# Plan and schedule for 2021

James Engleback

February 18, 2021

## Contents

# 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 mutantions with predicted activity. Prediction of BM3-mutant binding will be used as a fitness metric in virtual directed evolution experiments. The resulting mutants will be made and tested for predicted activity in the lab.

The other uses a deep learning model trained on screening data of BM3 mutant binding with herbicide-like compounds, with the aim of being able to predict the activity of any BM3 mutant with any herbicide. This work

involves high-throughput screening of BM3 mutants against herbicide-like compounds. Mutants resulting from virtual directed evolution experiments using the deep learning model for fitness estimation will be made in the lab and tested for predicted activity.

This document outlines each approach and a schedule for work to complete and write up each into a thesis by the September deadline.

## 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 hydroxylated 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, such as herbicides in development.

## 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 $K_d$, $K_{cat}$, $K_M$ 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 data of $K_d$ between 5-10 BM3 mutants and around 100 herbicide-like compounds, and will be used to make $K_d$ predictions for pairs of amino acid sequences and compound. The screening data will be generated in the lab using a high-throughput UV-Vis assay for measuring $K_d$.

The trained model can estimate $K_d$ rapidly and at scale, so will be used for large-scale virtual directed evolution experiments selecting for mesotrione affinity to generate pools of new BM3 mutants to be made in the lab and tested for predicted activity.

Both approaches generate pools of BM3 mutant sequences with predicted binding activity for mesotrione. The lab validation process is:

1. **Preparation of expression plasmids** will be done using site-directed mutagenesis. Several BM3 mutant plasmids are in stock and can serve as starting points. The search depth of the models will be limited to reduce the number of mutagenesis steps.
2. **Expression** of BM3 mutants in *E. coli* is fairly robust. Two mutants per shaker should yeild sufficient protein for testing. The full length protien will be expressed.
3. **Purification** by a single Nickel-affinity chromatography step is sufficient to reliably yeild BM3 mutants at a purity suitable for measuring $K_{cat}$, $K_d$, $K_m$ and product formation.
4. **Measurement**
    - $K_d$ will be measured by titration of compound into BM3 with UV-Vis measurements. This experiment typically lasts 20 minutes.
    - $K_m$ **and** $K_{cat}$ will be measured by monitoring NADPH consumption at 340 nm.
    - **Product formation** will be detected my LCMS of a turnover reaction. Indication of any mesotrione metabolite is the target. Further elucidation is future work.

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

Template x-ray crystal 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 be suffuciently accurate for use in fitness evaluation in a virtual directed evolution experiment.

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* Application Program Interface (API) I've made for this work that automates template-based structure prediction and docking. It can be used for virtual directed evolution by combination with a genetic algorithm. Docking calls *autodock 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 in a virtual directed evolution experiment using a custom docking score as a fitness metric and a simple geneteic algorithm to breed and mutate BM3 variants over around ten rounds. Improvements to *enz* can be made to imporove accuracy:

- **flexible docking** - currently side chains are modelled as rigid which limits docking accuracy. *vina* is capable of flexible docking, so an attempt to implement it should be made.
- **loop remodelling** - currently conformation of loops and flexible regions are not recalculated, this would use the cyclic coordinate descent (CCD) implementation in *pyrosetta* for all unstructured loops within a cutoff radius of a mutation. CCD is a relatively fast algorithm and there in BM3 some important active site mutations are in loops.

Ideally support would be added for these non-features before running the genetic algorithm with *enz*. I am allocating a day to attempt implementing each of the proposed improvements.

### 2.1.2 *enz* validation

Accuracy of *enz* will be assessed by replicating ligand-bound BM3 structures from the PDB in *enz*. In the experiment, a set of ligand-bound BM3 crystal structres are selected to be replicated using *enz* using structure prediction from a suitable template structure and docking to predict faseible ligand binding poses. RMSD ($C_\alpha$) between the predicted and experimentally determined BM3 structures will to indicate *enz*'s structure prediction accuracy and inform features top be added to *enz* in future work. RMSD between ligand positions in both the PDB structures and the *enz* predictions will indicate accuracy.

### 2.1.3 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 only 4 short (3-6 residues) flexible regions in the active site, the impact of loop remodelling may be small and despite not docking ligands with flexible protein side chains, 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

$$score = \frac{1}{n}\Sigma_n \Delta G_n \times d_n \tag{1}$$

where $\Delta G$ 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 but can be flexibly changed if needed. Improvements to this heuristic for future work would include RMSD to a favourably positioned reference pose and implementing a machine learning-based score function for more accurate binding energy predictions.

The runtime of the genetic algorithm should not exceed 2 days, even on a small virtual private server. A pool of mutants yielded by genetic algorithm-based virtual directed evolution 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.

### 2.1.4 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 $K_{cat}$ and $K_m$ via NAPDH consumption
- **LCMS** product detection. A +16 *m/z* or fragmentation of mesotrione is considered a hit.

### 2.1.5 Order of events

The expected order of events for this project is:

1. Final changes to *enz* - if possible - timescale: 2 days

2. Validate *enz* against known BM3 crystal structures - 1-2 days
3. Run genetic algorithm, select pool of mutants - runtime < 2 days
4. Prepare the mutants in the lab and test for the proposed metrics - timescale < 3 months

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

## 2.2 Follow-up work

Short-term (months) future work on structure prediction-based enzyme design using *enz* would see implementation of additional *pyrosetta* structure prediction methods in the `refold()` function and implementation of high-resolution docking in the `dock` function using a method provided by *pyrosetta*.

