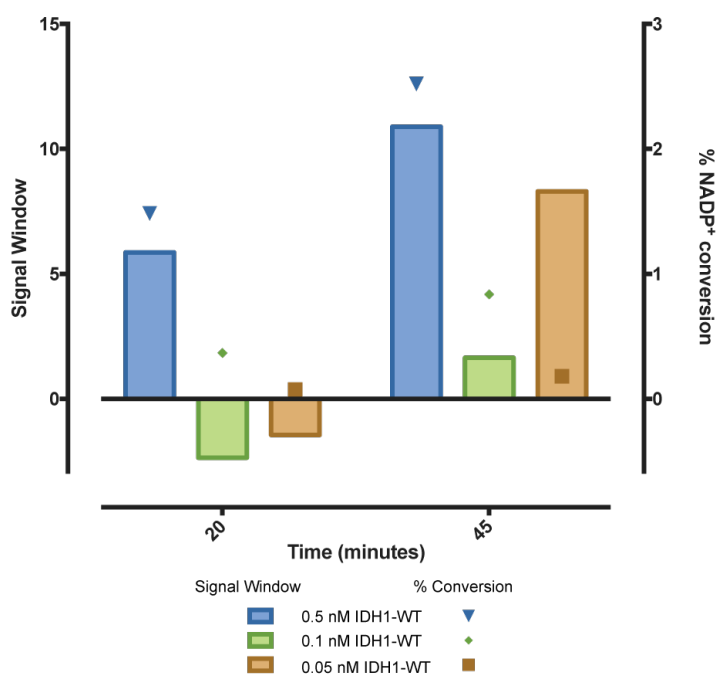


Figure 5.12: Comparison of signal windows and percentage NADP<sup>+</sup> conversion for three IDH1-WT concentrations after 20 minutes or 45 minutes incubation. Bars show the signal window, NADP<sup>+</sup> conversion is shown as points. Plot made in GraphPad Prism

### 5.2.2.3 IDH1-WT fluorescence interference assay

IDH1-WT produces fluorescent NADPH as the reaction progresses, and the percentage conversion is limited to maintain linearity. The raw fluorescence counts were lower in the IDH1-WT assay than the IDH1-R132H at both the start and end points (Figure 5.13). The IDH1-WT assay is therefore more sensitive to auto-fluorescent interferers, but less susceptible to quenchers. To investigate fragment fluorescence interference, the hit fragments were dispensed into plates in a 10-point twofold dilution curve starting at 3 mM. The low control was IDH1-WT without NADPH to mimic the inhibited reaction, while the high control was IDH1-WT with 5  $\mu$ M NADPH to mimic the uninhibited reaction.

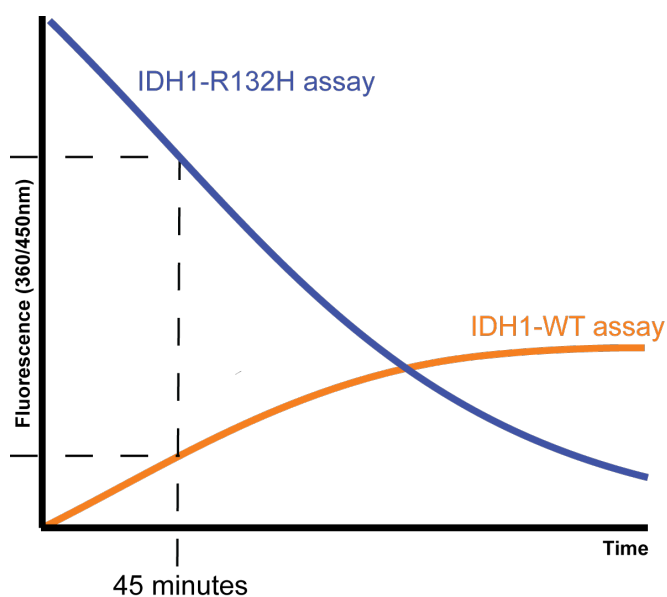


Figure 5.13: Comparison of raw fluorescent signals from IDH1-WT and IDH1-R132H assays. Due to the lower fluorescent signal in the IDH1-WT (orange) in comparison to the IDH1-R132H (blue) assay, the IDH1-WT biochemical assay is more sensitive to auto-fluorescent interferers, but less sensitive to quenchers.

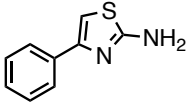
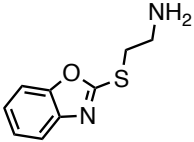
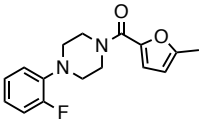
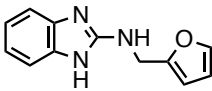
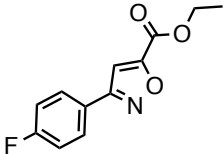
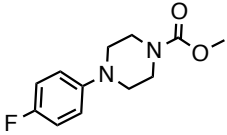
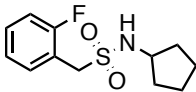
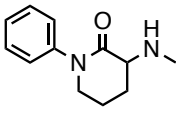
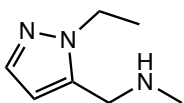
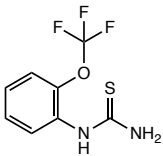
To investigate auto-fluorescent interferers, IDH1-WT without NADPH was then added to wells containing fragment, before the plate was incubated at room

temperature for 30 minutes. Buffer was then added to each well before the fluorescence measurements taken. Compounds that showed more than 10% increase or decrease in fluorescent signal in comparison to the control were considered to be fluorescent interferers. Due to compound availability at the time, CCT239544, CCT373604, CCT240772 and CCT37295 were not tested. Of the 15 fragments tested, 13 were found to interfere with the fluorescent signal at the maximum concentration investigated, with many interfering at multiple lower concentrations. CCT371095 and CCT370980 did not fluorescently interfere.

## **5.3 Inhibition of IDH1-R132H**

### **5.3.1 Inhibition of IDH1-R132H by hit fragments**

All 19 hit fragments identified by TSA and crystallographic fragment screening were investigated for their ability to inhibit IDH1-R132H activity (Table 5.3). Of these, ten fragments were found to inhibit IDH1-R132H activity with a range of potencies between 20% inhibition at 3mM and an  $IC_{50}$  of 84  $\mu$ M. In addition, two fragments, CCT370970 and CCT154567, were found to interfere and so the inhibitory effect could not be determined. The remaining seven showed less than 20% inhibition at 3 mM and were considered inactive. Fragment hits from all of the identified binding modes and from both fragment screening techniques were able to inhibit IDH1-R132H, including low occupancy PanDDA hits such as CCT240772 and CCT370982 that showed little to no density in normal  $2mF_o - DF_c$  maps. There was no clear correlation between the occupancy and activity. PanDDA therefore allowed identification of low occupancy fragments that would have otherwise been overlooked.

Compound	Structure	Biochemical IC50	Binding Mode	Screen
CCT242635		84.42 ± 12.17 μM	Loop remodelling	TSA
CCT239544		131.6 ± 13.4 μM	Loop remodelling	TSA
CCT242817		256.3 ± 30.5 μM	W205 Stack	TSA
CCT370971		1031.7 ± 49 μM	E361 stack	XChem
CCT373604		50% at 3mM	W205 Stack	XChem
CCT371098		37.8% at 3mM	W205 Stack	XChem
CCT240772		28% at 3mM	Loop remodelling	XChem
CCT370974		25.6% at 3mM	Loop remodelling	XChem
CCT370982		25.6% at 3mM	Loop remodelling	XChem
CCT370979		24% at 3mM	Loop remodelling	XChem

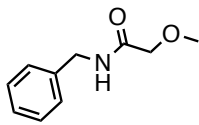
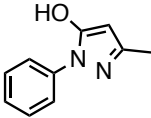
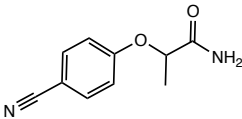
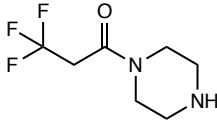
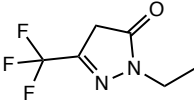
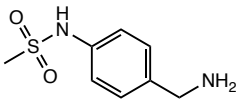
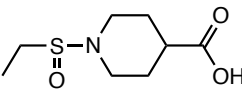
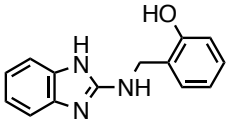
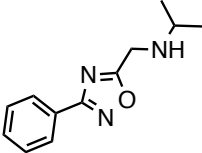
Compound	Structure	Biochemical IC <sub>50</sub>	Binding Mode	Screen
CCT372954		< 20% at 3mM	W205 Stack	XChem
CCT239686		< 20% at 3mM	Loop remodelling	XChem
CCT370980		<20% at 3mM	Loop remodelling	XChem
CCT370978		<20% at 3mM	Loop remodelling	XChem
CCT370973		<20% at 3mM	Other	XChem
CCT370977		<20% at 3mM	Loop remodelling	XChem
CCT371095		<20% at 3mM	Other	XChem
CCT370970		Fluorescent interferer	E361 stack	XChem
CCT154567		Fluorescent interferer	W205 Stack	XChem

Table 5.3: IC<sub>50</sub> values or maximum inhibition for hit fragments tested in the IDH1-R132H biochemical assay. Values reported are an average from three biological repeats with the standard deviation.

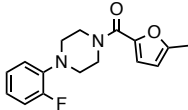
The ability of pocket-binding fragments to inhibit IDH1-R132H activity shows that the pocket has functional relevance, and marks a step towards confirming druggability in the novel secondary site.

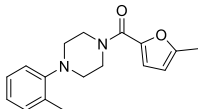
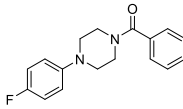
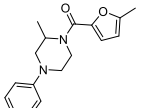
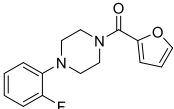
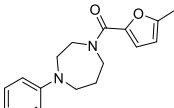
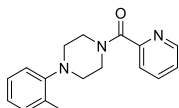
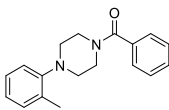
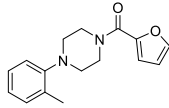
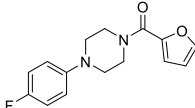
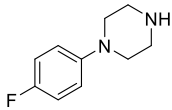
### **5.3.2 Investigation of fragment analogues against IDH1-R132H**

A series of 30 analogues were purchased and synthesised based on fragment hits CCT242635, CCT239554 and CCT242817 by Sandra Codony Gisbert and Dr Rosemary Huckvale. The ability of these fragments to inhibit IDH1-R132H was investigated, with the assay data collected together with Sandra Codony Gisbert. Of the 30 analogues, only one showed fluorescence interference, one was inactive, and 28 analogues inhibited IDH1-R132H activity. Full IC<sub>50</sub> curves could be measured for 24 compounds.

CCT242817 binds to the novel pocket through an edge-face  $\pi$ -stack on Trp205. In the crystal structure of IDH1-R132H bound to CCT242817, the electron density for the fluorophenyl-piperazine group is strong, but is weak for the terminal methyl-furanyl-ethanone group, preventing the modelling of this part of the fragment (Chapter 4.4.3). Despite this, it is the most potent fragment identified binding through an edge-face  $\pi$ -stack on Trp205. Therefore, 11 analogues of CCT242817 were tested (Table 5.4), of which six showed comparable potency and five showed reduced potency. Of the five weaker analogues, two were analogues of the fluorphenyl-piperazine moiety seen in the structure, CCT374321 and CCT374322. The fluorine group in CCT242817 is in the ortho position, which is the same as CCT374321, whereas in CCT374322 the fluorine is moved to the para position.



Parent Fragment Hit				
Compound	Structure	Biochemical IC50 (μM)		Change in potency between variants
		IDH1-R132H	IDH1-WT	
CCT242817		256.3 ± 30.5	-	non-comparable

Analogue				
Compound	Structure	Biochemical IC50 (μM)		Change in potency between variants
		IDH1-R132H	IDH1-WT	
CCT374036		152 ± 84.3	20% at 1.5 mM	> 10x decrease
CCT373807		241.9 ± 71.2	22% at 3 mM	> 10x decrease
CCT374320		260.4 ± 125.1	31% inhibition at 3mM	> 10x decrease
CCT373808		386 ± 61.9	20 % inhibition at 1.5 mM	~ 10x decrease
CCT374319		445.4 ± 162.3	2524 ± 673	5x decrease
CCT374035		622.3 ± 335.5	interferer	non-comparable
CCT299048		961 ± 92.3	20% at 3 mM	decrease
CCT304244		985.9 ± 1.8	2668 ± 470	2.5x decrease
CCT299915		1665 ± 110.7	28% at 3 mM	decrease
CCT374322		2912.8 ± 70.6	interferer	non-comparable

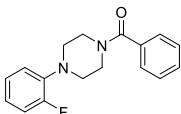
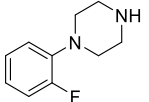
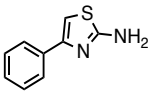
Compound	Structure	Biochemical IC <sub>50</sub> (μM)		Change in potency between variants
		IDH1-R132H	IDH1-WT	
CCT303854		22% at 375 μM	28% at 3 mM	decrease
CCT374321		25% at 3mM	inactive at 3 mM	decrease

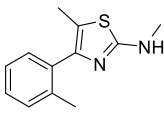
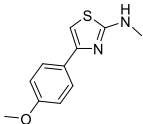
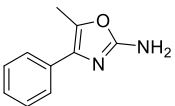
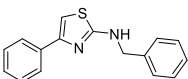
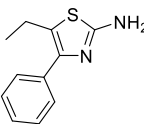
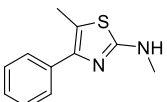
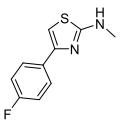
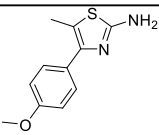
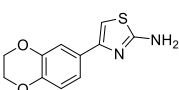
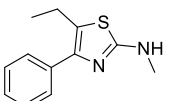
Table 5.4: Summary of biochemical assay data for analogues based on fragment hit CCT242817. Values reported are an average of two biological repeats with the standard deviation.

Both of these analogues are active, with IC<sub>50</sub> values of 2.8 mM and 25% inhibition at 3 mM for the para and ortho substitution respectively. However, both are significantly weaker than the parent fragment CCT242817, which has an IC<sub>50</sub> value of 256 μM.

With IC<sub>50</sub> of 84 μM, the most potent fragment hit was CCT242635, which binds with re-modelling of the pocket-forming loop. Of the 18 analogues designed based on CCT242635, 14 showed a decrease in potency, three maintained potency and one was more potent than the parent compound (Table 5.5). CCT242635 is a 2-aminothiazole, which is reported to be a promiscuous scaffold<sup>183</sup> and could lead to off-target interactions. In order to address this, an oxazole analogue of CCT242635 was synthesised. This new compound, CCCT374506, maintains potency, with an IC<sub>50</sub> of 152 μM, indicating that it is possible to move away from this potentially problematic thiazole scaffold.

The most potent compound tested across both fragment hits and analogues was CCT374509, with an IC<sub>50</sub> of 12.5 μM, which is approximately seven-fold more potent than the parent compound, CCT242635.

Parent Fragment Hit				
Compound	Structure	Biochemical IC50 (μM)		Change in potency between variants
		IDH1-R132H	IDH1-WT	
CCT242635		84.42 ± 12.17	interferer	non-comparable

Analogue				
Compound	Structure	Biochemical IC50 (μM)		Change in potency between variants
		IDH1-R132H	IDH1-WT	
CCT374509		12.5 ± 2	210 ± 69	13x decrease
CCT374037		114 ± 36.5	interferer	non-comparable
CCT374506		152 ± 65.4	interferer	non-comparable
CCT374554		250.1 ± 143.1	interferer	non-comparable
CCT373838		338 ± 21.7	interferer	non-comparable
CCT374447		385.6 ± 161.5	interferer	non-comparable
CCT374038		415 ± 267.2	interferer	non-comparable
CCT373840		711 ± 228.2	interferer	non-comparable
CCT017851		724.3 ± 240.9	interferer	non-comparable
CCT374449		800.7 ± 337.8	328 ± 101	2.5x increase

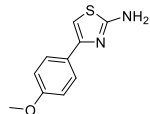
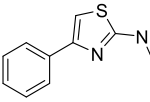
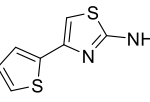
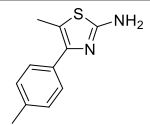
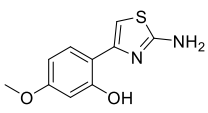
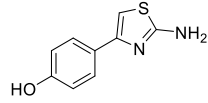
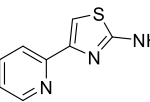
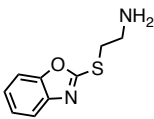
Compound	Structure	Biochemical IC <sub>50</sub> (μM)		Change in potency between variants
		IDH1-R132H	IDH1-WT	
CCT197091		993.2 ± 64.1	interferer	non-comparable
CCT374446		1043.8 ± 415.5	interferer	non-comparable
CCT374505		1783.7 ± 155.4	interferer	non-comparable
CCT373839		31% at 750 μM	interferer	non-comparable
CCT374503		25% at 750 μM	interferer	non-comparable
CCT374504		25 % 1.5 mM	interferer	non-comparable
CCT246301		> 3 mM	interferer	non-comparable

Table 5.5: Summary of biochemical assay data for analogues based on fragment hit CCT242635. Values reported for are an average of two biological repeats with the standard deviation.

In addition, one analogue of CCT239544 was designed, which showed similar activity to the parent fragment, with IC<sub>50</sub> values of 131.6 μM and 314.8 μM for CCT239544 and the analogue CCT374448 respectively (Table 5.6).

In general, the analogues of loop-moving fragments are more potent than those binding through a stack on Trp205. The ability to rapidly identify a range of compounds that bind to IDH1-R132H with similar or increased potency than the initial fragment hits further supports the druggability of the novel secondary site.

Parent Fragment Hit				
Compound	Structure	Biochemical IC50 (μM)		Change in potency between variants
		IDH1-R132H	IDH1-WT	
CCT239544		131.6 ± 13.4	-	non-comparable

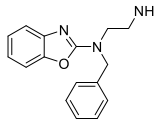
Analogue				
Compound	Structure	Biochemical IC50 (μM)		Change in potency
		IDH1-R132H	IDH1-WT	
CCT374448		314.8 ± 77.9	interferer	non-comparable

Table 5.6: Summary of biochemical assay data for analogue based on fragment hit CCT239544. Values reported are an average of two biological repeats with the standard deviation.

## 5.4 Inhibition of IDH1-WT

### 5.4.1 Inhibition of IDH1-WT by fragments and analogues

Of the 19 fragment hits, 15 were investigated for their ability to inhibit IDH1-WT activity, as CCT239544, CCT373604, CCT240772 and CCT37295 were not available at the time. Removal of concentration points found to fluorescently interfere with the biochemical assay points resulted in incomplete curves for 13 of these fragments. The two fragments that did not show interference, CCT371095 and CCT37098, were inactive against both IDH1-WT and IDH1-R132H. Based on the IDH1-WT NADPH fluorescence assay, no reliable selectivity between IDH1-WT and IDH1-R132H by targeting the novel secondary site could be identified for fragments binding to the novel secondary site.

The analogues designed by Sandra Codony Gisbert showed more activity against IDH1-R132H. I therefore investigated the ability of these analogues to inhibit IDH1-WT, to determine whether targeting the novel secondary site can

provide selectivity for the cancer-associated mutant IDH1-R132H. Reliable IC<sub>50</sub> curves could be generated for 12 of the 30 analogues investigated, (Table 5.4, Table 5.5, Table 5.6). The remaining 18 showed significant interference and IC<sub>50</sub> curves could not be calculated. CCT304224 and CCT374449 showed comparable potencies against IDH1-R132H and IDH1-WT, with measured IC<sub>50</sub> values showing less than a three-fold change between the variants. The remaining ten showed a fivefold to 13-fold decrease in potency when inhibiting IDH1-WT in comparison to IDH1-R132H. CCT374509, which was the most potent compound tested against IDH1-R132H, showed a 13-fold drop in potency when tested for activity against IDH1-WT. Although this may be due to errors in K<sub>m</sub> calculations rendering the assays non-comparable, it could indicate that some selectivity for IDH1-R132H over IDH1-WT can be obtained by targeting the novel secondary site.

## **5.5 Structural rationale for IDH1-R132H inhibition**

The two most potent fragments hits, CCT242635 and CCT239544, were both identified binding to the novel secondary site with remodelling of the pocket-forming loop. Overlaying these fragment-bound structures with the active conformation of IDH1-R132H shows a similar clash between the novel  $\alpha$ -helix and the regulatory segment Mg<sup>2+</sup>-stabilised  $\alpha$ -helix (Figure 5.14). Thus, fragment binding and subsequent helix formation may inhibit IDH1-R132H activity by hindering the formation of the active conformation. This could be a mechanism by which all fragments binding with remodelling of the pocket-forming loop inhibit IDH1-R132H activity.

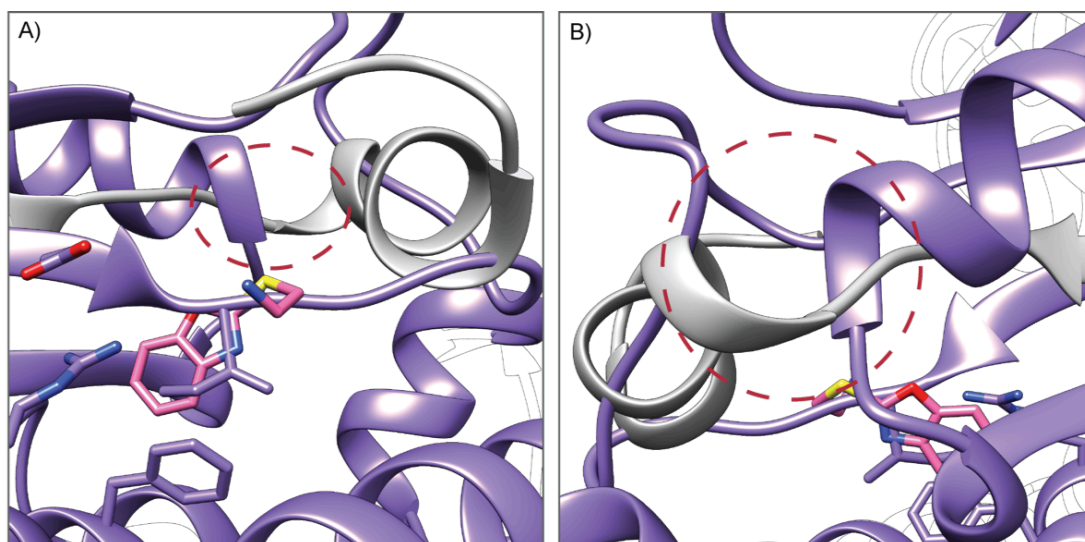


Figure 5.14: The novel  $\alpha$ -helical conformation may clash with the regulatory segment in the active conformation and offer structural rationale for inhibition of IDH1 activity. Front (A) and reverse (B) views of the active IDH1-R132H conformation (purple) showing the clash of the regulatory segment helix in the active conformation with the novel helical conformation (grey). The red circle highlights the location of the clash. The fragment shown is CCT239544, and the active conformation is from PDB 5YFM. Figure made in Chimera<sup>65</sup>.

In the novel helical conformation, Cys114 and Cys379 are in close proximity. In the structures of IDH1-R132H bound by CCT239686, a disulphide bond forms between these two cysteine residues. The presence of reducing agent in the crystallisation conditions used for XChem fragment screen prevents the formation of this bond, so adoption of the alpha-helix is observed without the disulphide bond. Cys379 is conserved in mammalian IDH1 enzymes (Appendix A.1.4), and can be reversibly modified by nitric oxide to form an S-nitrosothiol adduct resulting in inactivation of the enzyme<sup>184</sup>. It has been hypothesised that the S-nitrosothiol modification of Cys379 prevents the regulatory segment adopting the  $\alpha$ -helix required for activity through a steric clash<sup>130</sup>. This supports the hypothesis that a steric clash between the novel helix and the regulatory segment results in inhibition of IDH1-R132H activity.

## 5.6 Investigation of mutations in the novel pocket

During structural investigation of fragment screening hits, several side chain movements were consistently observed co-occurring with fragment binding. The salt bridge between Glu110 and Arg338 was repeatedly broken, both in structures with fragments binding through a stack on Trp205 and with fragments that caused remodelling of the pocket-forming loop. A patient-derived arginine to threonine mutation at position 338 was identified as part of the computational analysis, and is predicted to be both destabilising by Site Directed Mutator<sup>185</sup> and have functional impact by Mutation Assessor<sup>186</sup>. I designed two mutations in IDH1-R132H at this position – an arginine to threonine mutation to mimic the patient-derived mutation, IDH1-R132H-R338T, and an arginine to alanine mutation, IDH1-R132H-R338A.

Remodelling of the IDH1-R132H pocket-forming loop is also associated with a disrupted hydrophobic interaction between Phe334 and Ile112. In most structures with the loop remodelled, the fragment binds through a hydrophobic stack on top of Phe334, except for the most potent fragment-screening hit, CCT242635, in which the side chain of Phe334 rotates to form an edge-face stack with the thiazole moiety of this fragment (Chapter 4.4.4, Figure 4.14D). Breakage of the Phe334-Ile112 hydrophobic interaction therefore seems to be important during movement of the pocket-forming loop. To test this hypothesis, I designed an isoleucine to alanine mutation, IDH1-R132H-I112A, at this position to assess whether it would destabilise the interaction and facilitate formation of the novel  $\alpha$ -helix.



### 5.6.1 Characterisation of IDH1-R132H double mutants

IDH1-R132H variants were cloned, expressed and purified as described in Chapter 3.2.1. The purified proteins were characterised by native thermal shift as described in Chapter 3.2.2. All three of the IDH1-R132H variants maintained the ability to bind the native co-factor NADPH, with no significant differences observed in melting temperatures (Table 5.7, Appendix 8.2.4), indicating that they were properly folded.

NADPH Concentration (mM)	IDH1-R132H		IDH1-R132H-I112A		IDH1-R132H-R338T		IDH1-R132H-R338A	
	T <sub>m</sub>	ΔT <sub>m</sub>	T <sub>m</sub>	ΔT <sub>m</sub>	T <sub>m</sub>	ΔT <sub>m</sub>	T <sub>m</sub>	ΔT <sub>m</sub>
0	53.2°C	0°C	51.6°C	0°C	53.0°C	0°C	51.4°C	0°C
0.25	57.3°C	4.1°C	56.6°C	5°C	57.4°C	4°C	55.2°C	4°C
0.5	57.7°C	4.5°C	56.9°C	5°C	57.4°C	4°C	55.8°C	4°C
1	58.0°C	4.8°C	57.7°C	6°C	57.1°C	4°C	56.4°C	5°C
1.5	58.1°C	4.9°C	57.8°C	6°C	56.6°C	4°C	56.7°C	5°C
2	58.0°C	4.8°C	57.9°C	6°C	56.3°C	3°C	56.7°C	5°C

Table 5.7: Melting temperatures from label-free thermal shift data for IDH1-R132H double mutants with increasing concentrations of natural co-factor NADPH. IDH1-R132H melting temperatures are included for comparison.

### 5.6.2 Investigating the effects of secondary site mutations on IDH1-R132H activity

#### 5.6.2.1 Enzyme titration

A range of IDH1-R132H double mutant enzyme concentrations between 5 and 100 nM were investigated to allow measurement of  $V_0$  (Figure 5.15). The IDH1-R132H-I112A variant showed a much slower turnover than the IDH1-R132H and IDH1-R132H-R338X variants, indicating that Ile112 is important for enzymatic activity. I therefore increased the concentration of IDH1-R132H-I112A to 100 nM to measure  $V_0$ . The concentrations of other variants were maintained at 20 nM.

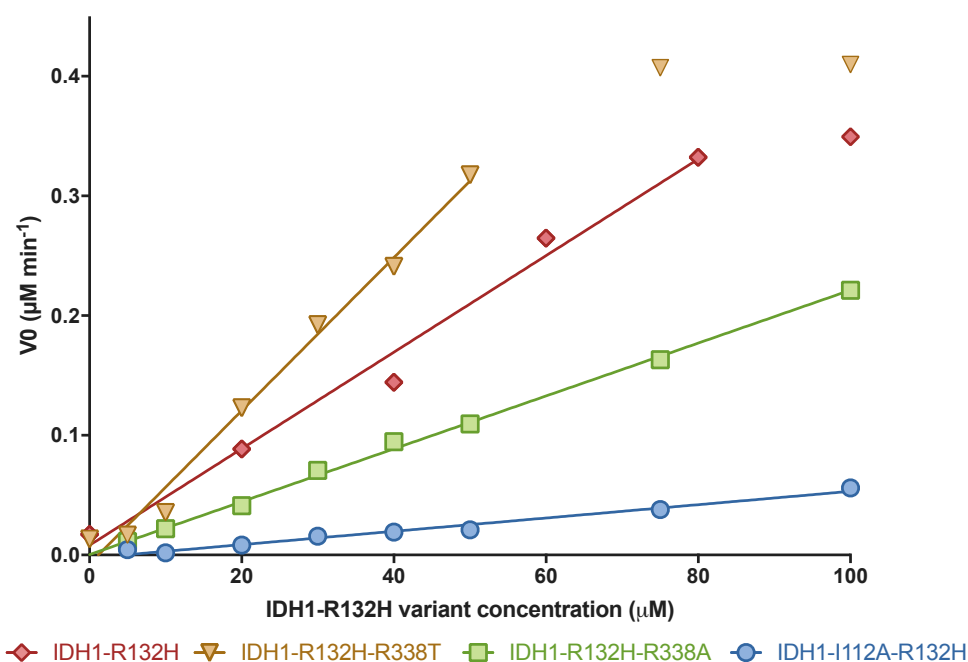


Figure 5.15: Initial rates of reaction for different IDH1-R132H variants at different enzyme concentrations, with the IDH1-R132H variant shown for comparison. IDH1-I112A-R132H was found to catalyse the reaction much more slowly than the other two variants, and I therefore selected a higher enzyme concentration to determine the kinetic parameters. Plot made in GraphPad Prism.

### 5.6.2.2 Determination of kinetic parameters

The Michaelis-Menten constants  $K_m$  and  $V_{max}$  for the double mutants were determined in the same way as for IDH1-R132H single mutant, with the exception of IDH1-R132H-I112A, which required 100 nM enzyme, in comparison to 20 nM enzyme for the other variants.

