Submission: Sophie Allen
Protein: 1BVR Enoyl-ACP Reductase (InhA) - Rozwarski D.A. et al. (1998a)

Structural Bioinformatics Coursework (December 2022)

Part A: Analysis and Description of the Structure

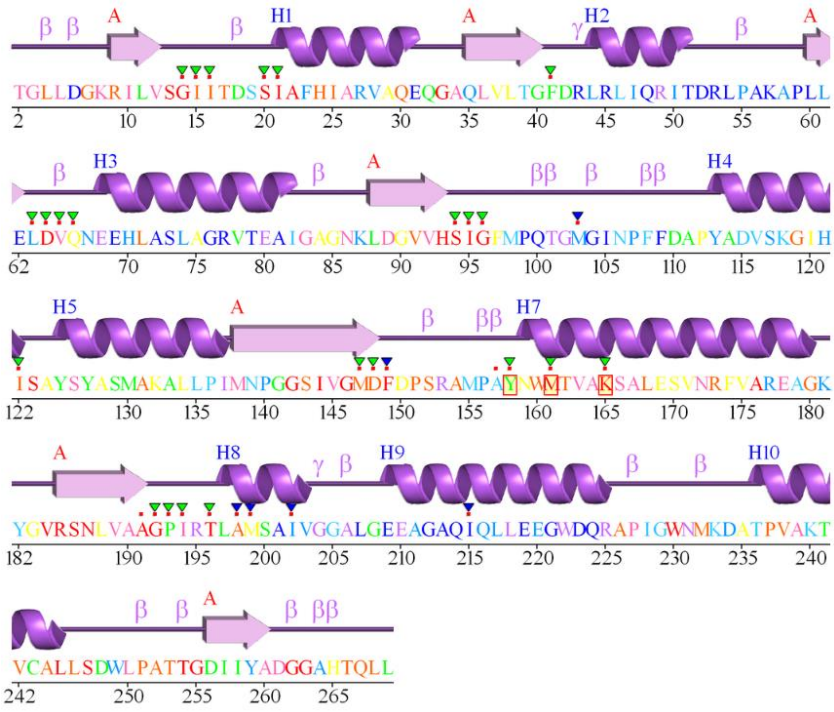
a) InhA catalyzes a two-step reaction that allows extension of long chain fatty acyl-CoA molecules in Mycobacterium Tuberculosis by using NADH as a cofactor to reduce one of the trans double bonds in a fatty-acyl at least 16 carbon atoms long, and this process produces a precursor molecule in the production of mycolic acid (Rozwarski *et al.*, 1999).

C16-Enoyl-ACP (fatty-acyl) + NADH → Enolate Anion Intermediate → C16-Acyl-CoA + NAD^+ (Vögeli *et al.*, 2018)

b) CATH Structural classification (http://cathdb.info/version/latest/superfamily/3.40.50.720/classification): NAD(P)-binding Rossmann-like Domain (Sillitoe et al., 2021).

c) InhA is a tetrameric enzyme with a buried surface area per subunit of 29%, a comparatively large proportion of each subunit involved in contact with the other subunits (Rozwarski *et al.*, 1999). Each subunit is 268 residues long, and forms one domain with a Rossmann fold that forms an unusually deep active site crevice for the enzyme class (Rozwarski *et al.*, 1999). The crevice has two entry points to accommodate binding first by NADH via the major entry portal, and then entry by the fatty acyl ligand via the minor entry portal. The Rossmann fold is a common fold for proteins binding to NAD, and forms from beta-alpha-beta motifs and tight turns between beta strands and alpha helices (Hanukoglu, 2015), which is clear from the PDBsum PROMOTIF analysis (Hutchinson and Thornton, 1996) (Figure1A), showing 5 clear motifs which form the Rossmann fold (Figure 1B).

A)



Submission: Sophie Allen
Protein: 1BVR Enoyl-ACP Reductase (InhA) - Rozwarski D.A. et al. (1998a)

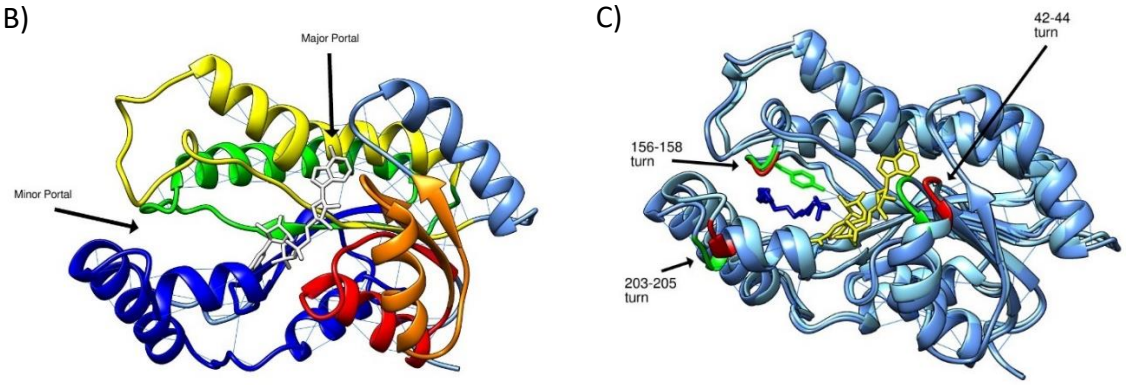


Figure 1: A) Plot of secondary structures for a subunit of InhA (taken from the PDBsum PROMOTIF analysis, (Hutchinson and Thornton, 1996), https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=1bvr&template=protein.html&o=RESCONS&l=1&chain=A&c=999&r=wiring)
B) Structure of active form of protein bound to NAD+. Each beta-alpha-beta motif involved in the Rossmann fold to hold NAD+ in place are coloured in red, orange, yellow, green, and dark blue, with NAD+ coloured in white. Also labelled are the major and minor entry portals for the crevice. Model produced using USCF Chimera (Pettersen *et al.*, 2004).
C) The three gamma turns highlighted. The position of each loop for an inactive form of the protein is shown in red, while the active form positions are shown in green. NAD is in yellow, and the THT fatty acyl ligand is in dark blue. Also shown for the active model is the confirmation of the Tyr158 residue positioned over/within the minor portal. Model produced using USCF Chimera (Pettersen *et al.*, 2004).

Subunits have one 7-stranded parallel beta sheet and 10 alpha helices, including two perpendicular helices forming the binding loop that the substrate binds to. Although 43.3% of the structure is made from unstructured loop regions and turns, (Hutchinson and Thornton, 1996), these loops are largely held in place by hydrogen bonds, such as the hydrogen bond between Ala157 and Gly104. In the active form of the enzyme, the binding loop (residues 196-219) moves 4Å to the left, widening the crevice opening and opening a new entryway leading from the minor entry portal of the crevice to the active site, allowing entry of the fatty acyl ligand (Rozwarski *et al*., 1999).

There are three gamma turns (42-44, 156-158, and 203-205) (Figure 1C). When compared to an inactive form of the protein (Dias *et al.*, 2007), the turn at 42-44 is formed in the active state of the protein over the major portal (Figure 1C), therefore this loop may hold the bound NAD in place while catalysis takes place. The 156-158 turn is also important to function, since the turn may be maintained but the Tyr158 residue rotates 60 degrees upon cofactor binding, therefore opening the minor portal to allow substrate entry (Rozwarski *et al*., 1999).

d) The Met1 residue and side-chain O-C in Thr2 residues for all subunits in the tetramer are missing from the structure, but no other atoms are missing. R and Rfree were good even before refinement in adding water molecules to artificially improve R, where R was 0.234 before and 0.217 afterwards, while Rfree was 0.356 (Rozwarski *et al*., 1999), both of which are considered good scores for solved structures and indicate good quality of the model.

B-factor of 51.7 was within expected range for a structure of overall 2.8Å resolution (Carugo, 2018), however B-factor varied greatly for each chain (Chain A had B-factor range of 18-104). Some residues may either have local poor quality or multiple residue conformations. Electron density is overall fairly poor, and the structure did not align well with the density until below 0.3 volume. Resolution varies between 2.8Å and 10Å considering all subunits (Rozwarski et al., 1999), which explains the poor-quality electron density map and indicates a poorer structure overall.

