Open Targets Platform: Multi-Target Search. A tool for the prediction of drug efficacy based on pharmacological targets.

Name: Edward Jonathan Chalstrey

Supervisor: Ioannis Melas (UCB)

Internal supervisor: Irilenia Nobeli (Birkbeck)

Degree course: MSc Bioinformatics with Systems Biology

Birkbeck, University of London

Malet St, Bloomsbury, London. WC1E 7HX

ABSTRACT

In silico prediction of drug efficacy is increasingly becoming an important part of the early stages of the drug development pipeline. Here I present the development and testing of "Open Targets Platform: Multi-Target Search", a tool that uses data from "Open Targets Platform" (developed by Koscielny et al., 2017) to generate a ranked list of disease treatment candidates from a list of drug targets. The target list can in principal, come from any drug (novel or approved) and the resulting disease candidates used to guide subsequent clinical trials. In this report, I describe methods used to take the association scores that Open Targets Platform assigns for a given disease with each target and create a combined association score for multiple targets with that disease. The utility of these methods in predicting/suggesting disease candidates for a drug are evaluated with a testing dataset of 192 approved drugs, for which the targets and disease indications are known. This analysis shows my tool to be effective; significantly higher scores are assigned to the true disease indications of drugs (indication category) than for those drugs with other diseases (non-indication category). This was shown with Welch's *t*-test for two samples of unequal variances (the combined association scores grouped by indication and non-indication categories) which gave the best performing scoring method $p =\sim 2.41e\text{-}22$. This suggests that "Open Targets Platform: Multi-Target Search" can be used as an *in silico* screening tool for novel drugs, and suggest realistic disease candidates for treatment with the drug, based on the drugs' targets.

CONTENTS

ACKNOWLEDGEMENTS

With thanks to my supervisor Ioannis who has been fantastic throughout the project and UCB for offering the external placement opportunity. Thanks also to all the lecturers on the MSc course for an intellectually stimulating 2 years, to my boss for letting me work part time to pursue a degree in an unrelated field and last, but not least, my parents, family and friends for putting up with all the bio-chat!

INTRODUCTION

Drug development is a long and costly process. New *in silico* screening methods for drugs are helping to speed up the development pipeline and have huge potential for the future of pharmacology. The faster that new drugs can be developed, the more lives can be saved from currently incurable diseases. Virtual screening is much faster and cheaper than any experimental method and offers the possibility to explore additional therapeutic uses for existing clinically approved drugs. A variety of computational resources for drug efficacy prediction have therefore been developed. These methods (and the software tools that use them) leverage the availability of a wide range of large biological datasets and have already yielded novel relationships between drugs and diseases that have gone on to inform further testing in the drug development pipeline (Katsila, Spyroulias, Patrinos, & Matsoukas, 2016).

Examples of biological "big data" being used in the pharmaceutical industry (and elsewhere) include gene expression profiles (e.g. from drug response data), chemogenomic data, protein structures and drug target to disease interaction data (Katsila et al., 2016). A drug target is defined as any molecule, usually a protein, that a drug interacts with. Many of these datatypes have associated databases that are freely available online and are used internationally, greatly speeding up the pace of scientific research, including drug development. Examples of these databases include PubChem (Kim et al., 2016) for chemogenomic data, the Protein Data Bank (Berman et al., 2000) for protein structures and Open Targets Platform (Koscielny et al., 2017) which compiles evidence for the association of known and potential drug targets with disease.

There are a variety of different software that have been developed for *in silico* drug screening, which vary in both the methodology and the kinds of data used. What unites them is the idea of comparing computational representations of disease mechanisms with drug mode of action, as a basis for predication of efficacy (prediction of whether a drug could be used to treat a disease). Disease phenotypes are caused by faulty biological processes, such as the genetic mutations that affect cell proliferation in cancers. Compounds that interfere with or block such disease mechanisms can be developed into drugs that effectively treat disease. It's this premise that underlies the motivation for developing computational methods for efficacy prediction, now that the aforementioned (large) biological datasets exist.