All of the IDH1-R132H double mutants show a decreased affinity for the substrate  $\alpha$ KG (Table 5.8), with IDH1-R132H-R338A showing a three-fold decrease in affinity and IDH1-R132H-I112A binding with approximately tenfold less affinity than the single mutant. For the IDH1-R132H-R338A and IDH1-R132H-I112A variants, the decrease in affinity at least partially drives the decreased  $k_{cat}$  in comparison to the IDH1-R132H single mutant.

	$\alpha$ KG $K_m^{app}$			
	IDH1-R132H	IDH1-R132H-R338A	IDH1-R132H-R338T	IDH1-R132H-I112A
<b>V<sub>max</sub> (<math>\mu</math>M min<sup>-1</sup>)</b>	0.244 $\pm$ 0.012	0.12 $\pm$ 0.0065	0.21 $\pm$ 0.006	0.35 $\pm$ 0.026
<b>K<sub>m</sub> (mM)</b>	803 $\pm$ 152	3035 $\pm$ 429	1893 $\pm$ 161	9099 $\pm$ 1191
<b>k<sub>cat</sub> (min<sup>-1</sup>)</b>	12.2	5.9	10.7	3.5

Table 5.8: Kinetic parameters for IDH1-R132H variants with respect to co-factor  $\alpha$ KG. All enzymes were tested at 20 nM, except for IDH1-R132H-I112A that was tested at 100 nM due to its lower catalytic activity. Data is an average of three biological repeats with standard deviations. Values calculated by GraphPad Prism

When determining the kinetic parameters for NADPH,  $\alpha$ KG concentrations should be present in saturating concentrations<sup>187</sup>. However, as all of the double mutants show an increased  $K_m$  for  $\alpha$ KG this was not technically feasible due to solubility limitations. For the double mutants, 5 mM  $\alpha$ KG only represents between 2.6-fold and 0.5-fold  $K_m$ . I maintained  $\alpha$ KG at 5 mM and report the NADPH  $V_{max}^{app}$  and  $K_m^{app}$  values (Table 5.9) allowing comparison between the different variants.

NADPH $K_m^{app}$				
	IDH1-R132H	IDH1-R132H-R338A	IDH1-R132H-R338T	IDH1-R132H-I112A
<b>V<sub>max</sub><sup>app</sup> (<math>\mu</math>M min<sup>-1</sup>)</b>	0.204 $\pm$ 0.009	0.03862 $\pm$ 0.0007	0.090 $\pm$ 0.002	0.044 $\pm$ 0.001
<b>K<sub>m</sub><sup>app</sup> (<math>\mu</math>M)</b>	3.17 $\pm$ 0.66	0.42 $\pm$ 0.13	0.725 $\pm$ 0.15	1.20 $\pm$ 0.23
<b>K<sub>cat</sub><sup>app</sup></b>	10.2	1.93	4.5	0.44

Table 5.9: Apparent kinetic parameters for IDH1-R132H variants with respect to co-factor NADPH. All enzymes were tested at 20 nM, except for IDH1-R132H-I112A that was tested at 100 nM due to its lower catalytic activity. Data is an average of three biological repeats with standard deviations. Values calculated by GraphPad Prism

The NADPH  $V_{max}$  and  $k_{cat}$  values for all three IDH1-R132H double mutants is lower than the IDH1-R132H, but this is likely due to the smaller excess of  $\alpha$ KG in comparison to  $K_m$ . The  $K_m^{app}$  for NADPH is broadly consistent across all of double mutants, indicating that the decrease in affinity for  $\alpha$ KG is driving

reduced catalytic activity of IDH1-I112A-R132H and IDH1-R132H-R338A. Although the absolute kinetic parameters could not be obtained, these results show that secondary site mutations do impact IDH1-R132H function, confirming that the novel secondary site is functionally relevant. No changes were observed in the overall protein fold for any of the IDH1-R132H variants, despite the impact on enzymatic activity

## 5.7 Conclusions

With the ligandability of the novel secondary site in IDH1-R132H confirmed in the previous chapter, this chapter describes the investigation of the secondary site's functional relevance. I established *in vitro* biochemical assays for both IDH1-R132H and IDH1-WT based on the intrinsic fluorescence of NADPH. Of the nineteen fragments identified through both thermal shift and crystallographic fragment screening, ten were able to inhibit IDH1-R132H activity, with IC<sub>50</sub> values down to 84  $\mu$ M, confirming that the pocket has functional relevance. For fragment hits binding with remodelling of the pocket-forming loop, a steric clash between the new helix with the regulatory segment may provide structural rationale for inhibition by these fragments.

Potency could be maintained and increased through exploration of the chemical space around the initial fragment hits. These analogues also showed selectivity for the IDH1-R132H over IDH1-WT. In addition, mutations in the novel secondary site negatively impact the ability of IDH1-R132H to bind  $\alpha$ KG and reduce enzymatic activity. Together, this shows that the novel secondary site is functionally relevant.

## Chapter 6: Conclusions, lessons learnt and remaining questions

---

### 6.1 Conclusions

This thesis described how combining *in silico* analysis with fragment screening could successfully identify novel, ligandable secondary sites in cancer-associated proteins. The first aim of the project was the adaptation of canSAR3D to identify novel secondary sites, which was then validated using known, ligandable secondary sites. The second aim was to use fragment screening to experimentally investigate the ligandability of the selected secondary site. I identified 19 fragments across two fragment screening approaches not only binding specifically to the novel secondary site in IDH1-R132H, but also showing activity in a biochemical assay.

Combining *in silico* analysis with fragment screening therefore allows rapid identification of ligandable secondary sites in cancer-associated proteins. The rest of this chapter discusses some of lessons learnt during the project, potential future plans and the outlook.

## 6.2 Lessons learnt

### 6.2.1 Limitations of the computational predictor

The reliability of computational predictions is dependent on the quality of data input. The PDB is an invaluable resource upon which to train and test predictors, but there are experimental limitations of structure-determination techniques that impact the accuracy and completeness of a given PDB structure. In flexible regions of proteins, the electron density may be absent or too weak to allow the modelling of side chains or loops. I found that the computational predictor was sensitive to missing residues and side chains, due to the impact on how pockets were defined. For example, in the published structure of IDH1-R132H, the side chain of Ly345 can be modelled in electron density, forming the edge of the novel secondary site. In the in-house structure, the lack of electron density means that the side chain was not modelled. Without this side chain, the pocket edge is not fully formed, leading to a decrease in the overall enclosure such that the novel secondary site was not predicted to be ligandable (Chapter 3.2.7).

The absence of loops due to the lack of electron density can also influence the prediction. For example, the known allosteric site in IDH1 is only predicted to be ligandable in the inactive conformation where the regulatory segment is modelled (Figure 6.1A, E-F). This segment is unstructured in most available IDH1 structures, which prevents complete formation of the pocket. The pocket is not predicted to be ligandable in these structures due to an increase in the accessible vertices (Figure 6.1B,E-F).

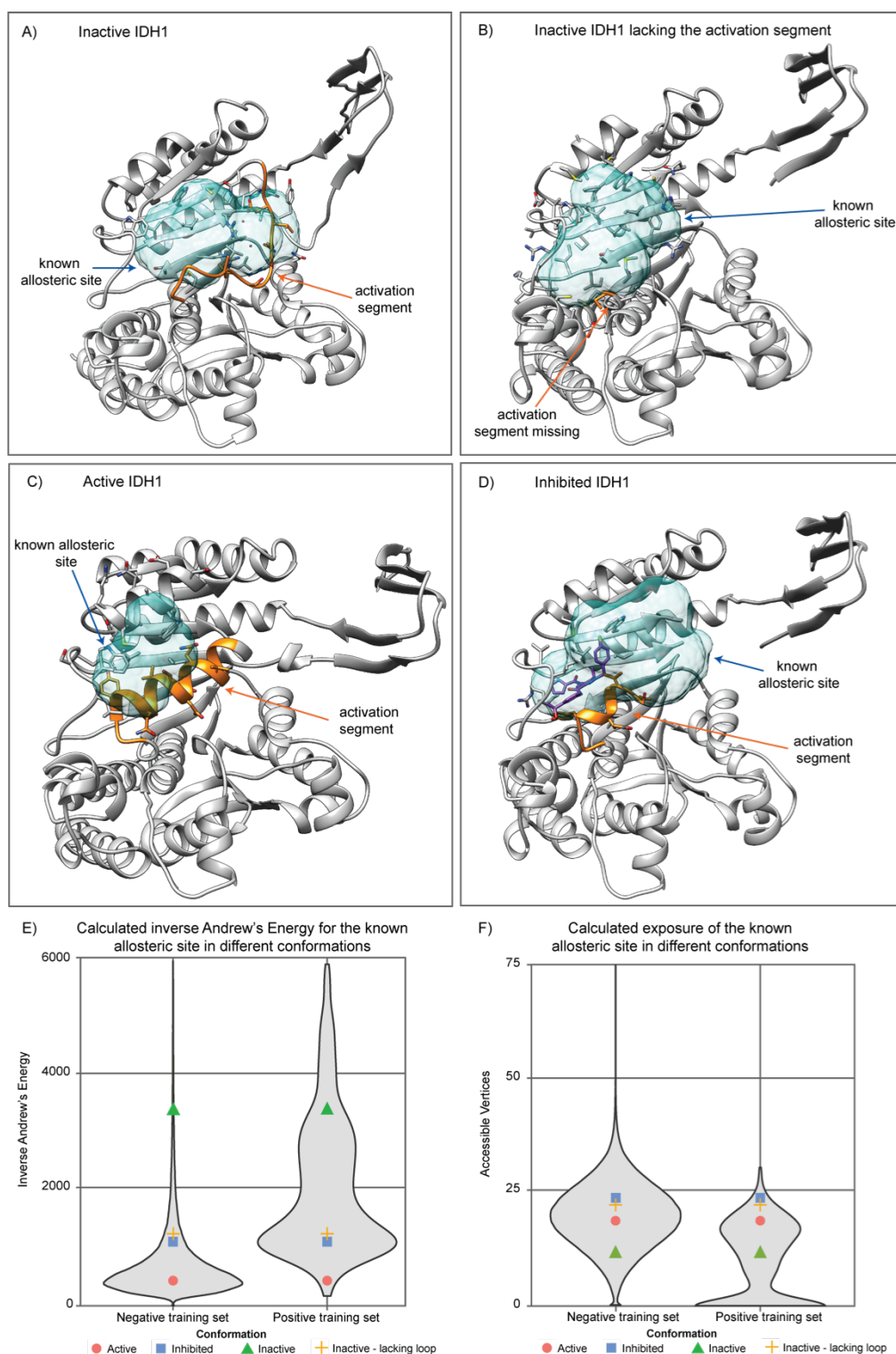


Figure 6.1: Definition of the known allosteric site in IDH1 is dependent on both the protein conformation and completeness of the regulatory segment. A) In the inactive conformation, the regulatory segment is a unstructured loop that forms the edge of the known allosteric site. B) Loss of the loop results in incomplete formation of the allosteric site. C) Formation of the helix in the active conformation occludes part of the pocket. D) Binding of inhibitors to the known allosteric site results in partial formation of the helix and partial formation of the allosteric site. E-F) Violin plots showing the population distributions of the positive and negative training sets for the inverse Andrews' energy and the accessible vertices. The corresponding values for the known allosteric site in each conformation are plotted as single points for comparison to the distribution of each training set. Structure figures made in Chimera<sup>65</sup>, violin plots made in R<sup>168</sup> with ggplot2<sup>147</sup>.

The known allosteric site in IDH1 is also an example of a conformation dependent pocket. The pocket was also identified in both the active and inhibited structures of IDH1, but was only predicted to be ligandable in the inactive conformation. In the active conformation the edge of the pocket is formed, the  $\alpha$ -helix encroaches into the pocket, reducing the volume and the inverse Andrew's energy (Figure 6.1C, E-F), while the incomplete formation of the helix in the inhibited conformation leads to an increase in the solvent accessibility (Figure 6.1D, E-F).

The protein construct can also impact the identification and analysis of pockets. Removal of flexible loops to promote crystallisation can result in pockets not being identified, while purification or solubilisation tags can form artificial pockets or even occlude real pockets. During the analysis, the secondary site in PIK3CA was identified and predicted to be ligandable in four structures. During triaging, I found that under some conditions the canSAR3D pipeline could not identify the pocket as it was occluded by the N-terminal HisTag (Figure 6.2). Computational removal of this tag from the input PDB and reanalysis allowed the pocket to be defined, and it was subsequently predicted to be ligandable in additional structures.

When protein structures are used for *in silico* screening or docking of small molecules in a given protein site, computational tools are used to add missing residues and side chains, with energy minimisation protocols to find the most likely conformer<sup>188, 189</sup>. However, as discussed above, analysis of a single conformation can prevent identification of conformation-dependent pockets. To generate an ensemble of structures that are representative of different possible



protein structures in solution, molecular dynamics simulations, or less computationally expensive Monte Carlo simulations, can be used to investigate dynamic changes in a protein structure. These ensembles could then be interrogated for the presence of ligandable pockets. This approach was unfeasible for my project, as I was analysing all publically available structures of human proteins. However, in the context of a drug discovery project where only a single or a few targets are under consideration, then this would be a valid approach to consider.

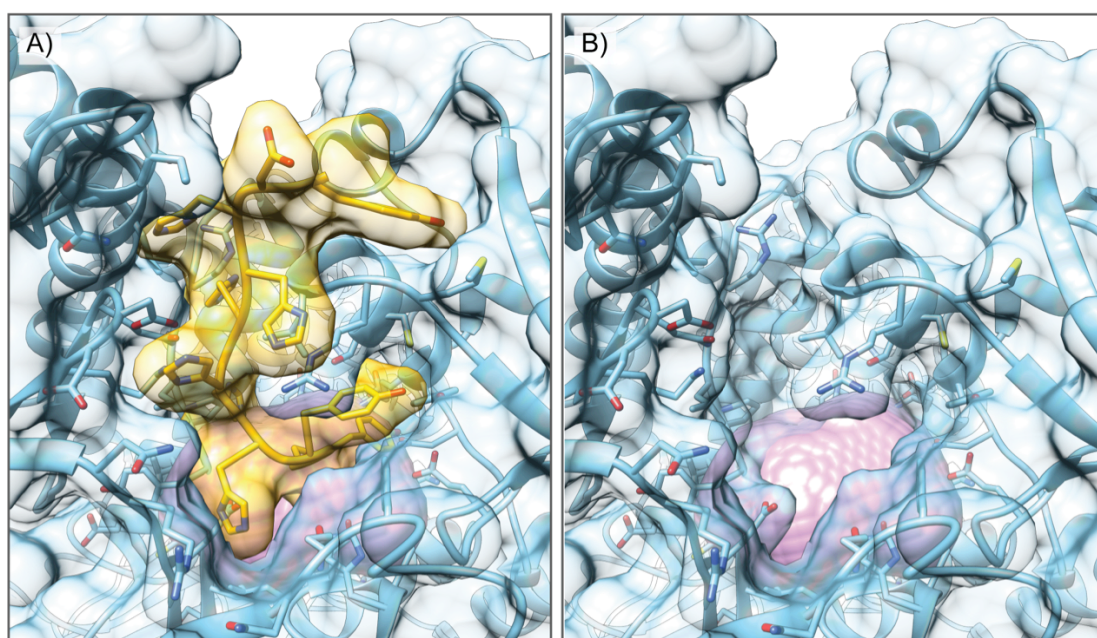


Figure 6.2: Structures showing the novel secondary site in PIK3CA, which since been validated by crystallographic fragment screening. A) The identified pocket in PIK3CA is occluded by an N-terminal HisTag in multiple constructs; canSAR3D does not identify the pockets under these conditions. B) Removal of the HisTag results in exposure of the pocket and subsequent identification by canSAR3D. Figures made in Chimera<sup>65</sup>

### **6.2.2 Training set bias**

The positive training set was formed of pockets known to be druggable, the majority of which were primary sites. As primary sites tend to be the largest pocket in a protein, this automatically introduces a bias for the largest pocket being predicted as the most ligandable. Consequently, the properties identified as showing statistical significance needed to be de-convoluted from those that were introduced by this training set bias.

With the increased availability of data from crystallographic fragment screening, building a new positive training set from pockets identified crystallographically may be feasible. While care would have to be taken to exclude sites that are less relevant in solution, such as those in crystal contacts, fragment screening hit rate is associated with ligandability, and so fragment binding hot spots identified through crystallographic fragment screening could be used to form a positive training set. The statistical analysis could then be repeated to identify pocket properties that show a statistically significant difference and where the thresholds are placed. In addition, rather than developing a binary discriminant, alternative approaches such as neural networks could be used to develop a sliding-scale predictor of ligandability.

### **6.2.3 Crystal form limitations**

Repeated fragment soaking experiments indicated that only a small proportion, approximately 5-10%, of protein crystals grown under the current conditions could permit remodelling of the pocket-forming loop to adopt the novel  $\alpha$ -helix, which was the conformation in which multiple hits were identified binding. There was no clear difference between either morphology or the unit cell

parameters of these crystals, nor was there clear structural rationale for this limitation. Because of this, it was very challenging to obtain structures for the fragments binding to this loop-remodelled conformation. Several residues in the pocket could have been targeted to explore this further (Figure 6.3), including a hydrophobic interaction between Ile112 and Phe334. The I112A mutation, designed to destabilise the pocket-forming loop impacted enzymatic activity (Chapter 5.6.2), but showed no effect on the percentage of crystal structures allowing loop remodelling. Phe334 could not be targeted by mutagenesis as many fragments were identified binding through a  $\pi$ - $\pi$  stack on this residue. The side chain of Arg338 forms multiple Hydrogen bonds in the normal loop conformation, but no interactions in the novel helical conformation. This could be further investigated with the aim of destabilising the pocket-forming loop in crystal structures.

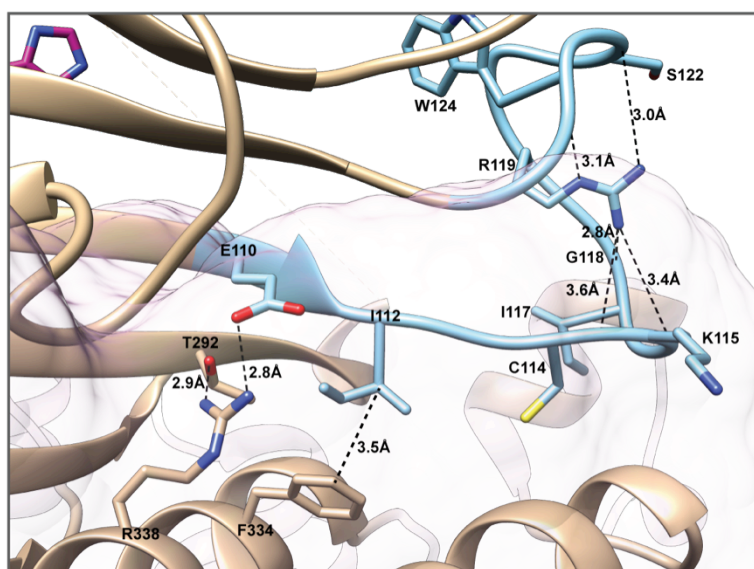


Figure 6.3: Key stabilising interactions in the IDH1-R132H novel secondary site. Mutations at positions 112 and 338 have already been generated, but showed no difference in loop conformation when investigated crystallographically. Arg119 is involved in an extensive hydrogen-bonding network, and could be targeted to destabilise the loop. Based on an in-house IDH1-R132H structure. Figure made in Chimera.

#### 6.2.4 Non-isomorphous crystals in PanDDA

Using PanDDA allowed identification of fragment hits that would have otherwise been overlooked. However, the non-isomorphism of the crystals presented a significant challenge. As discussed in Chapter 4.3.2, datasets showing movement between monomers or clasp movement in comparison to the MR model were removed from the ground state calculation to prevent blurring of the ground state map in this region. In the individual datasets that were subsequently analysed by PanDDA, these movements resulted in poorer phase estimations and lower quality initial maps. Further, because this movement is a significant deviation from the ground state, a large number of non-relevant events were identified in this region, which required manual inspection.

The movement of the pocket-forming loop presents a similar challenge. Reorganisation of the loop is a significant deviation from the ground state, and PanDDA identified these events very clearly. The Z-maps showed large regions of negative and positive density corresponding to loss of the normal conformation and adoption of the novel conformation, even when the occupancy was low. However, PanDDA could not always identify the additional density corresponding to the fragment. For example, the  $2mF_o - DF_c$  maps for a crystal soaked with TSA hit CCT242635 during the crystallographic fragment screen showed clear density for the fragment in a dual conformation, as well as the movement of the pocket-forming loop (Figure 6.4A). In comparison, the Z-map calculated by PanDDA clearly shows movement of the pocket-forming loop, but the corresponding fragment density was much more ambiguous, and the dual conformer wasn't seen (Figure 6.4B).

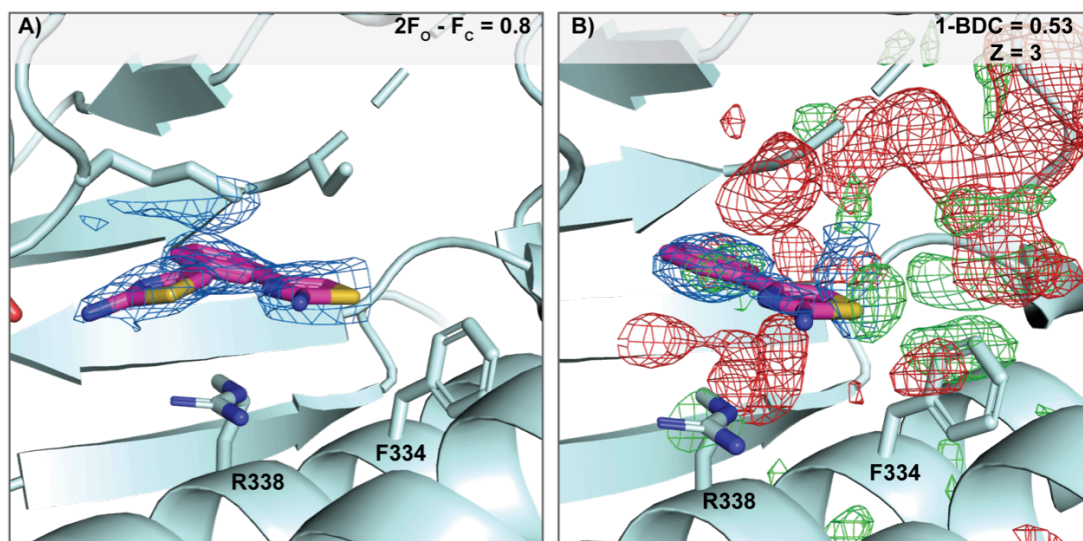


Figure 6.4: Comparison  $2mF_o - DF_c$  (A) and Z- and event (B) maps from the same dataset from an IDH1-R132H crystal soaked with CCT242635. Normal maps clearly show both loop remodelling and fragment density, while the PanDDA maps clearly show deviation from the ground state corresponding to the remodelling of the pocket-forming loop, but fragment density is ambiguous. Figures made in Pymol<sup>173</sup>.

An ideal solution would have been the generation of multiple MR models and corresponding ground states for each of conformations sampled (Figure 6.5). In the context of fragments that bind with remodelling of the pocket-forming loop, a molecular replacement model with the loop in the novel conformation would improve the initial phases in this region. This would in turn produce cleaner electron density maps for the PanDDA analysis and reduce ambiguity for these fragment hits. In addition, the remodelling of the pocket-forming loop itself would not constitute a large deviation from the ground state, and therefore would not obscure additional fragment density. However, calculation of each ground state requires approximately 50 isomorphous datasets, and too few datasets for each conformer was collected during XChem to achieve this.

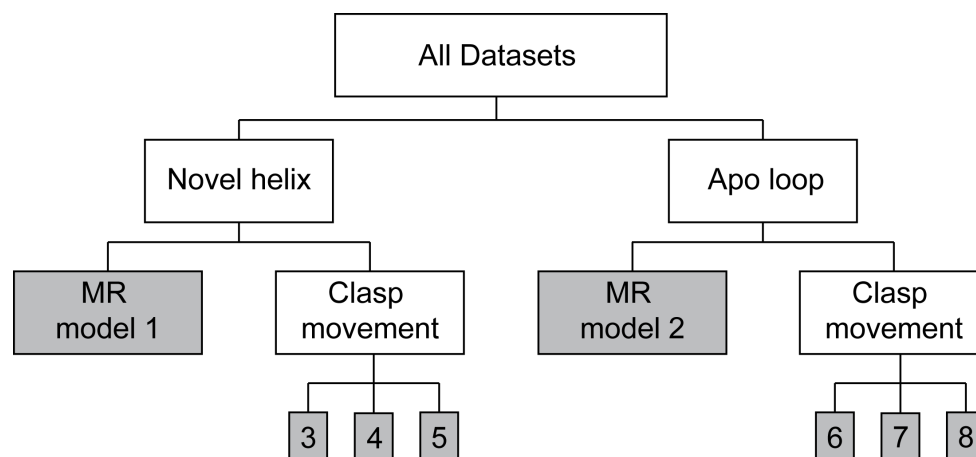


Figure 6.5: Example of class averages for ground state generation. In an ideal scenario, datasets would be classified into one of multiple possible states, similar to class averages routinely built during electron microscopy experiments. The shaded boxes indicate each separate class, for which a new molecular replacement model and ground state would be built and optimised and a separate PanDDA analysis run. The lack of datasets for each class collected during XChem made this unfeasible, as 50 datasets are required to build a robust ground state.

#### 6.2.4.1 *Fragment soaking concentration*

The concentration at which to soak fragments during crystallographic fragment screening is under discussion, and is system dependent. If fragments are soaked at too low a concentration, potentially interesting fragment hits binding through few, but specific, interactions may be overlooked. In contrast, soaking at higher concentrations of fragments can lead to the identification of fragment hits binding with weaker affinity, although they may be more challenging to progress. High fragment soaking concentration can also lead to the disruption of crystals. In addition, increasing fragment soaking concentration can lead to precipitation, or even the crystallisation of the fragment. For example, when soaks with TSA fragment hit CCT239686 were repeated with the fragment concentration increased to 200 mM in comparison to 50 mM in the initial soak, clusters of diffraction spots at high resolution corresponding to fragment crystals were found (Figure 6.6). No fragment density was observed in the



solved structures. This indicates that above a certain concentration, the planar fragment formed very small crystals on the surface of the protein crystals, which were not visible under the microscope. This resulted in a decrease in the fragment concentration in solution, and no fragment binding to the protein.

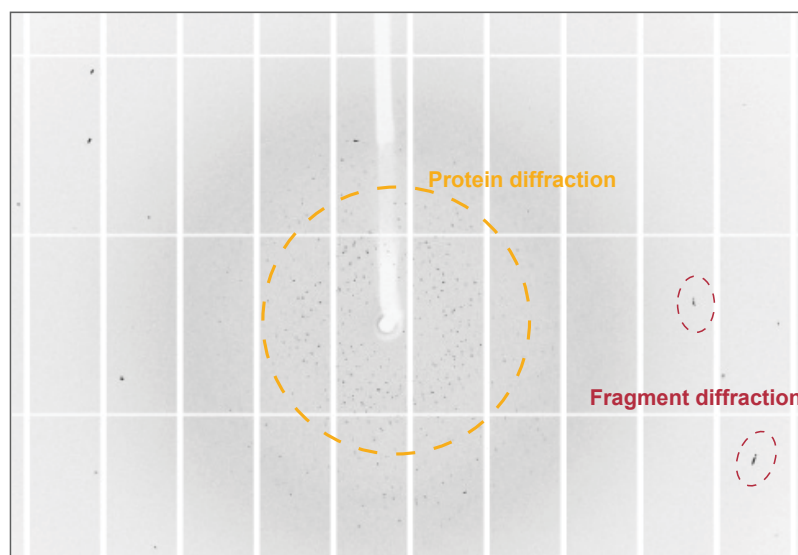


Figure 6.6: Diffraction pattern from an IDH1-R132H crystal soaked with high concentrations of CCT239686. High-resolution reflections corresponding to fragment crystals are circled in red.

### 6.3 Questions remaining

The major remaining question relates to the cellular validation of the novel secondary site in IDH1-R132H as a therapeutic target. The work described in this thesis shows that the novel secondary site is ligandable, and that fragment binding to this site can impact IDH1-R132H activity. In addition, for fragment hits that bind with remodelling of the pocket-forming loop, the steric clash between the novel  $\alpha$ -helix and the regulatory segment in the active conformation provides structural rationale for inhibition IDH1-R132H catalytic activity (Chapter 5.5). The impact of the I112A on IDH1-R132H catalytic activity indicates that the pocket-forming loop has functional importance (Chapter 5.6.2.2). However, it is not known if this loop, and its alternative  $\alpha$ -helical

conformer, has a role in the normal IDH1 catalytic cycle. The impact of perturbing IDH1-R132H through the novel secondary sites on cellular proliferation and viability is also unknown.

This could first be investigated using cellular assays, with either compounds that bind with remodelling of the pocket-forming loop, or through mutagenesis. The concentration of 2HG could also be used as to monitor IDH1-R132H activity in cellular assays<sup>169,39</sup>. The current fragment hits and analogues show micromolar potency, which is too weak to be expected to show cellular activity. Extensive synthetic chemistry would be required to develop sufficiently potent, cell penetrant compounds. Generation of cell lines expressing either IDH1-R132H or an IDH1-R132H variant with a mutation targeting the novel secondary site may be an alternative approach.

## 6.4 Outlook

This thesis described the combination of *in silico* analysis with fragment screening to rapidly identify and evaluate a novel ligandable secondary site. The computational approach identified two examples of novel secondary sites that have since been shown to be ligandable through fragment screening – the novel site in PIK3CA that was reported by another group, and the site in IDH1-R132H which I validated during this project. This approach is useful for rapid early evaluation of a potential target, before initiating a full drug discovery project.

IDH1-R132H is a clinically important cancer target. The first IDH1-R132X targeted drug, ivosidenib, was approved in July 2018<sup>190</sup> for treatment of IDH1-



R132X AML. It is currently in phase III trials for treatment of IDH1-R132H mutant cholangiocarcinomas<sup>191</sup>. Resistance mutations to Ivosidenib were reported in the literature in July 2018<sup>89</sup>. The reported S280F mutation is not located close to the novel secondary site. Targeting this novel site may provide a potential way to overcome the resistance already emerging against the approved therapeutic.

## Chapter 7: Materials and Methods

---

### 7.1 Computational Methods

#### 7.1.1 Building the ligandability predictor

Code was written in R<sup>168</sup>, using Rstudio (Boston, USA), unless otherwise stated.