Submission: Sophie Allen
Protein: 1BVR Enoyl-ACP Reductase (InhA) - Rozwarski D.A. et al. (1998a)

From the PDB file (Rozwarski *et al.*, 1999), 4 residues had bond angles deviating more than 6 RMSD from the Engh and Huber values (Engh and Huber, 1991) - Pro227, Pro107, Pro193, Leu217. Additionally, on average per chain only between 21-51% of residues were deemed to be high quality with respect to geometric quality tests. (https://files.rcsb.org/pub/pdb/validation_reports/bv/1bvr/1bvr_full_validation.pdf) (Rozwarski D.A. et al. (1998a)).

Table 1: Percentage of outlier residues per chain. Adapted from the PDB validation report for this structure (Rozwarski et al., 1999). The more outliers a residue has (from 0 up to >3), the more geometric quality tests that residue has failed.

Chain	0 Outliers (%)	1 Outlier (%)	2 Outliers (%)	3 or more Outliers (%)	Residues poorly fit to density map (%)
A	50	45	5	0	1
B	51	43	5	0	1
C	49	46	5	0	2
D	47	44	8	0	2
E	21	53	23	3	0
F	47	47	6	0	3

PROCHECK (http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=1bvr&template=procheck_summary.html) (Laskowski *et al.*, 1993) and the Ramachandran plots from UCSF Chimera (Pettersen *et al.*, 2004) were used to further analyse torsion angles. 64 residues were identified as outliers from the Ramachandran plot, several of these were Tyrosine residues (Figure 2A, 2B). These were Tyr158 residues, which are known to undergo a 60 degree rotational confirmation change on cofactor binding (Rozwarski *et al.*, 1999), and since this structure represents the active form of the protein, this is likely the reason for the outlying torsion angle for these residues. This only explains some residues however, and Ramachandran plot quality assessment from PROCHECK confirmed poor quality, as overall assessment was outside of the expected region (Figure 2C). Although some of the poor-quality features of the structure can be explained by the function of the active site, the model is still of general poor quality, which could be attributed to the resolution.

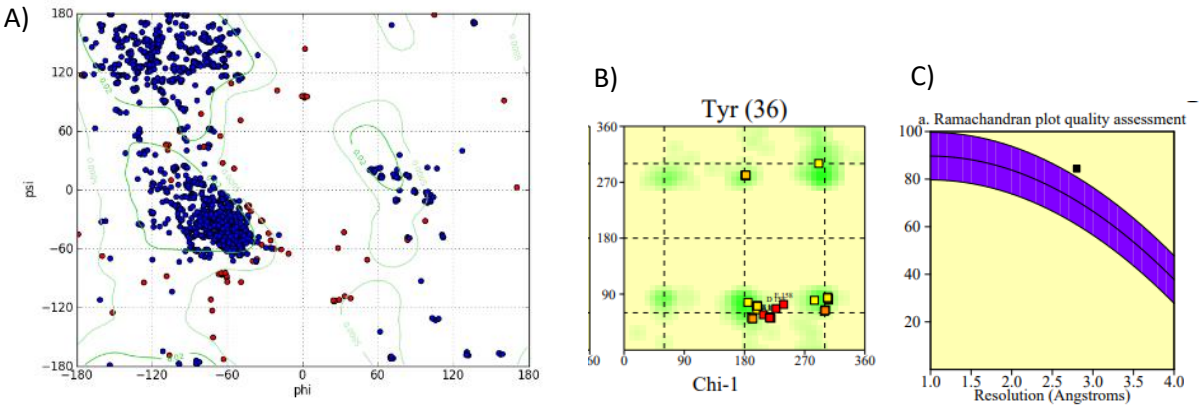


Figure 2: Bond analysis plots. A) Ramachandran plot, with potential outliers highlighted in red. B) Tyrosine bond angle analysis, with outliers highlighted in red, and C) overall plot quality from PROCHECK (actual plot quality plotted as the black square, with the curve region of expected quality in purple. Plots produced in PROCHECK (http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=1bvr&template=procheck_summary.html) (Laskowski *et al.*, 1993) and UCSF Chimera (Pettersen *et al.*, 2004).

e) NAD+ (cofactor to facilitate the *in vitro* reaction), and trans-2-hexadecenoyl-(N-acetylcysteamine)-thioester (THT).

f) The below are the identified contacts between the protein and the THT ligand:

Table 2: Groups in contact between THT ligand and protein active site. (adapted from Rozwarski et al., 1999)

Atom in Ligand	Residue	Atom in Residue	Distance of Contact	Interaction Type
O	Tyr^{158}	OH	3.7	Hydrogen Bond
C1	Met^{103}	CE	4.2	Hydrophobic Interaction
C1	Met^{199}	CG	4.9	Hydrophobic Interaction
C2	Met^{199}	CE	4	Hydrophobic Interaction
C3	Phe^{149}	CE2	4	Hydrophobic Interaction
C4	Phe^{149}	CE2	3.2	Hydrophobic Interaction
C4	Met^{199}	CE	3.9	Hydrophobic Interaction
C4	Pro^{193}	CG	4.2	Hydrophobic Interaction
C5	Phe^{149}	CZ	3.1	Hydrophobic Interaction
C5	Tyr^{158}	CE2	3.3	Hydrophobic Interaction
C5	Met^{199}	CE	4.5	Hydrophobic Interaction
C5	Pro^{193}	CG	4.8	Hydrophobic Interaction
C6	Met^{199}	CE	3.9	Hydrophobic Interaction
C6	Leu^{218}	CD2	4.4	Hydrophobic Interaction
C6	Pro^{193}	CG	4.8	Hydrophobic Interaction
C7	Tyr^{158}	CD2	3.6	Hydrophobic Interaction
C7	Leu^{218}	CD1	4.7	Hydrophobic Interaction
C8	Ile^{215}	CG1	3.4	Hydrophobic Interaction
C8	Ala^{157}	CB	3.8	Hydrophobic Interaction
C8	Leu^{218}	CD1	5	Hydrophobic Interaction
C9	Ile^{215}	CD1	3.4	Hydrophobic Interaction
C9	Met^{199}	CE	4.1	Hydrophobic Interaction
C9	Ala^{157}	CB	4.4	Hydrophobic Interaction
C10	Ala^{157}	CB	3.8	Hydrophobic Interaction
C10	Ile^{215}	CD1	3.8	Hydrophobic Interaction
C10	Met^{103}	CG	4.4	Hydrophobic Interaction
C10	Gly^{104}	CA	4.6	Hydrophobic Interaction
C11	Met^{103}	CE	3.4	Hydrophobic Interaction
C11	Met^{103}	CG	3.7	Hydrophobic Interaction
C11	Gly^{104}	CA	4.9	Hydrophobic Interaction
C12	Met^{103}	CE	3.9	Hydrophobic Interaction
C12	Ile^{202}	CG1	3.9	Hydrophobic Interaction
C12	Ile^{215}	CD1	4.5	Hydrophobic Interaction

C13	Ile^{202}	CG2	3	Hydrophobic Interaction
C13	Met^{103}	CE	4.5	Hydrophobic Interaction
C13	Leu^{207}	CD2	4.7	Hydrophobic Interaction
C14	Met^{103}	CE	3.8	Hydrophobic Interaction
C14	Ile^{202}	CG2	4	Hydrophobic Interaction
C14	Leu^{207}	CD1	5	Hydrophobic Interaction
C15	Ala^{198}	CB	3.7	Hydrophobic Interaction
C15	Ile^{202}	CG2	4.4	Hydrophobic Interaction
C16	Ala^{198}	CB	4.2	Hydrophobic Interaction
C16	Ala^{201}	CB	4.8	Hydrophobic Interaction
C16	Leu^{207}	CD1	5.1	Hydrophobic Interaction

The majority of these interactions are hydrophobic interactions between the non-polar carbon chain of the THT ligand, with one hydrogen bond being formed between Tyr158 and the C=O group within the enoyl 'head' of the ligand. Interactions were visualized using the Protein-Ligand Interaction Profiler PLIP (https://plip-tool.biotec.tu-dresden.de/plip-web/plip/index) (Adasme,M.F. et al. (2021)):

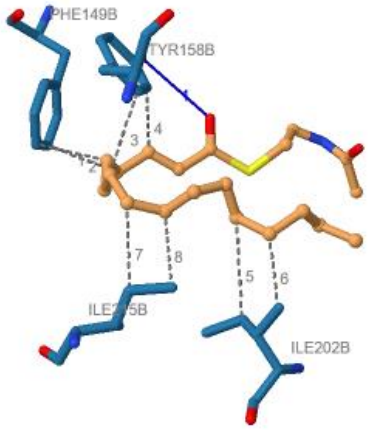


Figure 3: Visualization of recognised interactions between THT ligand and protein. Hydrogen bonds are shown in blue, and hydrophobic interactions shown as grey dashed lines. Produced with PLIP (https://plip-tool.biotec.tu-dresden.de/plip-web/plip/index) (Adasme *et al.*, 2021)

g) Binding affinity: The hydrophobic interactions along the carbon 'tail' of the ligand (C7-C16) are most important to binding affinity. As hydrophobic interactions such as van de Waals forces are weak, these are most likely to be important in increasing binding affinity of the fatty acid tail for the active site, and are shown to form with non-polar residues such as Ile202 and 215 in the active site (Figure 3). As these interactions are weak and easy to break, many forces form between the highly non-polar protein binding site and every carbon in the ligand to increase affinity as much as possible.

