Accepted Manuscript

Structural biology and the design of new therapeutics: from HIV and cancer to mycobacterial infections

Sherine E. Thomas, Vitor Mendes, So Yeon Kim, Sony Malhotra, Bernardo Ochoa-Montaño, Michal Blaszczyk, Tom L. Blundell

PII: S0022-2836(17)30315-7
DOI: doi:10.1016/j.jmb.2017.06.014
Reference: YJMBI 65445

To appear in: Journal of Molecular Biology

Received date: 23 May 2017
Accepted date: 19 June 2017

Please cite this article as: Thomas, S.E., Mendes, V., Kim, S.Y., Malhotra, S., Ochoa-Montaño, B., Blaszczyk, M. & Blundell, T.L., Structural biology and the design of new therapeutics: from HIV and cancer to mycobacterial infections, Journal of Molecular Biology (2017), doi:10.1016/j.jmb.2017.06.014

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
Structural biology and the design of new therapeutics: from HIV and cancer to mycobacterial infections

A Paper Dedicated to John Kendrew

Sherine E. Thomas, Vitor Mendes, So Yeon Kim, Sony Malhotra, Bernardo Ochoa-Montaño, Michal Blaszczyk & Tom L. Blundell*

Department of Biochemistry, University of Cambridge

Tennis Court Road, Cambridge, United Kingdom

*corresponding author

Key words: Structure-guided; Fragment-based; Drug design/discovery; Cancer; Infectious disease; Mycobacterium.

Abbreviations: FBDD, fragment-based drug design; SPR, Surface Plasmon Resonance; NMR, Nuclear Magnetic Resonance; TB, tuberculosis; Mtb, Mycobacterium tuberculosis; Mab, Mycobacterium abscessus; Mlep, Mycobacterium leprae; HIV, Human Immunodeficiency Virus; AIDS, Acquired Immune Deficiency Syndrome; SAR, structure-activity relationship, non-tuberculous mycobacteria; NTM; dpCoA, 3’-dephospho Coenzyme A
ABSTRACT

Interest in applications of protein crystallography to medicine was evident as the first high-resolution structures emerged in the 50s and 60s. In Cambridge Max Perutz and John Kendrew sought to understand mutations in sickle cell and other genetic diseases related to haemoglobin, while in Oxford the group of Dorothy Hodgkin became interested in long-lasting zinc-insulin crystals for treatment of diabetes and later considered insulin redesign as synthetic insulins became possible. The use of protein crystallography in structure-guided drug discovery emerged as enzyme structures allowed the identification of potential inhibitor-binding sites and optimisation of interactions of hits using the structure of the target protein. Early examples of this approach were the use of the structure of renin to design anti-hypertensives and the structure of HIV protease in design of AIDS antivirals. More recently, use of structure-guided design with fragment-based drug discovery, which reduces the size of screening libraries by decreasing complexity, has improved ligand efficiency in drug design and has been used to progress three oncology drugs through clinical trials to FDA approval. We exemplify current developments in structure-guided target identification and fragment-based lead discovery with efforts to develop new antimicrobials for mycobacterial infections.

Early Development of Structure-guided Drug Discovery

When Max Perutz, John Kendrew and their colleagues in Cambridge were solving the first protein structures of myoglobin and haemoglobin in the 1950s and 60s[1-3], they were already aware of the importance of their work to medicine. Understanding the impacts of mutations on oxygen affinity and subunit cooperativity in abnormal haemoglobins that resulted in inherited single-gene disorders, such as sickle-cell disease, was recognised as a major objective. Dorothy Hodgkin’s Oxford laboratory collaborated with Jørgen Schlichtkrull of Novo to understand how different crystalline forms of insulin could be exploited as slow-acting therapeutics for the treatment of diabetes[4]. This became a real possibility when the structure of insulin was solved[5, 6], as many insulin sequences had been defined in Fred Sanger’s laboratory in Cambridge[7]. Sequences and structures stimulated ideas not only about insulin storage and receptor binding, but also about producing more effective therapeutics. These speculations became real opportunities when groups in Aachen, New York and Shanghai completed the synthesis of insulin, encouraging ideas about the design of novel synthetic insulins.
Ideas about drug design were stimulated by the determination of the first enzyme structures – lysozyme, chymotrypsin and trypsin – in the 60s and an emerging understanding of the interactions that led to selectivity of enzyme substrate binding[8].

In the 70s and 80s, clinically important drug targets such as the aspartic protease renin[9, 10], which cleaves angiotensinogen to form angiotensin I, an essential step in regulating blood pressure, were modelled on less exciting enzymes such as fungal pepsins [11, 12]. The use of protein crystallography in drug discovery accelerated in the 1980s, especially by using a combination of protein structure and interactive computer graphics, such as the Evans and Sutherland machines[13]. The model of renin[14] was used widely in structure-guided drug design in the pharma industry in the 1980s. The high-resolution X-ray structures of apo-enzymes and complexes of renin and its close homologues followed much later[15, 16].

**New Paradigm of Structure-guided Drug Discovery: Targeting HIV Protease**

Probably the most influential development was the design of AIDS antivirals, based on the structure of HIV protease; these moved onto the market very quickly in the 1990s. Some basic science influenced this! In 1978, Jordan Tang who had sequenced pepsin together with crystallographic collaborators suggested that proteases had evolved from an ancestral dimer by gene duplication, fusion and divergence to give more effective enzymes[17]. A close relative of this dimeric ancestral aspartic protease was found in the retroviral proteases, first in Rous Sarcoma Virus, but then in HIV soon after the AIDS epidemic was recognised in the US and Europe. The genome of HIV, which encodes a polyprotein, was shown to include a protease[18], which was quickly recognised as a dimeric viral protease essential for the generation of infectious viral particles, and a model was produced based on the aspartic proteinase evolutionary relationship[19]. In 1989 structures followed for Rous Sarcoma Virus [20, 21] and HIV protease[22, 23], the structure of which was improved by further experimental structures determined independently by two labs[24] [25]. The subsequent development of new AIDS antivirals by 1997, including four very successful drugs (Roche Pharmaceuticals’ saquinavir, Abbot’s ritonavir, Merck’s indinavir, and Agouron’s nelfinavir) demonstrated the importance of understanding the genome not only in terms of functions of gene products, but also their architectures for use in structure-guided drug discovery, recorded recently in an excellent history of macromolecular crystallography and its fruits[26]
The identification of HIV protease as a target and the development of AIDS antivirals was a new paradigm in drug discovery. It demonstrated that there was value in computational analysis of genomes in order to identify targets. This was an “exploration of biological space”, an exciting challenge in the early 1990s as the sequence determination of human infectious agents such as Mycobacteriaceae that give rise to tuberculosis and leprosy, as well as the very much larger genome of Homo sapiens became real prospects. The HIV protease inhibitor story also illustrated the importance “exploring chemical space”: using protein structure to estimate the druggability of potential targets, followed by exploration of possible binding using screening libraries of chemical compounds. This idea of drug discovery can be summarised, as in Figure 1A.

