Plan and schedule for 2021

February 10, 2021

Contents

1 Overview **2**

1.1 Target . . . 3

1.2 Prior work . . . 3

1.3 Approaches . . . 3

1.3.1 Structure-based virtual screening - overview . . . 3

1.3.2 Artificial Intelligence-based design - overview . . . 3

2 Structure-based design **4**

2.1 Aim and overview . . . 4

2.1.1 *enz* - a *Python* package for enzyme design . . . 4

2.1.2 Designing new mutants with *enz* . . . 4

2.1.3 Validation of mutants . . . 5

2.1.4 Order of events . . . 5

2.2 Future work . . . 5

3 Machine learning-based design **6**

3.0.1 Aim . . . 6

3.0.2 Approach . . . 6

3.0.3 Model architecture . . . 6

3.0.4 Pretraining . . . 7

3.0.5 Pilot screen . . . 7

3.0.6 Screening Library Design . . . 8

3.0.7 Screening . . . 8

3.0.8 Training . . . 9

3.0.9 Mutant design and testing . . . 9

3.0.10 Future work: Active learning . . . 9

4 Syngenta Placement **10**

4.1 Future work . . . 10

5 schedule **10**

5.1 Feb . . . 10

5.1.1 Structure-based 10
5.1.2 Machine learning-based 10
5.1.3 Writing 11
5.1.4 Targets 11
5.2 March 12
5.2.1 Structure-based 12
5.2.2 Machine learning-based 12
5.2.3 Writing 12
5.2.4 Targets 12
5.3 April 12
5.4 Writing 12
5.4.1 Targets 13
5.5 May 13
5.5.1 Structure-based 13
5.5.2 Machine learning-based 13
5.5.3 Writing 13
5.5.4 Targets 13
5.6 June 14
5.6.1 Structure-based 14
5.6.2 Machine learning-based 14
5.6.3 Writing 14
5.6.4 Targets 14
5.7 July 14
5.7.1 Machine learning-based 14
5.7.2 Writing 14
5.7.3 Targets 14
5.8 August 15
5.8.1 Writing 15
5.8.2 Targets 15

1 Overview

This project has two peices of work both with the aim to engineer a P450 BM3 mutant with some sort of metabolic activity towards the herbicide mesotrione, ideally hydroxylation at carbon 5. One approach uses classical structure prediction and docking combined with a genetic algorithm to generate mutants with predicted activity. The other uses a deep learning model trained on a screening set of BM3 mutants and herbicide-like compounds, with the aim of being able to predict the activity of any BM3 mutant with any herbicide. This document outlines each approach and a schedule for work to complete and write up each into a thesis by the September deadline.

If it is computational (you don’t mention experiment here) you ‘predict’ or generate predictions

1.1 Target

Mesotrione is an important herbicide whose metabolism in plants is initiated by hydroxylation of carbon 5 by a P450. Given the promiscuity of glutathione-S transferases that help sequester a hydroxykated xenobiotic, alternative hydroxylation sites may be sufficient to initiate herbicide metabolism in plants. A BM3 mutant that hydroxylates mesotrione at any position may be suitable for engineering mesotrione resistance into a crop.

1.2 Prior work

The promiscuous BM3 A82F/F87V mutant shows no binding or catalytic activity towards mesotrione despite activity towards several less important herbicide classes. It may not be possible to create a BM3 mutant with desired mesotrione activity, so this work is motivated towards creating enzyme design systems that generalize to other herbicides.

1.3 Approaches

The two approaches developed here are based on traditional structure-based methods and one based on machine learning. Both approaches will produce a pool of BM3 variants with predicted binding activity towards mesotrione, for whom K_d, K_{cat} and product formation will be measured. In this document, this project is referred to as *enz*.

1.3.1 Structure-based virtual screening - overview

Structure based design relies on template-based structure prediction to generate mutant structures and molecular docking to evaluate the fitness of mutants for a genetic algorithm, which steers a virtual directed evolution process.