Catalysis: The only interaction which is not hydrophobic is the hydrogen bond between Tyr158 and the enoyl C=O oxygen group of the ligand. Tyr158 is within a highly conserved catalytic triad consensus sequence for the SDR family of enzymes (Tyr159 xxx Lys163 (Baker, 1995), shown in E.coli 7alpha-HSDH, where the Tyrosine residue is shown to interact with the fatty acid ligand). The consensus sequence within InhA is Phe149-Tyr158-Lys165 (Rozwarski et al., 1999), and as Tyr158 has interaction with the fatty acid as well, Tyr158 is involved with the catalytic process. It has been proposed that the hydrogen bond stabilizes the enolate intermediate in the reaction as it directly

hydrogen bonds with the carbonyl oxygen, which isn't involved in the reduction itself but Tyr needs to hold it in place to keep the fatty acid chain in place while the catalysis takes place.

The hydrophobic interaction between Phe149 and fatty acid is also important to the reaction rather than binding affinity. Aromatic residues are highly conserved at this position (Rozwarski et al., 1998b), indicating critical importance of this residue for the enzyme to function. During the catalytic reaction, a hydride ion transfers between the NAD ring of the cofactor and the enoyl carbon in the fatty acid chain – and as Phe149 is 4Å from both the NAD ring and the ligand, it is possible that the residue could direct ion transfer between NADH and the ligand, enabling cleavage of the acyl-CoA group (Rozwarski et al., 1998b).

Part B: Identifying potential inhibitors

a) A SMILES string for the THT ligand was available within the ZINC database (https://zinc15.docking.org/substances/ZINC000014880555/) (Sterling and Irwin, 2015), which was queried using the InChI Key for this molecule available on the PDB page for the complex (Rozwarski D.A. et al. (1998a)),(Berman *et al.*, 2000) (InChI used: GDVZALUOXPTSHD-UHFFFAOYSA-N).

SMILES string: CCCCCCCCCCCCCCCC(=O)SCCNC(C)=O

b) This molecule obeys Lipinski's rule of 5:

1. No more than 5 hydrogen bond donors: this molecule has 1 hydrogen bond donor (-NH), therefore obeys this rule.
2. No more than 10 hydrogen bond acceptors: this molecule can accept 4 hydrogen bonds (2 for each C=O bond), therefore obeys this rule.
3. A molecular weight less than 500 D: this molecule has a MW of 357.6 from the ZINC database (Sterling and Irwin, 2015), and therefore obeys this rule.
4. An octanol/water partition coefficient not > 5: Also from the ZINC database page (https://zinc15.docking.org/substances/ZINC000014880555/) (Sterling and Irwin, 2015), logP for this molecule is reported to be 5.864, which is above 5. However, while the molecule does not obey this rule, since molecules which violate only one of the rules are considered to be accepted, this molecule obeys Lipinski's rule of 5.

c) As there are multiple 'canonical' SMILES strings, similar molecules to the ligand were identified using a search for similar substructures. In this case, since the region of the ligand within the active site (ie in contact with the protein and the cofactor) is the enoyl group, this was chosen to be the substructure searched.

A list of candidate molecules were identified by searching the ChEMBL database (Mendez *et al.*, 2019)(https://www.ebi.ac.uk/chembl/) and performing a search for the substructure using the in-built chemical structure sketch API. SMILES string for the substructure was CC(=O)NCCSC(C)=O, and 236 molecules were identified from this search.

d) To generate the Tanimoto similarity scores for the 236 molecules from ChEMBL, first the full ligand SMILES string was written as a .smi file. Then, the following commands were used with the babel binary from OpenBabel (O'Boyle *et al.*, 2011) (The Open Babel Package v2.3.2, http://openbabel.org/wiki/Main_Page):

```
babel chembl_search.sdf chembl_id.sdf –append "chembl_id"
babel tht_ligand.smi chembl_id.sdf -ofpt tanimoto.sdf
```

Submission: Sophie Allen
Protein: 1BVR Enoyl-ACP Reductase (InhA) - Rozwarski D.A. et al. (1998a)

This appends the ChEMBL ID to each record for later retrieval, and generates a Tanimoto score of similarity for each of the ChEMBL molecules against the ligand using the default FP2 hashed fingerprint of each molecule (linear hashing algorithm used by babel that identifies fragments up to 7 atoms long and hashes into a 1024-bit vector (https://openbabel.org/wiki/FP2)).

To generate a histogram and other statistics about these scores, the output .sdf file was imported into R Studio (RStudio Team, 2020) and the following commands were used with R (v4.2.0, R Core Team 2021) and the Tidyverse package (Wickham et al., 2019):

```
library (tidyverse)

read_tsv ("tanimoto.sdf") -> tanimoto

# cleaning the imported dataframe for analysis

tanimoto %>%
  rename ("score" = ">") %>%
  separate (score, c("wording","Tanimoto_Score"), sep = "= ") %>%
  select (-1) %>%
  mutate (Tanimoto_Score = as.numeric (Tanimoto_Score)) %>%
  filter (!is.na (Tanimoto_Score)) -> tanimoto_cleaned

# generating histogram

ggplot (tanimoto_cleaned, aes (Tanimoto_Score)) +
geom_histogram (color = "black", fill = "aquamarine") +
  geom_vline (aes (xintercept = mean (Tanimoto_Score)), linetype = "dashed") +
  labs (title = "Histogram - Tanimoto Scores (ChEMBL Search)", y = "Count") +
  theme_classic()

# mean and standard deviation

mean (tanimoto_cleaned$Tanimoto_Score)
sd (tanimoto_cleaned$Tanimoto_Score)
```

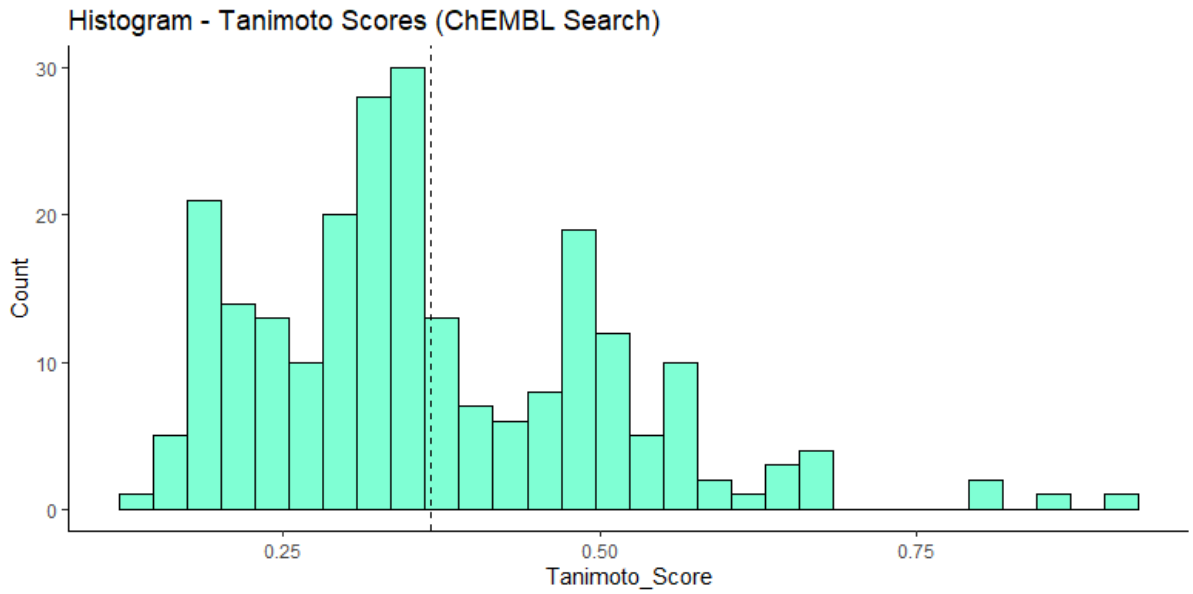


Figure 4: Histogram of Tanimoto similarity scores for the THT ligand against the identified 236 ChEMBL molecules. The dashed line represents the mean score for these molecules.