One avenue in this field of research has involved analysing the effect of drugs on gene expression profiles. An early example of this

was the development of Connectivity Map, a resource that relies on transcriptional expression data to determine connections between genes, diseases and drugs (Katsila et al., 2016). This has been expanded on by the development of tools specifically designed for efficacy prediction, such as DeSigN, a web tool that allows the user to query a database with lists of differentially expressed genes from experimental data (e.g. from a differential expression analysis between cancer cell line and normal tissues). The results of this query include rank ordered suggestions of approved drugs, which could be repurposed for the disease the experiment was testing (Lee et al., 2017). The user-friendly interface used by many online bioinformatics tools (like DeSigN), makes it easy for scientists without advanced programming or bioinformatics skills to take advantage of the underlying software in their research. This design methodology informed the design process of the software developed for this project.

Another path taken by researchers for *in silico* drug efficacy prediction research has involved exploring the relationships between drug targets and diseases. As an example, a method developed by Guney, Menche, Vidal, & Barábasi, 2016, involved the development of a proximity measure, which is based on the path lengths between drug targets and the proteins associated with a given disease (the disease proteins generally being clustered within a neighbourhood of the interactome). This measure was shown to work as a good proxy for therapeutic effect, meaning that novel (or repurposed) drugs with a high "proximity" to a given disease are suggested as candidates for effective treatment. This approach can be used to explore the potential utility of new drugs that are undergoing clinical trials and inform subsequent experimentation. Drug candidates in early stages of testing that would otherwise have wasted time and money in later stages, because of their low likelihood of being effective, can be eliminated by *in silico* screening.

Similar software tools to those described here (or the same tools) can be used for the repurposing of approved drugs to treat/cure other diseases than they are currently marketed for. Software tools that can compare one drug with another across whatever data they use to describe a drug's mode of action, can therefore be used to suggest alternative drugs to treat a given disease, when two drugs have a similar profile. Repurposing drugs in this fashion, described as drug repositioning, has become a vital part of the pharmaceutical industry due to the reduction in time and cost when compared with the development of novel compounds, which have yet to be evaluated for safety in a clinical context (Martínez, Navarro, Cano, Fajardo, & Blanco, 2015). In light of this, tools such as DrugNet have been developed for the express purpose of drug

repositioning. Like the aforementioned tools, DrugNet provides the user with a priority list of diseases in search results for a queried drug (or vice versa), that a biological interaction network based on proteins, drugs and diseases suggests as candidates (Martínez, Navarro, Cano, Fajardo, & Blanco, 2015).

Project aims

With the current state of drug efficacy prediction research in mind, my MSc research project carried out with Birkbeck, UoL and biopharmaceutical company UCB has focussed on the creation and testing of a new *in silico* screening tool for novel drugs being developed by UCB, that can also be used for repositioning approved drugs. To achieve this, the plan was to utilize the availability of existing online big data resources (accessing their data via API or download) and develop a new tool that compares this data with data from candidate drugs, generating suggestions of diseases those drugs could treat.

At the project outset, three key questions were raised that would determine the hypothesis that would be ultimately tested and the nature of the software tool created: (i) what kind of drug data should be used; (ii) what kind of data will need to be retrieved (and from which database(s)) for use by the tool; and (iii) what calculations need to be performed in order to generate and rank disease suggestions for a drug.

To answer the first question, inspiration was taken from some of the aforementioned tools and methodologies which use either: (a) gene or transcript expression data from specific experiments; or (b) drug target data. Both these kinds of data were available and provided by UCB. The decision to use target data was ultimately informed by the answer to the second question, which was to utilize the freely available open source software Open Targets Platform (available at https://www.targetvalidation.org/), which compiles and validates evidence for target-disease associations and drug-target interactions from multiple data sources and evidence types (Koscielny et al., 2017).

Open Targets Platform is designed to allow users to quickly retrieve diseases that are associated with a searched target (or vice versa) via a user-friendly website or API, the target being a protein (or gene or transcript) as defined by standard nomenclature (Koscielny et al., 2017). Of primary importance to this project, each piece of evidence supporting a target-disease association in Open Targets Platform (from source databases or literature) has an association score that represents the strength of association (a value between 0 and 1). Each unique target-disease association also has an overall score which is calculated with a harmonic sum function (see Figure 10 in Appendix I) of the scores for each piece of supporting evidence (Koscielny et al., 2017).

The diseases in Open Targets Platform are part of a controlled vocabulary of terms taken from the Experimental Factor Ontology (EFO). EFO was developed as an attempt to standardise terminology and description usage in bioinformatics (Malone et al., 2010). In the context of Open Targets Platform, this allows for ontology-enabled searches and the recognition of synonymous disease terms (Koscielny et al., 2017).