1.3.2 Artificial Intelligence-based design - overview

An artificial intelligence-based approach uses a deep neural network to predict the likelihood of binding between input amino acid sequences and chemical smiles and design lab experiments. The network is trained on a domain-specific BM3:herbicide-like compound screening set and uses adaptive learning to generate new screening experiments to improve its accuracy.

The trained model can be used to rapidly predict the binding likelihood for any given sequence-chemical pair, which enables large-scale directed evolution experiments using a genetic algorithm to generate candidate mutants. These mutants will be made and evaluated in the lab

In this document, this project is referred to as *rio*.

Both approaches are being built into *Python* packages to ensure portability to other enzyme engineering problems.

2 Structure-based design

2.1 Aim and overview

good to acknowledge that not all structures are considered equally. If these are experimental (Xtal) structures, then better to be explicit

Template structures for BM3 are plentiful. The BM3 active site is mostly helical and is fairly temperature stable, so low-resolution template-based structure prediction may generate a sufficiently accurate mutant for ligand docking with mesotrione.

For my own convenience and reusability, structure-prediction and molecular docking functions used in this work have been packaged into the *Python* module *enz*. Packaging has enabled large-scale virtual directed evolution experiments to be set up easily.

2.1.1 *enz* - a *Python* package for enzyme design

enz is a *Python* API made for this work that automates template-based structure prediction and docking. Docking calls *atuodock vina* and refolding calls *pyrosetta enz* is simple to use, is potentially useful to other protein engineers and has potential for further development. The package works and is ready to deploy with a genetic algorithm, however the following features are not yet supported:

- **flexible docking** - currently side chains are modelled as rigid which limits docking accuracy
- **loop remodelling** - currently conformation of loops and flexible regions are not recalculated, this would happen by using the cyclic coordinate descent implementation in *pyrosetta* for all loops within a cutoff radius of a mutation.

Ideally support would be added for these non-features before running the genetic algorithm with *enz*. Whilst I'm having trouble with flexible docking, loop remodelling may be solved with a day's focus. Solutions to either problem can be pushed to the main branch of *enz* in a matter of days.

Now is the time to focus on what can be done quickly. If you think this is worth doing, at it is only a day's work. Do it.
Leave flexible docking for enz 2.0

2.1.2 Designing new mutants with *enz*

enz works well enough to use and is ready to deploy with a simple genetic algorithm for mutant generation. Since the active site of BM3 has few flexible regions, the impact of loop remodelling may be small and despite not

How have you/are you going to benchmark this algorithm ?
As we discussed, can you benchmark against existing data, predict new mutants and then make a small number to test in the lab?

docking ligands with flexible protein side chaines, the results may be sufficient to generate a pool of mutants to be tested in the lab. However implementation of flexible docking and loop remodelling is very desirable.

The heuristic currently employed to estimate the desirability of each set of docking results is

$$score = \frac{1}{n}\Sigma_n \Delta G_n \times d_n \quad (1)$$

where ΔG is free energy of the interaction calculated by *autodock vina* (*kcal / mol*) and d is the distance between the heme iron and the C5 of mesotrione for n in binding poses. Where C5 is the target carbon for hydroxylation. It may be sufficient for use in the genetic algorithm, however could be improved.

easily extensible, so why not just make this an arbitrary distance between 2 atoms.
Future work would be to make this more flexible -
* shortest distance between set of atoms
* angles
* rmsd, etc

The runtime of the genetic algorithm should not exceed 2 days, even on a small virtual private server. A pool of mutants generated by the genetic algorithm will be made in the lab for testing activity towards mesotrione. The pool size of the mutants to make and test will be constrained by time and resource. The pool will be made in the lab and tested for K_d and K_{cat} with mesotrione and an incidation of product formation.

2.1.3 Validation of mutants

Mutants generated by the genetic algorith *enz* combination will be made and tested in the lab with three techniques:

- **Mesotrione titration** to get a K_d if any.
- **Steady state kinetics** for a k_{cat} via NAPDH consumption
- **LCMS** product detection. A +16 *m/z* is considered a hit.