The mean score was 0.3658743, and the standard deviation was 0.1395929.

e) Since the fingerprints used were binary hashed fingerprints, the Tanimoto score was used to measure similarity to the ligand.

Four molecules had a Tanimoto score above 0.75. The highest of these, CHEMBL244985, is a small molecule with a short carbon chain attached to the searched enoyl substructure (SMILES string CCCC(C)C(=O)SCCNC(C)=O), and while not as non-polar as the ligand, may be a candidate for further study to examine if this may bind to the active site of the enzyme since the enoyl group end heavily resembles that of the ligand.
(https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL244985/)(Mendez,D. et al. (2019))

Part C: Docking to a model structure
a) Sequence first found by searching UniProt, (https://www.uniprot.org/uniprotkb/P9WGR1/entry#structure) (Wang *et al.*, 2021), displaying the sequence published by Camus *et al.*, 2002.
Using GenTHREADER (http://bioinf.cs.ucl.ac.uk/psipred) (McGuffin and Jones, 2003), the template chosen was **5TF4 (PDB ID)**. GenTHREADER was chosen as in the scenario there are no high sequence identity homologs identified through UniProt or PDB searches. For this template, confidence was 'CERT' (p-value of below 0.0001, very high likelihood of homology). Looking at the alignment output, the first two hits were above 70% sequence identity and were excluded (100% and 89.9% identity respectively), the third hit had identity of 28.4% while retaining a p-value below 0.0001 and an alignment length covering most of the queried sequence (249/269, 93% of queried sequence aligned). Template was also functionally similar, being another enoyl-acyl carrier protein reductase (same enzyme as target) with very similar catalytic activity in a different bacterial species, and therefore may likely have similar structure.

b) Alignment was performed using ClustalOmega (https://toolkit.tuebingen.mpg.de/tools/clustalo) Sievers *et al.*, 2011), MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/), (Madeira *et al.*, 2022) and Psi-Coffee (https://tcoffee.crg.eu/apps/tcoffee/do:psicoffee) (Di Tommaso *et al.*, 2011), as each alignment produced is different and it was uncertain at this stage which alignment would produce the most accurate, or best-fitting, result. All alignments can be found in Appendix C for reference.

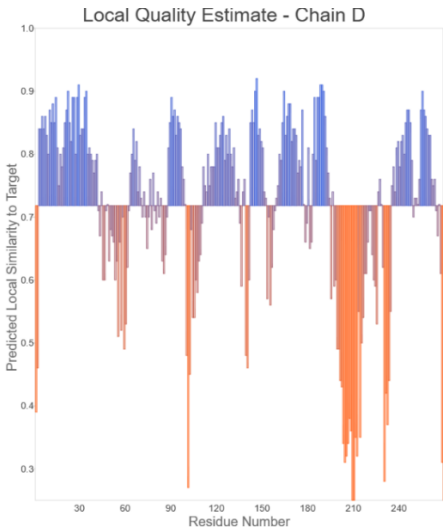
SWISS-MODEL (https://swissmodel.expasy.org/) (Waterhouse *et al.*, 2018) and MODELLER (Webb and Sali, 2016) were used to generate models from all three alignments, and Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) (Kelley *et al.*, 2015) was used to generate a separate model based on its own alignment algorithm, giving 7 different models in total.

The 'best' model was selected from the 7 options using QMEANDisCo (https://swissmodel.expasy.org/qmean/) (Studer *et al.*, 2020), a statistical potential-based method combined with a measure of how consistent the distance between all the carbon alpha backbone atoms are in the model when considering the average carbon alpha bond distance in homologous structures. This gives a global (overall) as well as a per-residue model quality estimate on a scale of 0-1, highlighting regions of lower quality in the model. For reference, the native experimental structure of the template 5tf4 attained a global QMEANDisCo of 0.90. The best global QMEANDisCo score was 0.72 from the Psi-Coffee alignment input into SWISS-MODEL – QMEANDisCo scores are provided for all models in Appendix C.

Submission: Sophie Allen
Protein: 1BVR Enoyl-ACP Reductase (InhA) - Rozwarski D.A. et al. (1998a)

The QMEANDisCo score also provides an indication of local areas of low model quality. In the chosen Psi-Coffee/SWISS-MODEL model, low scoring regions were highlighted in orange if the residue scored below the global average score of 0.72 (Figure 5). Of note, high quality predicted residues tended to be part of predicted beta sheet and alpha helix structures, while low quality regions such as residues 202-214 with QMEANDisCo scores between 0.25-0.55 were part of an extended loop region in the model that were aligned against a gap in the PsiCoffee alignment.

A)



B)

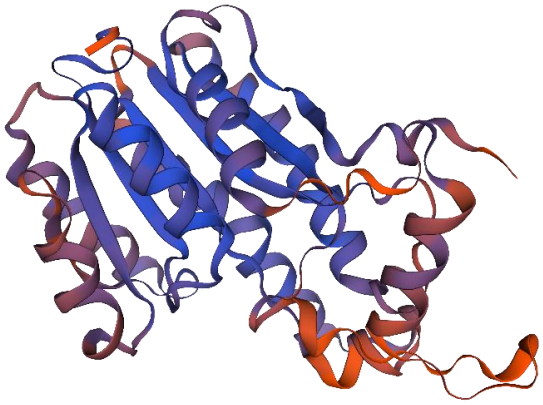


Figure 5: QMEANDisCo Local Scores. A) QMEANDisCo (https://swissmodel.expasy.org/qmean/) (Studer et al., 2020) local score for each residue in the model produced using the Psi-Coffee alignment and SWISS-MODEL modelling tool. Residues in orange scored below the global quality average of 0.72 and are considered to be of low quality in the predicted model. B) Predicted structure taken from SWISS-MODEL output (Waterhouse et al., 2018)

c) The overall structure closely resembles the experimental structure (successfully predicting most alpha helices and identifying the parallel beta sheet), represented well by the global RMSD of 0.937. This is considered to be a good superposition as the carbon-alpha backbones of the model and experimental structure are on average within less than 1 angstrom of each other.

For the 7 parallel beta sheet strands (residues 256-260, 185-191, 143-148, 90-93, 9-12, 35-40, 60-62), local RMSD remained around 1 angstrom, with the exception of the seventh strand, which had a slightly larger RMSD of 3.045 (Figure 6), showing that this beta sheet was a well-predicted region.

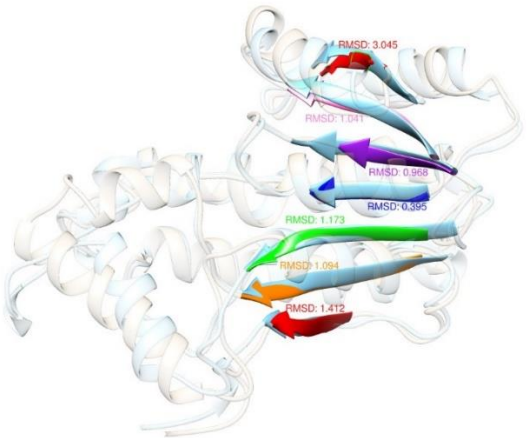


Figure 6: RMSD values of the 7-stranded parallel beta sheet (model vs experimental structure). Experimental structure residues from 'top' strand to 'bottom': 59-62 (RMSD: 3.045), 35-40 (RMSD: 1.041), 9-12 (RMSD: 0.968), 90-93 (RMSD: 0.395), 143-148 (RMSD: 1.173), 185-191 (RMSD: 1.094), 256-260 (RMSD: 1.412). Model produced using USCF Chimera (Pettersen *et al.*, 2004).