##### 7.1.1.1 Pocket definitions

Both the SURFNET-defined and Pickpocket-restrained pocket definitions, as well as their associated properties were obtained directly from canSAR3D (canSAR v3.0<sup>192</sup>). These form the basis of canSAR3D's druggability prediction, and will form the basis of my ligandability prediction. I used two canSAR3D identifiers for each pocket: the PDBID\_CHAIN and SITE\_NUMBER, such that PDBID\_CHAIN = 1KZY\_A and SITE\_NUMBER = 1 refers to the largest pocket identified in chain A of PDB structure 1KZY.

##### 7.1.1.2 Training set selection

The positive training set was formed of kinase primary sites, ligand-binding sites in nuclear receptors, and binding sites of FDA approved drugs. A list of kinase structures was taken from the PDB<sup>193</sup> (accessed 23/11/2015) using the E.C numbers 2.7.10-13 and 2.7.99. This was cross-referenced with the SURFNET properties using the PDB code. The primary site is the ATP binding site and was generally found to be SITE\_NUMBER = 1. Known allosteric sites, such as the ATP binding allosteric site in phosphofructose kinase<sup>194</sup>, were excluded.

Pfam<sup>195</sup> uses hidden Markov model sequence alignments to assign a probable domain to a given sequence, and clusters these domains into families. Nuclear receptors are multi-domain proteins, so the structures of the ligand-binding domain were selected by using the Pfam family classifier P00104. This was cross-referenced with the SURFNET properties using the PDB code. The ligand-binding site was identified as the pocket binding the natural ligand. The targets of FDA approved drugs were accessed directly from the PDB, and the pocket identified through cross-referencing the ligand bound to a given pocket with approved drug.

The positive training set was formed of 2,025 pockets. The ten pockets identified in each chain of each PDB structure, excluding those in the positive training set, were retained as the background set. The negative training set was randomly sampled from the background set.

#### **7.1.1.3 Statistical analysis**

All statistical analyses used to identify pocket properties that show statistically significant differences between the positive and negative training sets were completed in R<sup>168</sup>. Welch's t-test and Kolmogorov-Smirnov (KS) test were used to identify pocket properties that showed statistically significant differences between the positive and negative training sets. A Roc test was used to determine thresholds for the ligandability profile.

Both the Welch's T-Test and the Kolmogorov-Smirnov (KS) were bootstrapped 100,000 times, using the basic R package. The aggregate p-value is reported as the proportion of the time the outcome was found to be not significant ( $p \leq$

0.05). The cut-off for significance for the aggregate p-value was also set at  $p \leq 0.05$ .

For the Roc analysis, the package pROC<sup>196</sup> was used, bootstrapped 10,000 times. The greater processing power required for Roc tests limited the amount of bootstraps that could be achieved in a reasonable time frame. The cut-off for significance was set at AUC >80%, using the top-left method to determine the threshold.

The properties found to be significant were also investigated using randomisation. The positive and negative training sets were appended, and each half of the data randomly assigned to be either the positive or negative training set. The statistical tests were then repeated as before, and significance cut-offs maintained the same. None of the properties were found to be significant during randomisation, supporting their use for the ligandability profile.

The ligandability profile was defined as: Inverse Andrew's energy  $\geq 910$ , Accessible Vertices  $\leq 13.8$ , Buried Vertices  $\geq 70$ , Pocket Size  $\geq 750$ , Volume Ratio  $\leq 3$  and GAP  $\leq 3$ . Pockets that fit this profile were retained for triaging.

## **7.1.2 Triaging and target selection**

### **7.1.2.1 Cancer association**

To limit the targets to those associated with cancer, I correlated the proteins in which pockets are predicted to be ligandable with the Cancer Gene Census<sup>197</sup> (CGC, v75) and with those reported to be significantly mutated by Kandoth et

al<sup>13</sup>. The CGC is an expertly curated database of 564 cancer driving genes. Kandoth et al. report 127 significantly mutated genes (SGMs) following the systematic analysis of 3281 tumours across 12 cancer types<sup>13</sup>. From the two datasets, 80 genes were found in both, leaving 611 cancer-associated genes.

The cancer-associated genes and ligandable sites from the predictor were cross-referenced using UniProt<sup>198</sup> to Hugo Gene Nomenclature Committee (HGNC)<sup>199</sup> conversions. Pockets in proteins that were not cancer-associated were excluded.

#### **7.1.2.2 Conservation scores**

The conservation score is calculated as part of the canSAR3D pipeline. Each pocket is associated with both a PDBID\_CHAIN and a SITE\_NUMBER and the conservation score is reported. These scores were obtained directly from canSAR3D and correlated with the output from the ligandability predictor using PDBID\_CHAIN and SITE\_NUMBER.

#### **7.1.2.3 Mutation enrichment**

Patient-derived mutational information was acquired from The Cancer Genome Atlas (TCGA freeze.2015.213, <http://cancergenome.nih.gov/>). At the time of analysis, the TCGA contained 7,054,105 mutations from 8,855 patients across 32 cancer types. Only curated primary patient samples with matching normal tissue were included, as identified by sample\_type = 1 and background\_type = 1. This left 901,177 mutations from 585 patients and 11 cancer types. Of these, 38,090 mutations from 500 patients were found in cancer-associated genes.

The genomic coordinates were subsequently mapped to the amino acid sequence of Ensembl<sup>200</sup> canonical transcripts. Ensembl uses HGNC identifiers. The residues that make up the SURFNET-identified pockets in PDB structures are numbered based on the sequence submitted to the PDB. This is mostly in agreement with the UniProt canonical transcript, and was mapped using Structure Integration with Function, Taxonomy and Sequence (SIFTS)<sup>201, 202</sup>. Uniprot and Ensembl do not always consider the same transcript to be canonical. I compared the length of the canonical transcript as defined by Uniprot with the length of the canonical transcript as defined by Ensembl as a rapid indicator of agreement, and found that 13% reported a difference in transcript lengths. In these instances, the canonical transcript is likely to differ between Uniprot and Ensembl, which would lead to mismatched amino acid numbering. For these datasets, I manually mapped mutations to the pockets to avoid inappropriate exclusion of the pockets.

For pockets with associated mutations I also calculated a mutation enrichment score:

$$Score = \frac{cavity\ mutation\ rate}{protein\ mutation\ rate} \quad \text{Equation 7.1}$$

Where

$$cavity\ mutation\ rate = \frac{\#mutations\ in\ cavity}{\#residues\ in\ cavity}$$

and

$$protein\ mutation\ rate = \frac{total\ mutation\ frequency}{number\ of\ amino\ acids\ in\ protien}$$

#### **7.1.2.4 Manual Triaging**

Manual triaging involved exclusion of primary sites identified during the analysis, assessment of the structure quality around the pocket of interest, literature investigation for evidence of functional interest and assessment of technical feasibility.

## **7.2 Experimental investigation**

### **7.2.1 General**

#### **7.2.1.1 Reagent suppliers**

Analytical grade reagents were purchased from Sigma-Aldrich (Dorset, UK.) through Merck Millipore (Hertfordshire, UK), or from Melford (Ipswich, UK). Growth media, agar plates and antibiotic stocks, as well as HEPES and Tris stock solutions for assays, were purchased from the ICR Central Sterile Services Department (CSSD). Fragments tested were either part of the ICR's in house fragment library (1985 fragments) or from the 3D Fragment Consortium<sup>174</sup> (610 fragments), and stored as 100 mM stock solutions in 100% DMSO. Fragment hits were re-purchased as solid stocks from ChemBridge (San Diego, USA), Enamine (Monmouth, USA), MolPort (Riga, Latvia) or Fluorochem (Hadfield, UK) at > 95% purity. Both AGI-5198 and GSK-864 were purchased from Sigma-Aldrich at > 98% purity.

#### **7.2.1.2 Preparations of buffers and stock solutions**

The composition of buffers used is described in the respective protocols below. They were made by either reconstituting solid stocks or dilution of stock solutions in de-ionised water with approximately 18 MΩ.cm resistivity, except

for the buffers used for crystallisation which were made with OmniPur® Water for Injection (Calbiochem, Merck Millipore). Buffers were filtered with a 0.22 µm pore filter before use and kept no longer than one month. Dithiothreitol (DTT) was added on the day of use by diluting a 0.5 M stock stored as aliquots at -20 °C. NADP<sup>+</sup> and NADPH for assays were stored in single use aliquots at 50 mM at -80 °C. Both isocitrate and αKG for assays were stored as 100 mM stocks in single use aliquots also at -80 °C. NADPH for crystallography was stored at 250 mM in single use aliquots at -80 °C. NADP<sup>+</sup> for crystallography was stored at 500 mM in single use aliquots at -80 °C.

#### **7.2.1.3 Compound storage and dispensing**

Tool compounds and fragments were purchased as solid stocks and reconstituted in DMSO to a concentration of 10 mM for tool compounds and 500 mM for fragments. They were stored in Matrix™ 2D barcoded storage latch rack (Thermo Fisher, Wilmington, USA). For assays, fragments were diluted to 100 mM and 2 mM stocks in 100% DMSO (Fischer Scientific, Leicestershire, UK) and stored in sealed 384-well Echo HV polypropylene plates (Labcyte Inc., Sunnyvale, USA). Fragments and tool compounds for use in assays were dispensed into assay plates using an Echo 550 Liquid Handler (Labcyte Inc.). Assay plates were prepared no more than 48 hours in advance. All compound stocks were stored under compressed nitrogen gas in a Multipod™ system (Roylan Developments, Surry, UK).

Reagents were added to plates using a Tempest Liquid Handler (Formulatrix, Bedford, Massachusetts) unless otherwise stated.



#### **7.2.1.4 Assay data analysis**

Assay data was analysed using GraphPad Prism 7.0 (San Diego, CA), Studies and Vortex (domatics, Hertfordshire, UK) as stated in each section. The R script used to analyse the SYPRO Orange thermal shift data was written by Gary Nugent and was implemented in R.

### **7.2.2 Generation of IDH1 constructs**

#### **7.2.2.1 Construct design, synthesis and amplification**

The coding sequence for IDH1 (NCBI reference NM\_005896.3) including an N-terminal His<sub>6</sub>-tag and Tobacco Etch Virus (TEV)-cleavage site was codon optimized for expression in *E. coli* and generated by gene synthesis (Eurofin, Ebersberg, Germany). The plasmid was re-suspended in 10 mM Tris pH 8, and 1 µL was added to 9 µL of DH5α T1-resistant cells (Invitrogen, Carlsbad, California). The cells and DNA were incubated together on ice for approximately 15 minutes before being heat-shocked in a 42 °C water bath for 45 seconds. The cells were then incubated on ice for a further five minutes, before addition of 90 µL Super Optimal broth with Catabolite repression (SOC) medium (Sigma-Aldrich) and subsequently incubated at 37 °C, 220 rpm for one hour to recover. LB-agar plates containing kanamycin at 50 µg/mL were inoculated with 50 µL of cells and incubated over night at 37 °C. Single colonies were selected and grown overnight in 5 mL Luria Broth (LB) with 50 µg/mL kanamycin. The plasmid DNA was extracted using QIAprep spin Miniprep kit (Qiagen, Hilden, Germany).

#### **7.2.2.2 Subcloning into expression vector**

The gene of interest was excised from the commercial plasmid using NcoI-HF and BamHI-HF restriction enzymes (NEB, Ipswich, USA). After extraction, 75  $\mu$ L of the DNA was added to 10  $\mu$ L H<sub>2</sub>O, 10  $\mu$ L CutSmart buffer (NEB) and 2  $\mu$ L of each restriction enzyme. The sample was incubated at 37 °C for 90 minutes and analysed for size by migration on a 1% w/v agarose gel in 1x Tris-Acetate-EDTA (TAE) buffer. The corresponding band was excised from the gel and the DNA was extracted using Qiaex II Gel Extraction Kit (Qiagen). The gene insert was subsequently ligated into a linearized pET-28a plasmid using Quick Ligation kit (NEB) and transformed into RapidTrans™ TAM1 Competent *E. coli* (Active Motif, La Hulpe, Belgium). DNA was extracted from overnight cultures using 5Prime FastPlasmid mini kit (Qiagen). Sanger sequencing (Source Bioscience, Nottingham, UK) was used to confirm the sequence using T7 promoter and terminator primers. Plasmids were transformed into BL21-AI cells for expression using heat shock as described above.

#### **7.2.2.3 Site Directed Mutagenesis**

Primers for site directed mutagenesis were designed using the Eurofins PCR Primer design tool. The mutant constructs were generated by site-directed mutagenesis of the IDH1-pET28a plasmid previously produced. Reagents from the KOD Hot Start Master Mix (Novagen, Merck Millipore) and protocol from QuickChange II site-directed mutagenesis kit (Agilent, Santa Clara, USA) were used. Primers were reconstituted in nuclease-free water (Ambion®, Invitrogen) to 10  $\mu$ M, and 0.5  $\mu$ L of each forward and reverse primer was added to 12.5  $\mu$ L of QuickChange II Master Mix with 0.5  $\mu$ L of template DNA and 11  $\mu$ L nuclease free water (Ambion®, Invitrogen). The thermal cycling conditions were as

follows: 95 °C for 20 seconds, 50 °C for 20 seconds, 68 °C for 7 minutes, for 15 cycles, using a ProFlex PCR System (Life Technologies, Thermo Fischer). PCR products were incubated with 1 X Cutmart buffer (NEB) and 1 µL Dpn1 (NEB) at 37 °C for 90 minutes to digest methylated template DNA before transforming into RapidTrans™ TAM1 Competent *E. coli* (Active Motif) using heat shock as described previously. Single colonies were selected and grown overnight in 5 mL LB and plasmid DNA extracted using QIAprep spin Miniprep kit (Qiagen). Sequences were confirmed by Sanger sequencing (Source Bioscience).

### **7.2.3 Protein expression and purification**

#### **7.2.3.1 Expression**

All IDH1 constructs were expressed in BL21(DE3) AI cells (Thermo Fisher). Cells were transformed using heat shock as previously outlined and plated onto LB-agar plates containing 50 µg/mL kanamycin for overnight incubation at 37 °C. Starter cultures of 50 mL TB with 50 µg/mL kanamycin were inoculated with selected colonies from the LB-agar plates, and incubated at 37 °C and 220 rpm overnight.

Cultures of 1 L TB with 50 µg/mL kanamycin were inoculated with 20 mL of starter culture. Cells were grown at 37 °C, 220 rpm to an optical density at 600 nm (OD<sub>600</sub>) of 0.4. The temperature was then reduced to 18 °C and cells were allowed to grow to OD<sub>600</sub> = 0.8. Expression was induced by the addition of 1 mM IPTG (Sigma Aldrich) and 2 g/L L-arabinose (Melford, Ipswich, UK). After

16 hours incubation at 18 °C and 220 rpm, cells were harvested by centrifugation at 6,238 x g for 1 hour. Pellets were stored at -80 °C.

### **7.2.3.2 Purification**

All columns were purchased from GE Healthcare (Chicago, USA) unless otherwise stated. Purifications were carried out using an ÄKTApurifier UPC 10 (GE Healthcare). Purity was checked after each step by SDS PAGE analysis using NuPAGE® Novex 12% Bis-Tris gels (Thermo Fisher), using SeeBlue™ 2 Pre-stained Protein Standard (Invitrogen) and stained with expedeon InstantBlue™ Protein Stain (Sigma Aldrich).

Pellets were resuspended in Buffer A (50 mM HEPES, pH 8, 500 mM NaCl, 20 mM imidazole, 0.5 mM TCEP, 10% v/v glycerol) with EDTA-free protease inhibitor (Roche, Basel, Switzerland), 50 mg lysozyme (Sigma Aldrich) and 0.1% Triton X100 (Sigma Aldrich), and lysed by sonication. A 1:2000 dilution of benzonase (EMD Merck Millepore) was added before centrifugation at 53343 x g for 30 minutes. The soluble fraction was loaded onto a 5 mL HisTrap FF column equilibrated with Buffer A, and eluted with a Buffer B (50 mM HEPES, pH 8, 500 mM NaCl, 1 M imidazole, 0.5 mM TCEP, 10% v/v glycerol) gradient from 0-100% in 50 minutes, at 2 mL/min. Fractions containing IDH1 were pooled and 2% w/w TEV protease added to cleave the His<sub>6</sub>-Tag. The pooled fractions with TEV was dialysed overnight against a buffer of 50 mM HEPES, pH 8, 150 mM NaCl, 0.5 mM TCEP, 4 °C. Following this, imidazole and NaCl were added to a final concentration of 20 mM and 500 mM respectively. The sample was then loaded onto a HisTrap FF column equilibrated with Buffer A

and the flow through collected. Purified fractions were selected following SDS-PAGE analysis and pooled.

Pooled fractions were concentrated using an Amicon® Ultra 10 kDa concentrator (Merck Millipore) and loaded on to a Superdex200 16/60 size exclusion column equilibrated with 20 mM Tris, pH 7.5, 100 mM NaCl. Final purity was confirmed by SDS-PAGE. IDH1 was concentrated using an Amicon® Ultra 10 kDa concentrator (Merck Millipore) to 13.1 mg/mL and stored in SEC buffer at -80 °C. Mass Spectrometry confirmed that the protein mass was as expected from the sequence.

The optimised IDH1-R132H purification included an additional anion exchange step prior to SEC. HisTrap FF fractions were dialysed overnight against 50 mM HEPES, pH 8, 5 mM NaCl and subsequently loaded onto a ResourceQ ion exchange column equilibrated with 99.5% IEXA (50 mM HEPES pH 8) and 0.5% IEXB (50 mM HEPES, pH 8, 1 M NaCl) at 2 mL/min and the flow through collected. SDS-PAGE analysis allowed selection of fractions containing IDH1 but without the contaminant, which were subsequently pooled and subjected to size exclusion.

#### **7.2.3.3 Protein Mass Spectrometry**

Protein samples were submitted to Meirion Richards and Katia Grira, ICR, for intact LC-MS to confirm the molecular weight (Appendix 8.2.3). LC-MS CHROMASOLV solvents, formic acid, or alternative eluent modifiers were purchased from Sigma Aldrich unless otherwise stated.

A custom 8 step, 0.2  $\mu$ L injection program with water, methanol and acetonitrile washes of the samples were made onto a Security Guard C8 column cartridge (4 x 3 mm, AJO-4290, Phenomenex, Torrance, USA). The samples were refrigerated at 4 °C in a G1367B auto-sampler with G1330B thermostat module prior to injection. QuickShot chromatographic separation at 60°C was carried out using a 1200 Series HPLC (Agilent) over a 1 minute gradient elution. Sample was loaded onto the column cartridge using a G1312A binary pump dispensing a gradient from 95:5 to 10:90 water and acetonitrile (both modified with 0.1% v/v formic acid) at a flow rate of 3 mL/min. Between 0.3 and 0.6 minutes a ten port column selection valve (G1316A column module) was used to reverse eluent flow through the column cartridge. During this stage, a second binary pump (G1312B SL) was used to elute protein off the cartridge using a gradient from 60:40 to 10:90 water and acetonitrile (both modified with 0.1% formic acid) at a flow rate of 0.5 mL/min. The post column eluent flow was infused into a 6520 Series qToF mass spectrometer (G6520A) fitted with a dual ESI ionisation source (Agilent, Santa Clara, USA). LC eluent and nebulising gas was introduced into the grounded nebuliser with spray direction orthogonal to the capillary axis. The aerosol was dried by heated gas (10 L/min of nitrogen at 350 °C, 50 psi), producing ions by ESI. Ions entered the transfer capillary along which a potential difference of 4 kV was applied. The fragmentor voltage was set at 190 V and skimmer at 65 V. The signal was optimised by AutoTune.m. Profile mass spectrometry data was acquired in positive ionisation mode over a scan range of m/z 650-2000 (scan rate 1.0) with reference mass correction at m/z 922.009798 hexakis(1H,1H,3H-perfluoropropoxy)phosphazene. Raw data was processed using Agilent

MassHunter Qualitative Analysis B.07.00. Masses calculated were consistent with the expected weight from the sequence.

#### **7.2.4 Protein characterisation by label-free thermal shift**

The ability of IDH1 variants to bind to co-factors NADP<sup>+</sup> (IDH1-WT) or NADPH (IDH1-R132H variants) was shown by native thermal shift using Prometheus NT.48 (NanoTemper, München, Germany). IDH1 variants at 25 µM was mixed with co-factor at concentrations between 250 µM and 2 mM, in a buffer containing 75 mM HEPES pH 7.5, 100 mM NaCl and 1% v/v DMSO, and 10 µL added to nanoDSF Grade Standard Capillaries (NanoTemper). The temperature was increased from 20 °C to 95 °C over 75 minutes, and fluorescence measured at 330 nm and 350 nm.

IDH1 stabilisation by substrates isocitrate for IDH1-WT and αKG for IDH1-R132H was measured in the same way as above, using the same conditions and concentrations.

#### **7.2.5 SYPRO Orange thermal shift assays**

SYPRO Orange TSA experiments were completed using FrameStar 384-well PCR plates (4titude, Surrey, UK). Assay buffer was formed of 75 mM HEPES pH 7.5, 100 mM NaCl unless otherwise stated, and with varying DMSO concentrations as noted in individual sections. Plates were centrifuged at 172 x g after each addition to ensure proper mixing. The fluorescence signal was measured on a C1000 Thermal Cycler CFX384 Real-Time System (Bio-Rad, Hertfordshire, UK) between 10 °C and 95 °C in 0.5 °C steps. Data was analysed using Vortex (dotmatics, Hertfordshire, U.K.).

#### **7.2.5.1 TSA to show ligand binding**

To each well, 1  $\mu$ L ligand at fivefold the final concentration was added. Following this, 4  $\mu$ L IDH1-R132H at 12.5  $\mu$ M and 12.5 X SYPRO Orange (Sigma-Aldrich) with 2% v/v DMSO final subsequently added. NADPH and NADP<sup>+</sup> were tested at concentrations between 0 and 1 mM. Substrates isocitrate and  $\alpha$ KG were tested at concentrations between 0 and 1 mM, both in the presence and absence of 500  $\mu$ M NADP<sup>+</sup>/H. Tool compounds AGI-5918 and GSK-864 were tested at concentrations between 0 and 100  $\mu$ M, also in the presence and absence of 500  $\mu$ M NADPH for IDH1-R132H. GSK-864 was tested at against IDH1-WT at concentrations between 0 and 100  $\mu$ M, also in the presence and absence of 500  $\mu$ M NADP<sup>+</sup>.

#### **7.2.5.2 Fragment screen by SYPRO Orange TSA**

The primary fragment screen was performed with 7  $\mu$ M IDH1-R132H together with 500  $\mu$ M NADPH. Fragments were screened at 300  $\mu$ M, with 15 nL of fragments from a 100 mM stock added to wells and backfilled to 100 nL. A multichannel pipette was used to add 2.5  $\mu$ L of IDH1-R132H at 14  $\mu$ M with NADPH at 1 mM to the plate, which was incubated at room temperature for 30 minutes. Following this, 2.5  $\mu$ L of 20 X SYPRO Orange was added. Final assay conditions were 7  $\mu$ M IDH1-R132H, 500  $\mu$ M NADPH, 300  $\mu$ M fragment, 10 X SYPRO Orange.

Confirmation experiments were conducted in the same way, but the fragments were pre-dispensed to give 100  $\mu$ M, 300  $\mu$ M and 450  $\mu$ M final fragment concentrations in 2% v/v DMSO, and the IDH1-R132H was added in buffer with or without NADPH.



For investigation of IDH1-WT stabilisation, the assay was set up in the same way as described above, except for the 2.5  $\mu$ L addition of IDH1-R132H with NADPH, which was replaced with 2.5  $\mu$ L of IDH1-WT at 20  $\mu$ M with 1 mM NADP<sup>+</sup>.

### **7.2.5.3 Additive screen by SYPRO Orange TSA**

The 96 additives from the Solubility and Stability Screen (HR2-072, Hampton Research, Aliso Viejo, USA) were screened at a 1 in 10 dilution of the stock as recommended. To each well, 500 nL of additive was added with 2  $\mu$ L IDH1-R132H at 17.5  $\mu$ M with 1.25 mM NADPH using a multi-channel pipette. Following 30 minutes incubation at room temperature, 2.5  $\mu$ L of 20 X SYPRO Orange was added using a multi-channel pipette. Final conditions were 7  $\mu$ M IDH1-R132H with 500  $\mu$ M NADPH and 10 X SYPRO Orange. The additive concentration varied between 2 mM and 250 mM, or 0.5% and 15% v/v, depending on the initial formulation.

## **7.2.6 Crystallisation**

### **7.2.6.1 IDH1-R132H variants**

Crystals were grown using the sitting drop method in 96-well, 3-drop SwissCi plates (Molecular Dimensions, Newmarket, UK) with 30  $\mu$ L reservoir solution at 12 °C. Well buffer containing 100 mM Tris pH 7, 24.5% w/v PEG5000MME and 220 mM ammonium sulphate was dispensed using a Phoenix liquid handler (Art Robbins, Sunnyvale, USA). IDH1-R132H at 13.1 mg/mL was pre-incubated with 5 mM NADPH on ice for 4 hours in 20 mM Tris pH 7.5, 100 mM NaCl. Drops were formed of 0.15  $\mu$ L IDH1-R132H with NADPH and 0.15  $\mu$ L reservoir buffer and were made using a Mosquito (TTP Labtech, Hertfordshire, UK).

#### **7.2.6.2 IDH1-WT**

Crystals were grown using the sitting drop method in 96-well, 3-drop SwissCi plates (Molecular Dimensions) with 30  $\mu$ L buffer at 12  $^{\circ}$ C. Well buffer containing 100 mM BisTris pH 6.4, 21.5% w/v PEG5000MME and 220 mM ammonium sulphate was dispensed using a Phoenix liquid handler (Art Robbins). IDH1-WT at 12.9 mg/mL was pre-incubated with 5 mM NADP<sup>+</sup> on ice for 4 hours in 20 mM Tris pH 7.5, 100 mM NaCl. IDH1-R132H seeds were made by collection of IDH1-R132H drops containing crystals grown as described above. This was then transferred to a Seed Bead<sup>™</sup> (Hampton Research) microcentrifuge tube and vortexed to generate a seed stock, which was then diluted using reservoir buffer to a 1/10 and 1/100 stock. Drops were formed of 0.15  $\mu$ L IDH1-WT with NADP<sup>+</sup> with 0.15  $\mu$ L reservoir buffer and 25 nL of IDH1-R132H seeds, and were made using a Mosquito (TTP Labtech).

#### **7.2.6.3 IDH1-R132H crystals for fragment screening**

Crystals were grown in 96-well, 3-drop SwissCi plates with 30  $\mu$ L buffer at 12  $^{\circ}$ C using sitting drop vapour diffusion method. Well buffer contained 100 mM Tris pH 6.9, 26% w/v PEG5000MME and 220 mM ammonium sulphate and was dispensed using a Hydra liquid handler (Art Robbins). IDH1-R132H at 13 mg/mL was pre-incubated on ice with 5 mM NADPH and 0.1 mM TCEP for 4 hours in 20 mM Tris pH 7.5, 100 mM NaCl. Drops were formed of 0.15  $\mu$ L IDH1-R132H with NADPH and 0.15  $\mu$ L reservoir buffer and were made using a Mosquito (TTP Labtech).

### **7.2.7 Fragment soaking experiments with TSA fragment hits**

Fragment powders were dissolved in 100% v/v DMSO to 500 mM stocks, except for CCT242817, which was made to 250 mM. Fragment stocks were diluted with well buffer to 40% v/v DMSO final, and 0.1  $\mu$ L was added to 0.3  $\mu$ L crystal drops and incubated for one hour at room temperature before harvesting.

### **7.2.8 Data collection, processing and structure solution**

Crystals were cryo-protected using Perfluoropolyether Cryo Oil (Hampton Research). Data sets were collected at the Diamond Light Source (Oxford, UK) on the beamlines indicated. Reflections were integrated using XDS<sup>203</sup> or Xia2<sup>204</sup>, and scaled and merged with Aimless<sup>205</sup>. Molecular Replacement was carried out using PHASER<sup>206</sup> using the models indicated in Table 7.1. The structures were refined in BUSTER<sup>176, 207</sup> alternated with manual building rounds using COOT<sup>208</sup>. Ligand restraints were generated using Grade<sup>176</sup> and CSDS<sup>209</sup>.

IDH1 Variant / soak	Beamline	MR model
IDH1-R132H	I03	PDB 4UMY <sup>156</sup> with buffer and ligand molecules removed
IDH1- I112A-R132H	I03	In house IDH1-R132H structure with removal of buffer and ligand molecules, and residues 110 – 125.
IDH1-R132H-R338A	I03	In house IDH1-R132H structure with buffer and ligand molecules removed
IDH1-R132H-R338A	I03	In house IDH1-R132H structure with buffer and ligand molecules removed
IDH1-WT	I04-1	PDB 1T09 with buffer and ligand molecules removed
IDH1-R132H:CCT239686	I03	In house IDH1-R132H structure with removal of buffer and ligand molecules, and residues 110 – 125.
IDH1-R132H: CCT242817	I03	In house IDH1-R132H structure with buffer and ligand molecules removed
IDH1-R132H:CCT242544	I03	In house IDH1-R132H structure with removal of buffer and ligand molecules, and residues 110 – 125.

Table 7.1: Search models used for molecular replacement

### 7.2.9 XChem Crystallography based fragment screen

Crystals were harvested after two weeks growth. Plates were imaged using a ROCK IMAGER 1000 (FORMULATRIX, Bedford, USA) and drops ranked using TeXRan<sup>210</sup>. The 768 fragments from the DSi-Poised (DSiP) library were dispensed directly into crystallisation drops at a final concentration of 50 mM in 10% v/v DMSO with an ECHO acoustic dispenser (Labcyte). Plates were subsequently incubated at room temperature for 1 hour before harvesting using a Shifter (Oxford Lab Technologies Ltd., Oxford, UK). No cryo-protectant was used.

The XChem fragment screen is completed in three stages. The pre-run is used to investigate the robustness of the crystal system as well as DMSO tolerance to optimise soaking conditions and the requirements for cryo-protection.

During the pre-run, a total of 158 unique fragments were dispensed, 151 from DSiP as well as 7 TSA hit fragments and 39 soaked with DMSO only, with 222 datasets collected. In the second run, 316 fragments from the DSiP library and 2 TSA hits, CCT242635 and CCT239559, were soaked, and 474 datasets were collected. During the final visit, 214 unique fragments from DSiP were dispensed and 214 datasets collected. Dr Matthew Rodrigues (ICR) harvested crystals during the final visit. Data was collected un-attended on I04-1 at the Diamond Light Source. The pre-run utilised grid scanning to centre the crystals to the beam, but subsequent runs used loop centring to increase the throughput. Some datasets were collected twice, where loop centring failed.