Over the subsequent 25 years, several new approaches have been introduced that exploit knowledge of the architecture of the target and screening of chemical libraries. One of the most influential has been the development of structure-guided fragment-based drug discovery.

**Protein Crystallography, Fragment-Based Drug Discovery and Oncology**

In the 1980s and 90s meeting the challenge of the size and diversity of “chemical space” became a focus in the pharmaceutical industry with the realisation that chemical libraries of several thousand drug-like compounds explored only a tiny area of the chemical space. In order to estimate the number, Lipinski rules assuming a molecular weight limit of 500 daltons, the presence of carbon, hydrogen, oxygen, nitrogen and sulfur, and a maximum of 4 rings leads to an estimate of $10^{63}$[27]. Big pharma searched the world for new chemical diversity, often using the products of our natural environment in underdeveloped-forested areas. The chemical libraries grew to hundreds of thousands of compounds and screening was roboticized to cope with the challenge. But a solution to the challenge was also found in a different approach in which complexity of the chemicals screened was reduced by decreasing their molecular weights, which at the same time increased their promiscuity in binding targets. The innovation that allowed decrease of size of the chemical screening library was fragment-based drug discovery (FBDD).

In FBDD, a fragment library often of ~ 1000 compounds of $<300$ Da is screened against the target of interest, resulting in identification of initial hits. These are then moved to lead candidates by chemically growing or linking the fragments followed by optimisation of interactions, thereby exploring the chemical space available for binding to the target protein very effectively. A high-affinity lead molecule thus
developed from a fragment hit retains the key binding interactions of the original fragment with the “hotspot” on the target protein. Most of the fragments have lower potency than the more complex molecules found in typical high-throughput screening (HTS) compound libraries, however small fragments that bind, do so by making well defined and directional high-quality interactions and by displacing unhappy water molecules at the hotspots, giving rise to high ligand efficiency\[28\].

Early experiments used ligand-based NMR (Steve Fesik and his colleagues at Abbott) \[29\] and X-ray crystal screening\[30, 31\], developed at Astex initially by exploiting high-throughput analysis of cocktails of six to ten fragments soaked into apo-protein crystals. Knowledge of the structure of the complex of the fragment with target protein allowed the initial use of small, often non-chiral compounds, which were optimized using structure-guided approaches to make specific interactions and to introduce chirality in the molecules. The resulting fragment hits were capable of achieving high binding efficiency per atom and often better physicochemical properties in comparison to those from HTS approach which exploits much larger libraries of \(~10^6\) or more compounds\[30, 32\].

The relatively low affinities mean that a combination of biochemical, biophysical and structural techniques must be used to monitor hit identification, validation and subsequent elaboration into lead molecules, as summarised in Figure 1B. The choice of methods will depend on factors such as the availability of a sensitive biochemical assay, the solubility and stability of the protein, the existence of crystals of the apo-protein, and so on. Many groups use a two-stage approach of high-throughput screening of the fragment library using fluorescence-based thermal shift measurements\[32, 33\], ligand-based NMR, surface plasmon resonance (SPR) and increasingly with the roboticized screening facilities available on synchrotron beamlines, X-ray crystallographic screening. The fragment hits common between these techniques are then validated by optimisation of resolution of the structures by X-ray diffraction or structure determination by NMR if not possible to define by X-ray methods, as well as defining the kinetics with SPR and the thermodynamics of the binding using isothermal calorimetry. The combination of these techniques and others (Figure 1B) gives reassurance of the quality of the hit. The validated fragment hits are then elaborated iteratively by growing to a larger molecular weight or by linking using structure-guided techniques.

Where a crystal structure is not available, several labs have used fragment libraries to search chemical space using methods such as SPR\[34\], sometimes with many thousands of chiral compounds, often derived from analysis of successful drugs\[35\].
Where an experimental structure or computational homology model is available, the pipeline can be complemented by substructure searches on commercially available compound databases or \textit{in-silico} screening predictions of analogues\cite{36, 37}. An example of this approach was that of Winter \textit{et al.} \cite{38}, who had no suitable crystal structure to identify hits that disrupt the interaction between hepatocyte growth factor/scatter factor N-terminal fragment and the hepatocyte growth factor receptor (Met), and subsequently inhibit Met signalling. They carried out sub-structure searches on initial hits identified by biophysical methods such as SPR, and followed these by \textit{in-silico} docking predictions to identify compounds, with significant anti-tumorigenic and anti-migratory activity in cell-based assays.

In 2011 the first fragment-derived drug, Vemurafenib, was approved targeting a mutant form of BRAF, extending life for patients with skin cancer. The drug was discovered at Plexxikon and developed in partnership with Roche. The second drug, Venetoclax, developed by AbbVie and Genentech, binds to BCL-2 and blocks its interaction with other proteins; was approved by the US FDA in 2016 for chronic lymphocytic leukemia (CLL). In 2017 the third drug, Ribociclib, by Astex and Novartis was approved for targeting the protein kinase Cdk4 and will be used in combination with letrozole as a first-line treatment for advanced breast cancer. All of these campaigns featured an important role for structure-guided approaches with a combination of activity in relatively small companies, often founded by academics with an interest in protein structure, but developed as cancer therapeutics with strong scientific and financial involvement of large pharma.