In light of discovering this excellent resource, the idea for my project became to find a way to generate priority lists of diseases similar to those retrieved from an individual target search in Open Targets Platform, but for an entire list of targets associated with a drug being screened. The hypothesis being tested therefore was that a combined multi-target association score (calculated from the scores retrieved for each of a given drug's targets in Open Targets Platform), could be used as a proxy for therapeutic effect and inform likely disease candidates for effective treatment by the drug. An advantage this method has over other methods for drug efficacy prediction based on target data, is that by using Open Targets Platform, a wide range of evidence types and data sources are considered for ranking/scoring target-disease (and ultimately drug-disease) association (refer to Table 1 for details). Earlier ideas for the project that were not fully explored are described at the end of this report in Appendix III.

Table 1. List of the data types and sources used by "Open Targets Platform" and the tool developed for this project by extension. This figure is copied from (Koscielny et al., 2017).

Data type	Data sources
Genetic associations	GWAS Catalog, UniProt, European Variation Archive, Gene2Phenotype
Somatic mutations	Cancer Gene Census, European Variation Archive somatic, IntOGen
RNA expression	Expression Atlas
Drugs	ChEMBL
Affected pathways	Reactome
Text mining	Europe PMC
Animal models	PhenoDigm

MATERIALS AND METHODS

This project has involved developing and evaluating the functionality of "Open Targets Platform: Multi-Target Search", a web based bioinformatics tool for discovering applications for novel (or repositioned) drugs, of which the target interactions are known from experimental data. To that end, it uses the evidence for target-disease associations compiled by Open Targets Platform. The research that I will go on to present consists of three key steps: (i) the development of a module in the Python (3) programming language that leverages the Open Targets Platform Python client (available at http://opentargets.readthedocs.io/en/stable/) to enable searches of a list of targets (via individual target searches with Open Targets Platform's API) and provide combined association scores for diseases, based on the retrieved scores from each searched target; (ii) the creation of a functional web tool for performing "multi-target" searches that uses this Python module in tandem with standard web technologies such as a CGI script and the Bootstrap toolkit; (iii) testing the functionality of the combined association score approach using a testing dataset of approved drugs and their target interactions, for which we know the disease(s) the drugs are used to effectively treat in clinical practice.

Multiple target search functions

The Python module I have developed (see Appendix II of this report for details of how to view the code) consists of two key functions: `opentargets_associations` and `opentargets_evidence`; both of which pull data from the Open Targets Platform API on the associated diseases for each of the targets in a provided list, then compute a combined association score for the whole target list, for each disease. This is described further by Figure 1, in the form of pseudocode. Four separate methods for calculating a combined association score were tested as I shall go on to explain. The diseases covered by the results (output of either of the two key functions) can be limited to a list of provided disease terms, or include all the diseases from Open Targets Platform that have an association with at least one of the targets in the list. The targets in the target list required by either function can include any gene, transcript or protein searchable in the Open Targets API. As described previously, each disease term in Open Targets Platform is a part of the EFO.

Retrieval of the diseases associated with each target via the Open Targets Platform API was easy using functions available in freely available Python client. In `opentargets_associations`, the client function called `get_associations_for_target` is used to perform a search for each target. In

this case, what's returned from the API is an association score (a value in the range of 0 to 1) computed by Open Targets for each disease that their database has evidence of the target being associated with. This association score is computed using a harmonic sum of the association scores for each piece of evidence that supports a given target-disease association (Koscielny et al., 2017).

There are two separate methods I use in `opentargets_associations` to calculate a combined association score for a list of targets to a given disease, which I will refer to herein by the following definitions: (i) the **Target Weighted Association Score**, which multiplies the association score of each target by the total number of associations (the number of targets with an association) and sums the resulting values to create a combined score with no upper limit; (ii) the **Average Association Score**, which takes an average of the association scores for each target, where targets with no association to the disease are taken to have a score of zero, resulting in a combined score in the range of 0 to 1.