Km for mesotrione would be good too

2.1.4 Order of events

The expected order of events for this project is:

1. Final changes to *enz* - if possible - timescale: < 1 week
2. Run genetic algorithm, select pool of mutants - runtime < 2 days
3. Prepare the mutants in the lab and test for the mentioned metrics - timescale < 2 months

How will you know your mutants are sensible - perhaps prioritise short additional period of testing/ benchmarking first?

The timings of this operation will be expanded on in the **schedule** section.

2.2 Future work

Future work on this project would see the continued development of *enz* into an open-source python module that does not rely on *pyrosetta*. Folding and docking algorithms would be re-implemented in an open-source tensor math framework like *jax* or *pytorch*, which support parralelization with GPU processing.

3 Machine learning-based design

3.0.1 Aim

The aim of this project is to produce a general solution to BM3 mutant:herbicide metabolism prediction, and use that to generate BM3 mutants with activity towards mesotrione or any other important herbicide. The model aims to use amino acid sequence and chemical smiles alone.

extensible -
any possible
substrate

3.0.2 Approach

In this project, the approach to creating the model relies on creating a screening dataset of several BM3 mutants against many compounds.

define order of magnitude

To make the model faesible, it must be pre-trained on a large dataset before training on the domain-specific data. This approach is called transfer learning and enables learning from small domain-specific datasets.

again, sizes are
important here, so best
to quote order(s) of
magnitude / ranges

3.0.3 Model architecture

The model is based on leading techniques from chemical learning and natural language processing (NLP). The best performing models in chemical learning use graph convolutional neural networks to learn features of molecules, which are employed in this model for molecular input. The best known architecture for sequence learning is the transformer, a model that enjoys superiority in NLP and other fields so is also employed here.

In more detail, the model constructs a chemical graph from a smiles string, which is convolved by a graph convolutional network and downsampled to a fixed-size embedding vector using set2set. The amino acid sequence string is downsampled using three 1D-convolutional layers and processed to a short, variable-length tensor which is downsampled into a fixed-size vector using the final hidden-state from an LSTM layer. The learned embedding vectors of the sequence and conatenated into a combined representation vecor. A two layer perceptron provides a point estimate of binding likelihood.

At certain points in training, it may be necessary to replace the output head of the model with perceptrons that can generate multiple predictions. Replacing elements of a trained network is common in transfer learing to adapt a model to a new task.

again, my biggest concern is how you test this
it’s relatively easy to generate a model, but how will you know it works?
Do you have success criteria?

3.0.4 Pretraining

is that needed?
junk in = junk out comes to mind

The pre-training dataset has been mined from KEGG and currently holds roughly 0.2 million datapoints of enzyme sequence and smiles for reactants and products. I've identified changes to the data miner that will significantly increase the dataset size.

The sets of substrates for each sample in the KEGG dataset are positive examples only, so for the model to learn the distribution of valid sets of enzyme and compounds it is being trained with a generative adversary. For each datapoint input into the model, a seperate generator network generates a synthetic datapoint. Over training the model must discriminate between real and fake data, whilst the generator must create more realistic samples. The generator will be discarded after training, at which point the discriminative model will have learned the distibution of probable enzyme-substrate pairs.

this is a relatively hard problem to solve properly. How will you ensure this generates data that are realistically negative, yet still useful?

An additional dataset of P450:xenobiotics binding data has been filtered from *BindingDB*. The metrics for each P450:xenobiotic pair are one of K_d, K_{cat}, K_i and IC_{50}. For this task, the output layer of the model will be replaced with a perceptron with multiple output heads. This dataset conatins X data points and exposes the model to xenobiotics, which may assist with training on the in-house dataset.

one per xenobiotic?