Diamond auto-processing results from XDS<sup>203</sup> and Xia2<sup>204</sup> were imported into XChemExplorer<sup>164</sup>, and the best dataset was selected by XChemExplorer based on Rmerge, completeness,  $I/\sigma(I)$  and  $CC_{1/2}$ . Although the  $P4_32_12$  space group was pre-specified, the data was sometimes solved in  $P4_12_12$ , and could not be corrected even with re-processing within XChemExplorer. Structures were solved by molecular replacement using Dimple. The search model was an in-house IDH1-R132H structure with NADPH retained. Chlorine and sulphate ions that are found forming the same interactions across multiple IDH1-R132H structures were also retained. Ligand restraints were generated using Grade<sup>176</sup> and CSDS<sup>209</sup>, or aceDRG<sup>178</sup>.

For the PanDDA<sup>164</sup> analysis, a ground state map was optimised using 198 high resolution IDH1-R132H structures from the XChem screen experiment that did not have a fragment bound and did not significantly deviate from the MR model. A ground state model was refined to fit the ground state map using

Refmac<sup>165</sup> within the XChemExplorer interface and used to re-solve the soaked IDH1-R132H datasets.

PanDDA<sup>162</sup> was then used to identify bound fragments based on the deviation of the individual datasets from the ground state. Datasets are aligned in real space and the ground state is subtracted from each dataset. The deviation from the ground state at each voxel is calculated as a Z-score and visualised using a Z-map. Datasets with Z-scores greater than  $\pm 3$  are automatically reported, with subsequent de-convolution of  $2mF_o - DF_c$  maps to generate an event map. Fragments are modelled into the event map using PanDDA Inspect and refined using Refmac. A combination the RSCC, RMSD, B-ratio and RSZO/OCC, and the Z- and PanDDA event maps were used to evaluate the fragment density. Auto-processing and PanDDA results from datasets where the PanDDA maps and statistics indicate fragment binding were exported for local refinement as described in Chapter 7.2.7.

#### **7.2.10 NADPH fluorescence assay**

NADPH fluorescence assays used black ProxiPlate-384 Plus F plates (Perkin Elmer). Fluorescence was measured on either an Envision 2103 (Perkin Elmer, Waltham, USA) or PHERAstar FSX (BMG Labtech, Ortenberg, Germany) as stated for each experiment, at excitation/emission wavelengths of 350/460 nm. Buffer was formed of 100 mM HEPES pH 7.5, 100 mM NaCl, 10 mM MgCl<sub>2</sub>, 0.01% v/v Tween20 (Melford) and 0.5 mg/mL BSA (Sigma Aldrich) unless otherwise stated. DMSO concentration varied between experiments and is

noted individually. Plates were centrifuged at 172 x g after each addition to ensure proper mixing.

#### **7.2.10.1 Calibration Curve**

NADPH calibration curves were generated by diluting a 50 mM NADPH stock into a 12-point, three-fold concentration curve starting at 10 mM, and 10 µL of each concentration was added to wells in a in triplicate. Fluorescence was measured on Envision 2103 (Perkin Elmer). After the PHERAstar FSX (BMG Labtech) plate reader became available, this experiment was repeated and fluorescence measurements taken on both plate readers for comparison.

#### **7.2.10.2 IDH1-R132H assays**

##### **7.2.10.2.1 NADPH $K_m$**

For NADPH  $K_m$  measurement, 5 µL IDH1-R132H at 40 nM and αKG at 10 mM was dispensed into wells. The control points were made in the same way, but with a buffer lacking the catalytic  $Mg^{2+}$ . The reaction was initiated with 5 µL of NADPH between 0 and 200 µM. Final assay conditions were 20 nM IDH1-R132H with 5 mM αKG, 1 % v/v DMSO and NADPH concentration varying between 0 and 100 µM. Fluorescence was measured on an Envision 2103 plate reader (Perkin Elmer) every 30 seconds for the first 30 minutes to obtain  $V_0$ . The change in fluorescence was calculated by subtracting each time point from the control point. Kinetic parameters were calculated using GraphPad Prism 7.0a, by fitting to Michaelis-Menten model using equation 7.2.

$$Y = V_{max} * \frac{X}{K_m + X}$$

Equation 7.2

#### **7.2.10.2.2    $\alpha$ KG $K_m$**

For  $\alpha$ KG  $K_m$  measurements, 5  $\mu$ L IDH1-R132H at 40 nM with 100  $\mu$ M NADPH was dispensed into wells. The control points were made in the same way, but with a buffer lacking the catalytic  $Mg^{2+}$ . The reaction was initiated with 5  $\mu$ L of  $\alpha$ KG at concentrations between 0 and 20 mM. Final assay conditions were 20 nM IDH1-R132H with 100  $\mu$ M NADPH and  $\alpha$ KG between 0 and 10 mM, with 1% v/v DMSO. Data was collected and processed as described above.

#### **7.2.10.2.3    $Mg^{2+}$ $K_m$**

For  $Mg^{2+}$   $K_m$  measurements, 5  $\mu$ L IDH1-R132H at 40 nM with 200  $\mu$ M NADPH and  $\alpha$ KG at 10 mM was added to wells. The control points were made with the same IDH1-R132H and NADPH concentration, but without  $\alpha$ KG. The reaction was initiated with 5  $\mu$ L of  $Mg^{2+}$  between 0 and 250 mM. Final assay conditions were 20nM IDH1-R132H with 100  $\mu$ M NADPH and 5mM  $\alpha$ KG, with  $Mg^{2+}$  between 0 and 250 mM. Data was collected and processed as described above.

#### **7.2.10.2.4    DMSO tolerance**

Buffers were made as described in Section 7.2.10, with DMSO concentrations varying between 0% v/v and 5% v/v. These buffers were used to make stocks of IDH1-R132H at 40 nM with NADPH at 150  $\mu$ M, and  $\alpha$ KG at 1.6 mM. For each DMSO concentration, 5  $\mu$ L of IDH1-R132H with NADPH was added to wells in triplicate, and the reaction initiated with 5  $\mu$ L of  $\alpha$ KG with the same DMSO concentration. Data was collected and processed to calculate the  $V_0$  as described above.



#### **7.2.10.2.5 IDH1-R132H titration**

IDH1-R132H concentrations were investigated between 100 nM and 20 nM in the presence of 100  $\mu$ M NADPH and 800  $\mu$ M  $\alpha$ KG and 3% v/v DMSO. Reaction progression was monitored with measurements at 5 minute intervals on an Envision 2103 platereader (Perkin Elmer). The signal window was calculated as the reciprocal of the ratio of initiated to uninitiated reactions.

#### **7.2.10.2.6 NADPH titrations**

NADPH at a concentration of 75  $\mu$ M was incubated with 20 nM IDH1-R132H, 800  $\mu$ M  $\alpha$ KG and 3% v/v DMSO. Reaction progression was monitored with measurements at 5 minute intervals on an Envision 2103 plate reader (Perkin Elmer). Percentage conversion was calculated based on the calibration curve.

#### **7.2.10.2.7 Fragment IC50 measurements**

Fragments were dispensed into wells in a 10pt-concentration curve with twofold dilution starting at 3 mM. Subsequently, 5  $\mu$ L of IDH1-R132H at 40 nM with 150  $\mu$ M NADPH were added to the plates, which were then incubated for 30 minutes at room temperature. The reaction was then initiated with addition of 5  $\mu$ L  $\alpha$ KG at 1.6 mM. Final reaction conditions were 20 nM IDH1-R132H, 75  $\mu$ M NADPH (30-fold  $K_m$ ) and 800  $\mu$ M  $\alpha$ KG (equal to  $K_m$ ) and 3% v/v DMSO. Plates were incubated at room temperature for 45 minutes before reading on a PHERAstar FSX plate reader (BMG Labtech). The  $Z'$  value were routinely calculated to be between 0.7 and 0.9 by Studies (dotmatics) as shown in equation 7.3, with an average signal to background ratio of 3.

$$Z' = 1 - \frac{3(\sigma_p + \sigma_n)}{|\mu_p - \mu_n|} \quad \text{Equation 7.3}$$

where  $\sigma$  is the standard deviation,  $\mu$  is the mean,  $p$  is the positive control and  $n$  is the negative control.

The IC<sub>50</sub> values were also calculated by studies Studies (dotmatics) as shown in equation 7.4. The IC<sub>50</sub> values were normalised to the positive and negative controls, for which the respective uninitiated and uninhibited reactions were used. The minimum and maximum values were constrained to 0% and 100% respectively where appropriate.

Fragments were tested at 3 mM, 1.5 mM, 750  $\mu$ M, 375  $\mu$ M, 200  $\mu$ M, 100  $\mu$ M, 50  $\mu$ M, 25  $\mu$ M, 12  $\mu$ M and 6  $\mu$ M, and wells were backfilled to 300 nL.

$$Y = Bottom + \frac{(Top - Bottom)}{1 + \left( \frac{X^{HillSlope}}{IC50^{HillSlope}} \right)} \quad \text{Equation 7.4}$$

where the top is the maximal response (approximately 100%) and the bottom in the minimum response (approximately 0%).

#### 7.2.10.2.8 Fragment Interference

Fragments were dispensed into wells in the same concentration points as for the IC<sub>50</sub> measurements. Subsequently, 10  $\mu$ L of IDH1-R132H at 20 nM with 30  $\mu$ M NADPH were added to measure the impact of fragments on fluorescence at low signal, mimicking the end point of the reaction. Fluorescence was measured using a PHERAstar FSX plate reader (BMG Labtech). Concentrations for each fragment that showed greater than 110% of the control

signal were considered auto-fluorescent interferers and excluded during subsequent IC<sub>50</sub> calculations.

Interference was also measured in the presence of 75 µM NADPH to identify quenchers. Concentrations for each fragment that showed less than 90% of the control signal were considered quenchers and excluded during subsequent IC<sub>50</sub> calculations.

### **7.2.10.3 IDH1-WT fluorescence assay**

#### **7.2.10.3.1 Isocitrate K<sub>m</sub> measurements**

A 100 mM isocitrate stock was diluted into a 10-point, twofold dilution series starting at 1 mM, and 5 µL of each concentration was added to wells. The reaction was initiated with 5 µL of IDH1-WT at 1 nM with 1 mM NADP<sup>+</sup>. Final assay conditions were 0.5 nM IDH1-WT with 500 µM NADPH, 1 % v/v DMSO and isocitrate between 0 and 1 mM. Fluorescence measurements were taken every 30 seconds for 20 minutes on a PHERAstar plate reader (BMG Labtech). Kinetic parameters were calculated as described above.

#### **7.2.10.3.2 NADP<sup>+</sup> K<sub>m</sub> measurements**

A 50 mM NADP<sup>+</sup> stock was diluted into a 10-point, twofold dilution series starting at 1 mM, and 5 µL of each concentration was added to wells. The reaction was initiated with 5 µL of IDH1-WT at 1 nM with 500 µM isocitrate. Final assay conditions were 0.5 nM IDH1-WT with 250 µM isocitrate and 1% v/v DMSO, and NADP<sup>+</sup> between 0 and 1 mM. Data was collected and processed as described above.

#### **7.2.10.3.3 Fragment IC<sub>50</sub> measurements.**

Data was collected and measured in the same way as described for the IDH1-R132H fragment IC<sub>50</sub> measurements, with the following variations. The concentration of IDH1-WT used was 0.05 nM, with 500 µM NADP<sup>+</sup> and 5 µM isocitrate, with a 10 µL reaction volume. To each well, 5 µL of IDH1-WT at 0.1 nM and NADP<sup>+</sup> at 1 mM was added, and the plate incubated for 30 minutes. The reaction was initiated with the addition of 5 µL of isocitrate at 10 µM. Plates were incubated at room temperature for 45 minutes before reading.

Final reaction conditions were 0.05 nM IDH1-WT, 500 µM NADP<sup>+</sup> (30-fold K<sub>m</sub>) and 5 µM isocitrate (equal to K<sub>m</sub>) and 3% v/v DMSO.

#### **7.2.10.3.4 Fragment interference assay**

Data was collected and measured in the same way as described for the IDH1-R132H fragment interference measurements, with the following variations. To investigate quenchers, 10 µL of IDH1-WT at 0.1 nM with 490 µM NADP<sup>+</sup> and 10 µM NADPH and 3% v/v DMSO was added to mimic the end point of the reaction. Concentrations for each fragment that showed less than 90% of the control signal were considered quenchers and excluded during subsequent IC<sub>50</sub> calculations.

To investigate auto-fluorescent interferers, 10 µL of IDH1-WT at 0.1 nM with 500 µM NADP<sup>+</sup> and 3% v/v DMSO was added to fragments to mimic the start point of the reaction. Concentrations for each fragment that greater than 110% of the control signal were considered auto-fluorescent interferers and excluded during IC<sub>50</sub> calculations.

### **7.2.11 Kinetic characterisation of IDH1-R132H double mutants**

Kinetic parameters for NADPH and  $\alpha$ KG were collected in the same way for IDH1-R132H-R338A and IDH1-R132H-R338T as for the single mutant IDH1-R132H (see section 7.2.7.2.1 and 7.2.7.2.2). Corresponding kinetic parameters for IDH1-I112A-R132H were collected with 100 nM enzyme instead of 20 nM, but all other conditions were maintained the same.

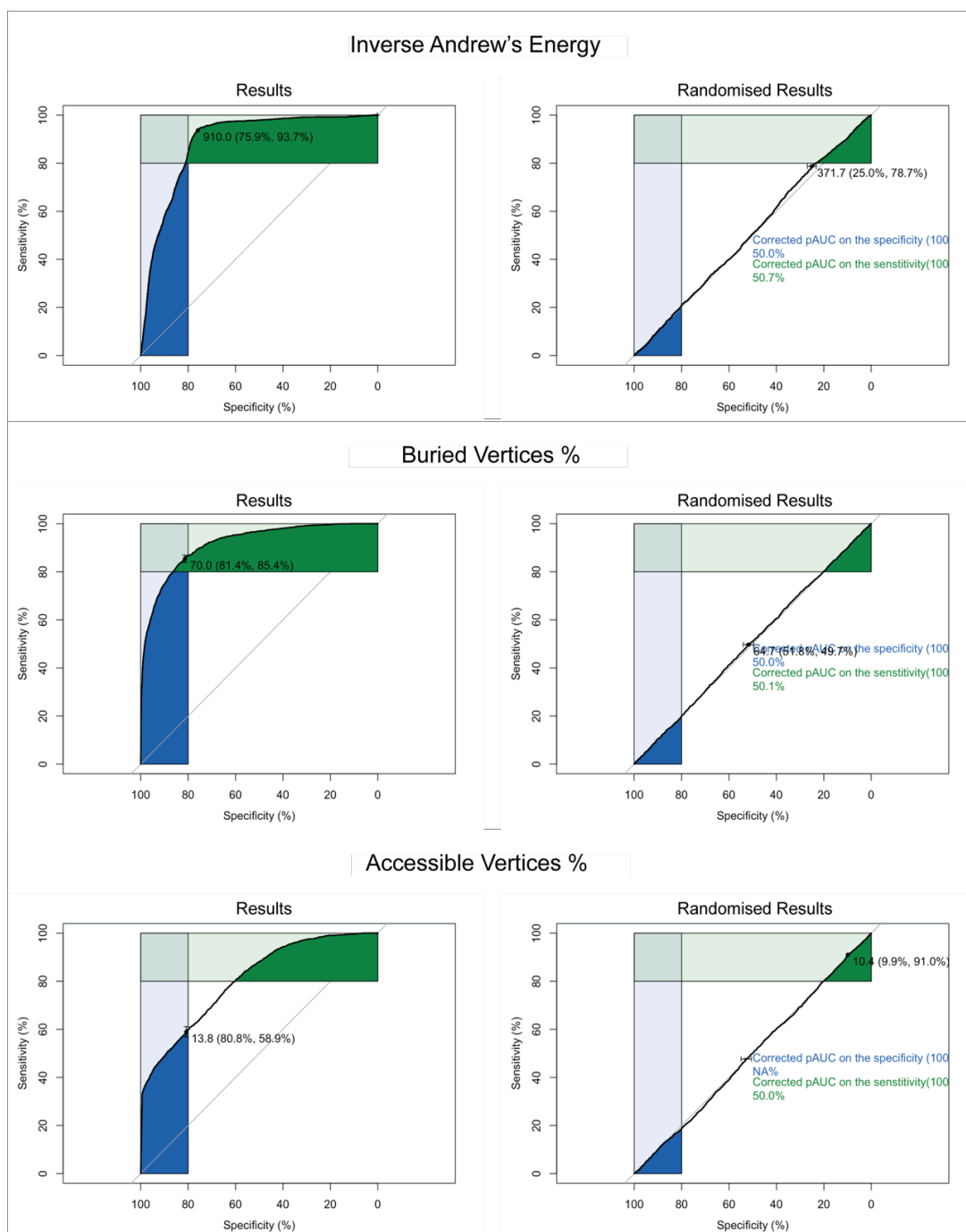
### **7.2.12 Compound Mass spectrometry**

LC/MS analysis was performed on an Agilent 1200 series HPLC and diode array detector coupled to a 6210 time of flight mass spectrometer with dual multimode APCI/ESI source. Analytical separation was carried out at 40 °C on a Merck Chromolith Flash column (RP-18e, 25 x 2 mm) using a flow rate of 1.5 mL/min in a 2 minute gradient elution with detection at 254 nm. The mobile phase was a mixture of methanol (solvent A) and water (solvent B), both containing formic acid at 0.1% v/v. Gradient elution was as follows: 5:95 (A/B) to 100:0 (A/B) over 1.25 min, 100:0 (A/B) for 0.5 min, and then reversion back to 5:95 (A/B) over 0.05 min, finally 5:95 (A/B) for 0.2 minutes.

## Chapter 8: Appendix

### 8.1 *in silico* analysis

#### 8.1.1 Roc curves and randomised Roc curves for properties identified as being significant



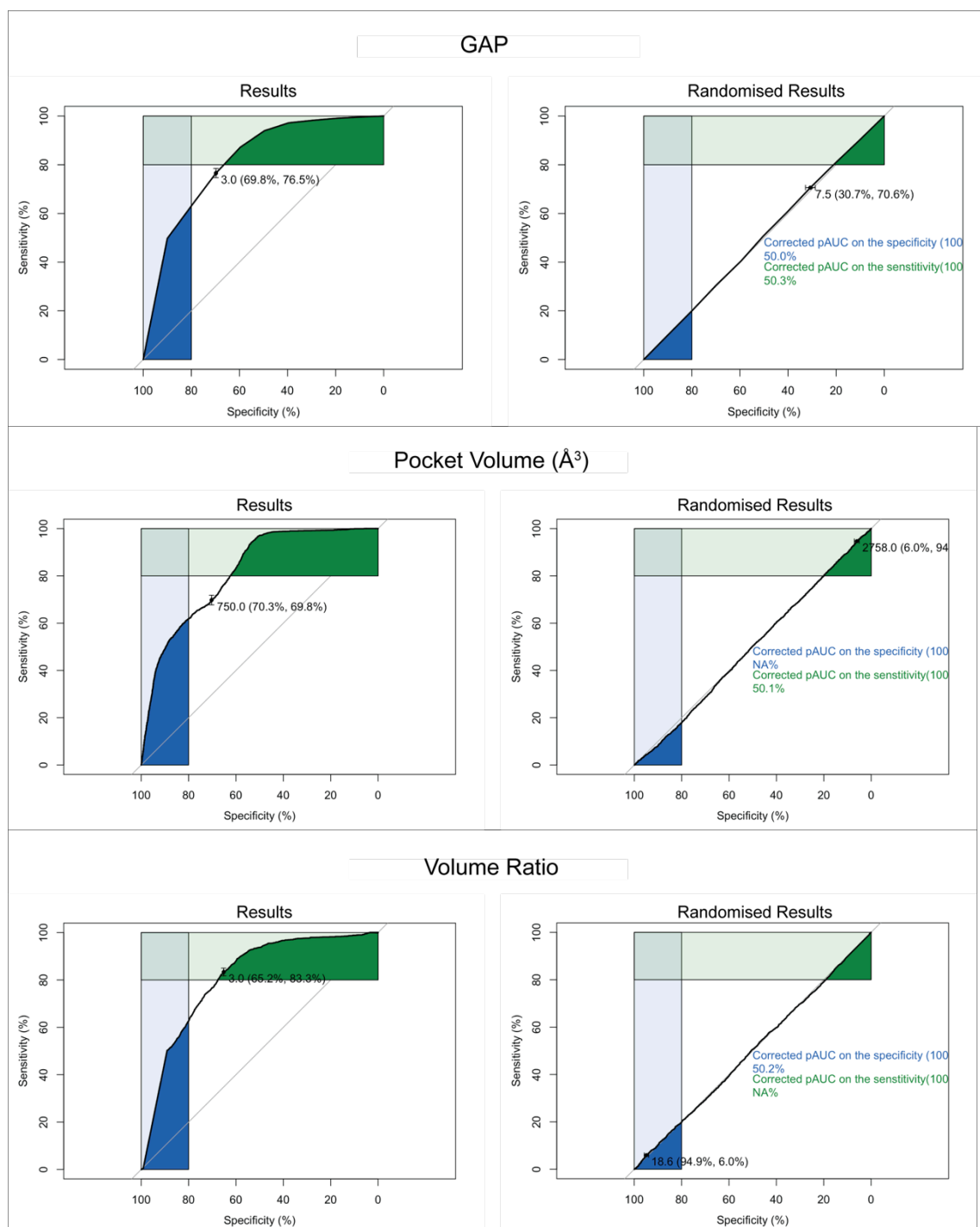


Figure A.1: ROC curves of properties identified as showing a statistically significant difference between test and training sets and used to build the ligandability predictor. ROC curves from randomisation shown the right. Plots made in R using pRoc.

## 8.1.2 Summary statistics for pocket properties

Property	Welch's T-test P-value	KS test P-value	ROC test AUC %
ACC_BUR_VERT_RATIO	0.96	0.39	54.35
ACC_RATIO	0.024	0.12	68.39
ACC_VERTICES	0.00009	0.0091	80.4
ANDREWS_ENERGY	0.014	0	88.94
BETA_SHEET	0.33	0.23	59.08
BUR_VERTICES	0	0	91.71
GAP	0.00018	0.0089	81.35
HP_RATIO	0.096	0.0121	79.86
HB_ACCEPTOR	0.29	0.57	64.27
HB_DONOR	0.31	0.81	63.45
HP_BOTH	0.31	0.81	62.05
HELIX	0.054	0.07	71.71
HOT_FRACTION	0.37	0.59	62.16
LONG_AXIS	0.53	0.0086	59.71
LOOP	0.28	0.55	61.91
MAX_DEPTH	0.28	0	65.56
MEAN_AXIS	0.74	0.0008	62.03
NRM_HYD_RATIO	0.69	0.72	56.56
NRM_POLAR_RATIO	0.00062	0.023	77.64
PCA_X	0.016	0.0373	76.44
PCA_Y	0.0061	0.0059	80.33
PCA_Z	0.00514	0.0035	81.44
POCKET_SIZE	0.059	0.0012	81.61
TEMP_RATING	0.79	0.71	61.54
TURN	0.29	0.46	65.23
VOL_RATIO	0.028	0	80.97

Table A.1: Summary of statistics for all investigated pocket properties. Properties highlighted in blue were used to define the ligandability profile



### 8.1.3 Pockets identified with ligandable secondary sites

Protein	Family	Average Resolution	Expression System	Druggable snapshots	Total number of structures	Mutation enrichment	Sequence conservation	Literature evidence and comments
ABL1	TYR KINASE	2	<i>Spodoptera frugiperda</i>	1	50	2.02	0.7	Myristoyl binding site - known secondary site
ABL2	TYR KINASE	2	<i>Spodoptera frugiperda</i>	1	9	0.49	1	No literature evidence.
BCL2	Bcl	22	<i>E. coli</i>	1	22	0.75	0.87	No literature evidence.
BLM	HELICASE	2.7	<i>E. coli</i>	1	8	3.22	0.76	No literature evidence. Close to DNA binding site; may be construct artefact
BRAF	SER/THR KINASE	3	<i>Spodoptera frugiperda</i> / <i>E. coli</i>	3	59	0.39	1	No literature evidence. Proximal to active site; pocket definition covers V600E
CARD11	CARD	1.79	<i>E. coli</i>	1	2	1.13	0.91	No literature evidence. May be Bcl10 interaction site
CBL	RING E3 LIGASE	2.1	<i>E. coli</i>	2	21	1.57	0.88	No literature evidence. Proximal to peptide binding site
CBL	RING E3 LIGASE	2.1	<i>E. coli</i>	1	21	0.47	1	No literature evidence. Proximal to peptide binding site
DDB2	DNA damage binding	2.4	<i>Trichoplusia ni</i>	3	4	1.63	1	No literature evidence.
EGFR	TYR KINASE	2.5	<i>Spodoptera frugiperda</i>	17	144	1.07	0.81	No literature evidence. Directly opposite active site; may be construct artefact
ESR1	NUCLEAR RECEPTOR	2.2	<i>E. coli</i>	50	193	1.68	0.96	No literature evidence. Close to ligand binding site; not AF-2. Also predicted druggable in other NR3
EZR	PEPTIDE BINDING	2	<i>E. coli</i>	1	4	0.34	0.72	No literature evidence.
FGFR2	RECEPTOR PROTEIN TYROSINE KINASE	2.3	<i>E. coli</i>	1	41	0.57	1	No literature evidence.
FHIT	PHOSPHATASE	2.2	<i>E. coli</i>	4	8	0.94	1	No literature evidence.
FLT3	RECEPTOR PROTEIN TYROSINE KINASE	3	<i>Trichoplusia ni</i>	1	7	0.81	0.87	No literature evidence.
FNBP1	MEMBRANE BINDING	2.6	None (cell free)	1	1	0.91	0.83	No literature evidence.
HRAS	GTPASE	1.7	<i>E. coli</i>	21	135	0.92	0.43	No literature evidence.
IDH1	DEHYDROGENASE	2	<i>E. coli</i>	1	18	1.54	0.82	Known secondary site. Changes ligation of catalytic Mg ion.
IDH1	DEHYDROGENASE	2	<i>E. coli</i>	1	18	0.70	0.93	Literature evidence: potential regulatory site. Changes conformation between active and inactive
IKBKB	SER/THR KINASE	2.6	<i>Spodoptera frugiperda</i>	1	5	1.66	0.92	No literature evidence.

Protein	Family	Average Resolution	Expression System	Druggable snapshots	Total number of structures	Mutation enrichment	Sequence conservation	Literature evidence and comments
ITK	TYR KINASE	1.6	<i>Spodoptera frugiperda</i>	2	9	0.56	0.94	Known secondary site
KEAP1	E3 LIGASE ADAPTOR	2.2	<i>E. coli</i>	5	35	1.20	0.95	No literature evidence.
KRAS	GTPASE	1.4	<i>E. coli</i>	33	41	0.20	0.94	Known secondary site
MAP2K1	SER/THR KINASE	2.2	<i>E. coli</i>	1	39	0.55	1	Known secondary site
MAP2K4	SER/THR KINASE	2.7	<i>E. coli</i>	1	3	0.60	1	No literature evidence.
MTOR	PI3/PI4 KINASE	3	<i>E. coli</i>	2	19	0.69	0.92	No literature evidence.
MYO5A	ATP-BINDING	2	<i>E. coli</i>	1	5	0.57	0.71	No literature evidence.
PIK3CA	PI3/PI4 KINASE	2.7	<i>Spodoptera frugiperda</i>	3	21	0.19	0.83	No literature evidence. Some structures of PIK3CA show pocket occluded by His-tag. Suggests peptide binding function.
PIK3CG	PI3/PI4 KINASE	2.7	<i>Spodoptera frugiperda</i>	24	89	0.87	0.96	No literature evidence, but same pocket as PIK3CA
PIM1	SER/THR KINASE	2.3	<i>E. coli</i>	1	119	1.32	0.87	No literature evidence. Proximal to primary site, includes DFG loop
RET	RECEPTOR PROTEIN TYROSINE KINASE	2.5	<i>Cricetulus griseus</i>	1	15	0.12	0.76	No literature evidence.
RUNX1	TRANSCRIPTION FACTOR	2.6	<i>E. coli</i>	1	5	0.53	1	No literature evidence.
SMAD4	DNA BINDING	2.5	<i>E. coli</i>	2	2	0.71	1	No literature evidence.
SPOP	BTB-POZ ADAPTOR	2	<i>E. coli</i>	1	14	0.32	0.94	No literature evidence. Cluster = 1 - can this be considered enrichment
STAT6	TRANSCRIPTION FACTOR	2.7	<i>E. coli</i>	1	6	0.71	0.95	No literature evidence.
TP53	TRANSCRIPTION FACTOR	2.5	<i>E. coli</i>	1	162	0.90	1	Involved in aggregation. Formed by steric zipper region required for aggregation
TRIM33	E3 LIGASE	2.7	<i>E. coli</i>	5	5	0.24	0.86	No literature evidence.
VHL	E3 LIGASE ADAPTOR	2.3	<i>E. coli</i>	5	25	1.01	1	No literature evidence. Interface of CUL2, ElonginC and VHL
WIF1	RECEPTOR BINDING	2	<i>H. Sapiens</i>	1	5	1.54	1	No literature evidence.

Table A.2: Summary of 56 cancer-associated targets with novel secondary sites predicted to be ligandable. Validated secondary sites are highlighted in blue, while those shortlisted for experimental investigation are highlighted in red.