In `opentargets_evidence`, the client function used is `get_evidence_for_target`, which retrieves all the evidence used by Open Targets Platform to support a target-disease association, each piece of evidence having its own association score. In this function, the two combined association score methods are: (iii) the **Evidence Weighted Association Score**, calculated by multiplying the score for each piece of supporting evidence (taken from all targets) by the total number of pieces of supporting evidence, then summing these values to create a combined score with no upper limit; and (iv) the **Evidence Association Score,** calculated by the sum of all the association scores for supporting evidence, divided by the number of targets (although note this division was not actually necessary as the value remains proportional to the sum of scores), creating a combined score that also has no upper limit.

These four methods for calculating the combined association score are summarised in Table 2. One other option would have been to take the average of the scores for all pieces of supporting evidence. This was not used as an alternative method for calculating a combined score, because I felt this method would unfairly penalize target list to disease associations where all the targets had some supporting evidence, relative to target lists where only one or a few of the targets were associated with the disease but there were many pieces of supporting evidence.

These functions were tested using the Python module `doctest`, to ensure that the computed combined association scores for target list searches conformed to the expected results calculated manually from API searches of targets (see the code, link provided in Appendix II).

Table 2. Methods used to create a combined association score for a list of targets with a given disease, based on the association scores for each target and that disease in Open Targets Platform. "Association scores" refer to the scores computed by Open Targets Platform for a target and a disease, whereas "Evidence scores" refer to the scores assigned to each piece of evidence supporting a target-disease association and *s* to either of these in the relevant case. In each case, *S* is the combined association score, *n* is the number of targets and *v* is the number of pieces of evidence.

Scoring method	Scores used in calculation	Calculation performed	Score range
Target Weighted Association Score	Association scores	$S = \sum_{i=1}^{n} s_i n$	0 - ∞
Average Association Score	Association scores	$S = \frac{\sum_{i=1}^{n} s_i}{n}$	0 - 1
Evidence Weighted Association Score	Evidence scores	$S = \sum_{i=1}^{v} s_i v$	0 - ∞
Evidence Association Score	Evidence scores	$S = \frac{\sum_{i=1}^{v} s_i}{n}$	0 - ∞

Figure 1. Pseudocode for the core functions of the Python module developed for this project, which is used to compute the combined association scores with one of the four methods described in Table 2. The input is a list of drug targets in standard nomenclature.

```
Create storage dictionary

For each target in target list
  search target in Open Targets Platform API and store retrieved association score for each disease in results to the dictionary

For each disease in storage dictionary
  calculate combined association score from stored association scores retrieved for each target
```

Web tool

Leveraging the capabilities of the Python module search functions for target lists described previously, a web interface has been designed that allows the user to perform a search for disease associations of a target list, ranked by one of the combined association score methods. Any number of targets can be entered in a search and the results can be filtered by evidence type and/or limited to specific diseases of interest. If a target entered is not found to be searchable in Open Targets, this will be skipped over when attempting to retrieve results from the API (although this still affects the combined association score calculated for each disease, as the bad target does count towards the total number of targets used in combined score calculations). By contrast, any disease terms entered that do not have an exact match in Open Targets Platform will be replaced by an EFO synonym, which can be used instead when generating the results. For example, if one of the disease terms by which to filter results is "Depression", the results will contain the combined association score of the targets for "Unipolar depression". A disease will only be omitted from the results if no EFO synonym can be found.

The default option for diseases to filter by is "All diseases", which ensures that combined association scores will be calculated for all the diseases from Open Targets Platform that have an association with at least one of the targets in the list.