3.0.5 Pilot screen

A 384 well plate-based UV-Vis assay for measuring K_d has been developed for this work. It can be set up using a multichannel pipette or an Echo acoustic liquid handling robot. One plate measures K_d between one mutants and 24 compounds at 8 concentrations, including 8 blanks to correct for UV-Vis absorbance of the compounds. One plate can be prepared and read in 20 minutes. Accuracy is limited compared to titration assays, but can identify clear binding signals. Attempts to optimize the assay were largely unsuccessful.

is the development work complete and ready to write up in your thesis?

this sort of statement can get you into trouble with examiners/ reviewers.
Another way to look at it is that you have optimised it (and further changes don't make it better by definition)

A pilot screening trial will take place in Feb, with the aim of proof of concept for this type of enzyme design approach. I have 5 purified BM3 mutants and an old FDA drug library on hand for this experiment. The pilot screen will take one week. Data from the pilot will be used to test the model which will have been pretrained by that point.

would be good to discuss the outcome of this trial.Are you still on track to have it done shortly?

One outcome of the pilot screen is the evaluation of whether attempting to measure K_d with 8 concentrations of compound gives better quality data than a qualitative hit or miss metric determined with only one concentration.

The pilot will also indicate a suitable dataset size for main screening. The training weights learned from the pilot dataset may also be progressed to training on the main screening set.

3.0.6 Screening Library Design

A suitable size of screening library can be inferred by results from the pilot study.A method for designing a herbicide-like screening library based on size constraints was programmed in this work. Library size is adjusted to budget and based on the *Molport* library - an aggregation of compounds from many suppliers (> 2M). *Molport* ship custom selections of compounds and have an API for generating quotes. The library selection program filters the entire library based on *herbicide likeness* rules and a Tanimoto distance cutoff between each compound and each herbicide. From the remaining subset, screening libraries of size n are selected using a diversity picking algorithm, *MaxMin* in this case. Since *MaxMin* is randomly seeded, it can be re-run many times, generating many candidate screening libraries. The program genertes a quote for each and the cheapest library can be selected from this.

I have been working with a sales manager from *Molport* for this, who tells me that the lead time for a library to arrive in the UK is X weeks and that they can provide a COSHH form that covers all compounds. The pandemic has not affected their ability to operate.

how many weeks?

The size of the library can be adjusted according to budget constraints and the volume of each compound is determined by the number of mutants I will make for the first round of screenning. I expect that 96 compounds will be parctical and sufficient. Purchasing may be difficult, but I'm told that I have > £8,000 in my budget and that a tender exemption form will be required to make a large purchase like this.

not sure about this.
The value may be low enough

3.0.7 Screening

The 5 mutants I have on hand will be used in the pilot assay. Whilst the autoclave is down for servicing in March, I will prepare as many BM3 mutant dna stock as practical, where the choice of mutants will be based on the expected information gain of near-mutants of BM3. Choice of mutants is constrained by the number of mutations from the nearest plasmid stock in my inventory. Practically, I'll prepare as many mutants as I can and express and screen as many as I have time for. Likely around 4 and up to around 10.

This stage of the project is the most labour intensive so with time-efficiency in mind, I'll fit the screening size to available time.

if your data set is that small, you should also do a manual rationale analysis of your data - look at the position of the mutation(s), generate the homology /in silico model and then try to infer behaviour based on conventional structure-function thinking

3.0.8 Training

The required dataset size for domain-specific data is not known, but transfer learning gives the best chance of sucess with a small dataset size and the best chance that the model can extrapolate to predictions outside of the screening dataset. The model will either be trained to directly predict K_d between sequence and compound, or continue to predict likelihood of binding interaction in the case that the data is only accurate enough to yield qualative results.

I have sufficient access to hardware for this task. Training and parameter selection may last one week.

After training the model will be evaluated on a validation set left out of the training data. Uncertainty of model predictions can be evaluated and used to determine which areas of input space the model is unsure of and next experiments can be designed accordingly.

3.0.9 Mutant design and testing

The trained model can deliver predictions very fast and in parralel. To generate mutants a genetic algorith will be used to generate mutants due to its simplicity. In this case, the genetic algorithm will generate amino acid strings of BM3 mutants based on constrained mutation types and a search depth constraint to reduce the number of PCR steps in synthesis.