Looking at the predicted loop region of low quality (residues 202-214), local RMSD is much larger, and the experimental structure shows this region as a helix, turn, helix, which is a more structured region that occupies a difference arrangement in space, validating the predicted low quality of these residues structures (see Figure 7).

Within one loop region of the experimental model, the predicted model identified two short alpha helices, highlighted in green in Figure 7. Despite this region having a low RMSD comparable to the higher quality prediction of the beat sheet (1.332), the prediction of the residues arranged in a helix structure does not match the experimental structure, potentially highlighting a drawback of using RMSD in this way (since RMSD in Chimera only considers position of the carbon alpha atoms rather than the residue side-chains). Some of the helices, while being predicted to occur correctly, also have a different position in space to the experimental structure (Figure 7).

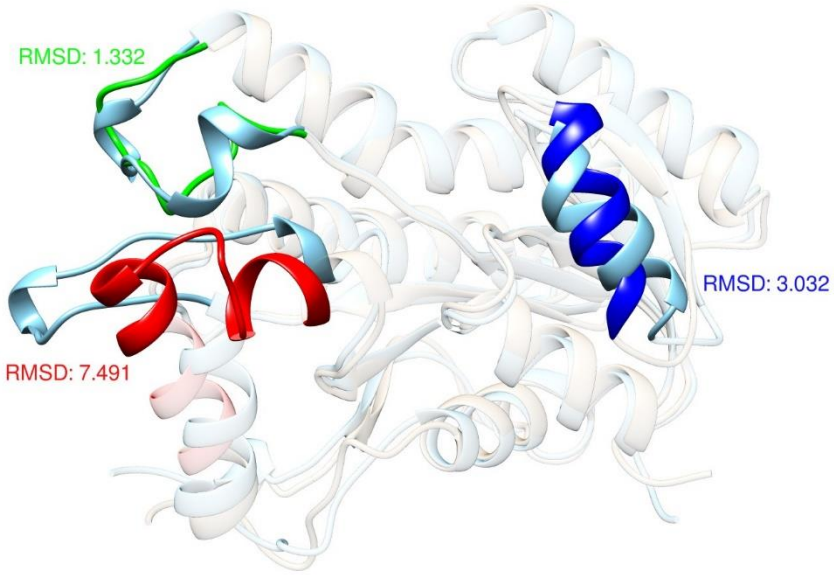


Figure 7: Local differences between experimental and predicted models, predicted model colored in light blue. In red: The experimental structure helix-turn-helix alongside the predicted loop, residues 202-214 (RMSD: 7.491). In green: The experimental loop region alongside predicted helix-like turn regions in the predicted model, residues 99-112 (RMSD: 1.332). In dark blue: One alpha helix with different arrangement in space to the predicted model, residues 44-53 (RMSD: 3.032). Model produced using UCSF Chimera (Pettersen *et al.*, 2004).

To account for these local differences, the TM-score (Zhang and Skolnick, 2004) (https://seq2fun.dcmb.med.umich.edu//TM-score/) was calculated and was output as 0.8960, scoring highly on the TM-score scale of between 0-1, again showing that the model is very similar to the experimental structure.

d) The ligand identified in Part A(e), THT, was identified within the ZINC database as ZINC14880555 (https://zinc15.docking.org/substances/ZINC000014880555/) (Sterling and Irwin, 2015). The model was prepared using the Dock Prep Chimera tool (Lang *et al.*, 2009) and the docking modelled using SwissDock (Grosdidier *et al.*, 2011). Table 3 shows the ten docked models with the lowest predicted energy confirmations, and Model 1 with the lowest energy confirmation was chosen (note: one

model (not shown) had a slightly higher FullFitness score, however was not chosen since the model had a much higher energy and deltaG, and therefore a model with the lowest energy was chosen instead).

Table 3: The 10 predicted structures from SwissDock (Grosdidier et al., 2011) with the lowest Energy outcome.

Model	deltaG	FullFitness	Energy
1	**-8.883539**	**-1378.516**	**-70.5968**
2	-8.550922	-1377.5789	-68.9529
3	-8.550922	-1377.5789	-68.9529
4	-8.550922	-1377.5789	-68.9529
5	-8.469255	-1376.683	-68.1576
6	-8.469255	-1376.683	-68.1576
7	-8.469255	-1376.683	-68.1576
8	-8.855798	-1375.5188	-67.8335
9	-8.350707	-1373.7324	-67.4806
10	-8.930396	-1376.3048	-67.3309

e) The predicted model has docked THT in the cofactor NADH binding site (Figure 8). This is a limitation of producing a model from sequence, since the model produced cannot account for cofactor binding and shape confirmation changes, therefore the predicted docking model can only represent the inactive form of the protein. NADH binding changes the confirmation of the binding site, moving the binding loop 4 angstroms to the left and rotating the key Tyr158 residue 60 degrees to the right to widen crevice to allow large fatty acid chains to enter the crevice (Rozwarski *et al*., 1999). Without NADH binding, the correct binding site is physically unavailable, preventing the ligand from entering the site.

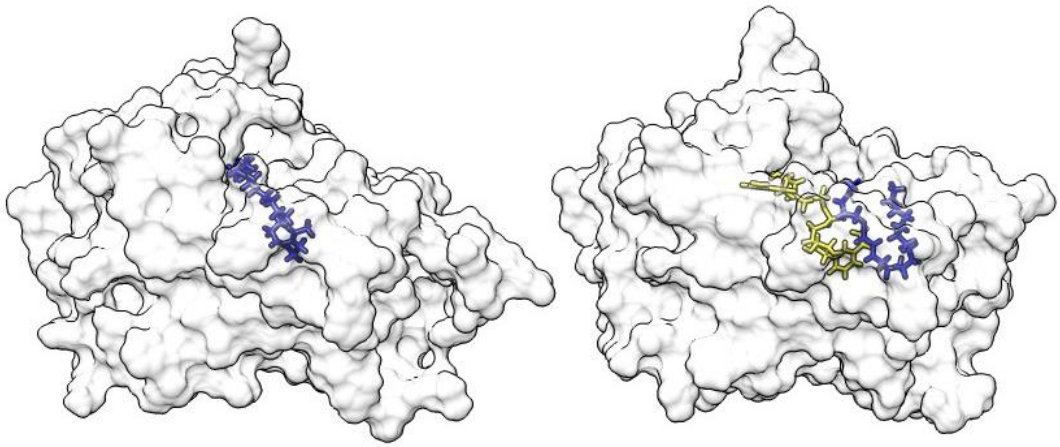


Figure 8: Overall side-by-side comparison of the experimental structure (left) and the SwissDock (Grosdidier *et al.*, 2011) structure with ligand THT docked (right). THT is highlighted in blue, and the cofactor NAD+ is highlighted in yellow. Models produced using UCSF Chimera (Pettersen *et al.*, 2004).

Looking at electrostatic potential of the two models, the experimental model has a fairly neutral electrostatic potential in the NADH and ligand binding site to increase binding affinity for hydrophobic carbon chains (Rozwarski *et al*., 1999). Two regions in the NADH binding site show charge – a section of positive charge, accommodating partial negative charges of -OH groups of the NAD+ ring, and a section of negative charge which accommodates the partial positive charges of the nitrogen atoms in the adenine ring (Figure 9). The predicted model also reflects these regions of charge within the NADH binding site, with the polar end of the THT ligand the positively charged region and the longer non-polar chain in the more neutral part.