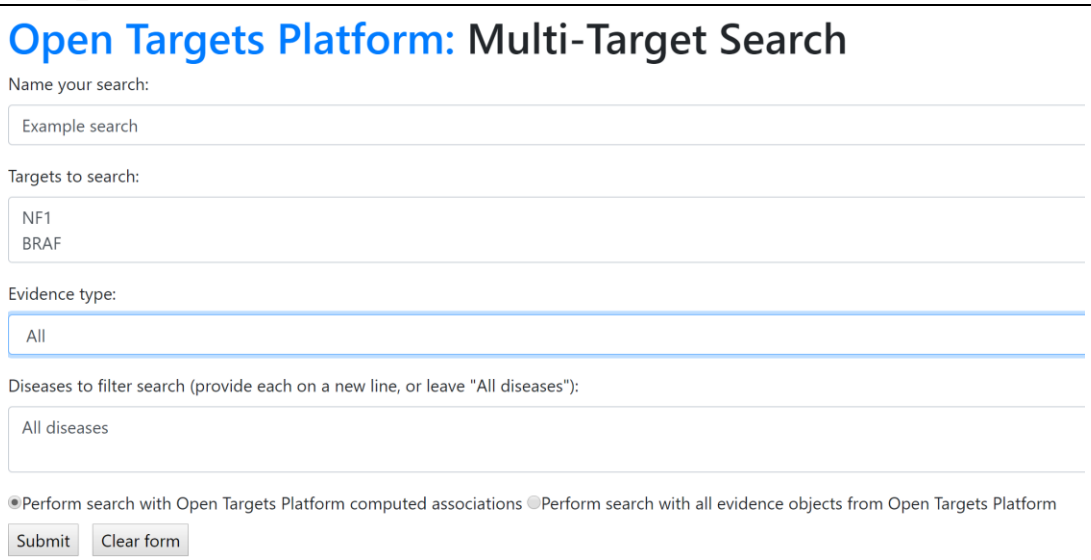
The option is available in the web tool to use either a) the association scores for each target computed by Open Targets Platform or b) each piece of individual supporting evidence for a target-disease association, when calculating the combined association scores for each disease with the target list. This selection informs whether the multiple target search is performed using `opentargets_associations` or `opentargets_evidence`, as described previously. The scoring methods used to rank the results for each of these selections are the Average Association Score and the Evidence Association Score respectively, which the results presented and discussed later in this report showed to be better methods than the Target Weighted Association Score and the Evidence Weighted Association Score.

Selecting "Perform search with all evidence objects from Open Targets Platform" will mean that for any result that includes evidence of the association from drugs (Open Targets Platform gets this data from ChEMBL), the names of the drugs that support this association are included (this only applies if the Evidence type is set to "All" or "Drugs").

Examples of the web tool in action and a flowchart describing how it works can be seen in Figures 2 and 3.

Figure 2. Screenshots of example searches with "Open Targets Platform: Multi-Target Search", using the following target list: `NF1, BRAF, MAP2K1, KRAS, MAP2K3, STAG2, RAS, NRAS, G6PD, RAC1, PIK3CA, EGFR, MAP2K7, MAP2K6, MAP2K5, MAP2K2, ATXN1L, TP53, EZH2, MAP2K4`. **(A)** and **(B)** show the user input and results pages for a search with all Open Targets Platform evidence types and all diseases that are associated with at least one target using the `opentargets_associations` Python function. **(C)** and **(D)** show the user input and results pages for a search with only drug (ChEMBL) evidence considered, filtered by these disease terms: `Carcinoma, Melanoma, Leukemia, Pain, Unipolar depression, Hypertension and Osteoporosis`, using the `opentargets_evidence` function. Since the association score in **(B)** is a value from 0 to 1, the combined association is represented as a percentage. This is not possible for an unlimited score, as shown in **(D)**, which also displays drugs when the evidence type is "Drugs" or "All" and `opentargets_evidence` is used as the search function.

(A)



(B)

Target search results for: Example search

Disease	Association	Number of associated targets/ 20
Neoplasm	90.06%	19: NF1, BRAF, MAP2K1, KRAS, MAP2K3, STAG2, RAS, NRAS, G6PD, RAC1, PIK3CA, EGFR, MAP2K7, MAP2K6, MAP2K5, MAP2K2, TP53, EZH2 and MAP2K4
Cancer	89.89%	19: NF1, BRAF, MAP2K1, KRAS, MAP2K3, STAG2, RAS, NRAS, G6PD, RAC1, PIK3CA, EGFR, MAP2K7, MAP2K6, MAP2K5, MAP2K2, TP53, EZH2 and MAP2K4
Epithelial neoplasm	89.63%	19: NF1, BRAF, MAP2K1, KRAS, MAP2K3, STAG2, RAS, NRAS, G6PD, RAC1, PIK3CA, EGFR, MAP2K7, MAP2K6, MAP2K5, MAP2K2, TP53, EZH2 and MAP2K4
Carcinoma	89.63%	19: NF1, BRAF, MAP2K1, KRAS, MAP2K3, STAG2, RAS, NRAS, G6PD, RAC1, PIK3CA, EGFR, MAP2K7, MAP2K6, MAP2K5, MAP2K2, TP53, EZH2 and MAP2K4
Urogenital neoplasm	83.01%	19: NF1, BRAF, MAP2K1, KRAS, MAP2K3, STAG2, RAS, NRAS, G6PD, RAC1, PIK3CA, EGFR, MAP2K7, MAP2K6, MAP2K5, MAP2K2, TP53, EZH2 and MAP2K4
Adenocarcinoma	82.37%	19: NF1, BRAF, MAP2K1, KRAS, MAP2K3, STAG2, RAS, NRAS, G6PD, RAC1, PIK3CA, EGFR, MAP2K7, MAP2K6, MAP2K5, MAP2K2, TP53, EZH2 and MAP2K4
Reproductive system disease	81.90%	19: NF1, BRAF, MAP2K1, KRAS, MAP2K3, STAG2, RAS, NRAS, G6PD, RAC1, PIK3CA, EGFR, MAP2K7, MAP2K6, MAP2K5, MAP2K2, TP53, EZH2 and MAP2K4