### 8.1.4 Roc curves for sequence conservation

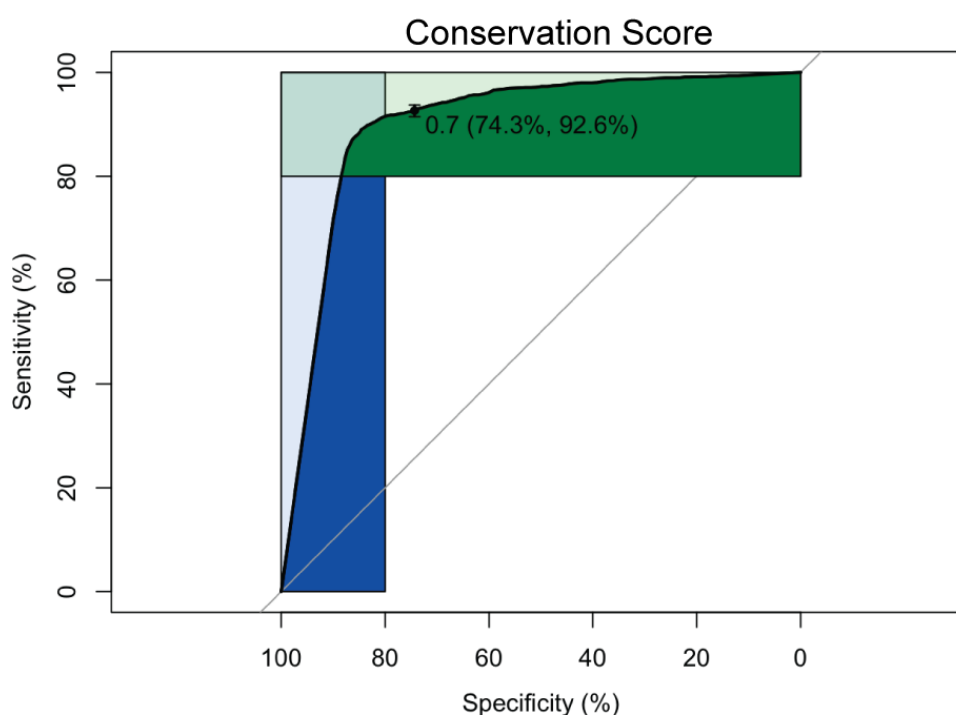


Figure A.2: Roc curve for conservation score as calculated the canSAR3D pipeline. The initial cut-off identified was at 0.9. However, the primary site is known to show high levels of sequence conservation, with functional secondary sites thought to show lower levels of conservation. A threshold of 0.7 (70%) was selected. Plot made in R using pROC.

### 8.1.5 Publically available IDH1 structures by variant and conformation

Construct	Number of chains		
	Active	Inactive	Inhibited
Wild type	5 (7)	2 (8)	0 (20)
R132H	9 (10)	4 (4)	12 (35)
G97N	6	0	0
G97D	3	0	0
Y139D	6	0	0

Table A.3: Breakdown of crystal structures deposited in the PDB when triaging was completed. There are now 95 chains from 35 crystal structures available in the PDB; the updated breakdown is in parenthesis, showing the recent interest in inhibition of the IDH1-R132H mutant protein. The novel secondary site in IDH1-R132H was predicted to be ligandable in one chain of the inactive conformation, representing 25% of the available structures.

## 8.1.6 Sequence conservation of mammalian IDH enzymes

	1	11	21	31
Consensus	M S k K I q G G S V	V E M Q G D E M T R	I I W E L I K E K L	I I P Y V E L D L H
Conservation				
IDHC_HUMAN	M S K K I S G G S V	V E M Q G D E M T R	I I W E L I K E K L	I F P Y V E L D L H
IDHC_PONAB	M S K K I S G G S V	V E M Q G D E M T R	I I W E L I K E K L	I F P Y V E L D L H
IDHC_RAT	M S R K I H G G S V	V E M Q G D E M T R	I I W E L I K E K L	I L P Y V E L D L H
IDHC_MICOH	M S K K I H G G S V	V E M Q G D E M T R	I I W E L I K E K L	I L P Y V E L D L H
IDHC_MICME	M S K K I H G G S V	V E M Q G D E M T R	I I W E L I K E K L	I L P Y V E L D L H
IDHC_MOUSE	M S R K I Q G G S V	V E M Q G D E M T R	I I W E L I K E K L	I L P Y V E L D L H
IDHC_BOVIN	M S Q K I Q G G S V	V E M Q G D E M T R	I I W E L I K E K L	I F P Y V E L D L H
IDHC_SHEEP	M S H K I Q G G S V	V E M Q G D E M T R	I I W E L I K E K L	I F P Y V D L D L H
	41	51	61	71
Consensus	S Y D L G I E N R D	A T N D Q V T K D A	A E A I K K Y N V G	V K C A T I T P D E
Conservation				
IDHC_HUMAN	S Y D L G I E N R D	A T N D Q V T K D A	A E A I K K H N V G	V K C A T I T P D E
IDHC_PONAB	S Y D L G I E N R D	A T N D Q V T K D A	A E A I K K Y N V G	V K C A T I T P D E
IDHC_RAT	S Y D L G I E N R D	A T N D Q V T K D A	A E A I K K Y N V G	V K C A T I T P D E
IDHC_MICOH	S Y D L G I E N R D	A T N D Q V T K D A	A E A I K K Y N V G	V K C A T I T P D E
IDHC_MICME	S Y D L G I E N R D	A T N D Q V T K D A	A E A I K K Y N V G	V K C A T I T P D E
IDHC_MOUSE	S Y D L G I E N R D	A T N D Q V T K D A	A E A I K K Y N V G	V K C A T I T P D E
IDHC_BOVIN	S Y D L G I E N R D	A T N D Q V T K D A	A E A I K K Y N V G	V K C A T I T P D E
IDHC_SHEEP	S Y D L S I E N R D	A T N D Q V T K D A	A E A I K K Y N V G	V K C A T I T P D E
	81	91	101	111
Consensus	K R V E E F K L K Q	M W K S P N G T I R	N I L G G T V F R E	A I I C K N I P R L
Conservation				
IDHC_HUMAN	K R V E E F K L K Q	M W K S P N G T I R	N I L G G T V F R E	A I I C K N I P R L
IDHC_PONAB	K R V E E F K L K Q	M W K S P N G T I R	N I L G G T V F R E	A I I C K N I P R L
IDHC_RAT	K R V E E F K L K Q	M W K S P N G T I R	N I L G G T V F R E	A I I C K N I P R L
IDHC_MICOH	K R V E E F K L K Q	M W K S P N G T I R	N I L G G T V F R E	A I I C K N I P R L
IDHC_MICME	K R V E E F K L K Q	M W K S P N G T I R	N I L G G T V F R E	A I I C K N I P R L
IDHC_MOUSE	K R V E E F K L K Q	M W K S P N G T I R	N I L G G T V F R E	A I I C K N I P R L
IDHC_BOVIN	K R V E E F K L K Q	M W K S P N G T I R	N I L G G T V F R E	A I I C K N I P R L
IDHC_SHEEP	K R V E E F K L K Q	M W K S P N G T I R	N I L G G T V F R E	A I I C K N I P R L
	121	131	141	151
Consensus	V s G W V K P I I I	G R H A Y G D Q Y R	A T D F V V P G P G	K V E I t Y T P s D
Conservation				
IDHC_HUMAN	V S G W V K P I I I	G R H A Y G D Q Y R	A T D F V V P G P G	K V E I T Y T P S D
IDHC_PONAB	V S G W V K P I I I	G R H A Y G D Q Y R	A T D F V V P G P G	K V E I T Y T P S D
IDHC_RAT	V T G W V K P I I I	G R H A Y G D Q Y R	A T D F V V P G P G	K V E I T Y T P K D
IDHC_MICOH	V T G W V K P I I I	G R H A Y G D Q Y R	A T D F V V P G P G	K V E I T Y T P K D
IDHC_MICME	V T G W V K P I I I	G R H A Y G D Q Y R	A T D F V V P G P G	K V E I T F T P K D
IDHC_MOUSE	V T G W V K P I I I	G R H A Y G D Q Y R	A T D F V V P G P G	K V E I T Y T P K D
IDHC_BOVIN	V S G W V K P I I I	G R H A Y G D Q Y R	A T D F V V P G P G	K V E I S Y T P S D
IDHC_SHEEP	V S G W V K P I I I	G R H A Y G D Q Y R	A T D F V V P G P G	K V E I C Y T P S D
	161	171	181	191
Consensus	G s q K v t Y L V H	n F e E g G G V A M	G M Y N Q D K S I E	D F A H S S F Q M A
Conservation				
IDHC_HUMAN	G T Q K V T Y L V H	N F E E G G G V A M	G M Y N Q D K S I E	D F A H S S F Q M A
IDHC_PONAB	G T Q K V T Y L V H	N F E E G G G V A M	G M Y N Q D K S I E	D F A H S S F Q M A
IDHC_RAT	G S Q K V T Y L V H	D F E E G G G V A M	G M Y N Q D K S I E	D F A H S S F Q M A
IDHC_MICOH	G S Q K V T Y L V H	S F E E G G G V A M	G M Y N Q D K S I E	D F A H S S F Q M A
IDHC_MICME	G S Q K V T Y L V H	S F E E G G G V A M	G M Y N Q D K S I E	D F A H S S F Q M A
IDHC_MOUSE	G T Q K V T Y M V H	D F E E G G G V A M	G M Y N Q D K S I E	D F A H S S F Q M A
IDHC_BOVIN	G S P K T V Y L V H	N F T E S G G V A M	G M Y N Q D K S I E	D F A H S S F Q M A
IDHC_SHEEP	G S P K T V Y L V H	N F T E S G G V A M	G M F N Q D K S I E	D F A H S S F Q M A

201	211	221	231
Consensus L S K g W P L Y L S	T K N T I L K K Y D	G R F K D I F Q E I	Y D K Q Y K S q F E
Conservation			
IDHC_HUMAN L S K G W P L Y L S	T K N T I L K K Y D	G R F K D I F Q E I	Y D K Q Y K S Q F E
IDHC_PONAB L S K G W P L Y L S	T K N T I L K K Y D	G R F K D I F Q E I	Y D K Q Y K S Q F E
IDHC_RAT L S K G W P L Y L S	T K N T I L K K Y D	G R F K D I F Q E I	Y D K Q Y K S K F E
IDHC_MICOH L S K G W P L Y L S	T K N T I L K K Y D	G R F K D I F Q E I	Y D K Q Y K S Q F E
IDHC_MICME L S K G W P L Y L S	T K N T I L K K Y D	G R F K D I F Q E I	Y D K Q Y K S Q F E
IDHC_MOUSE L S K G W P L Y L S	T K N T I L K K Y D	G R F K D I F Q E I	Y D K Q Y K S Q F E
IDHC_BOVIN L S K N W P L Y L S	T K N T I L K K Y D	G R F K D I F Q E I	Y D K Q Y K S E F E
IDHC_SHEEP L S K N W P L Y L S	T K N T I L K K Y D	G R F K D I F Q E I	Y D K Q Y K S Q F E
241	251	261	271
Consensus A Q k I W Y E H R L	I D D M V A Q A M K	S E G G F I W A C K	N Y D G D V Q S D S
Conservation			
IDHC_HUMAN A Q K I W Y E H R L	I D D M V A Q A M K	S E G G F I W A C K	N Y D G D V Q S D S
IDHC_PONAB A R K I W Y E H R L	I D D M V A Q A M K	S E G G F I W A C K	N Y D G D V Q S D S
IDHC_RAT A Q K I W Y E H R L	I D D M V A Q A M K	S E G G F I W A C K	N Y D G D V Q S D S
IDHC_MICOH A Q K I W Y E H R L	I D D M V A Q A M K	S E G G F I W A C K	N Y D G D V Q S D S
IDHC_MICME A Q K I W Y E H R L	I D D M V A Q A M K	S E G G F I W A C K	N Y D G D V Q S D S
IDHC_MOUSE A Q K I C Y E H R L	I D D M V A Q A M K	S E G G F I W A C K	N Y D G D V Q S D S
IDHC_BOVIN A Q N I W Y E H R L	I D D M V A Q A M K	S E G G F I W A C K	N Y D G D V Q S D S
IDHC_SHEEP A Q N I W Y E H R L	I D D M V A Q A M K	S E G G F I W A C K	N Y D G D V Q S D S
281	291	301	311
Consensus V A Q G Y G S L G M	M T S V L i C P D G	K T V E A E A A H G	T V T R H Y R M y Q
Conservation			
IDHC_HUMAN V A Q G Y G S L G M	M T S V L V C P D G	K T V E A E A A H G	T V T R H Y R M Y Q
IDHC_PONAB V A Q G Y G S L G M	M T S V L V C P D G	K T V E A E A A H G	T V T R H Y R M Y Q
IDHC_RAT V A Q G Y G S L G M	M T S V L I C P D G	K T V E A E A A H G	T V T R H Y R M Y Q
IDHC_MICOH V A Q G Y G S L G M	M T S V L I C P D G	K T V E A E A A H G	T V T R H Y R M H Q
IDHC_MICME V A Q G Y G S L G M	M T S V L I C P D G	K T V E A E A A H G	T V T R H Y R M H Q
IDHC_MOUSE V A Q G Y G S L G M	M T S V L I C P D G	K T V E A E A A H G	T V T R H Y R M Y Q
IDHC_BOVIN V A Q G Y G S L G M	M T S V L V C P D G	K T V E A E A A H G	T V T R H Y R M Y Q
IDHC_SHEEP V A Q G Y G S L G M	M T S V L V C P D G	K T V E A E A A H G	T V T R H Y R M Y Q
321	331	341	351
Consensus K G Q E T S T N P I	A S I F A W s R G L	A H R A k L D N N k	E L s F F A k A L E
Conservation			
IDHC_HUMAN K G Q E T S T N P I	A S I F A W T R G L	A H R A K L D N N K	E L A F F A N A L E
IDHC_PONAB K G Q E T S T N P I	A S I F A W T R G L	A H R A K L D N N K	E L A F F A N A L E
IDHC_RAT K G Q E T S T N P I	A S I F A W S R G L	A H R A K L D N N T	E L S F F A N A L E
IDHC_MICOH K G Q E T S T N P I	A S I F A W S R G L	A H R A R L D N N T	E L S F F A K A L E
IDHC_MICME K G Q E T S T N P I	A S I F A W S R G L	A H R A R L D N N T	E L S F F A K A L E
IDHC_MOUSE K G Q E T S T N P I	A S I F A W S R G L	A H R A K L D N N T	E L S F F A K A L E
IDHC_BOVIN K G Q E T L T N P I	A S I F A W T R G L	A H R A K L D N N K	E L S F F A K A L E
IDHC_SHEEP K G Q E T S T N P I	A S I F A W T R G L	A H R A K L D N N K	E L S F F A K A L E
361	371	381	391
Consensus E V c I E T I E A G	F M T K D L A A C I	K G L P N V Q R S D	Y L N T F E F M D K
Conservation			
IDHC_HUMAN E V S I E T I E A G	F M T K D L A A C I	K G L P N V Q R S D	Y L N T F E F M D K
IDHC_PONAB E V S V E T I E A G	F M T K D L A A C I	K G L P N V Q R S D	Y L N T F E F M D K
IDHC_RAT E V C I E T I E A G	F M T K D L A A C I	K G L P N V Q R S D	Y L N T F E F M D K
IDHC_MICOH E V C I E T I E A G	F M T K D L A A C I	K G L P N V Q R S D	Y L N T F E F M D K
IDHC_MICME E V C I E T I E A G	F M T K D L A A C I	K G L P N V Q R S D	Y L N T F E F M D K
IDHC_MOUSE D V C I E T I E A G	F M T K D L A A C I	K G L P N V Q R S D	Y L N T F E F M D K
IDHC_BOVIN E V C I E T I E A G	F M T K D L A A C I	K G L P N V Q R S D	Y L N T F E F M D K
IDHC_SHEEP E V C I E T I E A G	F M T K D L A A C I	K G L P N V Q R S D	Y L N T F E F M D K
401	411		
Consensus L G E N L k a K L A	Q A K L		
Conservation			
IDHC_HUMAN L G E N L K I K L A	Q A K L		
IDHC_PONAB L G E N L K I K L A	Q A K L		
IDHC_RAT L G E N L K A K L A	Q A K L		
IDHC_MICOH L G E N L K A K L A	Q A K L		
IDHC_MICME L G E N L K A K L A	Q A K L		
IDHC_MOUSE L G E N L K A K L A	Q A K L		
IDHC_BOVIN L G E N L Q L K L A	Q A K L		
IDHC_SHEEP L G E N L Q L K L A	Q A K L		

Figure A.3: IDH1 sequence alignments were generated using curated mammalian homologues in SwissProt<sup>4</sup>. Overall sequence identity between the eight homologues is 90%.

## 8.2 *in vitro* investigation

### 8.2.1 Coding sequences and primers

#### 8.2.1.1 *His<sub>6</sub>-IDH1-WT* sequence

MGHHHHHHSSGVDLGTENLYFQGMSKKISGGSVVEMQGDEMTRIIEWELIKEKLIFPYVELDLH  
SYDLGIENRDATNDQVTKDAAEAIKKHNVGVKCATITPDEKRVEEFKLKQMWKSPNGTIRNILG  
GTVFREAIICKNIPRLVSGWVKPIIIGRHAYGDQYRATDFVVPGPVKVEITYTPSDGTQKVITYLV  
HNFEEGGGVAMGMYNQDKSIEDFAHSSFQMALSKGWPLYLSTKNILKKYDGRFKDIFQEYD  
KQYKSQFEAQKIWYEHRLIDDMVAQAMKSEGGFIWACKNYDGDVQSDSVAQGYGSLGMMTS  
VLVCPDGKTVEAAAHGTVTRHYRMYQKGQETSTNPIASIFAWTRGLAHRAKLDNNKELAFFA  
NALEEVSITIEAGFMTKDLAACIKGLPNVQRSYDLNTFEFMDKLGLENLKIKLAQAKL

Hexa-histidine tag and TEV cleavage site are underlined.

Construct		Primer
R132H	F	5' CCAATCATTATTGGCCATCACGCATACGGCGACCAATACCGC 3'
	R	5' GCGGTATTGGTCGCCGTATGCGTGATGGCCAATAATGATTGG 3'
R338A	F	5'-GCGAGCATTTTTGCATGGACT <u>GCC</u> GGTCTGGCCC-3'
	R	5'-GGGCCAGACCGGCAGTCCATGCAAAAATGCTCGC-3'
R338T	F	5'-GCGAGCATTTTTGCATGGACT <u>ACC</u> GGTCTGGCCC-3'
	R	5'-GGGCCAGACCGGTAGTCCATGCAAAAATGCTCGC-3'
I112A	F	5'-CGTGGAATATTCTTACAAAT <u>GGC</u> AGCTTCGCGGAATACCGTTCC-3'
	R	5'-GCAACGGTATTCCGCGAAGCT <u>GCC</u> ATTTGTAAGAATATTCCACG -3'

Table A.4: Primers used for site directed mutagenesis. Site of mutation is underlined.

### 8.2.2 Expression tests

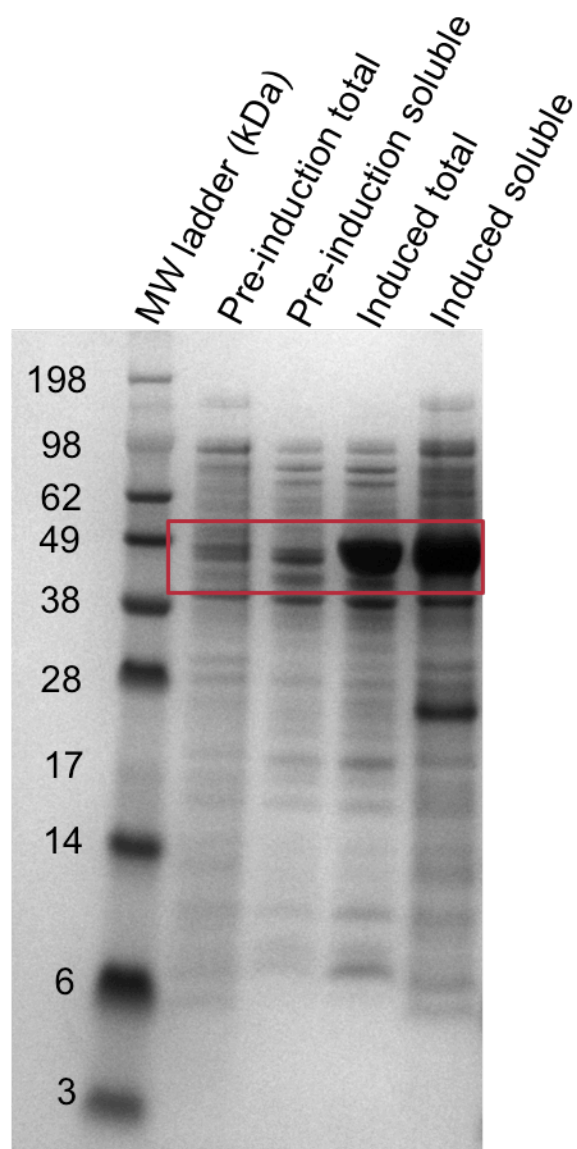


Figure A.4: Representative expression tests of IDH1 variants as described in section 7.2.3.1, but with BugBuster® (Merck millipore) used to lyse samples instead of sonication. Pre-induced samples were taken before the addition of IPTG, while the induced samples were taken following 18 hours expression at 18°C.

## 8.2.3 Mass spectrometry of IDH1 variants

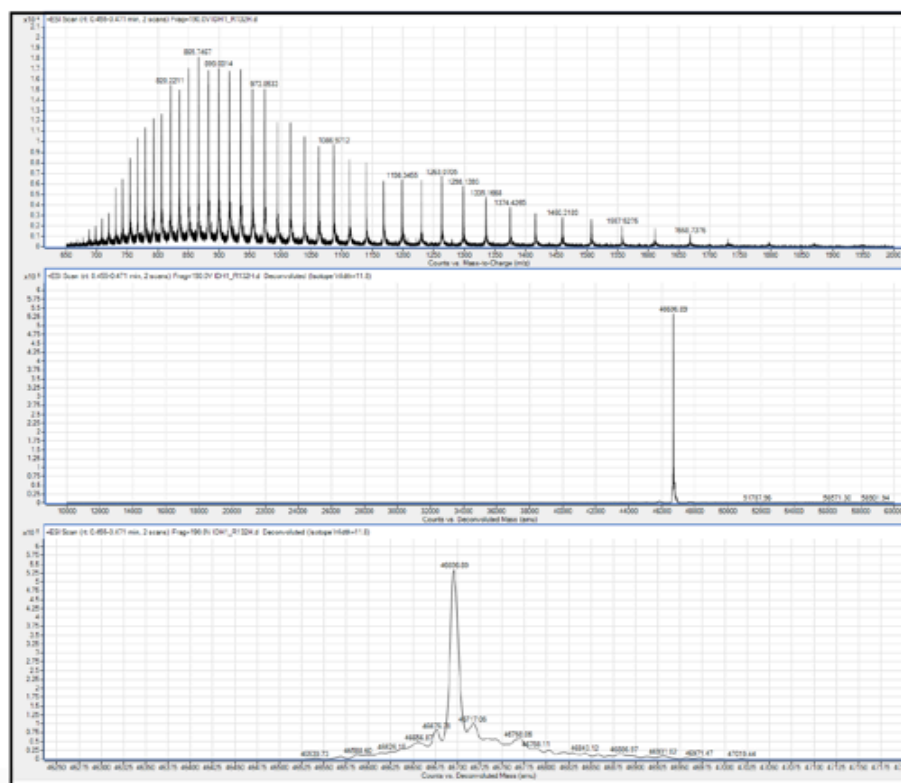


Figure A.5: De-convoluted spectrum of IDH1-R132H; molecular weight of major peak is measured at 46,697 Da, as expected based on the primary sequence.

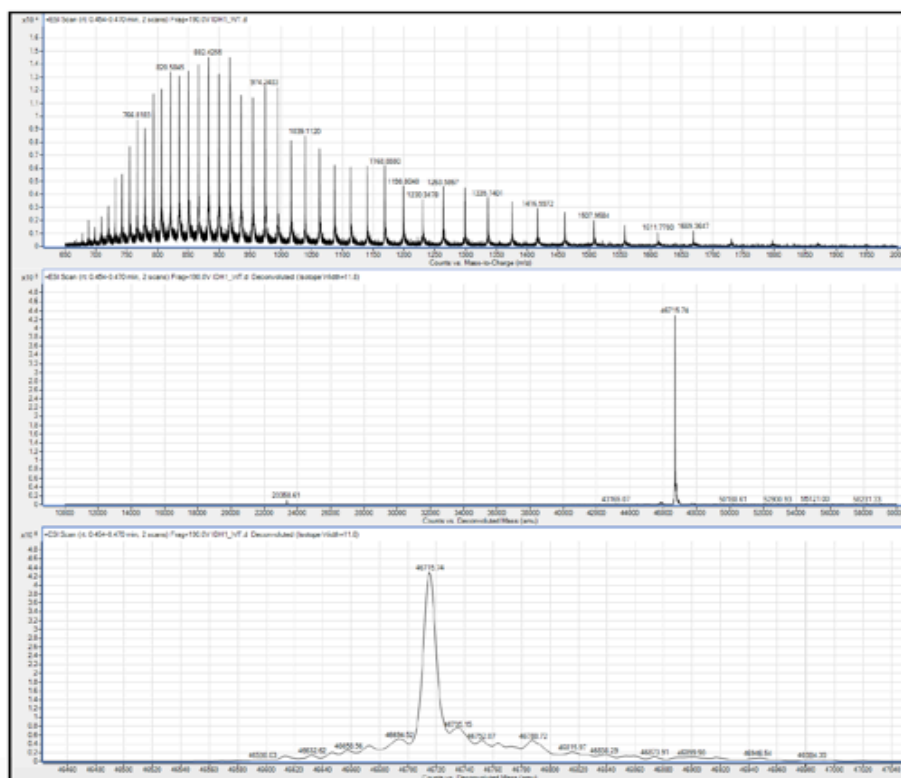
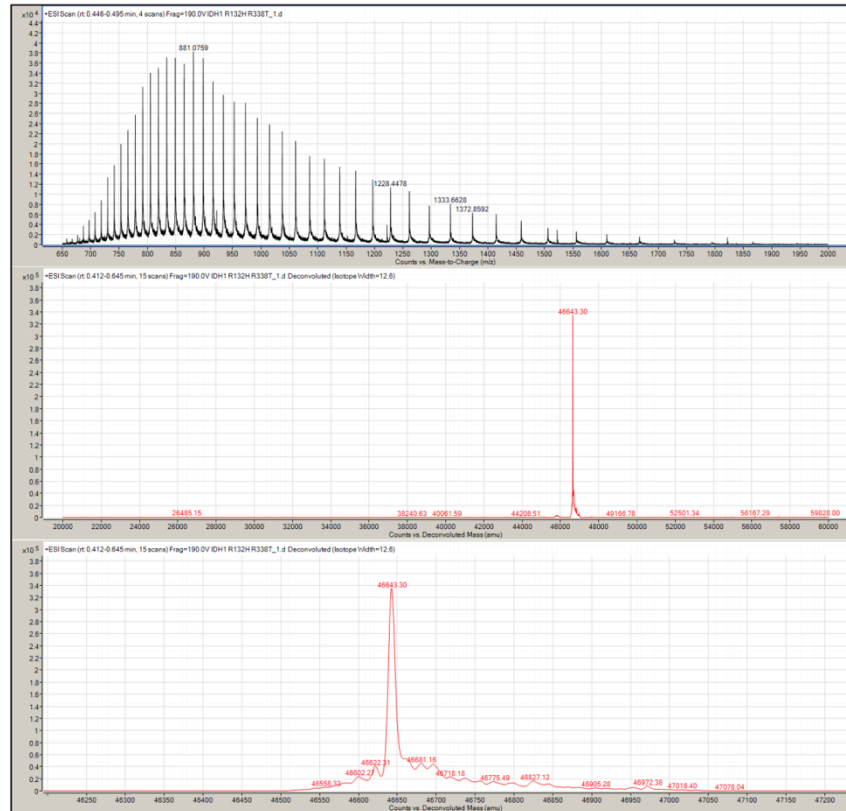


Figure A.6: De-convoluted spectrum of IDH1-WT; molecular weight of major peak is measured at 46,716 Da, as expected based on the primary sequence.





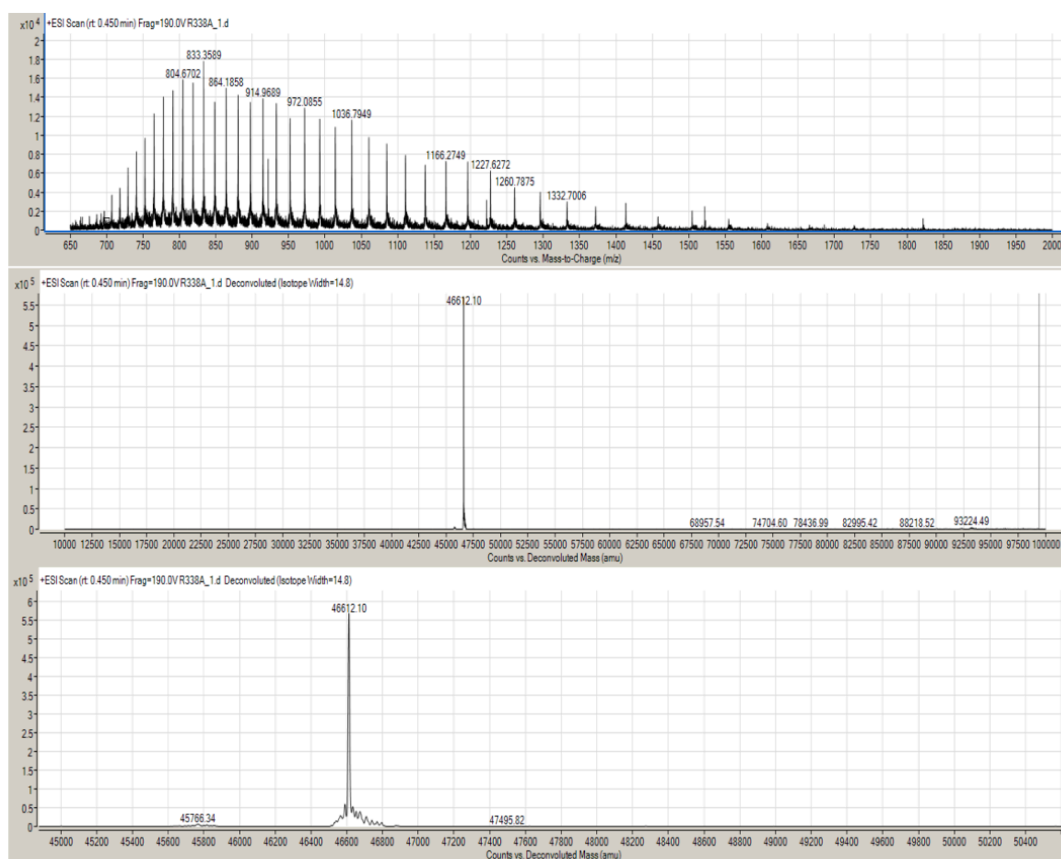


Figure A.9: De-convoluted spectrum of IDH1-R132H-R338A; molecular weight of major peak is measured at 46,612 Da, as expected based on the primary sequence.

## 8.2.4 Label-free TSA using Prometheus for IDH1 variant characterisation

These experiments were run the same way as for the IDH1-R132H variant characterisation, as described in Chapter 3.2.2 and Chapter 7.2.4 .