\textbf{Moving Structure-guided Discovery from Oncology to other Diseases}

We have shown that the HIV protease paradigm exploiting a combination of efficient exploration of biological space by understanding the genome and structural proteome, together with exploration of chemical space using structure-guided and fragment-based approaches, works well in cancer. As described above, FBDD has not only provided useful leads but also led to drugs that have been approved by the FDA. Much of the work has been done in biotech and small pharma companies in collaboration with large pharma.

We now address the question as to how far this structure-guided approach can be used to target diseases that are rare in the West, where big pharma finds it more difficult to get returns on its investment. These include genetic diseases occurring in a few families, for example cystic fibrosis or black bone disease (alkaptonuria), or those
infectious diseases that are mainly found in developing countries, such as tuberculosis and leprosy. Large pharma finds these diseases more challenging to address, as the financial returns on investment are limited either by the small numbers of patients or by the difficulty in finding a market in relatively less prosperous countries in Africa, South America and the Indian subcontinent.

Structure-guided approaches will depend on knowledge of the genomes and strain variation of the causative organism in order to develop a fruitful/successful target-based approach. This is being addressed worldwide, mainly by charitable initiatives such as the Wellcome Trust at their Genome Campus or the Bill & Melinda Gates Foundation in initiatives such as HIT-TB and Shorten-TB. There is also considerable activity in Government-funded institutes such as NIH as well as academia, for example the London School of Hygiene and Tropical Medicine in the UK. Some companies, such as GlaxoSmithKline, have active research programmes in infectious disease, while several are involved with academic institutions in the TB Drug Accelerator (TBDA) partnership, supported by the Bill & Melinda Gates Foundation.

Here we focus on infectious diseases caused by mycobacteria to illustrate the way that structure-guided lead discovery in academia can use its experience in early lead discovery built on biochemistry, medicinal chemistry and structural biology, together with microbiology and human medicine. We show how knowledge of protein structure is central not only to select targets (exploring biological space) but also to develop new leads (exploring chemical space).

**Genomes of Mycobacteriaceae**

The *Mycobacteriaceae* family comprises 174 known species of acid-fast bacteria with high GC content and much thicker hydrophobic outer cell walls (containing mycolic acids) than most other bacteria [http://www.bacterio.net/mycobacterium.html]. The family includes several pathogenic bacteria such as *Mycobacterium tuberculosis* (*Mtb*, the causative agent of tuberculosis), *Mycobacterium leprae* (*Mlep*, causes leprosy), *Mycobacterium abscessus* (*Mab*, rapid-growing opportunistic pathogen that causes chronic lung diseases) and free-living non-pathogenic bacteria such as *Mycobacterium smegmatis*.

*Mtb*, found in *Homo sapiens*, is an obligate pathogen, which has evolved with humans over many thousands of years. According to the most recent WHO report, 10.4 million people were diagnosed with tuberculosis (TB) and 1.4 million died because of the infection in 2015, and an additional 0.4 million TB deaths have been reported among
HIV-positive patients (WHO global TB reports 2016) [39]. Current first-line therapy for drug-sensitive TB involves a combination of four drugs (isoniazid, rifampicin, pyrazinamide and ethambutol) for two months followed by four months of a combination of isoniazid and rifampicin. The rise of multi-drug resistant TB (MDR-TB) and extensively drug resistant TB (XDR-TB) further necessitates prolonged therapy and the addition of second and third line antibiotics such as aminoglycosides, fluoroquinolones and macrolides. These drugs are often poorly tolerated, ineffective against drug-resistant TB or have toxic side effects due to the required prolonged treatments (up to 24 months), thereby leading to significant morbidity and mortality[40].

Although previously thought to be benign environmental microbes, Non-Tuberculous mycobacteria (NTM) are increasingly observed over the past decade, as a cause of infections worldwide[41, 42]. This paradigm shift in NTM prevalence is presumably due to higher environmental exposure rates to such mycobacteria, increased selective pressure imparted by intensive antibiotic use in patients, autophagy-inhibition-associated reduction in host immunity after prolonged antibiotic treatment or patient-to-patient transmission[43, 44]. Mab, a distant relative of Mtb and Mlep[45, 46] is a free-living nontuberculous mycobacterium commonly found in water and soil, which causes skin and soft-tissue infections and pulmonary diseases, particularly in patients with lung disease (such as cystic fibrosis) or a previous history of tuberculosis[47, 48]. A single-centre retrospective study of CF patients showed an increase in incidence of NTM infections from 0 to 9% over a period from 2002 to 2011[49].

After several changes in nomenclature and subspecies differentiation over recent years, the M.abcenss complex is now considered to comprise three distinct subspecies: M.abcenss subsp. abscessus, M.abcenss subsp. massiliense and M.abcenss subsp. bolletii [47, 50, 51]. High virulence and resistance to chemotherapy of Mab complex is attributed to a combination of intrinsic and acquired resistance mechanisms. The intrinsic resistance of Mab and other mycobacteria to drugs results from factors such as low permeability of the mycobacterial cell envelope acting synergistically with antibiotic inducible internal systems like drug-efflux pumps, antibiotic-inactivating enzymes, target-modifying enzymes and genes providing metal resistance [52-54]. Acquired resistance is frequently associated with spontaneous mutations of antibiotic target genes, although alteration in function of one or more other genes may also be involved[52, 53].

The current clinical management of Mab infection in CF is highly challenging due to the high rates of resistance of the bacterium to antibiotics and most of the classical
anti-tuberculosis drugs[45]. Further, owing to its resistance to most commonly used disinfectants, various outbreaks of Mab infections in clinical and post-surgical settings have been reported[55]. Current treatment involves prolonged therapy using antibiotic combinations such as clarithromycin, amikacin, and cefoxitin, which are often poorly tolerated and lead to toxic side effects[47]. Hence, there is an urgent need of effective drugs to treat Mab infection.