(C)

Open Targets Platform: Multi-Target Search

Name your search:

Example search 2

Targets to search:

NF1
BRAF

Evidence type:

Drugs

Diseases to filter search (provide each on a new line, or leave "All diseases"):

carcinoma
melanoma

Perform search with Open Targets Platform computed associations Perform search with all evidence objects from Open Targets Platform

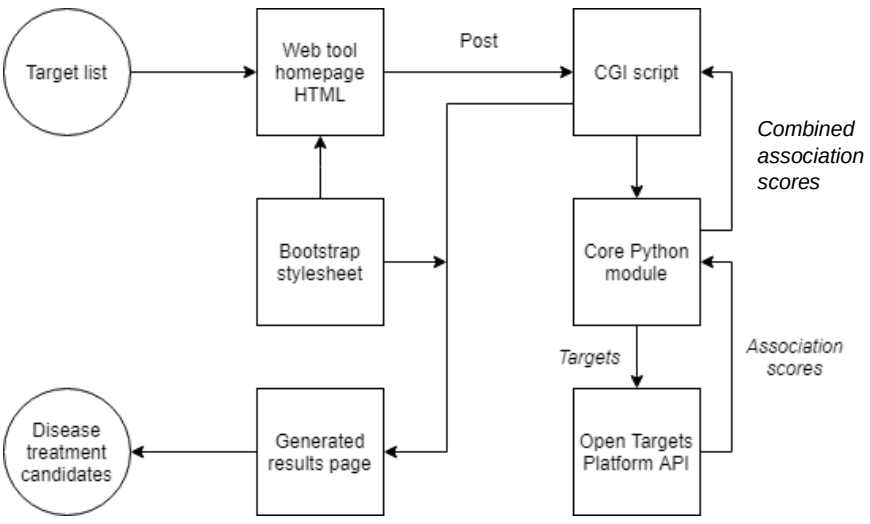
Submit Clear form

(D)

Target search results for: Example search 2

Disease	Combined association score	Number of associated targets/ 20	Drugs that interact with the searched targets and treat these diseases
Melanoma	3.17	5: BRAF, MAP2K1, PIK3CA, EGFR and MAP2K2	Vemurafenib, Dabrafenib, Sorafenib, Encorafenib, Regorafenib, Ly-3009120, Xl-281, Plx-8394, Trametinib, Cobimetinib, Binimetinib, Selumetinib, Pimasertib, Buparlisib, Erlotinib and Lapatinib
Hypertension	0.71	1: EGFR	Captopril
Carcinoma	0.24	4: BRAF, MAP2K1, EGFR and MAP2K2	Sorafenib, Regorafenib, Binimetinib, Erlotinib, Dacomitinib, Panitumumab, Cetuximab and Neratinib
Leukemia	0.1	5: BRAF, MAP2K1, PIK3CA, EGFR and MAP2K2	Sorafenib, Trametinib, Binimetinib, Buparlisib, Pixantrone and Erlotinib

Figure 3. Flowchart to show the interacting components of "Open Targets Platform: Multi-Target Search" that enable the display of disease treatment candidates based on the searched target list, ranked by combined association score (one of the methods in Table 2). The Core Python module is what I have developed, containing the `opentargets_associations` and `opentargets_evidence` functions.