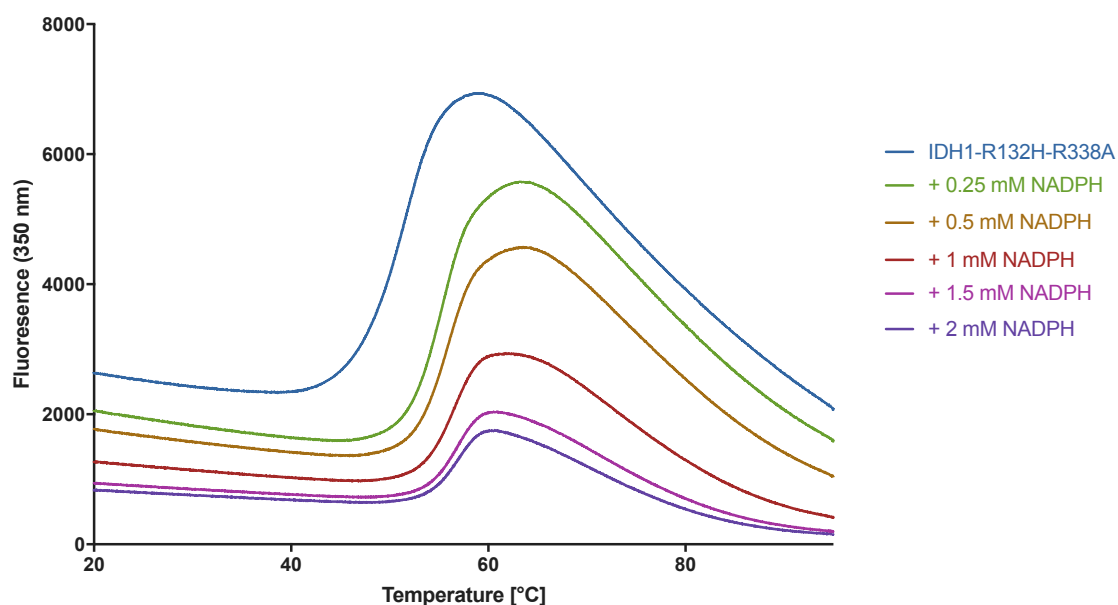


Figure A.11: Native thermal shift of IDH1-R132H-R338A double mutant with increasing concentrations of NADPH. NADPH absorbance is maximum at 350 nm, leading to a decrease in overall fluorescence with increasing NADPH concentrations

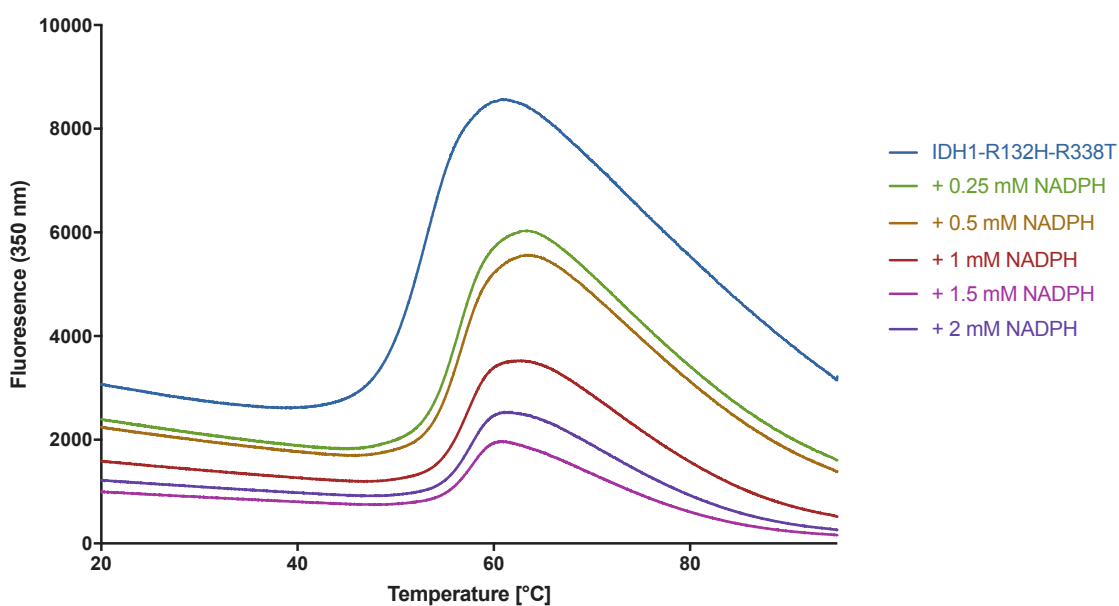


Figure A.10: Native thermal shift of IDH1-R132H-R338T double mutant with increasing concentrations of NADPH. NADPH absorbance is maximum at 350 nm, leading to a decrease in overall fluorescence with increasing NADPH concentrations

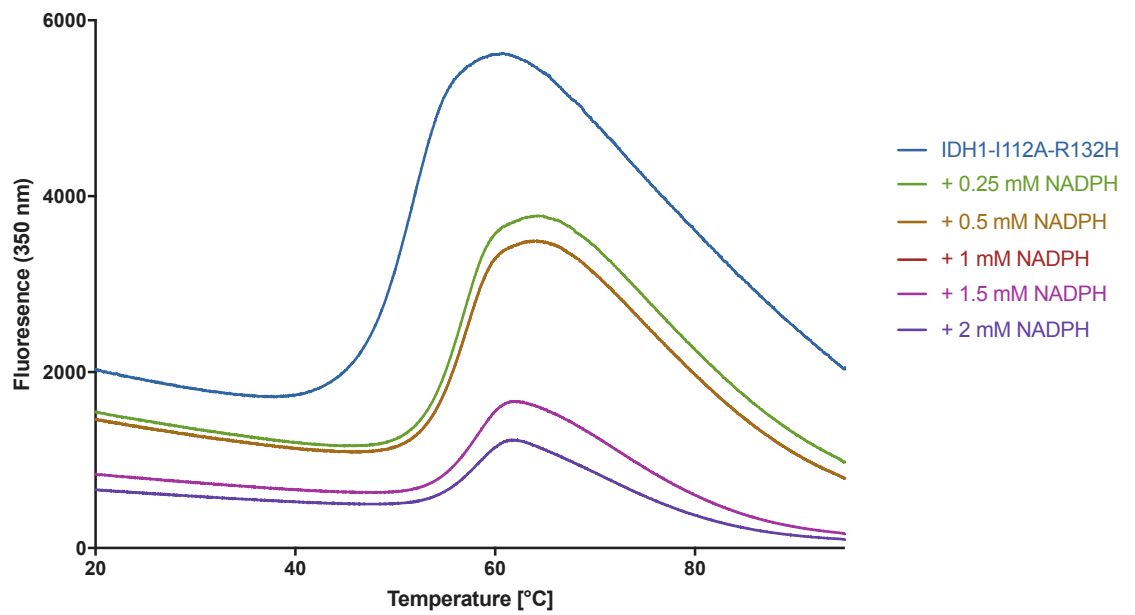


Figure A.12: Native thermal shift of IDH1-I112A-R132H double mutant with increasing concentrations of NADPH. NADPH absorbance is maximum at 350 nm, leading to a decrease in overall fluorescence with increasing NADPH concentrations

## 8.2.5 SYPRO Orange TSA

### 8.2.5.1 Investigation of the impact of salt concentration on IDH1-R132H stability

To investigate the impact of salt concentration on IDH1-R132H for optimisation of ion exchange purification, a KCl,  $(\text{NH}_4)_2\text{SO}_4$  and NaCl at a range of concentrations was investigated by SYPRO Orange TSA. The experiment was carried out as described in Chapter 7.2.5, and as described below. An IDH1-R132H stock at 70  $\mu\text{M}$  with 2 mM NADPH and 100 X SYPRO Orange was made in 100 mM HEPES with 75 mM NaCl. A series of buffers were made with 100 mM HEPES and either KCl,  $(\text{NH}_4)_2\text{SO}_4$  or NaCl at a range of concentrations between 0 and 500 mM. To each well, 0.5  $\mu\text{L}$  of the IDH1-R132H stock was added, followed by 4.5  $\mu\text{L}$  of varying buffer.

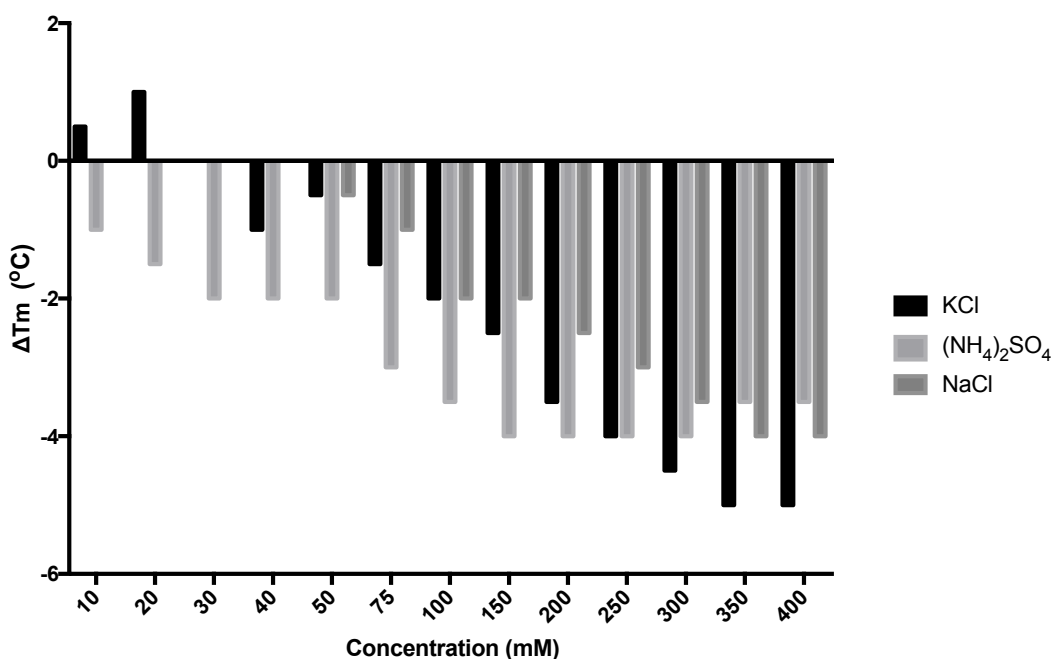


Figure A.13: SYPRO-Orange thermal shift investigation of effect of salt on IDH1-R132H stability. Increasing concentrations of all salts tested resulted in a decrease in protein stability. This observation aided in the design of the anion exchange step during optimisation of IDH1-R132H purification.

### 8.2.5.2 Investigating differences in $\Delta T_m$ for IDH1 variants by $\text{NADP}^+/\text{H}$

The ability of both oxidised and reduced co-factor to stabilise IDH1-WT and IDH1-R132H was also investigated (Section 7.2.4.2). Both variants show larger  $\Delta T_m$  values for NADPH than  $\text{NADP}^+$ . The reason for this is unclear. Despite this, I chose to use  $\text{NADP}^+$  with IDH1-WT, as this is the natural co-factor.

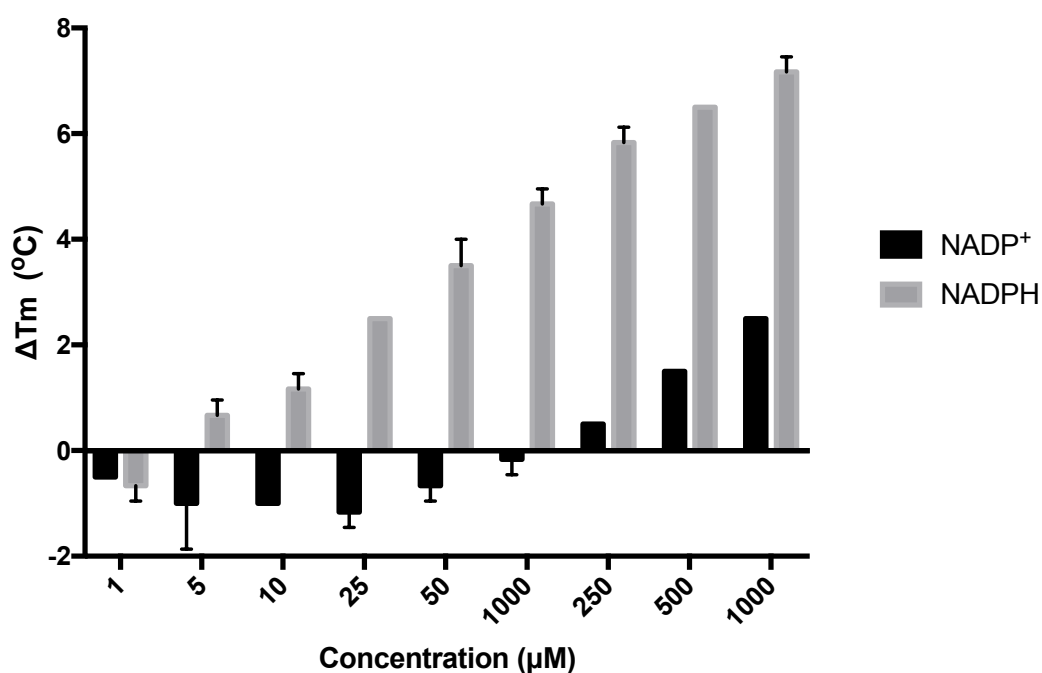


Figure A.14: SYPRO-Orange thermal shift assay investigating stabilisation of IDH1-WT by oxidised and reduced cofactor

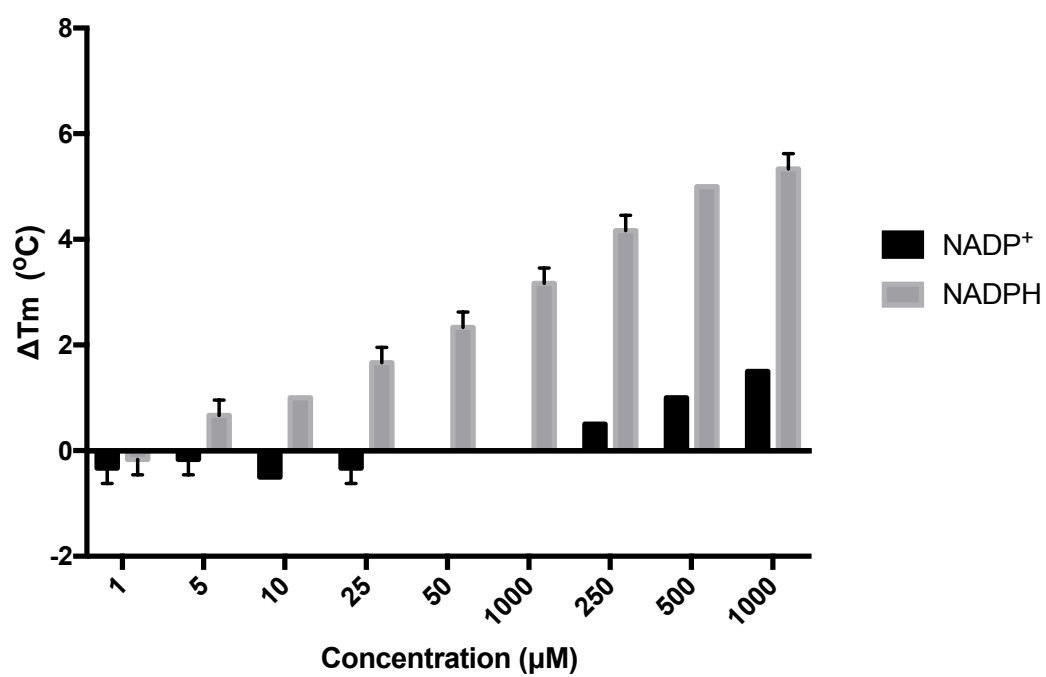


Figure A.15: SYPRO-Orange thermal shift assay investigating stabilisation of IDH1-R132H by oxidised and reduced co-factor

## 8.2.6 Compound mass spectrometry

Selected compounds were analysed by LC/MS according to a general protocol described in Section 7.2.11.

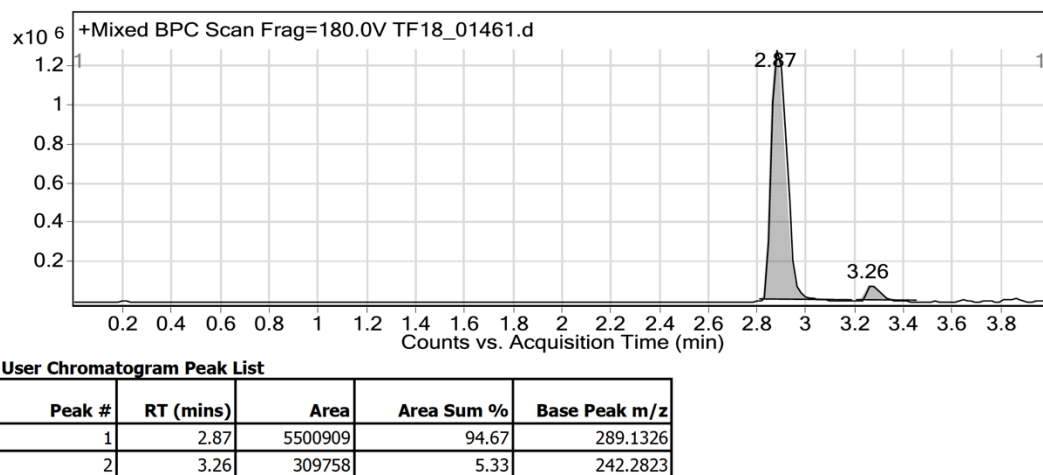


Figure A.16: Mass spectrometry analysis of CCT242817 shows a base peak at 289 that is consistent with the expected mass of the fragment. The lower molecular weight species (242) is too large to correspond to the fluorophenyl-piperazine ring that is observed in the crystal structure.

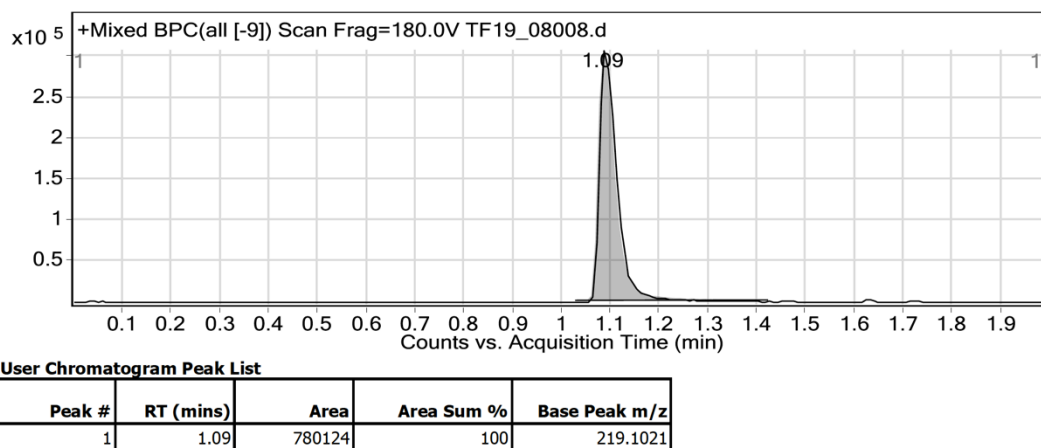


Figure A.17: Mass spectrometry analysis of potent analogue CCT374509, synthesised by Sandra Codony Gisbert and Rosemary Huckvale. This compound is pure, giving confidence in the biochemical IC<sub>50</sub>. Reaction 34



## 8.2.7 Crystallographic refinement statistics tables

Protein construct	IDH1-WT: NADPH	IDH1-R132H: NADPH
<i>Crystal</i>		
Space group	P 4 <sub>3</sub> 2 <sub>1</sub> 2	P 4 <sub>3</sub> 2 <sub>1</sub> 2
Unit cell dimensions (a/b/c in Å)	81.85/81.85/308.79	81.08/81.08/305.09
Unit cell angles (α/β/γ in °)	90.00/90.00/90.00	90.00/90.00/90.00
<i>Data collection and processing</i>		
Beamline	DLS I04-I	DLS I03
Wavelength (Å)	0.92819	0.97625
Integration program	XDS	XDS
Reduction program	AIMLESS	AIMLESS
Resolution range	49.30 – 1.85	48.75 – 1.89
Number of unique reflections <sup>a</sup>	90934 (4410)	82691 (4468)
Completeness <sup>a</sup>	99.2 (100)	99.3 (100)
Redundancy <sup>a</sup>	10.3 (8)	10.7 (9.2)
R <sub>merge</sub> (%) <sup>a</sup>	6.0 (155.3)	9.4 (211.6)
I/σ(I) <sup>a</sup>	60.2 (1.2)	40.6 (1.1)
CC <sub>1/2</sub> <sup>a, b</sup>	1.000 (0.401)	0.999 (0.335)
<i>Refinement</i>		
Program	BUSTER	BUSTER
R <sub>work</sub> (%)	18	16.9
R <sub>free</sub> (%)	20.7	19.8
Number of residues	1631	1554
Number of water molecules	838	749
Average B-factor (Å <sup>2</sup> )	46.22	47.24
Ramachandran favoured (%)	97.8	97.05
Ramachandran outliers (%)	0	0
RMSD bonds (Å)	0.01	0.01
RMSD angles (°)	1.57	1.62

Table A.5: Statistics for NADP+/H-bound IDH1-WT and IDH1-R132H

<sup>a</sup> Values in parentheses are for the highest resolution shell.

<sup>b</sup> Half-dataset correlation coefficient, see: Karplus, P. A.; Diederichs, K. Linking crystallographic model and data quality. Science 2012, 336, 1030–1033

Protein construct	IDH1-R132H: NADPH: CCT242817	IDH1-R132H: NADPH: CCT239686
<i>Crystal</i>		
Space group	P 4 <sub>3</sub> 2 <sub>1</sub> 2	P 4 <sub>3</sub> 2 <sub>1</sub> 2
Unit cell dimensions (a/b/c in Å)	80.73/80.73/306.64	82.88/82.88/305.95
Unit cell angles (α/β/γ in °)	90.00/90.00/90.00	90.00/90.00/90.00
<i>Data collection and processing</i>		
Beamline	DLS I03	DLS I03
Wavelength (Å)	0.9762	0.9762
Integration program	XIA2 DIALS	XIA2 DIALS
Reduction program	AIMLESS	AIMLESS
Resolution range	48.83 – 2.50	49.02 – 2.50
Number of unique reflections <sup>a</sup>	36358 (6523)	37937 (4193)
Completeness <sup>a</sup>	98.7 (100)	99.5 (100)
Redundancy <sup>a</sup>	7.1 (7.2)	20.6 (16.6)
R <sub>merge</sub> (%) <sup>a</sup>	18.3 (199)	22 (226.5)
I/σ(I) <sup>a</sup>	19.4 (1.3)	19.2 (1.7)
CC <sub>1/2</sub> <sup>a, b</sup>	0.997 (0.321)	0.992 (0.639)
<i>Refinement</i>		
Program	BUSTER	BUSTER
R <sub>work</sub> (%)	0.201	0.2
R <sub>free</sub> (%)	0.247	0.25
Number of residues	778	1017
Number of water molecules	103	222
Average B-factor (Å <sup>2</sup> )	71.35	72.24
Ramachandran favoured (%)	96.07	96.22
Ramachandran outliers (%)	0	
RMSD bonds (Å)	0.014	0.014
RMSD angles (°)	1.74	1.86
Fragment RSCC	0.799	0.856

Table A.6: Statistics for NADPH-bound IDH1-R132H in complex with CCT242817 and CCT239686,

<sup>a</sup> Values in parentheses are for the highest resolution shell.

<sup>b</sup> Half-dataset correlation coefficient, see: Karplus, P. A.; Diederichs, K. Linking crystallographic model and data quality. *Science* **2012**, 336, 1030–1033

<sup>c</sup> Fragment RSCC is calculated by MolProbity<sup>5</sup>

Protein construct	IDH1-R132H:NADPH: CCT239544	IDH1-R132H:NADPH: CCT242635
<i>Crystal</i>		
Space group	P 4 <sub>3</sub> 2 <sub>1</sub> 2	P 4 <sub>3</sub> 2 <sub>1</sub> 2
Unit cell dimensions (a/b/c in Å)	80.82/80.82/306.52	82.01/82.01/305.85
Unit cell angles (α/β/γ in °)	90.00/90.00/90.00	90.00/90.00/90.00
<i>Data collection and processing</i>		
Beamline	DLS I03	DLS I04-I
Wavelength (Å)	0.97623	0.92819
Integration program	XIA2 DIALS	DIALS
Reduction program	AIMLESS	AIMLESS
Resolution range	80.82-2.13	49.30 – 2.3
Number of unique reflections <sup>a</sup>	58192 (2692)	61752 (4857)
Completeness <sup>a</sup>	100 (95.8)	100 (100)
Redundancy <sup>a</sup>	12.4 (13)	7.5 (6.8)
R <sub>merge</sub> (%) <sup>a</sup>	0.1471 (0.8691)	5.2 (208.2)
I/σ(I) <sup>a</sup>	9.8 (1.8)	33.7 (1.3)
CC <sub>1/2</sub> <sup>a, b</sup>	0.995 (0.6)	0.989 (0.788)
<i>Refinement</i>		
Program	BUSTER	BUSTER
R <sub>work</sub> (%)	18.9	20.3
R <sub>free</sub> (%)	22.1	23.3
Number of residues	778	785
Number of water molecules	522	373
Average B-factor (Å <sup>2</sup> )	48.75	77.5
Ramachandran favoured (%)	97.04	96.11
Ramachandran outliers (%)	0	0
RMSD bonds (Å)	0.014	0.014
RMSD angles (°)	1.701	1.754
Fragment RSCC <sup>c</sup>	0.779	0.856

Table A.7: Statistics for NADPH-bound IDH1-R132H in complex with CCT239544 and CCT242635

<sup>a</sup> Values in parentheses are for the highest resolution shell.

<sup>b</sup> Half-dataset correlation coefficient, see: Karplus, P. A.; Diederichs, K. Linking crystallographic model and data quality. *Science* **2012**, 336, 1030–1033

<sup>c</sup> Fragment RSCC is calculated by MolProbity<sup>5</sup>

Protein construct	IDH1-R132H: NADPH: CCT370970	IDH1-R132H: NADPH: CCT370971
<i>Crystal</i>		
Space group	P 4 <sub>3</sub> 2 <sub>1</sub> 2	P 4 <sub>3</sub> 2 <sub>1</sub> 2
Unit cell dimensions (a/b/c in Å)	82.59/82.59/306.59	82.62/82.62/305.59
Unit cell angles (α/β/γ in °)	90.00/90.00/90.00	90.00/90.00/90.00
<i>Data collection and processing</i>		
Beamline	DLS I04-I	DLS I04-1
Wavelength (Å)	0.92819	0.91587
Integration program	XDS	DIALS
Reduction program	AIMLESS	AIMLESS
Resolution range	49.23 (2.55)	49.14 (2.50)
Number of unique reflections <sup>a</sup>	35889 (4310)	37902 (4164)
Completeness <sup>a</sup>	100 (100)	100 (100)
Redundancy <sup>a</sup>	13.0 (13.3)	19.0 (19.8)
R <sub>merge</sub> (%) <sup>a</sup>	14.3 (258.6)	11.7 (232.5)
I/σ(I) <sup>a</sup>	11.1 (1.0)	14.7 (1.5)
CC <sub>1/2</sub> <sup>a, b</sup>	0.997 (0.611)	0.999 (0.714)
<i>Refinement</i>		
Program	BUSTER	BUSTER
R <sub>work</sub> (%)	18.9	20.3
R <sub>free</sub> (%)	23.8	24.3
Number of residues	772	772
Number of water molecules	273	257
Average B-factor (Å <sup>2</sup> )	79.5	81.06
Ramachandran favoured (%)	95.85	96.50
Ramachandran outliers (%)	0	0
RMSD bonds (Å)	0.015	0.014
RMSD angles (°)	1.807	1.792
Fragment RSCC <sup>c</sup>	0.73	0.70

Table A.8: Statistics for NADPH-bound IDH1-R132H in complex with CCT370971 and CT370970.

<sup>a</sup> Values in parentheses are for the highest resolution shell.

<sup>b</sup> Half-dataset correlation coefficient, see: Karplus, P. A.; Diederichs, K. Linking crystallographic model and data quality. *Science* **2012**, 336, 1030–1033

<sup>c</sup> Fragment RSCC is calculated by MolProbity<sup>5</sup>

Protein construct	IDH1-R132H: NADPH: CCT371098	IDH1-R132H: NADPH: CCT154567
<i>Crystal</i>		
Space group	P 4 <sub>3</sub> 2 <sub>1</sub> 2	P 4 <sub>3</sub> 2 <sub>1</sub> 2
Unit cell dimensions (a/b/c in Å)	81.47/81.47/305.56	82.38/82.38/299.39
Unit cell angles (α/β/γ in °)	90.00/90.00/90.00	90.00/90.00/90.00
<i>Data collection and processing</i>		
Beamline	DLS I04-1	DLS I04-I
Wavelength (Å)	0.91587	0.9159
Integration program	XDS	XDS
Reduction program	AIMLESS	AIMLESS
Resolution range	48.75 – 2.20	49.90 – 2.75
Number of unique reflections <sup>a</sup>	52034 (4417)	27957 (3945)
Completeness <sup>a</sup>	100 (99.9)	100 (100)
Redundancy <sup>a</sup>	12.7 (12.6)	7.5 (6.8)
R <sub>merge</sub> (%) <sup>a</sup>	15.8 (283.1)	19.0 (268.6)
I/σ(I) <sup>a</sup>	8.6 (0.9)	9.0 (0.9)
CC <sub>1/2</sub> <sup>a, b</sup>	0.998 (0.472)	0.998 (0.541)
<i>Refinement</i>		
Program	BUSTER	BUSTER
R <sub>work</sub> (%)	19.9	20.4
R <sub>free</sub> (%)	23.9	24.8
Number of residues	776	777
Number of water molecules	367	75
Average B-factor (Å <sup>2</sup> )	60.48	80.14
Ramachandran favoured (%)	97.94	96.27
Ramachandran outliers (%)	0	0
RMSD bonds (Å)	0.014	0.014
RMSD angles (°)	1.745	1.835
Fragment RSCC <sup>c</sup>	0.83	0.85

Table A.9: Statistics for IDH1-R132H+NADPH in complex with Trp205 stacking fragments CCT371098 and CCT154567

<sup>a</sup> Values in parentheses are for the highest resolution shell.

<sup>b</sup> Half-dataset correlation coefficient, see: Karplus, P. A.; Diederichs, K. Linking crystallographic model and data quality. *Science* **2012**, 336, 1030–1033

<sup>c</sup> Reported fragment RSCC was calculated by MolProbity.