For the Mab proteome there are greater challenges, due to the poor annotation of the Mab genome leading to 4920 predicted protein coding sequences[54], nearly one thousand more than for Mtb (3959 coding sequences), and the availability of only 39 experimentally defined structures in the Protein Data Bank (PDB)[56, 57]. A preliminary modelling exercise has predicted structures of over 3000 of these genes[58] & [Malhotra, S & Blundell, TL: unpublished].

**Fragment Screening for Mycobacterium abscessus**

Our approach to lead discovery for Mab targets has been centred on the experience gained with the Gates Foundation HIT-TB and EU-FP7 MM4TB Mtb programmes. To develop the Mab fragment-screening programme it was fundamental to identify essential gene products in this organism that have either a known crystal structures or alternatively an orthologue of a known structure, where we already had an Mtb FBDD programme. This strategy led to targeting PurC (SAICAR synthase), an enzyme involved in bacterial de-novo Purine biosynthesis and TrmD (tRNA-(N'G37) methyl transferase), an essential tRNA modifying enzyme in bacteria, by reproducing the known Mab 3D structures, and selecting CoaD/PPAT (Phosphopantetheine adenyltransferase), where we had a programme for Mtb funded by the Gates Foundation HIT-TB.

Phosphopantetheine adenyltransferase (PPAT) catalyzes the penultimate step in the biosynthesis of CoA in prokaryotes. Coenzyme A (CoA), an essential co-factor for many cellular enzymes involved in biosynthetic, degradative and metabolic pathways, acts as an important acyl group carrier for all organisms[59]. The biosynthesis of CoA consists of a five-step reaction catalysed by different enzymes and utilizing pantothenate (Vitamin B5), cysteine and ATP[60] as starting substrates. PPAT enzyme is a member of the nucleotidylyltransferase α/β phosphodiesterases superfamily and catalyses the reversible transfer of an adenyl group from ATP to 4’-phosphopantetheine (Ppant) to yield 3’-dephospho-CoA (dPCoA) and pyrophosphate,
as shown in Figure 2A. dPCoA in turn is phosphorylated by dephospho-CoA kinase (DPCK) to generate CoA[61, 62].

Although some bacteria are capable of de novo pantothenate synthesis, others utilize extracellular pantetheine for the biosynthesis of CoA. Nonetheless, all intermediate pathways converge at the penultimate step of CoA biosynthesis catalysed by PPAT[63]. Studies of CoA pathway intermediates in Escherichia coli showed that pantetheine and 4’-phosphopantetheine (Ppant) accumulated in the cell, suggesting an important rate-limiting role of PPAT in the pathway that may be crucial to its potential as an antibiotic target[63-65]. In higher eukaryotes, the final two steps of CoA biosynthesis are catalysed by a single bi-functional enzyme called CoA synthase, containing a PPAT like domain. The marked structural differences between bacterial and human PPATs further suggest the potential for this bacterial enzyme as a novel antibiotic target[66-68].

The previously determined PPAT structures from various bacterial species such as Escherichia coli, Staphylococcus aureus, Pseudomonas aeruginosa, Enterococcus faecalis, Burkholderia pseudomallei and Mab demonstrate the presence of a canonical Rossmann fold as seen in many dinucleotide binding proteins. Attempts to develop PPAT inhibitors targeting E-coli PPAT enzyme, using HTS and combinatorial synthesis were only partially successful as none of the resulting candidate compounds showed inhibitory activity at the cellular level[69, 70]. A high throughput screening of the AstraZeneca compound library and subsequent structure-guided lead optimization led to the development of a set of compounds that significantly reduced the bacterial burden of gram-positive bacteria such as S. pneumoniae and S. aureus both in vitro and in animal models of infection[71]. However, these compounds were not progressed into clinical development owing to unfavorable toxicity and pharmacokinetic profiles.

We have built on the fragment-based drug discovery campaign against Mtb PPAT, which has led to the development of potential lead compounds having binding affinities to Mtb PPAT in the low micromolar range (Jamal, E. B., Blaszczyk, M., Blundell, T.L. & Abell, C. unpublished). We solved the crystal structures of Mab PPAT in apo form (PDB code: 5O06), and in a binary complex with the enzyme product, 3’-dephospho Coenzyme A (PDB code: 5O08). Having a working crystallization system, we then initiated a structure-guided fragment-based drug discovery campaign against Mab focused on this target.

The crystal structure shows that the biologically active hexameric form of Mab apo-PPAT consists of two trimers, as seen with all known PPATs from bacteria[72-77].
Each subunit consists of a canonical Rossmann fold[78] made of five parallel β-strands surrounded by six α-helices (Figure 2B & C).

The three-dimensional structure of Mab PPAT in complex with the reaction product, 3’-dephospho Coenzyme A (dpCoA) shows a similar binding mode to the one previously observed for Mtb PPAT, where dpCoA binds to all six sub-units of the hexamer[79]. The reaction product spans the large active-site cavity of Mab PPAT and adopts a ‘bent’ conformation upon binding, as has been previously observed with PPAT: dpCoA structures from other bacterial species[72-77] (Figure 2C & D).

A structural superposition of Mab PPAT:Apo and Mab PPAT:dpCoA shows large conformational changes in the flexible loop II (residue range 36-46) that connects β-strands β2 and α-helix α2 as well as loop IV (residues 87-99) at the base of the active site (Figure 2D & E). In Mab PPAT:dpCoA structure, loop II is found to have shifted towards the active site by about 7Å, when measured at Cα atom of Lys41, as compared to its position in Mab PPAT:Apo structure (Figure 2D & E). This large shift in loop II positions the invariant catalytic Lys41 residue favorably within H-bonding distance of adenylate α-phosphate (2.8 Å) and 4’-phosphate (2.9Å) of dpCoA respectively.