Further work 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.

A notable advantage of a tensor math backend is compatibility with machine learning-based local structure prediction, which outperforms many traditional methods. The proposed machine learning-based structure predicion method will replicate that of Alphafold 1 and 2, which won CATEGORY in CASPX (2018) and CASPY (2020) respectively. This method relies on a conformation-dependent learned score function of protein structure likelihood and gradient descent to steer confomational changes to maximise the structure likelihood score. This step is notably fast, taking XXXX to XXx with XXXXX.

# 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 compound. The model aims to use amino acid sequence and chemical smiles alone to make $pK_d$ predictions. This is a proof of concept method and models in follow-up work would can be trained to estimate additional metrics.

[**?**] is

### 3.0.2 Approach

In this project, the approach to creating the model relies on creating a screening dataset of several BM3 mutants against a planned 100 compounds.

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. The Pre-training dataset contains pairs of P450 amino acid sequences and SMILES codes for reactants for the corresponding Enzyme Comission number. This has been gathered using data mining programs made for this work and currently contains the order of $10^5$ datapoints. The hope is that quantity of data overcomes quality during pre-training.

The sucess of pre-training will be measured by final model performance on the in-house screening dataset. Search for a suitable pre-training end poind can be automated using meta-learning techniques such as Bayesian optimization.

### 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 ayer 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.

### 3.0.4 Pretraining

The pre-training dataset has been mined from KEGG and currently holds roughly 0.2 million datapoints of enzyme sequence and smiles for reac-

tants and products. I've identified changes to the data miner that will could increase the dataset size close to $10^6$.

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 will discriminate between real and fake data, whilst the generator will 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.

There is scope to monitor training by assessing the quality of the generated sequences and molecular graphs. The extent of pre-training will be deterimend by trails that measure accuracy of the model on screening data when retrained.

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 an output head for each target metric. This dataset conatins $X$ data points and exposes the model to xenobiotics, which may assist with training on the in-house dataset.

Saved pre-trained models can then be transferred to the screening data task for their accuracy to be assessed.

#### 3.0.5 Pilot screen

A 384 well plate-based UV-Vis assay for measuring $K_d$ has been developed for this work. Development is ready to write into a *methods* section. 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.

An alternative assay format that measures the UV-Vis response of BM3 to a single high concentration can be readily adopted. This would save X compound and X protein per plate and increase the compound capacity of each plate to 192. This alternative may produce data with similar value to $pK_d$ with either a hit or miss end point, or a quantification of BM3 response. It will also increase the number of compounds that can be tested.

A pilot screening trial of 100 in-house compounds aginst 5 BM3 mutants has taken place and preliminary results indicate X. The 5 BM3 mutants in-

cluded:

- wild type
- a82f/f87v

The screening compounds contained a herbicide-like and structurally diverse set of herbicides and compounds selected from an in-house FDA-approved drug screening library. The library was filtered according to herbicide-likeness rules (ref) and a set of $n$ compounds was selected using a compound diversity picking algorithm (MaxMin). The screening compounds contained an additional set of $n$ herbicides, including mesotrione.

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.

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 20 days if the stock compounds and that they can provide a COSHH form that covers all compounds. *Molport* has an API that I have been using to autogenerate quotes and their lead times. The compound selection method from the *Molport* library is the same as that used to select $n$ structurally diverse and herbicide-like compounds.

$n$ compounds are ready to be selected from the *Molport* database, where $n$ and $volume$ of compounds an be adjusted to fit buget constraints. $n$ and $volume$ will also depend on the decided assay format, either hit or miss or $pK_d$ measurements.

BM3 mutants for screening currently comprise X variants who have crystal structures in the PDB. Purified variants that are ready to screen are:

- we

X mutants have been expressed in *E. coli* and are ready to be purified:

- cell pellets

Mutants planned for expression, purification and screening are:

- planned

### 3.0.7 Screening

The mutant BM3 $n$ cell pellets can be purified immediately over the course of two weeks and screened against either in-house herbicide-like FDA-approved compounds or a new libraryselected from *Molport* over the course of another week.

Further rounds of mutant screening involve BM3 mutants with known altered substrate specificity can proceed as soon as the mutants are purified. The proposed mutants for screening are listed in order of priority:

1. a333s

Structure prediction and docking can be used to indicate altered substrate specificity before synthesis to probe the usefulness of each mutant in a screen.

### 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 and parralelizability. The genetic algorithm will generate amino acid strings of BM3 mutants based on constrained mutation sites, using predicted $pK_d$ towards mesotrione as a selection criteria.

Members of the mutant pool generated by the genetic algorithm will be selected for validation in the lab. A major factor is the number of mutations W.R.T existing plasmid stocks.

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 Follow-up 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.

## 4.1 Follow-up 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 this work in this follow-up work, it will be important to include a spray test.

# 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.

### 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. By the end of the month I will need a peice of work ready to submit.

### 5.8.2 Targets

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

# References

[1] Shuya Li, Fangping Wan, Hantao Shu, Tao Jiang, Dan Zhao, and Jianyang Zeng. Monn: A multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Systems*, 10:308–322.e11, 04 2020.