Protein construct	IDH1-R132H: NADPH: CCT373604
<i>Crystal</i>	
Space group	P 4 <sub>3</sub> 2 <sub>1</sub> 2
Unit cell dimensions (a/b/c in Å)	
Unit cell angles (α/β/γ in °)	90.00/90.00/90.00
<i>Data collection and processing</i>	
Beamline	DLS I04-I
Wavelength (Å)	0.92819
Integration program	XDS
Reduction program	AIMLESS
Resolution range	48.50 - 2.80
Number of unique reflections <sup>a</sup>	25756 (3364)
Completeness <sup>a</sup>	100 (100)
Redundancy <sup>a</sup>	12.8 (12.9)
R <sub>merge</sub> (%) <sup>a</sup>	24.3 (298.6)
I/σ(I) <sup>a</sup>	9.5 (1.0)
CC <sub>1/2</sub> <sup>a, b</sup>	0.997 (0.380)
<i>Refinement</i>	
Program	BUSTER
R <sub>work</sub> (%)	19.1
R <sub>free</sub> (%)	24.1
Number of residues	778
Number of water molecules	136
Average B-factor (Å <sup>2</sup> )	75.07
Ramachandran favoured (%)	96.41
Ramachandran outliers (%)	0
RMSD bonds (Å)	0.014
RMSD angles (°)	1.75
Fragment RSCC <sup>c</sup>	0.72

Table A.10: Statistics for IDH1-R132H+NADPH in complex with Trp205 stacking fragment CCT373604

<sup>a</sup> Values in parentheses are for the highest resolution shell.

<sup>b</sup> Half-dataset correlation coefficient, see: Karplus, P. A.; Diederichs, K. Linking crystallographic model and data quality. *Science* **2012**, 336, 1030–1033

<sup>c</sup> Reported fragment RSCC was calculated by MolProbity.

## 8.2.8 XChem fragment screening

### 8.2.8.1 PanDDA hit statistics for the novel secondary site

Fragment	Resolution	PanDDA maps		Fit to normal maps				Binding Mode
		1-BDC	Z-peak	RSCC	RSZO/OCC	B-factor Ratio	RMSD	
CCT370971	2.42	0.45	4.3	0.68	1.33	1.4	0.19	Singlet
CCT370970	2.47	0.37	4	0.67	0.81	1.5	0.77	Singlet
CCT371095	2.63	0.34	4.7	0.82	0.5	1.28	0.4	Glu361
CCT370974	2.52	0.31	5.6	0.81	0.81	0.96	0.2	Glu361
CCT371098	2.22	0.42	6.4	0.5	0.57	1.24	0.27	Trp205
CCT154567	2.72	0.55	6.3	0.64	0.92	1.0	0.45	Trp205
CCT373604	2.8	0.51	3.8	0.57	1.1	1.02	0.21	Trp205
CCT372954	2.82	0.54	5.4	0.63	0.8	0.9	0.26	Trp205
CCT370974	2.82	0.38	8.6	0.87	0.8	1.81	5.87	Loop
CCT370982	2.38	0.46	5.4	0.58	0.29	0.97	0.37	Loop
CCT370980	2.67	0.4	4.3	0.59	1.2	1.4	2	Loop
CCT370978	2.02	0.44	6.1	0.59	0.38	1.65	1.01	Loop
CCT370979	2.67	0.42	4.8	0.78	0.48	1.1	1.47	Loop
CCT370972	2.22	0.46	4	0.62	1.36	1.31	1.57	Loop

Table A.11: PanDDA statistics for XChem fragments identified binding to the novel secondary site

### 8.2.8.2 PanDDA hits for the known allosteric site

Fragment	Resolution	PanDDA maps		Fit to normal maps				IC <sub>50</sub> (μM)
		1-BDC	Z-peak	RSCC	RSZO/OCC	B-factor Ratio	RMSD	
CCT333387	2.62	0.22	5.2	0.83	0.3	1.4	0.53	640 ± 33
CCT370976	2.47	0.39	5	0.91	0.5	1.33	0.65	2592 ± 280
CCT370981	2.47	0.39	5.1	0.74	0.6	1.5	0.61	1612 ± 118
CCT371096	2.57	0.32	6	0.86	0.2	1.48	0.74	Interferer
CCT371097	2.57	0.32	3.9	0.83	0.25	1.76	0.49	21% at 3mM

Table A.12: PanDDA statistics and measured IC<sub>50</sub> values for XChem fragment hits targeting the known allosteric site. The IC<sub>50</sub> values reported are against IDH1-R132H and are the mean of three biological repeats, with the standard deviation.

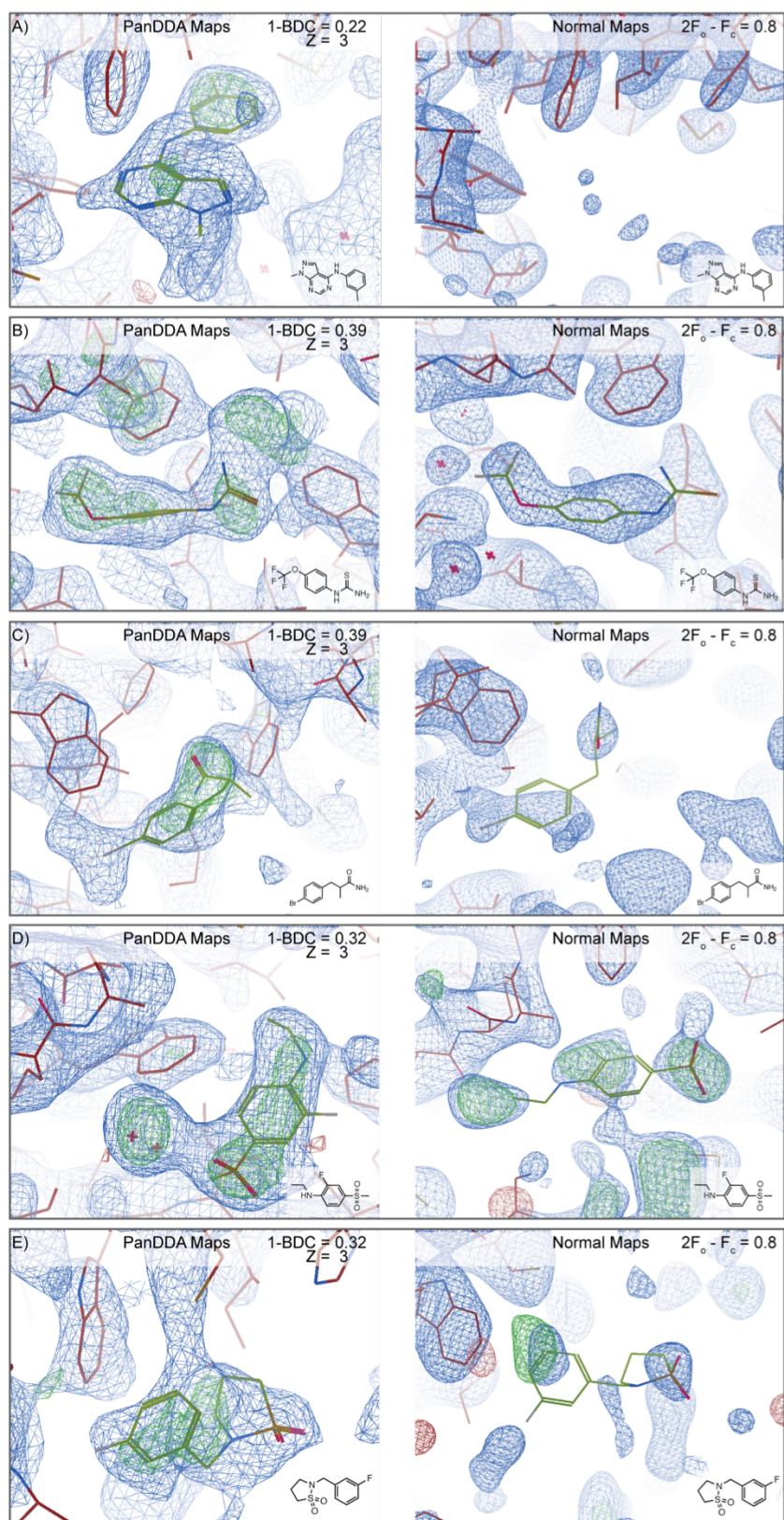


Figure A.18: Comparison of PanDDA maps (left) and  $2F_o - DF_c$  maps (right). A) CCT333387; B) CCT370976; C) CCT370981; D) CCT371096; E) CCT371097. PanDDA maps are contoured to the 1-BDC (absolute)



## 8.3 Introduction to Crystallography

X-ray crystallography is currently the most commonly used technique for protein structure determination<sup>211</sup>. It is the preferred method for investigation of protein-ligand interactions as it provides structural information for both receptor and ligand at near atomic resolution. Solving a protein crystal structures has multiple steps, as outlined below, and in many introductory texts and reviews<sup>3</sup>,

212, 213

### 8.3.1 Crystallisation

Production of well ordered, diffracting crystals is a pre-requisite for protein crystallography, and remains a major bottleneck<sup>214</sup>. Vapour diffusion is one of the main techniques used to produce crystals. Crystallisation drops are formed of a mix of concentrated protein and crystallisation solution, placed in a sealed container next to a well of crystallisation solution. As the concentration of precipitant is lower in the drop, water diffuses from the drop to the well as

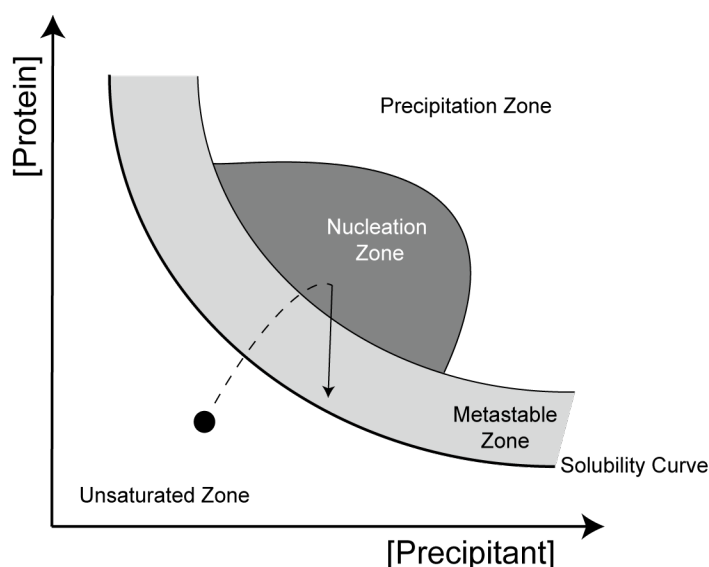


Figure A.19: Protein crystallisation phase diagram based on protein and precipitating concentration, two of the most commonly varied parameters. The protein is initially in the unsaturated zone; as vapour diffusion occurs, both protein and precipitant concentrations increase, until it reaches the nucleation zone. Nucleation reduces the protein concentration but maintains the precipitant concentration such that the system reaches the metastable zone in which crystals can form. Adapted from Chayen, N. E., 1992<sup>1-3</sup>

vapour, progressively increasing the concentration of both protein and precipitant in the well. The process of vapour diffusion can be described using a phase diagram.

The phase diagram in Figure A.19 is based on two of the most commonly varied parameters, protein and precipitant concentrations. Initially, the protein is in the unsaturated zone, where the protein is soluble. As vapour diffusion occurs, protein and precipitant concentrations are both increased until the nucleation zone is reached. Nucleation of protein crystals results in a drop in protein concentration whilst maintaining the precipitant concentration, allowing access to the metastable zone and the subsequent growth of crystals from the nuclei. Buffer type, pH, salt concentration, precipitant concentration, presence of additives, drop size, well volume and drop ratio are all common parameters varied to optimise crystallisation. Once protein crystals have been obtained, they are harvested using crystal-mounting loops, cryo-cooled in liquid nitrogen, and mounted on a goniometer head for data collection. As the cryo-cooling process can induce the formation of ice crystals, protein crystals often have to be cryo-protected before freezing, whether by addition of cryoprotectant directly into the crystallisation drops, or by transfer to a cryoprotectant-containing solution before freezing. The most common cryoprotectants are Glycerol and Ethylene-glycol, though many other soluble chemicals such as alcohols, polymers and sugars can also be used, as well as oils to get rid of the aqueous solvent surrounding the crystals.

### 8.3.2 X-ray diffraction

For data collection, crystals are exposed to X-rays – high-energy electromagnetic radiation. Whilst most incident X-rays will pass directly through the crystals, some are diffused by the electrons of atoms in the crystal. As crystals are constituted of 3D repetitions of the same molecules with the same orientation, this creates sets of parallel planes of diffusing electrons in the crystal, which in turn give rise to constructive and destructive interference. This results in the creation of diffraction patterns on the detector, where each diffraction spot is called a reflection. The planes in the crystal lattice can be defined by the Miller indices  $(h, k, l)$ , the sizes of which are inversely proportional to scattering angle. Planes that are closer together in the crystal will cause spots that are further from the centre of the detector. The resultant wave from diffraction of X-rays can be defined by the structure factor equation:

$$F_{hkl} = \sum_j F_j e^{2\pi i(xh_j + ky_j + lz_j)} \quad \text{Equation A.1}$$

where  $F_j$  is the scattering factor of atom  $j$ , and  $x$ ,  $y$  and  $z$  are the atomic coordinates of atom  $j$ . The structure factor of a given reflection is defined as the sum of the individual amplitudes and phases of X-rays scattered by each atom in the unit cell reaching this point of the detector. Therefore, movement of a single atom will result in a change in amplitude of all reflections in the diffraction pattern.

To calculate the electron density, a large number of reflections, often hundreds of thousands for medium to large proteins, are collected as the crystal is rotated. The structure factor is a complex waveform, so an inverse Fourier

transform can then be used to calculate the electron density in each point of the unit cell:

$$p(x, y, z) = \frac{1}{v} \sum_j F_{hkl} e^{-2\pi i(xh_j + ky_j + lz_j)} \quad \text{Equation A.2}$$

$$F_{hkl} = |F_{hkl}| e^{i\phi_{hkl}} \quad \text{Equation A.3}$$

To calculate the inverse Fourier Transform, information about both the amplitude and the phase of the resultant wave is required. The amplitude is proportional to the measured intensity of the reflections, which is collected during the diffraction experiment, but the phase information can't be collected directly. This lack of phase information is termed **the phase problem**<sup>215</sup>. Phases can be derived from the positions of intrinsic or added heavy atoms using MIR, MIRAS, SIR, SIRAS, MAD or SAD<sup>216</sup>. Another possibility is to use the phases calculated from a previously solved structure of a homologous protein, which is termed Molecular Replacement (MR, see 8.1.5).

### 8.3.3 Data collection strategies

After crystals are grown, they are mounted and exposed to X-rays, with rotation in small increments to collect datasets. Two sources of X-rays are commonly used: rotating CuK anode tubes, which emit radiation at 1.54 Å and are commonly used as home-sources<sup>217</sup>; and synchrotrons, which usually have tuneable wavelengths but are often centred around 0.97 Å, the absorption edge of selenium, which is the heavy metal most commonly used for phasing. The short wavelength and high flux of X-rays (number of photons per second) in combination with highly sensitive detectors<sup>218, 219</sup> at synchrotron sources allows

collection of datasets within a few minutes, with data at higher resolutions than could be achieved using low flux home-sources.

X-ray radiation damages crystals through interaction of energetic photons with atoms within the crystal to produce free electrons, free radicals and charged species that result in degradation of the protein crystals. Although collection of data under cryo-conditions significantly slows radiation damage, the hard X-rays produced by modern synchrotron sources can still induce significant radiation damage<sup>220</sup>. This leads to a reduction in reflection intensity and loss of high-resolution data across the course of a collection. A good data collection strategy is therefore a balance between increasing exposure to high-flux radiation to collect high-resolution data, and limiting exposure to prevent radiation damage and deterioration of datasets. A number of parameters can be varied to limit radiation damage, including the beam transmission, the exposure time and the angle of incremental rotation between image collections. Other strategies such as helical scans or wedged scan can also be used to reduce radiation damage by moving the crystal during data collection.

#### **8.3.4 Data processing and assessing data quality**

After collection, data is processed to allow structure solution. Initially, reflections are identified and indexed to give estimates of unit cell dimensions and crystal symmetry. The reflections are then integrated to produce a list of reflection intensities. These steps can be completed by programs such as XDS<sup>221</sup> and DIALS<sup>222</sup>.

Following integration, data collected at different angles around the crystal is often on different scales due to factors like anisotropy of diffraction power of the crystal in different directions, as well as radiation damage. Data is therefore scaled by programs such as Aimless<sup>223</sup> such that all symmetry-related reflections have comparable intensities. The data are then merged, which means that the average intensities of symmetry-related reflections are calculated, and outliers are discarded to increase the precision. Data quality is assessed after merging by several statistics, and pathologies within the dataset such as radiation damage can be identified. These statistics are also used to determine the cut-off for resolution.

$R_{\text{merge}}$  is a merging R-factor that has historically been used to provide an indication of data quality. It is a measure of internal consistency of the data, and reports the spread of independent measurements of intensity of a reflection around the average intensity for that reflection, with automatic outlier rejection. It is an indicator of the precision of the reflections' intensity measurement. Yet,  $R_{\text{merge}}$  tends to increase with redundancy, and therefore favours crystals with lower symmetry and datasets with lower multiplicity. With the development of alternative similar statistics such as  $R_{\text{meas}}$ ,  $R_{\text{merge}}$  is less frequently used.

Another historic metric of data quality is the  $I/\sigma(I)$ , which correspond to an average signal to noise ratio for the measured intensities, and can be used to gauge the meaningfulness of measured intensities. High and low resolution data tends to have lower intensities, while medium-resolution have higher intensities. The diffraction limit used to be defined at a resolution where the

$I/\sigma(I)$  decreases to about a value of two<sup>224</sup>. This was especially true with CCD detectors, which were associated with a lot of noise. Modern detectors such as Pilatus and Eiger systems use HPC approaches and are associated with significantly less noise. This allows the collection of good quality datasets with  $I/\sigma(I)$  values of one, or even slightly below. As accurate estimation of the uncertainties of the measurements is not trivial, calculation of  $I/\sigma(I)$  is not always accurate.

The half dataset correlation coefficient, or  $CC_{1/2}$ , is a more recently developed metric for evaluation of data quality that compares the correlation of intensities between randomly selected halves of the dataset<sup>225</sup>. Good datasets are generally expected to give a  $CC_{1/2}$  value of 0.3 or above, with higher resolution reflections that reduce  $CC_{1/2}$  generally excluded.  $CC_{1/2}$  is a more modern, more accurate metric to determine resolution and tends to be used in preference to  $R_{\text{merge}}$ <sup>225</sup>.

Following data processing, the merged datasets can be used to solve the structure. With over 140,000 protein structures deposited in the PDB (from various techniques), many proteins structures can be solved using Molecular Replacement (MR).

### **8.3.5 Molecular Replacement to solve crystal structures**

MR is based on the assumption that proteins with sufficient sequence identity (>30%) will have sufficient structural homology to allow initial estimation of the phases<sup>226</sup>. Patterson maps are derived for both the search model and the observed data and superimposed in Fourier Space, using rotation and

translation functions. Modern programs, such as Phaser<sup>227</sup>, use maximum likelihood methods to identify the best possible solution, and the accuracy of the superimposition is reported as a Log Likelihood Gain (LLG) score, with a higher LLG indicating a greater probability that the solution is correct. Following a successful molecular replacement, an initial set of phases can be calculated from the model and applied to the amplitudes derived from the measured intensities to calculate initial electron density maps.

### 8.3.6 Model correction and refinement

Model correction and refinement is an iterative process involving manual corrections of the initial model, or *de novo* building of parts of the protein that were not present in the initial model, to improve the fit to the electron density, and smaller, computational corrections and periodical re-calculation of the phases by a refinement program. This leads to improvement of the electron density maps due to the improved estimation of phases, such that new features can be observed in the maps, and the model further corrected. The overall process is repeated until the best possible fit of the model into electron density maps is obtained.

Two sigma-A weighted maps are commonly used for inspection and correction of the model. The  $2mF_o - DF_c$  map, where  $F_o$  is the observed structure factor and  $F_c$  is the calculated structure factor, shows the electron density in which the whole model should fit as perfectly as possible. In contrast, the  $mF_o - DF_c$  difference map shows negative density where an atom is placed that is not accounted for in the data, and hence should be removed from the model, and positive density where the data indicate the presence of atoms not currently



modelled, which should be added<sup>224</sup>. M, the Figure of Merit, is an approximate measure of phase quality and is calculated for each reflection, while D is the sigma-weighting value. Sigma-A weighted maps account for errors in the model-based amplitudes and amplify portions of the map where parts of the model are missing<sup>228, 229</sup>.

In order to keep the geometry correct, additional restraints are used during refinement. These are experimentally derived features such as bond lengths, bond angles and Ramachandran angles. Refinement programs also use some supplementary restraints to aid in convergence, such as non-crystallographic symmetries (NCS) which applies some partial localisation restraints for comparable atoms of different chains in the asymmetric unit that are not linked by crystallographic symmetry but are highly similar<sup>230</sup>. Use of NCS increases the observations-to-parameters ratio and therefore helps to prevent over-refinement. It can be especially useful at lower resolutions where the number of reflections, and therefore the number of observations, is lower<sup>207</sup>. TLS (Torion Liberation Screw) restraints allows movement of groups to modelled more accurately through ellipsoidal B-factors rather than the classical spherical ones<sup>224</sup>, but is an additional factor to refine and can promote over-refinement. To follow that the correction/refinement process is improving the agreement between the model and the experimental data,  $R_{work}$  is calculated after each round of refinement to measure the agreement between the observed and calculated structure factor amplitudes<sup>231</sup>:

$$R_{work} = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|} \quad \text{Equation A.4}$$

Whilst the initial  $R_{work}$  following MR can be up to around 40%, depending on how similar the MR model is to the new structure, it should decrease as refinement progresses, indicating greater agreement between the observed and calculated data.

To prevent over-refinement,  $R_{free}$  is calculated alongside  $R_{work}$ .<sup>232, 233</sup> A proportion of the data, typically 5%, is set aside at the merging step of data processing and is then not used during refinement. The excluded data is then used to calculate  $R_{free}$  in the same way as the  $R_{work}$ . While  $R_{work}$  will usually always decrease during refinement after a cycle of correction,  $R_{free}$  will only decrease if corrections are applied following real signal in maps. It will plateau or even increase if atoms are placed in noise, indicating over-refinement. Hence, an increase in the difference between  $R_{work}$  and  $R_{free}$  is a clear sign of over-refinement. Conventionally, this gap should be less than 5%, although it will be larger at lower resolution or with crystals giving dirty diffraction.

### 8.3.7 Validation of model

Crystal structures are refined until no further improvement of the agreement between the measured and calculated data can be achieved, which is highly dependant on the skills and experience of the crystallographer. Also, the advancement in processing and refinement software means that structures solved many years ago could be re-solved and re-refined to improve their overall quality. In addition, some crystallographers prefer to publish structures with some incompleteness, with un-modelled side chains or loops due to lack of electron density, which is more common in flexible regions of proteins and at lower resolutions. Other crystallographers prefer to publish models where no

atom is missing, even if they have to take the risk of placing side chains and loops at potentially wrong locations due to the absence of signal. Hence, various statistics are used to evaluate model quality before deposition into the PDB, or to evaluate the quality of previously deposited structures. Two important statistics,  $R_{work}$  and  $R_{free}$ , have already been discussed in the previous section.

In addition, validation software such as MolProbity<sup>5</sup> are used to evaluate both global and local structure quality<sup>234</sup>. This includes the deviation of bond lengths and angles from the ideal, reported as root mean square deviations (RMSDs), as well as Ramachandran outliers and cis-peptides<sup>235</sup>. More deviations from ideal geometries are accepted at lower resolution. While both Ramachandran outliers and cis-peptides can occur in crystal structures and reflect a particular fold, only approximately 0.5 % or less of residues will be real Ramachandran outliers<sup>236</sup> and should be supported by clear electron density.

## Chapter 9: Bibliography

---

1. Chayen, N.E. Comparative Studies of Protein Crystallization by Vapour-Diffusion and Microbatch Techniques. *Acta Crystallographica Section D* **54**, 8-15 (1998).
2. Kobayashi, S. et al. An Alternative Inhibitor Overcomes Resistance Caused by a Mutation of the Epidermal Growth Factor Receptor. *Cancer Research* **65**, 7096 (2005).
3. Giacovazzo, C. et al. Fundamentals of Crystallography (Oxford University Press, 2002).
4. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45**, D158-D169 (2017).
5. Davis, I.W., Murray, L.W., Richardson, J.S. & Richardson, D.C. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic acids research* **32**, W615-W619 (2004).
6. WHO. Cancer Fact Sheet (2018).
7. Sawyers, C. Targeted cancer therapy. *Nature* **432**, 294 (2004).
8. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* **2**, 401-404 (2012).
9. O'Brien, S.G. et al. Imatinib Compared with Interferon and Low-Dose Cytarabine for Newly Diagnosed Chronic-Phase Chronic Myeloid Leukemia. *New England Journal of Medicine* **348**, 994-1004 (2003).
10. Gaulton, A. et al. The ChEMBL database in 2017. *Nucleic acids research* **45**, D945-D954 (2017).
11. International Cancer Genome, C. et al. International network of cancer genome projects. *Nature* **464**, 993-998 (2010).
12. Forbes, S.A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research* **45**, D777-D783 (2017).
13. Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333 (2013).
14. Ashburn, T.T. & Thor, K.B. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery* **3**, 673 (2004).
15. Verdine, G.L. & Walensky, L.D. The Challenge of Drugging Undruggable Targets in Cancer: Lessons Learned from Targeting BCL-2 Family Members. *Clinical Cancer Research* **13**, 7264 (2007).
16. Hoekstra, E., Peppelenbosch, M.P. & Fuhler, G.M. Meeting Report Europhosphatase 2015: Phosphatases as Drug Targets in Cancer. *Cancer Research* **76**, 193 (2016).
17. Gao, T., Furnari, F. & Newton, A.C. PHLPP: A Phosphatase that Directly Dephosphorylates Akt, Promotes Apoptosis, and Suppresses Tumor Growth. *Molecular Cell* **18**, 13-24 (2005).
18. Chen, Y.-N.P. et al. Allosteric inhibition of SHP2 phosphatase inhibits cancers driven by receptor tyrosine kinases. *Nature* **535**, 148 (2016).
19. Massey, A.J. ATPases as Drug Targets: Insights from Heat Shock Proteins 70 and 90. *Journal of Medicinal Chemistry* **53**, 7280-7286 (2010).
20. Rajalingam, K., Schreck, R., Rapp, U.R. & Albert, Š. Ras oncogenes and their downstream targets. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1773**, 1177-1195 (2007).
21. Shao, H. et al. Exploration of Benzothiazole Rhodacyanines as Allosteric Inhibitors of Protein-Protein Interactions with Heat Shock Protein 70 (Hsp70). *Journal of Medicinal Chemistry* **61**, 6163-6177 (2018).

22. Ferraro, M. et al. Allosteric Modulators of HSP90 and HSP70: Dynamics Meets Function through Structure-Based Drug Design. *Journal of Medicinal Chemistry* **62**, 60-87 (2019).
23. Ostrem, J.M., Peters, U., Sos, M.L., Wells, J.A. & Shokat, K.M. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature* **503**, 548-551 (2013).
24. Davis, M.I. et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotech* **29**, 1046-1051 (2011).
25. Norman, R.A., Toader, D. & Ferguson, A.D. Structural approaches to obtain kinase selectivity. *Trends in Pharmacological Sciences* **33**, 273-278 (2012).
26. Reddy, A.S. & Zhang, S. Polypharmacology: drug discovery for the future. *Expert review of clinical pharmacology* **6**, 41-47 (2013).
27. Ahmad, T. & Eisen, T. Kinase Inhibition with BAY 43-9006 in Renal Cell Carcinoma. *Clinical Cancer Research* **10**, 6388S (2004).
28. Ludlow, R.F., Verdonk, M.L., Saini, H.K., Tickle, I.J. & Jhoti, H. Detection of secondary binding sites in proteins using fragment screening. *Proceedings of the National Academy of Sciences* **112**, 15910 (2015).
29. Housman, G. et al. Drug Resistance in Cancer: An Overview. *Cancers* **6**, 1769-1792 (2014).
30. Carter, T.A. et al. Inhibition of drug-resistant mutants of ABL, KIT, and EGF receptor kinases. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 11011-11016 (2005).
31. Pao, W. et al. Acquired Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib Is Associated with a Second Mutation in the EGFR Kinase Domain. *PLoS Med* **2**, e73 (2005).
32. Reddy, E.P. & Aggarwal, A.K. The Ins and Outs of Bcr-Abl Inhibition. *Genes & Cancer* **3**, 447-454 (2012).
33. Rudolph, U. & Möhler, H. GABA-based therapeutic approaches: GABAA receptor subtype functions. *Current Opinion in Pharmacology* **6**, 18-23 (2006).
34. Jeffrey Conn, P., Christopoulos, A. & Lindsley, C.W. Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nat Rev Drug Discov* **8**, 41-54 (2009).
35. Nickols, H.H. & Conn, P.J. Development of allosteric modulators of GPCRs for treatment of CNS disorders. *Neurobiology of Disease* **61**, 55-71 (2014).
36. Wylie, A.A. et al. The allosteric inhibitor ABL001 enables dual targeting of BCR-ABL1. *Nature* **543**, 733 (2017).
37. Caunt, C.J., Sale, M.J., Smith, P.D. & Cook, S.J. MEK1 and MEK2 inhibitors and cancer therapy: the long and winding road. *Nat Rev Cancer* **15**, 577-592 (2015).
38. Cho, Y.S. et al. Discovery and Evaluation of Clinical Candidate IDH305, a Brain Penetrant Mutant IDH1 Inhibitor. *ACS Medicinal Chemistry Letters* **8**, 1116-1121 (2017).
39. Dhillon, S. Ivosidenib: First Global Approval. *Drugs* **78**, 1509-1516 (2018).
40. Rath, O. & Kozielski, F. Kinesins and cancer. *Nat Rev Cancer* **12**, 527-539 (2012).
41. Wylie, A. et al. ABL001, a Potent Allosteric Inhibitor of BCR-ABL, Prevents Emergence of Resistant Disease When Administered in Combination with Nilotinib in an *in Vivo* Murine Model of Chronic Myeloid Leukemia. *Blood* **124**, 398 (2014).
42. Hughes, J.P., Rees, S., Kalindjian, S.B. & Philpott, K.L. Principles of early drug discovery. *British journal of pharmacology* **162**, 1239-1249 (2011).
43. Gibbs, J.B. Mechanism-Based Target Identification and Drug Discovery in Cancer Research. *Science* **287**, 1969 (2000).