The initial fragment screening of the Cambridge in-house library, consisting of 960 fragments, was performed on Mab PPAT, using differential scanning fluorimetry. Thirty fragments from the library showed a difference in unfolding temperature greater than 1°C upon PPAT binding. For the validation of these fragment hits, the structure of each was investigated by X-ray crystallography, resulting in crystal structures of PPAT in complex with fragments at resolutions ranging from 1.4 to 1.9 Å for 18 out of 30 fragment hits studied. The fragments were found to occupy four distinct binding sub-pockets within the PPAT active site, as shown in (Figure 3C & D). These fragments included examples that occupy the binding sites for adenylyl and 4’-phosphopantetheine regions of dpCoA, thereby spanning the entire active site groove of PPAT. Sub-pockets I & II correspond to the indole and ribose phosphate binding regions of dpCoA respectively whereas III & IV correspond to pantothenate and β-mercaptoethylamine sites of dpCoA respectively. Binding of fragments is mediated by several strong and weak hydrogen bonds, π-interactions as well as hydrophobic and ionic contacts. Figure 4 shows the detailed binding mode of 6 representative fragment hits at sites I to IV.
Many fragments engage key residues, such as His17, Arg90, Val125, Ser9, Leu73 involved in substrate binding and transition state stabilization of PPAT[72], in addition to making new interactions. Fragments binding to sites I & II, for example, mimic some of the strong π-interactions to side chains of Arg 90 and His 17 and hydrogen bonding interactions to the back bone amide of Gly 88, mediated by the adenylyl part of dpCoA (Figures 3A & 4). PPAT binding to dpCoA was found to result in considerable ordering and rearrangement of the loop IV to two complete α-helical turns at the N-terminal of helix4. Several fragments also lead to a similar conformational switching of loop IV near the base of the catalytic site of PPAT as shown in Figures 2E & 3B. Fragments at site III and IV recapitulate interactions of 4′-phosphopantetheine arm of dpCoA, such as hydrogen-bonding contacts mediated by the side chains of Lys41, Ser9, Asn38 and Glu132 as well as hydrophobic contacts to Leu 73, Met101 and Leu129, in addition to several water-mediated hydrogen bonds at the PPAT active site (Figures 3A & 4).

The fragments were further characterized by isothermal titration calorimetry. The thermodynamics of the fragment-binding events showed that many were associated with very low heat changes, which may be attributed to a lower enthalpic contribution to binding, as previously reported for Mtb PPAT-ATP, where interactions are likely driven by entropic component of displacement of water molecules in the active site[80]. For fragment interactions that resulted in a reasonable binding isotherm, the measured dissociation constants were greater than or equal to 500µM (example shown in Figure 5B), while fragment 2 afforded an average $K_d$ of 66µM ($K_{d1}$= 37.3 µM, $K_{d2} = 83.3$ µM, $K_{d3}= 72.5$ µM, $K_{d4} = 76.3$ µM, $K_{d5} =13.9$ µM and $K_{d6} =114$ µM), where the data were fitted using a six-site sequential binding model (Figure 5C). Interestingly, fragment 2 adopts a dual binding mode where two molecules each of the fragment bind to each subunit of PPAT hexamer, at sub-pockets I and II respectively (Figures 3D & 4B).
Fragment Binding & Hotspot Maps

We further analyzed the binding propensities of PPAT protein to ligands with the help of the hotspot-mapping program developed by Radoux and co-workers [81]. Hotspots are often defined as areas within the protein that provide relatively large contributions to the overall binding affinity of ligands[82, 83]. These regions not only satisfy the minimum binding requirement for fragments but also maintain the original fragment binding interactions when elaborated. The importance of hydration thermodynamics is becoming increasingly recognized as a driving force for fragment binding to target proteins[84]. It has been hypothesized that hotspot regions are characterized by the presence of water molecules having restricted translational and rotational freedom owing to their location within a hydrophobic cavity or close to a patchwork of hydrogen bonds and lipophilic amino-acid sidechains. Ligand binding that displaces such water molecules leads to a favorable free-energy change mainly due to an entropic contribution [84]. The hotspot-mapping program identifies interactions that drive fragment binding by a global search of the protein structure, calculating grid-based atomic propensities for specific donor, acceptor and hydrophobic interactions. This is followed by weighting the propensity scores based on buriedness of the grid points and sampling of the weighted atomic scores using three simple molecular probes to create corresponding donor, acceptor and hydrophobic hotspot maps[81].

While the observed fragment hits satisfy many of the predicted protein-hotspot interactions, the map contoured at a lower level suggests potential ways to elaborate the fragments to reach those hotspot regions that are not yet explored (Figure 5A). For example, fragment 2 is found to be mediating the donor and acceptor hotspots via hydrogen bonds to backbone carbonyl of Gly88 by the indazole NH and to active site Ser9 and Phe10 via the carboxylic acid moiety, at sub-pockets I and II respectively. Similarly, fragment 4 is found to satisfy the predicted H-bond acceptor interactions through the mandelic acid moiety forming H-bond to Ser9 and active site water molecules and the phenoxy group of the fragment making hydrophobic contact to Leu73. These fragments may be further elaborated in the direction of the hotspots corresponding to interactions with Asn105, Thr113 and Glu98 at the base of the active site, as shown in Figure 5A.
‘Fragment-like’ Drugs against *Mycobacterial species*

Many critical drugs in current chemotherapy regimen against *Mtb* are indeed smaller than ‘drug-like’ having molecular weights in the range of 100-300 daltons. These include anti-tuberculosis drugs like isoniazid (INH), ethionamide (ETH) and para-aminosalicylic acid (PAS), which are still found to be useful against drug-susceptible TB, as well as pyrazinamide (PZA) a key relapse-reducing drug. Though these drugs require fairly long treatment periods and are less effective against drug-resistant *Mtb*, it is intriguing how these small molecules have retained their killing effect over nearly a century from the time they were developed. Gopal et al[85] in a recent review suggest that molecules that are small, reactive and promiscuous have certain advantages over larger, single-target drugs that may render these molecules effective against mycobacterial species. Compounds like INH, ETH and PZA for example, owing to their small size and promiscuity are more likely to be accepted as substrate analogues and metabolized by bacterial enzymes turning these prodrugs into more reactive compounds capable of hitting multiple-targets, resulting in irreversible damage to the cell. Being small (MW<300Da) and moderately lipophilic (clogP <3) also allows easy penetration of these molecules through the otherwise highly impermeable mycobacterial cell envelope, leading to higher *in-vivo* exposure and favorable pharmacokinetic properties[86, 87][85].