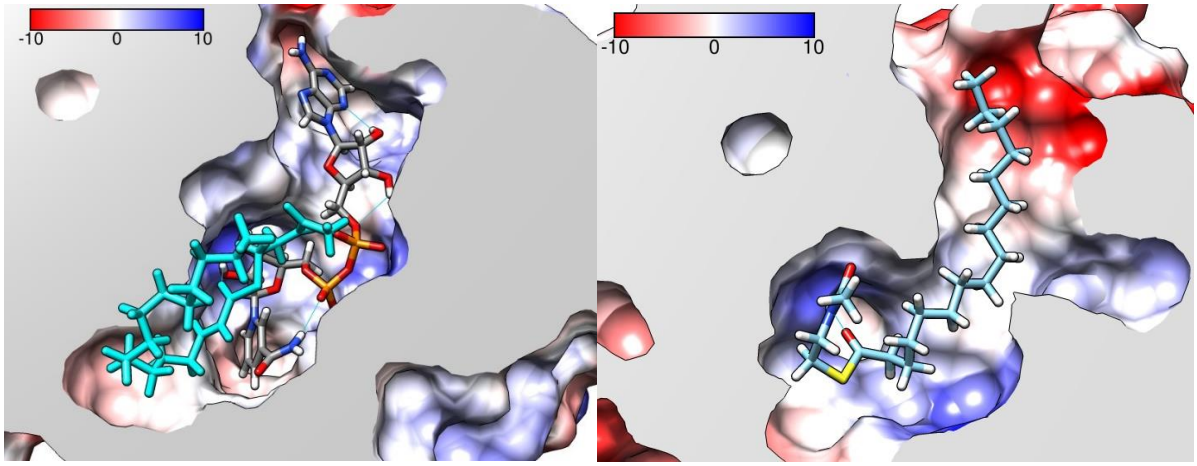


Figure 9: NADH binding site in each model, coloured by electrostatic potential (Red corresponds to negative charge, Blue corresponds to positive charge). On the left, NAD+ and THT (light blue) bound to the experimental model. On the right, THT bound in the NADH binding site in the predicted model. Binding site is smaller in the predicted model since the model does not represent the confirmation change allowing access to the second binding site as shown for the experimental model. Solvent-excluded molecular surfaces were created using the MSMS package of UCSF Chimera (Sanner *et al.*, 1996) (Pettersen *et al.*, 2004).

In the experimental structure, NAD+ in the NADH binding site forms hydrogen bonds with Thr196 and Ile194, residues important in catalysis (Rozwarski *et al*., 1999), as well as at Ser20, Val65, and Asp64 at regions towards the binding site opening, to provide stability and additional binding affinity for NADH. However, in the predicted model, the only hydrogen bond predicted between ligand and protein was at Lys165, which is logical since the ligand is highly non-polar and would be unable to create many hydrogen bonds. This, coupled with the predicted docking of the non-polar chain in a region of high electronegative charge, means this model is unlikely to be stable in reality and is not very useful in understanding binding of this ligand.

REFERENCES

Adasme,M.F. et al. (2021) PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. Nucleic Acids Research, 49, W530–W534.

Baker,M.E. (1995) Enoyl-acyl-carrier-protein reductase and Mycobacterium tuberculosis InhA do not conserve the Tyr-Xaa-Xaa-Xaa-Lys motif in mammalian 11 beta- and 17 beta-hydroxysteroid dehydrogenases and Drosophila alcohol dehydrogenase. Biochem J, 309, 1029–1030.

Berman,H.M. et al. (2000) The Protein Data Bank. Nucleic Acids Research, 28, 235–242.

Camus,J.-C. et al. (2002) Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. Microbiology (Reading), 148, 2967–2973.

Carugo,O. (2018) How large B-factors can be in protein crystal structures. BMC Bioinformatics, 19, 61.

Di Tommaso,P. et al. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res, 39, W13–W17.

Dias,M.V.B. et al. (2007) Crystallographic studies on the binding of isonicotinyl-NAD adduct to wild-type and isoniazid resistant 2-trans-enoyl-ACP (CoA) reductase from Mycobacterium tuberculosis. J Struct Biol, 159, 369–380.

Engh,R.A. and Huber,R. (1991) Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallographica Section A, 47, 392–400.

Hanukoglu,I. (2015) Proteopedia: Rossmann fold: A beta-alpha-beta fold at dinucleotide binding sites. Biochemistry and Molecular Biology Education, 43, 206–209.

Hutchinson,E.G. and Thornton,J.M. (1996) PROMOTIF--a program to identify and analyze structural motifs in proteins. Protein Sci, 5, 212–220.

Grosdidier,A. et al. (2011) SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic Acids Res, 39, W270-277.

Kelley,L.A. et al. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc, 10, 845–858.

Lang,P.T. et al. (2009) DOCK 6: combining techniques to model RNA-small molecule complexes. RNA, 15, 1219–1230.

Laskowski,R.A. et al. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. Journal of Applied Crystallography, 26, 283–291.

Madeira,F. et al. (2022) Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res, gkac240.

McGuffin,L.J. and Jones,D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics, 19, 874–881.

Mendez,D. et al. (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Research, 47, D930–D940.

O'Boyle,N.M. et al. (2011) Open Babel: An open chemical toolbox. Journal of Cheminformatics, 3, 33.

Pettersen,E.F. et al. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem, 25, 1605–1612.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rozwarski D.A. et al. (1998a) M.TB. ENOYL-ACP REDUCTASE (INHA) IN COMPLEX WITH NAD+ AND C16-FATTY-ACYL-SUBSTRATE. DOI: 10.2210/pdb1BVR/pdb

Rozwarski,D.A. et al. (1998b) Modification of the NADH of the isoniazid target (InhA) from Mycobacterium tuberculosis. Science, 279, 98–102.

Rozwarski,D.A. et al. (1999) Crystal structure of the Mycobacterium tuberculosis enoyl-ACP reductase, InhA, in complex with NAD+ and a C16 fatty acyl substrate. J Biol Chem, 274, 15582–15589.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. URL http://www.rstudio.com/.

Sanner,M.F. et al. (1996) Reduced surface: an efficient way to compute molecular surfaces. Biopolymers, 38, 305–320.

Sievers,F. et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology, 7, 539.

Sillitoe,I. et al. (2021) CATH: increased structural coverage of functional space. Nucleic Acids Res, 49, D266–D273.

Sterling,T. and Irwin,J.J. (2015) ZINC 15 – Ligand Discovery for Everyone. J. Chem. Inf. Model., 55, 2324–2337.

Studer,G. et al. (2020) QMEANDisCo—distance constraints applied on model quality estimation. Bioinformatics, 36, 1765–1771.

Wang,Y. et al. (2021) A crowdsourcing open platform for literature curation in UniProt. PLOS Biology, 19, e3001464.

Waterhouse,A. et al. (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Research, 46, W296–W303.

Webb,B. and Sali,A. (2016) Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics, 54, 5.6.1-5.6.37.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4, 1686.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins, 57, 702–710.

Submission: Sophie Allen
Protein: 1BVR Enoyl-ACP Reductase (InhA) - Rozwarski D.A. et al. (1998a)

APPENDIX 1: ADDITIONAL CODE USED FOR PART A

```
# retrieve mmCIF from PDB
open cifID:1bvr

# removing the subunits from the second incomplete tetramer (retain only the whole tetrameric structure)
delete :.A,.B
# retain only 'best quality' subunit per validation report (51% residues with 0 outliers)
delete :.A,.C,.D,.E,.F

c) # Rossmann fold figure
color red :9-40
color orange :35-62
color yellow :88-148
color green :138-191
color blue :185-260

show :NAD
color white :NAD

# Gamma loops figure
open cifID:2IED
delete :.C,.D,.A
mm #0 #1

color blue :THT
color yellow :NAD
color green #0:42-44,156-158,203-205
color red #1:42-44,156-158,203-205
select #0 :158
Actions -> Atoms/Bonds -> show

# Ramachandran Plot Creation (highlighted phi angle of residues with suspicious torsion angles)
select
:5,21.e,23.e,41.a,41.b,41.c,41.d,41.f,42.b,48.e,55,56.e,64.e,65.e,74.a,74.b,74.c,74.d,84.e,94.d,98.a,98.b,98.c,98.d,100,103.
d,103.e,105.e,107.d,108.d,112.e,114.e,123.e,124.e,125.e,142.e,150,157.a,157.b,157.c,157.d,157.e,159,193.e,197.d,202.e,
203.d,203.e,205.e,207.e,208.e,209.a,209.b,209.c,209.e,209.f,210.a,210.b,210.c,210.d,210.e,221.e,247.e@N,CA,C,O
ramachandran

d) Contacts between protein and ligand

PLIP: 1bvr entered as PDB ID. All other options kept as default.
```