44. Wyatt, P.G., Gilbert, I.H., Read, K.D. & Fairlamb, A.H. Target validation: linking target and chemical properties to desired product profile. *Current topics in medicinal chemistry* **11**, 1275-1283 (2011).
45. Zhang, J., Yang, P.L. & Gray, N.S. Targeting cancer with small molecule kinase inhibitors. *Nature Reviews Cancer* **9**, 28 (2009).
46. Santos, R. et al. A comprehensive map of molecular drug targets. *Nature reviews. Drug discovery* **16**, 19-34 (2017).
47. Bhullar, K.S. et al. Kinase-targeted cancer therapies: progress, challenges and future directions. *Molecular cancer* **17**, 48-48 (2018).
48. Agüero, F. et al. Genomic-scale prioritization of drug targets: the TDR Targets database. *Nature reviews. Drug discovery* **7**, 900-907 (2008).
49. Coker, E.A. et al. canSAR: update to the cancer translational research and drug discovery knowledgebase. *Nucleic acids research* **47**, D917-D922 (2019).
50. Broomhead, N.K. & Soliman, M.E. Can We Rely on Computational Predictions To Correctly Identify Ligand Binding Sites on Novel Protein Drug Targets? Assessment of Binding Site Prediction Methods and a Protocol for Validation of Predicted Binding Sites. *Cell Biochemistry and Biophysics* **75**, 15-23 (2017).
51. Mitsopoulos, C., Schierz, A.C., Workman, P. & Al-Lazikani, B. Distinctive Behaviors of Druggable Proteins in Cellular Networks. *PLOS Computational Biology* **11**, e1004597 (2015).
52. Hajduk, P.J., Huth, J.R. & Tse, C. Predicting protein druggability. *Drug Discovery Today* **10**, 1675-1682 (2005).
53. Hajduk, P.J., Huth, J.R. & Fesik, S.W. Druggability Indices for Protein Targets Derived from NMR-Based Screening Data. *Journal of Medicinal Chemistry* **48**, 2518-2525 (2005).
54. Erlanson, D.A., Fesik, S.W., Hubbard, R.E., Jahnke, W. & Jhoti, H. Twenty years on: the impact of fragments on drug discovery. *Nat Rev Drug Discov* **15**, 605-619 (2016).
55. Bohacek, R.S., McMartin, C. & Guida, W.C. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews* **16**, 3-50 (1996).
56. Blum, L.C. & Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *Journal of the American Chemical Society* **131**, 8732-8733 (2009).
57. Ruddigkeit, L., van Deursen, R., Blum, L.C. & Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling* **52**, 2864-2875 (2012).
58. Kim, S. et al. PubChem 2019 update: improved access to chemical data. *Nucleic acids research* **47**, D1102-D1109 (2019).
59. Chemistry, R.S.o.
60. Pence, H.E. & Williams, A. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education* **87**, 1123-1124 (2010).
61. Hubbard, R.E. & Murray, J.B. in *Methods in Enzymology* (ed. Kuo, L.C.) 509-531 (Academic Press, 2011).
62. Leach, A.R. & Hann, M.M. Molecular complexity and fragment-based drug discovery: ten years on. *Current Opinion in Chemical Biology* **15**, 489-496 (2011).
63. Lee, S.M. et al. Cytosolic NADP<sup>+</sup>-dependent isocitrate dehydrogenase status modulates oxidative damage to cells. *Free Radical Biology and Medicine* **32**, 1185-1196 (2002).
64. Hanukoglu, I. Proteopedia: Rossmann fold: A beta-alpha-beta fold at dinucleotide binding sites. *Biochemistry and Molecular Biology Education* **43**, 206-209 (2015).

65. Pettersen, E.F. et al. UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605-1612 (2004).
66. Itsumi, M. et al. Idh1 protects murine hepatocytes from endotoxin-induced oxidative stress by regulating the intracellular NADP<sup>+</sup>/NADPH ratio. *Cell Death And Differentiation* **22**, 1837 (2015).
67. Wahl, D.R. et al. Glioblastoma Therapy Can Be Augmented by Targeting IDH1-Mediated NADPH Biosynthesis. *Cancer research* **77**, 960-970 (2017).
68. Yan, H. et al. IDH1 and IDH2 Mutations in Gliomas. *New England Journal of Medicine* **360**, 765-773 (2009).
69. Amary, M.F. et al. Ollier disease and Maffucci syndrome are caused by somatic mosaic mutations of IDH1 and IDH2. *Nature Genetics* **43**, 1262 (2011).
70. Khurshed, M. et al. IDH1-mutant cancer cells are sensitive to cisplatin and an IDH1-mutant inhibitor counteracts this sensitivity. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **32**, fj201800547R-fj201800547R (2018).
71. Sulkowski, P.L. et al. 2-Hydroxyglutarate produced by neomorphic IDH mutations suppresses homologous recombination and induces PARP inhibitor sensitivity. *Science Translational Medicine* **9**, eaal2463 (2017).
72. Labussiere, M., Sanson, M., Idbaih, A. & Delattre, J.-Y. IDH1 gene mutations: a new paradigm in glioma prognosis and therapy? *The oncologist* **15**, 196-199 (2010).
73. McKenney, A.S. & Levine, R.L. Isocitrate dehydrogenase mutations in leukemia. *The Journal of clinical investigation* **123**, 3672-3677 (2013).
74. Dang, L. et al. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* **462**, 739-744 (2009).
75. Al-Khallaf, H. Isocitrate dehydrogenases in physiology and cancer: biochemical and molecular insight. *Cell & bioscience* **7**, 37-37 (2017).
76. DiNardo, C.D. et al. Serum 2-hydroxyglutarate levels predict isocitrate dehydrogenase mutations and clinical outcome in acute myeloid leukemia. *Blood* **121**, 4917-4924 (2013).
77. Lu, C. et al. IDH mutation impairs histone demethylation and results in a block to cell differentiation. *Nature* **483**, 474-478 (2012).
78. Koivunen, P. et al. Transformation by the (R)-enantiomer of 2-hydroxyglutarate linked to EGLN activation. *Nature* **483**, 484-488 (2012).
79. Xu, W. et al. Oncometabolite 2-Hydroxyglutarate Is a Competitive Inhibitor of  $\alpha$ -Ketoglutarate-Dependent Dioxygenases. *Cancer Cell* **19**, 17-30 (2011).
80. Wang, P. et al. Oncometabolite D-2-Hydroxyglutarate Inhibits ALKBH DNA Repair Enzymes and Sensitizes IDH Mutant Cells to Alkylating Agents. *Cell Reports* **13**, 2353-2361 (2015).
81. Figueroa, M.E. et al. Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. *Cancer Cell* **18**, 553-567.
82. Losman, J.-A. et al. R-2-Hydroxyglutarate Is Sufficient to Promote Leukemogenesis and Its Effects Are Reversible. *Science* **339**, 1621 (2013).
83. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* **6**, pl1-pl1 (2013).
84. Bielski, C.M. et al. Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer. *Cancer Cell* **34**, 852-862.e4 (2018).
85. Singh, A., Gurav, M., Dhanavade, S., Shetty, O. & Epari, S. Diffuse glioma – Rare homozygous IDH point mutation, is it an oncogenetic mechanism? *Neuropathology* **37**, 582-585 (2017).
86. Stein, E.M. et al. Enasidenib in mutant IDH2 relapsed or refractory acute myeloid leukemia. *Blood* **130**, 722-731 (2017).

87. Yang, H., Ye, D., Guan, K.-L. & Xiong, Y. IDH1 and IDH2 mutations in tumorigenesis: mechanistic insights and clinical perspectives. *Clinical cancer research : an official journal of the American Association for Cancer Research* **18**, 5562-5571 (2012).
88. Leonardi, R., Subramanian, C., Jackowski, S. & Rock, C.O. Cancer-associated isocitrate dehydrogenase mutations inactivate NADPH-dependent reductive carboxylation. *The Journal of biological chemistry* **287**, 14615-14620 (2012).
89. Intlekofer, A.M. et al. Acquired resistance to IDH inhibition through trans or cis dimer-interface mutations. *Nature* **559**, 125-129 (2018).
90. Fauman, E.B., Rai, B.K. & Huang, E.S. Structure-based druggability assessment—identifying suitable targets for small molecule therapeutics. *Current Opinion in Chemical Biology* **15**, 463-468 (2011).
91. Laurie, A.T.R. & Jackson, R.M. Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics* **21**, 1908-1916 (2005).
92. Nayal, M. & Honig, B. On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. *Proteins: Structure, Function, and Bioinformatics* **63**, 892-906 (2006).
93. Laskowski, R.A. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics* **13**, 323-330 (1995).
94. Krasowski, A., Muthas, D., Sarkar, A., Schmitt, S. & Brenk, R. DrugPred: A Structure-Based Approach To Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set. *Journal of Chemical Information and Modeling* **51**, 2829-2842 (2011).
95. Borrel, A., Regad, L., Xhaard, H., Petitjean, M. & Camproux, A.-C. PockDrug: A Model for Predicting Pocket Druggability That Overcomes Pocket Estimation Uncertainties. *Journal of Chemical Information and Modeling* **55**, 882-895 (2015).
96. Vukovic, S. & Huggins, D.J. Quantitative metrics for drug–target ligandability. *Drug Discovery Today* **23**, 1258-1266 (2018).
97. Lewis, R.A. in *Methods in Enzymology* 126-156 (Academic Press, 1991).
98. Larkin, M.A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).
99. Zuccotto, F. Protein Binding Pocket Chemogenomics. *Pharmagenomics*, 32 (2002).
100. Valdar, W.S.J. & Thornton, J.M. Protein–protein interfaces: Analysis of amino acid conservation in homodimers. *Proteins: Structure, Function, and Bioinformatics* **42**, 108-124 (2001).
101. Halling-Brown, M., Bulusu, K., Patel, M., E Tym, J. & Al-Lazikani, B. canSAR: An integrated cancer public translational research and drug discovery resource (2012).
102. Davies, M. et al. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic acids research* **43**, W612-20 (2015).
103. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research* **28**, 45-48 (2000).
104. Schmidtke, P. & Barril, X. Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. *Journal of Medicinal Chemistry* **53**, 5858-5867 (2010).
105. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
106. Erlanson, D.A., Jahnke, W., Mannhold, R., Kubinyi, H. & Folkers, G. *Fragment-based Drug Discovery: Lessons and Outlook* (Wiley, 2016).
107. Böhm, H.J., Schneider, G., Mannhold, R., Kubinyi, H. & Folkers, G. *Protein-Ligand Interactions: From Molecular Recognition to Drug Design* (Wiley, 2006).



108. Andrews, P.R., Craik, D.J. & Martin, J.L. Functional group contributions to drug-receptor interactions. *Journal of Medicinal Chemistry* **27**, 1648-1657 (1984).
109. Wellek, S. A critical evaluation of the current "p-value controversy". *Biometrical Journal* **59**, 854-872 (2017).
110. Nuzzo, R. in nature (2014).
111. Hayden, E.C. in nature (npb, 2013).
112. Johnson, V.E. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences* **110**, 19313 (2013).
113. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer* **18**, 696-705 (2018).
114. Hardy, J.A., Lam, J., Nguyen, J.T., O'Brien, T. & Wells, J.A. Discovery of an allosteric site in the caspases. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 12461-12466 (2004).
115. Zarzycki, M. et al. Structure of E69Q mutant of human muscle fructose-1,6-bisphosphatase. *Acta Crystallographica Section D* **67**, 1028--1034 (2011).
116. Kitas, E. et al. Sulfonylureido thiazoles as fructose-1,6-bisphosphatase inhibitors for the treatment of Type-2 diabetes. *Bioorganic & Medicinal Chemistry Letters* **20**, 594-599 (2010).
117. Watterson, S.H. et al. Small Molecule Antagonist of Leukocyte Function Associated Antigen-1 (LFA-1): Structure-Activity Relationships Leading to the Identification of 6-((5S,9R)-9-(4-Cyanophenyl)-3-(3,5-dichlorophenyl)-1-methyl-2,4-dioxo-1,3,7-triazaspiro[4.4]nonan-7-yl)nicotinic Acid (BMS-688521). *Journal of Medicinal Chemistry* **53**, 3814-3830 (2010).
118. Yang, J. et al. Discovery and Characterization of a Cell-Permeable, Small-Molecule c-Abl Kinase Activator that Binds to the Myristoyl Binding Site. *Chemistry \ Biology* **18**, 177-186 (2011).
119. Huang, H. et al. E2 enzyme inhibition by stabilization of a low-affinity interface with ubiquitin. *Nat Chem Biol* **10**, 156--163 (2014).
120. Talapatra, S.K., Anthony, N.G., Mackay, S.P. & Kozielski, F. Mitotic Kinesin Eg5 Overcomes Inhibition to the Phase I/II Clinical Candidate SB743921 by an Allosteric Resistance Mechanism. *Journal of Medicinal Chemistry* **56**, 6317-6329 (2013).
121. Marzinzik, A.L. et al. Discovery of Novel Allosteric Non-Bisphosphonate Inhibitors of Farnesyl Pyrophosphate Synthase by Integrated Lead Finding. *ChemMedChem* **10**, 1884--1891 (2015).
122. Jahnke, W. et al. Allosteric non-bisphosphonate FPPS inhibitors identified by fragment-based discovery. *Nat Chem Biol* **6**, 660--666 (2010).
123. Burgin, A.B. et al. Design of phosphodiesterase 4D (PDE4D) allosteric modulators for enhancing cognition with improved safety. *Nat Biotech* **28**, 63--70 (2010).
124. Han, S. et al. Selectively targeting an inactive conformation of interleukin-2-inducible T-cell kinase by allosteric inhibitors. *Biochemical Journal* **460**, 211--222 (2014).
125. Saalau-Bethell, S.M. et al. Crystal Structure of Human Soluble Adenylate Cyclase Reveals a Distinct, Highly Flexible Allosteric Bicarbonate Binding Pocket. *ChemMedChem* **9**, 823--832 (2014).
126. Nedyalkova, L.T., Y.; Rabeh, W.; Tempel, W.; Landry, R.; Arrowsmith, C.H.; Edwards, A.M.; Bountra, C.; Weigelt, J.; Bochkarev, A.; Park, H. Crystal structure of the C-terminal Hexokinase domain of human HK3.
127. Chen, H. et al. Discovery of a novel allosteric inhibitor-binding site in ERK5: comparison with the canonical kinase hinge ATP-binding site. *Acta Crystallographica Section D* **72**, 682--693 (2016).

128. Wallace, M.B. et al. Structure-based design and synthesis of pyrrole derivatives as MEK inhibitors. *Bioorganic & Medicinal Chemistry Letters* **20**, 4156-4158 (2010).
129. Joerger, A.C. et al. Exploiting Transient Protein States for the Design of Small-Molecule Stabilizers of Mutant p53. *Structure* **23**, 2246-2255 (2015).
130. Xu, X. et al. Structures of Human Cytosolic NADP-dependent Isocitrate Dehydrogenase Reveal a Novel Self-regulatory Mechanism of Activity. *Journal of Biological Chemistry* **279**, 33946-33957 (2004).
131. Shen, Q. et al. ASD v3.0: unraveling allosteric regulation with structural mechanisms and biological networks. *Nucleic Acids Research* **44**, D527-D535 (2015).
132. Jacob, R.E., Zhang, J., Gray, N.S. & Engen, J.R. Allosteric interactions between the myristate- and ATP-site of the Abl kinase. *PloS one* **6**, e15929-e15929 (2011).
133. Shen, Q. et al. Proteome-Scale Investigation of Protein Allosteric Regulation Perturbed by Somatic Mutations in 7,000 Cancer Genomes. *American journal of human genetics* **100**, 5-20 (2017).
134. Soussi, T. & Wiman, K.G. TP53: an oncogene in disguise. *Cell Death And Differentiation* **22**, 1239 (2015).
135. Boeckler, F.M. et al. Targeted rescue of a destabilized mutant of p53 by an <em>in silico</em> screened drug. *Proceedings of the National Academy of Sciences* **105**, 10360 (2008).
136. Williams, C.J. et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* **27**, 293-315 (2018).
137. Hoe, K.K., Verma, C.S. & Lane, D.P. Drugging the p53 pathway: understanding the route to clinical efficacy. *Nat Rev Drug Discov* **13**, 217-236 (2014).
138. Liu, X. et al. Small molecule induced reactivation of mutant p53 in cancer cells. *Nucleic Acids Research* **41**, 6034-6044 (2013).
139. Soragni, A. et al. A Designed Inhibitor of p53 Aggregation Rescues p53 Tumor Suppression in Ovarian Carcinomas. *Cancer Cell* **29**, 90-103 (2016).
140. Thomas, C. & Gustafsson, J.-Å. Estrogen receptor mutations and functional consequences for breast cancer. *Trends in Endocrinology & Metabolism* **26**, 467-476 (2015).
141. Brzozowski, A.M. et al. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **389**, 753--758 (1997).
142. Shiau, A.K. et al. The Structural Basis of Estrogen Receptor/Coactivator Recognition and the Antagonism of This Interaction by Tamoxifen. *Cell* **95**, 927-937 (1998).
143. Fruman, D.A. & Rommel, C. PI3K and cancer: lessons, challenges and opportunities. *Nat Rev Drug Discov* **13**, 140-156 (2014).
144. Vanhaesebroeck, B., Stephens, L. & Hawkins, P. PI3K signalling: the path to discovery and understanding. *Nat Rev Mol Cell Biol* **13**, 195-203 (2012).
145. Miller, M.S. et al. Identification of allosteric binding sites for PI3Kα oncogenic mutant specific inhibitor design. *Bioorganic & Medicinal Chemistry* **25**, 1481-1486 (2017).
146. Wu, F. et al. Inhibition of Cancer-Associated Mutant Isocitrate Dehydrogenases by 2-Thiohydantoin Compounds. *Journal of Medicinal Chemistry* **58**, 6899-6908 (2015).
147. Wickham, H. ggplot2: Elegant Graphics for Data Analysis (Springer Publishing Company, Incorporated, 2009).
148. Bauman, J.D. et al. Detecting Allosteric Sites of HIV-1 Reverse Transcriptase by X-ray Crystallographic Fragment Screening. *Journal of Medicinal Chemistry* **56**, 2738-2746 (2013).

149. Niesen, F.H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature Protocols* **2**, 2212 (2007).
150. Miyazaki, Y., Doi, N., Koma, T., Adachi, A. & Nomaguchi, M. Novel In Vitro Screening System Based on Differential Scanning Fluorimetry to Search for Small Molecules against the Disassembly or Assembly of HIV-1 Capsid Protein. *Frontiers in Microbiology* **8** (2017).
151. Vivian, J.T. & Callis, P.R. Mechanisms of tryptophan fluorescence shifts in proteins. *Biophysical Journal* **80**, 2093-2109 (2001).
152. Lakowicz, J.R. Principles of Fluorescence Spectroscopy (Springer, 2006 ).
153. Silvestre, H.L., Blundell, T.L., Abell, C. & Ciulli, A. Integrated biophysical approach to fragment screening and validation for fragment-based lead discovery. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 12984-12989 (2013).
154. Casu, B., Arya, T., Bessette, B. & Baron, C. Fragment-based screening identifies novel targets for inhibitors of conjugative transfer of antimicrobial resistance by plasmid pKM101. *Scientific Reports* **7**, 14907 (2017).
155. Lo, M.-C. et al. Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Analytical Biochemistry* **332**, 153-159 (2004).
156. Deng, G. et al. Selective Inhibition of Mutant Isocitrate Dehydrogenase 1 (IDH1) via Disruption of a Metal Binding Network by an Allosteric Small Molecule. *Journal of Biological Chemistry* **290**, 762-774 (2015).
157. Huynh, K. & Partch, C.L. Analysis of protein stability and ligand interactions by thermal shift assay. *Current protocols in protein science* **79**, 28.9.1-28.9.14 (2015).
158. Okoye-Okafor, U.C. et al. New IDH1 mutant inhibitors for treatment of acute myeloid leukemia. *Nat Chem Biol* **11**, 878-886 (2015).
159. Lamoree, B. & Hubbard, Roderick E. Current perspectives in fragment-based lead discovery (FBLD). *Essays In Biochemistry* **61**, 453 (2017).
160. Patel, D., Bauman, J.D. & Arnold, E. Advantages of crystallographic fragment screening: Functional and mechanistic insights from a powerful platform for efficient drug discovery. *Progress in Biophysics and Molecular Biology* **116**, 92-100 (2014).
161. Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. Features and development of Coot. *Acta crystallographica. Section D, Biological crystallography* **66**, 486-501 (2010).
162. Pearce, N.M. et al. A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. *Nature Communications* **8**, 15123 (2017).
163. .
164. Krojer, T. et al. The XChemExplorer graphical workflow tool for routine or large-scale protein-ligand structure determination. *Acta Crystallographica Section D* **73**, 267-278 (2017).
165. Murshudov, G.N. et al. REFMAC5 for the refinement of macromolecular crystal structures. *Acta crystallographica. Section D, Biological crystallography* **67**, 355-367 (2011).
166. Glasel, J. Validity of nucleic acid purities monitored by 260 nm/280nm absorbance ratios. *BioTechniques* **18**, 2 (1995).
167. Gasteiger, E.H., C.; Gattiker, A.; Duvaud, S.; Wilkins, M.R.; Appel, R.D.; Bairoch, A.; John M. Walker (ed): The Proteomics Protocols Handbook; Identification and Analysis Tools on the ExPASy Server; (Humana Press, 2005).
168. Team, R.C. R: A Language and Environment for Statistical Computing. (2015).
169. Rohle, D. et al. An inhibitor of mutant IDH1 delays growth and promotes differentiation of glioma cells. *Science (New York, N.Y.)* **340**, 626-630 (2013).

170. Urban, D.J. et al. Assessing inhibitors of mutant isocitrate dehydrogenase using a suite of pre-clinical discovery assays. *Scientific reports* **7**, 12758-12758 (2017).
171. Deller, M.C., Kong, L. & Rupp, B. Protein stability: a crystallographer's perspective. *Acta crystallographica. Section F, Structural biology communications* **72**, 72-95 (2016).
172. Rupp, B. in Biomolecular Crystallography (ed. Scholl, S.) 97 (Garland Science, New York, 2010).
173. Schrodinger, LLC. (2015).
174. Morley, A.D. et al. Fragment-based hit identification: thinking in 3D. *Drug Discovery Today* **18**, 1221-1227 (2013).
175. Thomas Sherine, E. et al. Structure-guided fragment-based drug discovery at the synchrotron: screening binding sites and correlations with hotspot mapping. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **377**, 20180422 (2019).
176. Bricogne G., B.E., Brandl M., Flensburg C., Keller P., Paciorek W., & Roversi P, S.A., Smart O.S., Vonnrhein C., Womack T.O. . (Global Phasing Ltd, Cambridge, United Kingdom, 2017).
177. Groom, C.R., Bruno, I.J., Lightfoot, M.P. & Ward, S.C. The Cambridge Structural Database. *Acta Crystallographica Section B* **72**, 171-179 (2016).
178. Long, F. et al. AceDRG: a stereochemical description generator for ligands. *Acta crystallographica. Section D, Structural biology* **73**, 112-122 (2017).
179. Xie, X. et al. Allosteric Mutant IDH1 Inhibitors Reveal Mechanisms for IDH1 Mutant and Isoform Selectivity. *Structure* **25**, 506-513 (2017).
180. Wu, G., Yuan, Y. & Hodge, C.N. Determining Appropriate Substrate Conversion for Enzymatic Assays in High-Throughput Screening. *Journal of Biomolecular Screening* **8**, 694-700 (2003).
181. Okoye-Okafor, U.C. et al. New IDH1 mutant inhibitors for treatment of acute myeloid leukemia. *Nature chemical biology* **11**, 878-886 (2015).
182. Sittampalam GS, G.A., Brimacombe K, et al. Assay Guidance Manual (Eli Lilly and the National Center for Advancing Translational Sciences, 2004).
183. Devine, S.M. et al. Promiscuous 2-Aminothiazoles (PrATs): A Frequent Hitting Scaffold. *Journal of Medicinal Chemistry* **58**, 1205-1214 (2015).
184. Yang, E.S. et al. Inactivation of NADP<sup>+</sup>-dependent isocitrate dehydrogenase by nitric oxide. *Free Radical Biology and Medicine* **33**, 927-937 (2002).
185. Worth, C.L., Preissner, R. & Blundell, T.L. SDM--a server for predicting effects of mutations on protein stability and malfunction. *Nucleic acids research* **39**, W215-W222 (2011).
186. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research* **39**, e118-e118 (2011).
187. Brooks HB, G.S., Kahl SD, et al. . in Assay Guidance Manual (ed. Sittampalam GS, G.A., Brimacombe K, et al.) (2004).
188. Moretti, R., Bender, B.J., Allison, B. & Meiler, J. Rosetta and the Design of Ligand Binding Sites. *Methods in molecular biology (Clifton, N.J.)* **1414**, 47-62 (2016).
189. Elokely, K.M. & Doerksen, R.J. Docking challenge: protein sampling and molecular docking performance. *Journal of chemical information and modeling* **53**, 1934-1945 (2013).
190. Popovici-Muller, J. et al. Discovery of AG-120 (Ivosidenib): A First-in-Class Mutant IDH1 Inhibitor for the Treatment of IDH1 Mutant Cancers. *ACS Medicinal Chemistry Letters* (2018).
191. Lowery, M.A. et al. Safety and activity of ivosidenib in patients with IDH1-mutant advanced cholangiocarcinoma: a phase 1 study. *The Lancet Gastroenterology & Hepatology* **4**, 711-720 (2019).

192. Tym, J.E. et al. canSAR: an updated cancer research and drug discovery knowledgebase. *Nucleic Acids Research* **44**, D938-D943 (2015).
193. Berman, H.M. et al. The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).
194. Berg, J.M., Tymoczko, J.L. & Stryer, L. Biochemistry, Fifth Edition (W.H. Freeman, 2002).
195. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic acids research* **47**, D427-D432 (2019).
196. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 1-8 (2011).
197. Futreal, P.A. et al. A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183 (2004).
198. UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* **47**, D506-D515 (2019).
199. Yates, B. et al. Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic acids research* **45**, D619-D625 (2017).
200. Aken, B.L. et al. The Ensembl gene annotation system. *Database : the journal of biological databases and curation* **2016**, baw093 (2016).
201. Dana, J.M. et al. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Research* **47**, D482-D489 (2018).
202. Velankar, S. et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research* **41**, D483-D489 (2012).
203. Kabsch, W. XDS. *Acta Crystallographica Section D: Biological Crystallography* **66**, 125-132 (2010).
204. Winter, G. xia2: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography* **43**, 186-190 (2010).
205. Winn, M.D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography* **67**, 235-242 (2011).
206. Adams, P.D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D* **66**, 213-221 (2010).
207. Smart, O.S. et al. Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. *Acta Crystallographica Section D* **68**, 368-380 (2012).
208. Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. Features and development of Coot. *Acta Crystallographica Section D* **66**, 486-501 (2010).
209. Groom, C.R., Bruno, I.J., Lightfoot, M.P. & Ward, S.C. The Cambridge Structural Database. *Acta crystallographica Section B, Structural science, crystal engineering and materials* **72**, 171-179 (2016).
210. Ng, J.T., Dekker, C., Kroemer, M., Osborne, M. & von Delft, F. Using textons to rank crystallization droplets by the likely presence of crystals. *Acta Crystallographica Section D: Biological Crystallography* **70**, 2702-2718 (2014).
211. Burley, S.K. et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research* **47**, D464-D474 (2018).
212. Blow, D. Outline of Crystallography for Biologists (OUP Oxford, 2002).
213. .
214. Giegé, R. What macromolecular crystallogenesis tells us - what is needed in the future. *IUCrJ* **4**, 340-349 (2017).
215. Taylor, G. The phase problem. *Acta Crystallographica Section D* **59**, 1881-1890 (2003).

216. Taylor, G.L. Introduction to phasing. *Acta crystallographica. Section D, Biological crystallography* **66**, 325-338 (2010).
217. Pflugrath, J.W. Practical macromolecular cryocrystallography. *Acta crystallographica. Section F, Structural biology communications* **71**, 622-642 (2015).
218. Casanas, A. et al. EIGER detector: application in macromolecular crystallography. *Acta Crystallographica Section D* **72**, 1036-1048 (2016).
219. Kraft, P. et al. Performance of single-photon-counting PILATUS detector modules. *Journal of synchrotron radiation* **16**, 368-375 (2009).
220. Duke, E.M.H. & Johnson, L.N. Macromolecular crystallography at synchrotron radiation sources: current status and future developments. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **466**, 3421-3452 (2010).
221. Kabsch, W. XDS. *Acta crystallographica. Section D, Biological crystallography* **66**, 125-132 (2010).
222. Winter, G. et al. DIALS: implementation and evaluation of a new integration package. *Acta crystallographica. Section D, Structural biology* **74**, 85-97 (2018).
223. Evans, P.R. & Murshudov, G.N. How good are my data and what is the resolution? *Acta crystallographica. Section D, Biological crystallography* **69**, 1204-1214 (2013).
224. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *The FEBS journal* **275**, 1-21 (2008).
225. Karplus, P.A. & Diederichs, K. Assessing and maximizing data quality in macromolecular crystallography. *Current opinion in structural biology* **34**, 60-68 (2015).
226. Evans, P. & McCoy, A. An introduction to molecular replacement. *Acta crystallographica. Section D, Biological crystallography* **64**, 1-10 (2008).
227. McCoy, A.J. et al. Phaser crystallographic software. *Journal of applied crystallography* **40**, 658-674 (2007).
228. Read, R.J. Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallographica Section A* **42**, 140-149 (1986).
229. PHENIX.
230. Kleywegt, G.J. Use of Non-crystallographic Symmetry in Protein Structure Refinement. *Acta Crystallographica Section D* **52**, 842-857 (1996).
231. Crystallography, I.U.o. in Online Dictionary of Crystallography (2017).
232. Brünger, A.T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472-475 (1992).
233. Brünger, A.T. in *Methods in Enzymology* 366-396 (Academic Press, 1997).
234. Karplus, P.A. & Diederichs, K. Linking crystallographic model and data quality. *Science (New York, N.Y.)* **336**, 1030-1033 (2012).
235. Kleywegt, G. Validation of protein crystal structures. *Acta Crystallographica Section D* **56**, 249-265 (2000).
236. Gore, S. et al. Validation of Structures in the Protein Data Bank. *Structure* **25**, 1916-1927 (2017).