Based on this idea, Moreira and co-workers in a recent study[88] identified 38 fragment hits having bactericidal activity against *Mtb* (MIC< 500 µM), from screening a fragment library of 1725 compounds. Interestingly, 23 out of the 38 identified hits were shown to have bactericidal activity against *M. avium* and 13 hits against *Mab* as well. These hits retained their favourable pharmacokinetic properties and a good fraction of the hits were found to be selective against mammalian cells, though no structure-activity relationship could be established.

Fragment-like molecules were found to potentiate the effect of ethionamide (ETH), a second line TB drug, by 10-fold and showed significant intrinsic bactericidal activity in the absence of ethionamide in a study done by Nikiforov and co-workers[89]. Here, Fragment 1 (MEC = 3 ± 1.8 µM) was identified from screening an in-house library of 1250 fragments against *Mtb* EthR using a cascade of biochemical, biophysical and structural biology approaches. EthR, a transcriptional repressor belonging to TetR family, controls the expression of Flavin-dependent, monooxygenase enzyme EthA which in turn activates ethionamide., a second-line anti-TB drug.
Subsequent structure activity relationship (SAR) exploration around fragment 1 and its analogue 2 led to the development of the potent ethionamide boosters, the amide 14 (MEC = 0.4 ± 0.2 μM) and the carbamate 28 (MEC = 0.4 ± 0.2 μM), which represent approximately 7-fold improvement in boosting the effect of ethionamide in Mtb culture from the starting fragment 1 (MEC = 3 ± 1.8 μM). Further, compounds 14 and 28 showed significant intrinsic bactericidal effect in the absence of ethionamide with (IC50 ≈ 1 μM) and exhibited low nanomolar activity in macrophage assays in the presence of ethionamide[89].

**Chemical Elaboration of Fragments for M. tuberculosis**

For the second stage of FBDD involving elaboration of fragments by growing, merging or linking strategies, we seek to develop a structure-guided approach using the range of biochemical, structural, thermodynamics and kinetic methods. Chemical optimization of fragments involves an iterative process where analysis of binding interactions as well as biophysical, *in-vitro* and *in-vivo* functional testing of the optimized analogues are done to inform further chemical modifications until a significantly potent lead compound is obtained.

Because the fragment-elaboration for the *Mab* PPAT target is ongoing, we use examples of the fragment-merging approach that was successfully applied in our lab and elsewhere to develop small molecule inhibitors of *Mtb*.

**(i) Fragment merging to develop inhibitors against Mtb EthR.**

A fragment-merging approach to develop small molecule inhibitors of *Mtb* EthR led to the development of a potent lead candidate in a recent study[90]. EthR is a physiological dimer having a large hydrophobic binding cavity but containing a number of H-bond acceptors and aromatic residues capable of directed π interactions; the starting fragments 1 & 2 were each found to occupy two distinct sites thus filling the entire cavity as shown in Figure 6A & B. Merging of Fragments 1 & 2 further resulted in a significantly potent lead molecule, compound 15 (Figure 6C) affording an IC50 of 3μM[90].
(ii) Fragment elaboration to inhibit a \textit{Mtb} cytochrome P450 (CYP) enzyme

A fragment elaboration approach by Hudson & co-workers led to the development of high affinity ligands against CYP121, a cytochrome P450 (CYP) enzyme[91]. The mycobacterial genome encodes for at least 20 different CYPs that play an important role in virulence and survival of the bacteria[92]. CYPs belong to a family of mono-oxygenase enzymes containing heme as a co-factor that catalyze the oxidation of organic molecules and thereby mediate diverse metabolic processes. CYP121 is found exclusively in \textit{Mtb} and helps in the production of mycocyclosin, a secondary metabolite, by catalyzing the formation of a C-C bond between the two tyrosine residues of the substrate cyclodityrosine (cYY). Rv2276, the gene encoding for CYP121, was found to be an essential \textit{Mtb} gene, suggesting that the product mycocyclosin has an important cellular role or an overproduction of the substrate cYY is toxic to the organism[93, 94]. A fragment screening approach that combines fluorescence-based thermal shift assay, NMR (STD and WaterLOGSY) and X-ray crystallography helped in the identification of 4 hits from a fragment library of 665 compounds adopting two distinct binding modes[95].

Elaboration of these fragments hits resulted in several promising derivatives (Figure 6D). In particular, the two different binding conformations of fragment 10 ($K_d = 1.7 \text{mM}$) was exploited to synthesize compound 12. Although this was less active than the starting fragment, further replacement of its five-membered ring with an aminopyrazole led to compound 13 with a $K_d$ of 40 $\mu \text{M}$. Growth of compound 13 into the water-filled cavity of CYP121 resulted in compound 14, with an additional phenol ring which mediated further H-bonds with the active site water molecules, improving the affinity by over 100-fold from the starting fragment ($K_d$ of 15 $\mu \text{M}$)[91, 95, 96].

Elaboration of the lead compound 14 by Kavanagh and colleagues and the addition of a 3amino-phenyl group at the meta position of the aromatic ring attached to five-membered ring, led to compound 15 ($K_d$ of 0.015$\mu \text{M}$, LE=0.41), the highest affinity inhibitor of CYP121 to date as shown in Figure 6D [97].
(iii) Inhibiting *Mtb* Thymidylate kinase

Thymidylate kinase (TMK) is an enzyme that belongs to the DNA synthesis pathways and that produces thymidine 5’-diphosphate using ATP and thymidine 5’-monophosphate. This enzyme is an established drug target and has been the focus for many drug discovery campaigns against different pathogens, in both eukaryotes[98] and prokaryotes, including *Mtb*[99, 100] and also viruses[101]. The essentiality of this enzyme is established in *Mtb* and drug discovery efforts with this target identified a number of inhibitors some of them with active in cell based assays. However, the vast majority of the compounds obtained are analogues of thymidine and contain a thymidine moiety[100]. To move away from these scaffolds, Naik and colleagues carried out a fragment screening of TMK, combining biochemical assay and NMR, which resulted in the identification of a 3-cyanopyridone containing fragment for chemical elaboration[102]. The mode of binding of this fragment hit was determined by molecular docking. Subsequent development of the fragment, supported by X-ray crystal structures of key compounds, allowed the development of nanomolar-potency lead compounds against TMK that retained high ligand efficiency. These compounds were found to have different binding modes from the substrate analogues and some were found to inhibit *Mtb* in whole cell assays.