I don't think this is a GA. Are you doing multiple steps of 'evolution'?

As in the structure-based design, the testing process for mutants for a pool of n mutants is:

1. expression and purification using nickel affinity chromatography
2. K_d measurement with mesotrione using titration and UV-Vis spectroscopy
3. K_{cat} measurement with mesotrione by monitoring reaction NADPH consumption by UV-Vis spectroscopy
4. product analysis with LCMS where a hydroxylation at any position is considered a hit

3.0.10 Future work: Active learning

Continuation of this project would see the model built into an artificial intelligence system that can design optimal screening experiments using adaptive learning. This could allow full automation of enzyme design using this approach.

4 Syngenta Placement

I plan to re-establish communication with Syngenta by sending a report with proof of concept for each project. The *enz* project can be presented in its current state and the *rio* project can be presented after the pilot screen due to complete in Feb. Given this, the deadline for sending this report is Feb 28th. The most useful placement to me is to work remotely with Syngenta computer scientist who have had involvement with this project. Nathan Kidley is a virtual screening specialist who can advise on *enz*-related work and Kostas Papachristos is a machine learining engineer who can advize on *rio* work. Richard Dale and Christian Noble can advise on lab work if required.

please run this past us first

4.1 Future work

Syngenta have the facilities to transform a P450 into a model crop and spray test it against mesotrione in a glass house. The turn around time for this process is 1-2 months. To publish any work in this thesis, it will be important to include a spray test. This may have to happen after thesis submission.

no, this would be follow-up work

5 schedule

5.1 Feb

5.1.1 Structure-based

A genetic algorithm using *enz* as is ready to be deployed using the heuresic described. Improvements to *enz* to improve accuracy by enabling flexible ligand docking and loop remodelling will be added this month. I would benefit from some examples of flexible docking in *autodock vina* using *Openbabel*. My deadline for running the program is Feb 16th, whatever the state of *enz*. Primer design for site-directed mutagenesis will be completed by the 17th.

5.1.2 Machine learning-based

A pretraining dataset has been mined from KEGG and the model has been constructed and is ready for generative training. A dry-run of generative pre-training will be complete by Feb 16th.Final model enhancements will be added prior to the pre-training wet-run, scheduled for Feb 16th.

A proof of concept lab screening experiment is ready to be done. I'm expecting this will take 1 week for all five mutants. Data will be processed into a training data set. Training the model on the proof-of-concept screening dataset can start imediately after it becomes available. The proof-of-concept pilot can be used to inform compound library and pool of mutants for screening. Alternatively a set of mutants and compounds can be designed right away.

A program for library selection can be ordered based on the results of the pilot or if purchasing restrictions permit it can be ordered right

5.1.3 Writing

I am currently writing chapters for both approaches as papers. I will need some advice on overall features of the papers, depth of literature reveiws and data to include and data to exclude which will guide my experimental work. I will send you the draft PDFs for comment by Feb 16th.

I will expand the draft papers into thesis chapters from here up until submission. By the end of the month I will have sent template papers for each project for comments and have a clear direction of where to expand into a thesis.

I think you need to get closer to haveing results before this would be useful. Writing as you go is good, but you can't write the results and discussion until you have results.

5.1.4 Targets

- Feb 16th - pre-train model dry run
- Feb 16th - run enz evolution run
- Feb 17th - design primers for enz mutants
- Feb 16th - send draft papers for comment
- Feb 16th - Start pilot screen
- Feb 16th - Finish pilot screen
- Feb 21st - implement flexible docking and loop remodelling in *enz*
- Feb 28th - Generate mutants for both projects - order primers
- Feb 28th - Order compound library
- Feb 28th - Get writing advice
- Feb 28th - Set up LCMS pipeline
- Feb 28th - Finish template papers for thesis writing

5.2 March

5.2.1 Structure-based

In March I will prepare DNA stocks for mutants predicted using *enz*. The maximum number of rounds of site-directed mutagenesis will be three, so this work can be fit into one month. Any available slots for incubator and centrifuge time will be booked for next month. I will create benchmarks for mesotrione product detection by LCMS using BM3 A82F/F87V I have on hand. This establishes a pipeline for upcoming product detection experiments.