APPENDIX 2: CODE AND OUTPUT FROM PART B

c) ChEMBL .sdf list of identified substructure matches – first two structures identified in list of 236:

```
  RDKit          2D

 25 25  0  0  0  0  0  0  0  0999 V2000
   -0.3918    0.5872    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3918    1.4122    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -1.1063    0.1747    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.6214   -0.4927    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.9569   -1.2464    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.4720   -1.9138    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.7774   -1.3326    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.1991   -0.4065    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -1.7194   -0.3773    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   -2.4338    0.0352    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -3.1875   -0.3004    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -2.2623    0.8421    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.8143    1.4552    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4418    0.9284    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.0293    1.6429    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.3227    0.1747    0.0000 S   0  0  0  0  0  0  0  0  0  0  0  0
    1.0371    0.5872    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.7516    0.1747    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.7516   -0.6503    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.4661   -1.0628    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.0371   -1.0628    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    2.4661    0.5872    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    3.1805    0.1747    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.1805   -0.6503    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.8950    0.5872    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0
  3  1  1  6
  3  4  1  0
  3  9  1  0
 14  3  1  0
  4  5  1  0
  4  8  1  1
  5  6  1  0
  5  7  1  0
 10  9  1  0
 10 11  2  0
 12 10  1  0
 12 13  1  6
 14 12  1  0
 14 15  1  6
  1 16  1  0
 17 16  1  0
 18 17  1  6
 18 19  1  0
 19 20  2  0
 19 21  1  0
 18 22  1  0
 23 22  1  0
 23 24  1  0
 23 25  2  0
M  END

> <chembl_id>
CHEMBL374308

> <chembl_pref_name>
```

Submission: Sophie Allen
Protein: 1BVR Enoyl-ACP Reductase (InhA) - Rozwarski D.A. et al. (1998a)

LACTACYSTIN

$$$$

RDKit 2D

```
 31 34  0  0  0  0  0  0  0  0999 V2000
   15.4626  -27.3663    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   14.7925  -26.8820    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   14.1267  -27.3663    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   15.2029  -28.1506    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   14.3779  -28.1506    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   14.1242  -28.9354    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   14.7925  -29.4222    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   15.4607  -28.9354    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   13.5488  -28.1465    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
   16.0238  -28.1506    0.0000 S   0  0  0  0  0  0  0  0  0  0  0  0
   16.4385  -28.8656    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   17.2635  -28.8656    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   16.0239  -29.5807    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   16.2478  -27.1136    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   13.3416  -27.1129    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   13.3393  -29.1891    0.0000 S   0  0  0  0  0  0  0  0  0  0  0  0
   12.7252  -28.6398    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   11.9403  -28.8934    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   12.9001  -27.8328    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   17.6739  -28.1505    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   11.7696  -29.7006    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   10.9864  -29.9513    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   10.8155  -30.7576    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   11.4300  -31.3119    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   12.2181  -31.0545    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   12.3852  -30.2488    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   18.4959  -28.1532    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   18.9104  -27.4390    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   18.4957  -26.7230    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   17.6665  -26.7257    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   17.2599  -27.4405    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  4  1  1  0
  1  2  1  0
  2  3  1  0
  3  5  1  0
  4  5  1  0
  5  6  1  0
  6  7  1  0
  7  8  1  0
  8  4  1  0
  5  9  1  6
  4 10  1  6
 10 11  1  0
 11 12  1  0
 11 13  2  0
  1 14  2  0
  3 15  2  0
  6 16  1  0
 16 17  1  0
 17 18  1  0
 17 19  2  0
 12 20  1  0
 18 21  1  0
```

```
 21 22  2  0
 22 23  1  0
 23 24  2  0
 24 25  1  0
 25 26  2  0
 26 21  1  0
 20 27  2  0
 27 28  1  0
 28 29  2  0
 29 30  1  0
 30 31  2  0
 31 20  1  0
M  END

> <chembl_id>
CHEMBL4078511

> <chembl_pref_name>
None

$$$$
```

d) .smi file created for ligand using nano (thoth.cryst.bbk.ac.uk > nano). SMILES string used:
CCCCCCCCCCCCCCCC(=O)SCCNC(C)=O

e) The top ten ChEMBL molecules with a substructure matching the ligand THT:

ChEMBL ID	Tanimoto Score
CHEMBL244985	**0.921569**
CHEMBL243567	**0.854545**
CHEMBL244833	**0.810345**
CHEMBL244204	**0.810345**
CHEMBL1179301	0.678571
CHEMBL1179279	0.678571
CHEMBL67577	0.666667
CHEMBL68560	0.666667
CHEMBL501082	0.653846

Submission: Sophie Allen
Protein: 1BVR Enoyl-ACP Reductase (InhA) - Rozwarski D.A. et al. (1998a)

APPENDIX C: CODE AND OUTPUT FROM PART C

a) GenTHREADER used to identify structural homologs, settings: GenTHREADER (Rapid Fold Recognition) selected, no other options selected.
Protein sequence input:
MTGLLDGKRILVSGIITDSSIAFHIARVAQEQGAQLVLTGFDRLRLIQRITDRLPAKAPLLELDVQNEEHLASLAGRVTEAIGAGN
KLDGVVHSIGFMPQTGMGINPFFDAPYADVSKGIHISAYSYASMAKALLPIMNPGGSIVGMDFDPSRAMPAYNWMTVAKSAL
ESVNRFVAREAGKYGVRSNLVAAGPIRTLAMSAIVGGALGEEAGAQIQLLEEGWDQRAPIGWNMKDATPVAKTVCALLSDW
LPATTGDIIYADGGAHTQLL

b) Template model downloaded from PDB (https://www.rcsb.org/structure/5TF4), experimental validation report checked.
Most complete subunit (least missing residues from model) = Chain D (6 residues missing)
Chimera:
retrieve mmCIF from PDB
open cifID:5tf4

remove all chains except chain D
sel :.D
select invert
Actions -> Atoms/Bonds -> delete

delete substrate/cofactor and identify sequence
delete solvent
delete :GOL,NAD
~show
File -> Save PDB

sequence :.D
In Sequence dialog: File -> Save.as (used 'Aligned FASTA' file type)

Alignment inputs:
>sequence
MTGLLDGKRILVSGIITDSSIAFHIARVAQEQGAQLVLTGFDRLRLIQRITDRLPAKAPLLELDVQNEEHLASLAGRVTEAIGAGN
KLDGVVHSIGFMPQTGMGINPFFDAPYADVSKGIHISAYSYASMAKALLPIMNPGGSIVGMDFDPSRAMPAYNWMTVAKSAL
ESVNRFVAREAGKYGVRSNLVAAGPIRTLAMSAIVGGALGEEAGAQIQLLEEGWDQRAPIGWNMKDATPVAKTVCALLSDW
LPATTGDIIYADGGAHTQLL
>5tf4 (#0) chain D/-1-272
GSMAKGNGLLYGKRGLILGLANNRSIAWGIAKTASSAGAELAFTYQGEAMKKRVEPLAEEVKGFVCGHCDVSDSASIDAVFN
TIEKKWGKLDFLVHAIGFSDKEELSGRYVDISESNFMMTMNISVYSLTALTKRAEKLMSDGGSILTLTYYGAEKVVPNYNVMG
VAKAALEASVKYLAVDLGPKHIRVNAISAGPIKTLAASGIGDFRYILKWNEYNAPLRRTVTIEEVGDSALYLLSDLSRSVTGEVH
HVDSGYNIIGMKAVDAPDISVVKE

ClustalOmega Input: no changes to default parameters
MUSCLE Input: Pearson/FASTA, all other parameters default
Psi-Coffee Input: no changes to default parameters (Alignment Length = 80, Alignment Format = clustalw_aln, fasta_aln, phylip, score_ascii, score_html. Case = Upper, Residue number = off, outorder = input)

ClustalOmega Output:
>sequence
-----MTGLLDGKRILVSGIITDSSIAFHIARVAQEQGAQLVLTGFDRLRLIQRITDRLPAKAPL---
LELDVQNEEHLASLAGRVTEAIGAGNKLDGVV
HSIGFMPQTGMGINPFFDAPYADVSKGIHISAYSYASMAKALLPIMNPGGSIVGMDF-
DPSRAMPAYNWMTVAKSALESVNRFVAREAGKYGVRSNLVAA
GPIRTLAMSAIVGGALGEEAGAQIQLLEEGWDQRAPIGWNMKDAT--------PVAKTVCALLSDWLPATTGDIIYADGGAHTQLL--
----
>5tf4
GSMAKGNGLLYGKRGLILGLANNRSIAWGIAKTASSAGAELAFTYQGEA-MKKRVE-PLAEEVKGFVCGHCDVSDSA---
SIDAVFNTIEKKWGKLDFLV
HAIGFSDKEELS-
GRYVDISESNFMMTMNISVYSLTALTKRAEKLMSDGGSILTLTYYGAEKVVPNYNVMGVAKAALEASVKYLAVDLGPKHIRVN
AISA