CONCLUSIONS

The use of protein structure, defined by experimental and computational approaches for renin and HIV protease, developed for target identification, validation and development of new leads in the 1980s and 1990s, has provided a paradigm for use of structural biology to underpin drug discovery. The progress in structure-guided drug discovery has been impressive, but there remain many challenges for the future. They lie not only in more efficient exploration of biological and chemical space at the molecular level, but also in understanding the many aspects of cell and whole organism biology that affect ligand transport, permeability and metabolism, as well as the changes in complex assemblies over space and time.

Progress in defining biological space requires more efficient approaches to improve experimental structures. This may come from X-ray methods using conventional crystallisation techniques and from X-ray small angle scattering. However, some of the most exciting new developments have come from single-particle cryo-EM methods, for example those developed with the new powerful electron microscopes.
such as the FEI Titan Krios transmission electron microscope (TEM) combined with direct electron detectors[103], and new software such as RELION[104], which carries out 2D image classification followed 3D reconstruction. These revolutionary changes have allowed binding of antibiotics to be envisaged on ribosome particles[105] [106]. This approach will be particularly powerful at defining structures of the complex systems involved in cell regulation and signalling, where interfacial inhibitors or stabilisers are of interest.

Experimental methods of defining biological and chemical space will likely continue to be challenging even for organisms with relatively small genomes. Structures are required to define interactions of macromolecular and small ligands with proteins. The impacts of strain variation, including mutations that are selectively advantageous in evolution as well as disadvantageous in disease, will need to be understood at the level of protein structure. Thus, it is inevitable that computational methods will need to be improved not only to predict the structures and dynamics of monomers and their interactions with small molecules, but also the multicomponent systems often involving proteins that are intrinsically disordered in parts or the whole of their polypeptide chains; these are still very challenging.

As we have seen, fragment-based drug discovery has improved the exploration of chemical space, especially where targets are well described in terms of structure. Fragments that are stably bound and maintain their positions during elaboration bind at hotspots. Some progress has been made in recognising these from the structure as we have described, for example Radoux et al., 2016 [81]. It is now becoming clear that there is wide range of interaction types in addition to H-bonds and simple lipophilic interactions, including $\pi$-$\pi$, $\pi$-$\delta$ -positive and $\pi$-$\delta$ -negative charge, and these can be displayed using software such as Arpeggio [107] & Intermezzo [Ochoa-Montaño, B., Blundell, T.L., unpublished].

Fragment-based approaches have proved successful in getting drugs to the market in cancer, and the lessons learnt are proving helpful in the design of antimicrobials, as we have described. The FBDD approach allows efficient exploration of the chemical space against targets like those of mycobacteria, where features, such as the bacterial cell wall, demand the investigation of molecules that are likely different - in size, lipophilicity and other properties - than those optimised for human protein targets.

Advances in protein crystallography using high-throughput, roboticized approaches at synchrotrons offer further opportunities. Frank von Delft and colleagues have developed automated methods to soak crystals with fragments and mount them in the
X-ray beam. Perhaps the most impressive advance has been in PanDDA[108], a method that reveals subtle differences when the ligand is partially occupied. The PanDDA method reveals electron density for only the changed state, even from poor models and inaccurate maps, by subtracting a proportion of the apo state, accurately estimated by averaging many apo-protein crystals.

Fragment based approaches can screen a vast chemical space using a relatively small library of fragments with a set of desired physiochemical properties that can be retained and controlled during the structure-guided optimization of the fragment hits. Finding inhibitors against mycobacteria is a highly challenging task due to the impermeability of cell wall, multidrug efflux-pumps and drug modifying enzymes that these organisms possess. Small hydrophilic molecules are known to cross the mycobacterial cell wall through porins but large macrocycles and small hydrophobic molecules also diffuse through the cell wall, resulting in TB drugs being widely distributed across chemical space[109]. Solving the structure of efflux pumps in these organisms will surely provide insights into the efflux mechanisms and help identify chemical scaffolds that are extruded from cells, thereby allowing FBDD approaches to optimise the chemistry of new lead molecules.

The structure-guided fragment-based approach avoids the requirement for large screening libraries and provides a route to the discovery of new therapeutics for diseases that lead to many deaths throughout the world and gives hope for effective medicines for rare diseases and those that affect the large populations of less affluent parts of the world. Knowledge of protein structure has contributed to the development of early stages of drug discovery and is widely exploited in both industry and academia.

Acknowledgements: We thank Dr.Andres Floto and team for kindly providing \textit{M.\textit{abcessus}} genomic DNA. Prof. Chris Abell and Dr.Anthony Coyne for access to the in-house fragment library collection. The authors would like to thank Diamond Light Source for beam-time (proposals mx9537, mx14043) and the staff of beamlines I03, I04, I04-1 and I24 for assistance with data collection. SET is funded by the Cystic Fibrosis Trust (Registered as a charity in England and Wales (1079049) and in Scotland (SC040196); VM and BOM are funded by the Bill and Melinda Gates Foundation; SM is funded by the Medical Research Council (MRC Newton/DBT Grant: RG78439); TLB is funded by the Wellcome Trust (Wellcome Trust Investigator Award: 200814/Z/16/Z).
**ACCESSION NUMBERS:** Coordinates and structure factors have been deposited in the Protein Data Bank with accession numbers: 5O06; 5O08; 5O0A; 5O0B; 5O0C; 5O0D; 5O0F; 5O0H.