5.2.2 Machine learning-based

At this point, a compound library must have been ordered. Allowing a lead time of one month, the screen will be ready for April. Primers should arrive early in the month. DNA work here will be batched with DNA work for the *enz* project.

5.2.3 Writing

Computational methods: I will write a detailed description for the *rio* model, including relevant background. I will write a detailed description of the *enz* / genetic algorithm combination. Methods can be written with input from Syngenta.

5.2.4 Targets

- March 1st - order compound library
- March 1st - order primers for both sets of mutant
- March 30th - finish DNA work
- March 30th - Benchmark LCMS experiment of mesotrione and BM3 A82F/F87V
- March 30th - Write computational methods for *enz* and *rio*

5.3 April

April will be dedicated to expression and purification of mutants from both projects. By batching expressions and purifications I can work on several mutants concurrently. *rio* mutants have higher priority.

5.4 Writing

Write all lab methods.

5.4.1 Targets

- April 30th - Finish mutant expression
- April 30th - Have purified most *rio* mutants
- April 30th - have started all purifications
- April 30th - finish writing all lab methods

5.5 May

5.5.1 Structure-based

enz mutants finish purification in May. By the end of the month all *enz* mutants should be purified and ready for testing. Testing of these mutants takes place in batches of 3-4, to be completed by end of June.

5.5.2 Machine learning-based

All *rio* mutants are to be purified in the first weeks of May with higher priority than the *enz* mutants. The screen will happen in May. The screen can be done in 1 week, but I'm allowing 1 month for up to 8 mutants and 96 compounds for leeway. It'll be done in one batch as far as posible. Model training can begin as soon as the data is available. Some parameter searches will take place here, lasting one week. The trained model can be deployed with an existing genetic algorithm to generate fit mutants immediately with indicated model certainty for each. Mutant generation will be constrained to within 3 mutations of existing DNA templates. Primers for site-directed mutagenesis to be ordered immediately.

5.5.3 Writing

Make any necessary retrospective changes to methods. Start preparing results section based on screen.

5.5.4 Targets

- May 30th - Finish screening
- May 30th - Train model on screening data
- May 30th - Predict new *rio* mutants and order primers
- May 30th - Finish purification of all *enz* mutants, start testing

5.6 June

5.6.1 Structure-based

Test K_d, K_{cat} and indication of metabolite production for all *enz* mutants by end of month. Finish *enz* lab work.

5.6.2 Machine learning-based

Express all *rio* mutants by end of month. Begin purifications for all *rio* mutants. Begin testing K_d, K_{cat} and indication of metabolite production for all mutants by end of month.

5.6.3 Writing

Write results for *enz* mutants. Write results so far for *rio*.

5.6.4 Targets

- End of month - Finish all lab testing for *enz*
- End of month - Have begun all purifications for *rio* mutants
- End of month - Finish writing results for *enz*

5.7 July

5.7.1 Machine learning-based

Purification and testing of *rio* mutants must proceed as fast as possible in July, after which all lab work will finish.

5.7.2 Writing

I will write a discussion for *enz* and results and discussion for *rio*.I will also start writing introductions for each peice of work. I aim to reach a minumum viable product (MVP) for each introduction by the end of the month.

5.7.3 Targets

- Finish lab testing of *rio* mutants
- End of month - End all lab work
- End of month - Finish *enz* results and discussion to MVP
- End of month - Finish *rio* results
- End of month - MVP for both introductions

5.8 August

5.8.1 Writing

At this point, I should have finished writing methods and results sections for both projects and the discussion for *enz*. I will write a discussion for *rio*. I will finish introductions for each section and iteratively apply suggestions from readers.

5.8.2 Targets

- Finish *rio* discussion
- Finish *rio* introduction
- Finish *enz* introduction
- Ready to submit