Submission: Sophie Allen
Protein: 1BVR Enoyl-ACP Reductase (InhA) - Rozwarski D.A. et al. (1998a)

GPIKTLAASGIGDF----------
RYILKWNEYNAPLRRTVTIEEVGDSALYLLSDLSRSVTGEVHHVDSGYNIIGMKAVDAPDISVVKE

MUSCLE Output:
>sequence
-----MTGLLDGKRILVSGIITDSSIAFHIARVAQEQGAQLVLTGFDRLRLIQRITDRLP
AKAPLL--ELDVQNEEHLASLAGRVTEAIGAGNKLDGVVHSIGFMPQTGMGINPFFDAPY
ADVSKGIHISAYSYASMAKALLPIMNPGGSIVGMD-FDPSRAMPAYNWMTVAKSALESVN
RFVAREAGKYGVRSNLVAAGPIRTLAMSAIVGGALGEEAGAQIQLLEEGWDQ-RAPIGWN
MKDATPVAKTVCALLSDWLPATTGDIIYADGGAHTQLL----------
>5tf4 (#0) chain D/-1-272
GSMAKGNGLLYGKRGLILGLANNRSIAWGIAKTASSAGAELAFT-YQGEAMKKRVEPLAE
EVKGFVCGHCDVSDSASIDAVFNTIEKKWG---KLDFLVHAIGFSDKEELS-GRYVDISE
SNFMMTMNISVYSLTALTKRAEKLMSDGGSILTLTYYGAEKVVPNYNVMGVAKAALEASV
KYLAVDLGPKHIRVNAISAGPIKTLAASGI---------GDFRYILK--WNEYNAPLRRT
VT-IEEVGDSALYLLSDLSRSVTGEVHHVDSGYNIIGMKAVDAPDISVVKE

PsiCoffee Output:
>sequence _R_ sequence.prf
-----MTGLLDGKRILVSGIITDSSIAFHIARVAQEQGAQLVLTGFDR--LRLIQRITDRLPAKAPLLELDVQNEEHLAS
LAGRVTEAIGAGNKLDGVVHSIGFMPQTGMGINPFFDAPYADVSKGIHISAYSYASMAKALLPIMNPGGSIVGMDF-DPS
RAMPAYNWMTVAKSALESVNRFVAREAGKYGVRSNLVAAGPIRTLAMSAIVGGALGEEAGAQIQLLEEGWDQRAPIGWNM
KDATPVAKTVCALLSDWLPATTGDIIYADGGAHTQLL----------
>5tf4 (#0) chain D/-1-272 _R_ 5tf4.prf
GSMAKGNGLLYGKRGLILGLANNRSIAWGIAKTASSAGAELAFTYQGEAMKKRVEPLAEEVKG-FVCGHCDVSDSASIDA
VFNTIEKKWG---KLDFLVHAIGFSDKEELS-GRYVDISESNFMMTMNISVYSLTALTKRAEKLMSDGGSILTLTYYGAE
KVVPNYNVMGVAKAALEASVKYLAVDLGPKHIRVNAISAGPIKTLAASGIGD----------FRYILKWNEYNAPLRRTV
T-IEEVGDSALYLLSDLSRSVTGEVHHVDSGYNIIGMKAVDAPDISVVKE

Target-Template Modelling Inputs:
SWISS-MODEL: Biounit '5tf4.1.D' chosen to build model for all three alignments. No other changes to default parameters.
Phyre2: Expert Mode Login required. One-to-One Threading, Secondary Structure scoring = yes, weight = 0.1, alignment method = local.
MODELLER: Open both alignment and template model in Chimera. In the alignment window: Structure -> Modeller (homology), Target sequence = sequence, Template = 5tf4, Run Modeller via a web service, Advanced Options -> Number of output models = 1

SWISS-MODEL QMEANDisCo v4.3.0 Inputs: Selected 'QMEANDisCo', all other parameters left as default.

QMEANDisCo (Studer et al., 2020) Global Scores for each of the 7 output models:

Alignment + Model	**QMEANDisCo**
PSI-COFFEE, SWISS-MODEL	0.72 ±0.05
MUSCLE, SWISS-MODEL	0.66 ±0.05
PSI-COFFEE, MODELLER	0.66 ±0.05
ClustalO, SWISS-MODEL	0.63 ±0.05
MUSCLE, MODELLER	0.62 ±0.05
Phyre2	0.61 ±0.05
ClustalO, MODELLER	0.58 ±0.05

c) In chimera:
to get one subunit and none of the ligands/non-standard residues for better comparison to model

open cifID:1bvr
delete :NAD
delete :THT
delete :HOH
delete :.A,.C,.D,.E,.F
select

~show

Run MatchMaker, superimpose the predicted model onto the actual structure model, do not recalculate secondary structure predictions. All other settings as default.

mm #0 #1 computeSS false

Highlight the 7 beta sheet strands and their RMSD values (local RMSD)

color red #0:256-260
color orange #0:185-191
color green #0:143-148
color blue #0:90-93
color purple #0:9-12
color hot pink #0:35-40
color red #0:59-62

select
transparency 90,r
transparency 0,r #0:256-260,185-191,143-148,90-93,9-12,35-40,59-62 # experimental model beta sheet
transparency 0,r #1:257-260,184-191,142-148,90-93,9-14,35-40,57-62 # predicted model beta sheet
background solid white

rmsd #0:256,260 #1:256,260
rmsd #0:185,191 #1:185,191
rmsd #0:143,148 #1:143,148
rmsd #0:90,93 #1:90,93
rmsd #0:9,12 #1:9,12
rmsd #0:35,40 #1:35,40
rmsd #0:59,62 #1:59,62

Actions -> Label -> Residue -> Custom (for each beta sheet strand, to represent RMSD in figure format)

Highlight the differences in local structure

color red #0: 200-214
color green #0: 99-112
color blue #0:44-53

select
transparency 90,r
transparency 0,r #0:200-214,99-112,44-53   # experimental structure
transparency 0,r #1:200-214,99-112,44-53   # model structure
background solid white

rmsd #0:200,214 #1:200,214
rmsd #0:99,112 #1:99,112
rmsd #0:44,53 #1:44,53

Actions -> Label -> Residue -> Custom

d) Chimera:
prepare model protein for docking
select
addh hbond true   # hbond true = considers hydrogen bonds in adding hydrogen atoms
Tools -> Structure Editing -> Dock Prep (Remove: Delete Solvent and Write Mol2 file, all other options remain as default)

SwissDock Settings: Target (model) was uploaded as a .pdb file as prepared above. Ligand (THT) was added using the ZINC ID for THT (ZINC14880555). All parameters left as default.

Model 1 preparation for figure
select
~show
represent stick
show :LIG
color blue :LIG

select invert
surface #0
color white,s #0
transparency 75,s #0

e) Chimera:
prepare experimental docked model for figure – general structure comparison (binding sites)
Note: leave Model 1 from above step open

open cifID:1bvr
delete #2:.A,.C,.D,.E,.F

color white #2
color blue :THT
color yellow :NAD
color red :HOH

addh spec #2:THT hbond true
surface #2
color white,s #2
transparency 75,s #2
~surface :NAD,THT,HOH

mm #0 #2 computeSS false
Tools -> General Controls -> Model Panel -> Deactivate model #0 and #1.6, move model #2 to the right, re-activate models #0 and #1.6.

select #2
~show
show: THT,NAD
Actions -> Ribbon -> Hide

electrostatic potential models

open cifID:1bvr
delete :.A,.C,.D,.E,.F
delete #2:HOH
select :LIG,NAD,THT
Actions -> Color -> by element
transparency 0,s #0

addh hbond true
addcharge # Charge model: AMBER ff14SB
coulombic key -10 red 0 white 10 blue

~show
show :THT,NAD,LIG
~ribbon

hbonds
display :196,194,20,65,64
Tools -> General Controls -> Model Panel -> clipping