**Figure Legends:**

**Figure 1:** A) Schematic to illustrate the concept of drug discovery by exploring biological space encompassing genome information to identify new targets, followed by exploration of chemical space using screening libraries based on knowledge of the target. Target protein in the illustration: HIV-1 proteinase (PDB code: 3PHV). Drug: HIV proteinase inhibitor: saquinavir (Roche) B) Fragment Based Drug Discovery workflow illustrated as four stages involving 1. Target Identification and protein production 2. Fragment library screening 3. Hit validation and 4. Iterative fragment elaboration cycle through to a potential lead candidate. Various tools and techniques involved at each stage of the process are also shown.

**Figure 2:** A) Schematic depicting bacterial Coenzyme A biosynthetic pathway. The penultimate step catalyzed by PPAT enzyme is boxed and the gene names corresponding to each enzyme is given in brackets. Many organisms are capable of exogenous uptake of the starting substrate Pantothenate, while some bacteria and fungi undertake de-novo Pantothenate synthesis from β-alanine and Pantoate, catalysed by Pantothenate synthetase (panC). PPCS* and PPCDC* activities reside within a single polypeptide chain in bacteria. B) Crystal structure of *Mab* PPAT:Apo solved at 1.5Å resolution showing a hexameric assembly. C) Protomer of *Mab* PPAT bound to dpCoA (purple ball & stick representation) forming a ‘bent’ conformation at active site. The secondary structure elements are coloured with β-strands in yellow, α-helices in red and loop regions in green. D) Structural superposition of *Mab* PPAT:Apo (green) and *Mab* PPAT:dpCoA (blue) at Cα atoms showing conformational changes in the flexible loop II and loop IV. dpCoA, represented as spheres, is seen bound to all the three protomers in the asymmetric unit. E) Enlarged view showing conformational changes at loop II positioning the invariant catalytic Lys41 residue close to adenylyl α-phosphate of dpCoA (shown in pink stick representation) and ordering of loop IV to two complete α-helical turns at the N-terminal of helix4 at the base of catalytic site along with inward flipping of the Arg90. Other invariant residues interacting with DpCoA, His17 and Ser9 are also shown. The figures were prepared using Pymol (PyMOL Molecular Graphics System, Version 1.7 Schrödinger, LLC). PDB codes: *Mab* PPAT:Apo: 5O06, *Mab* PPAT:dpCoA: 5O08.
**Figure 3:** A) Active site diagram showing binding interactions of dpCoA shown in pink ball and stick model and the active site residues as green lines. Hydrogen bonds are represented with blue dotted lines, \(\pi\)-interactions in black dotted lines, hydrophobic contacts in red dotted lines. B) Overlay of *Mab* PPAT:Apo (green) and *Mab* PPAT: fragment 1(orange) structures showing conformational changes in loop IV invariant Arg90, His17, and Ser127 are represented in blue lines. D & C) Structural Superposition of six *Mab* PPAT structures bound to fragments 1(purple), 2(gold), 3(green), 4(blue), 5(salmon) and 6(turquoise), showing fragments occupying four distinct regions (I-IV) in the PPAT active site shown in surface electrostatic representation. (PDB codes: 5O0A, 5O0B, 5O0C, 5O0D, 5O0F, 5O0H). The figures were prepared using Pymol (PyMOL Molecular Graphics System, Version 1.7 Schrödinger, LLC) and Intermezzo (Ochoa-Montaño, B., Blundell, T.L. et al unpublished; Jubb, H.C., et al, 2017 [105]).

**Figure 4:** Active site diagram of *Mab* PPAT showing binding interactions of fragments 1-6 (shown in orange ball and stick model) from A to F respectively and the active site residues as blue lines. Fragment contacts are represented as dotted lines, hydrogen bonds are represented in blue, \(\pi\)-interactions in black, hydrophobic contacts in red and ionic contacts in green dotted lines respectively. The figures were prepared using Pymol (PyMOL Molecular Graphics System, Version 1.7 Schrödinger, LLC) and Intermezzo (Ochoa-Montaño, B., Blundell, T.L. et al unpublished; Jubb, H.C. et al, 2017[105]). (PDB codes 5O0A, 5O0B, 5O0C, 5O0D, 5O0F, 5O0H)

**Figure 5:** A) Hotspot map (Radoux et al, 2015) [79] of *Mab* PPAT:Apo protein with *Mab* PPAT structures bound to fragments 1(purple), 2(gold), 3(green), 4(blue), 5(salmon) and 6(turquoise), overlaid. The hydrophobic hotspot map is shown in yellow, donor and acceptor hotspots are shown as blue and red surfaces respectively. The major residues mediating interactions at the hotspots are shown in white stick representation. B) Representative thermodynamic profile for fragment binding to *Mab* PPAT. The figures were prepared using Pymol (PyMOL Molecular Graphics System, Version 1.7 Schrödinger, LLC). (PDB codes: 5O0A, 5O0B, 5O0C, 5O0D, 5O0F, 5O0H)
Figure 6: Fragment elaboration to develop inhibitors of *Mtb*. **A)** Fragment merging approach to develop small molecule inhibitors of *Mtb* EthR. The region corresponding to the hydrophobic binding cavity of EthR dimer is boxed. **B)** Binding pocket of EthR where the starting fragments 1 (in gold) & 2 (in grey) were each found to occupy two distinct sites in two copies thereby filling the entire cavity involving sub-pockets I, II, III and IV (marked in blue). **C)** Merging of Fragments 1 (in gold) & 2 (in grey) results in a significantly potent lead, compound 15 (green). The corresponding thermal shift and IC\textsubscript{50} values for each compound are also illustrated. PDB codes: 5FUJ, 5F27, 5F0F. **D)** Fragment elaboration to develop compounds that inhibit *Mtb* CYP121. Two molecules of Fragment 10 (green) binds CYP121 near the Heme (yellow) binding pocket. Subsequent merging and elaboration of fragment 10, via compounds 13 and 14 (blue), led to the development of potent compound 15 (pink). The corresponding *K*\textsubscript{d} and Ligand efficiency values for each compound are also illustrated. PDB codes (4G47, 4KTL, 5IBE). The figures were prepared using Pymol (PyMOL Molecular Graphics System, Version 1.7 Schrödinger, LLC).
References


27


Figure 1
Figure 2
Figure 3
Figure 4
Figure 5
Figure 6 a-c
Figure 6 d
Graphical abstract