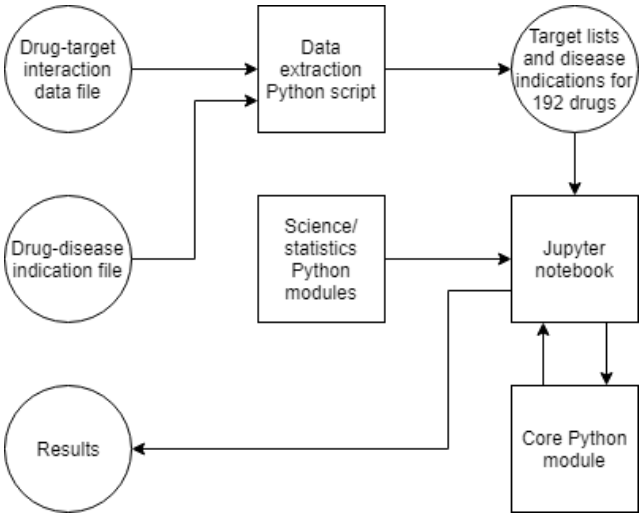


Testing the functionality of the combined association score approach

To evaluate the usefulness of the search results for multiple targets (retrieved using the Python module I developed, as described previously), a sample of 192 approved drugs for which the targets and indications (diseases treated by the drugs) are known, were used as a testing dataset. Each drug in the set had at least one (several had more than one) disease indication, resulting in a list of 196 diseases. By collecting the computed combined association scores for each disease in that list with every drug's target list (using the previously described Python functions: `opentargets_associations` and `opentargets_evidence`), the aim was to show that search results give high scores for the diseases treated by each drug (indications) and lower scores for the other diseases in the list (non-indications). Crucially, all the results I will go on to present were gathered by setting the `datatype` parameter of one of `opentargets_associations` or `opentargets_evidence` to `overall`, which considers all evidence types in Open Targets Platform (see Table 1) when calculating the combined association scores (in the web tool this parameter can be set to a specific evidence type, as can be seen in Figure 2).

The dataset sampled from was provided by UCB and is manually curated. It consisted of two data files, one of which included disease indications for a range of approved drugs and the other containing the targets of drugs. To generate the sample needed, a custom data extraction Python script was written that finds drugs from the indication file with a) unique disease(s) (i.e. a drug with the same indication as one already sampled is not also sampled) and b) diseases that are searchable in Open Targets Platform (i.e. are found in EFO ontology), which was tested using the Open Targets Platform API Python client. The target list for each sampled drug was extracted from drug-target interactions file and all targets were checked for searchability (in Open Targets Platform) in the same manner as the diseases. The Python script saved this information in a dictionary structure that would enable easy manipulation of the data once extracted. To achieve this, results were generated using Jupyter notebooks that first imported the data extraction Python script as a module and used it to generate the sample data as outlined. The Python module developed for multiple target searches (described previously) was also imported and used to generate results for the targets of each drug, filtering each search with the list of 196 diseases that the data indicated were each treatable by one drug in the sample. This process is summarised in Figure 4 and was repeated to test each of the scoring methods shown in Table 2.

Figure 4. Flowchart showing the interacting components/technologies used to evaluate the performance of the core Python module developed for this project by generating the output shown in the Results section of this report.



RESULTS

The results obtained in testing the functionality of the combined association score methods (see Table 2), included scores for each of the 192 drugs in the testing dataset, with each of the 196 diseases. This created four separate matrices of association scores to be analysed (one for each method). Since the expectation was that drugs would have higher scores for their indications than for other diseases, a comparison was needed between all the combined association scores for drugs with their indications, against the scores for those drugs with the other diseases in the dataset.

Each association score matrix was plotted as a heatmap, presented in Figure 5 for the Average Association Score matrix (equivalent heatmaps for the other three methods are presented in Figure 8 in Appendix I). Horizontal and vertical lines revealed by this heatmap show that some diseases have consistently high combined association scores across many of the drugs (e.g. Cancer) and the targets of some drugs have consistently high scores across many of the diseases (e.g. Deflazacort). The drugs are arranged alphabetically on the x-axis and the diseases they treat are arranged on the y-axis, creating a diagonal line of cells that represent the scores of drugs with their disease indication(s). Figure 5 does not show the highlighted cells as being consistently the highest scoring (darkest shaded), across all the drugs and diseases tested.

To further explore how the combined association scores vary across the testing dataset, the scores were categorised by indication and non-indication diseases and compared with two sample t-tests. Results of this analysis are displayed in Figure 6 and Table 3, showing three key findings: 1) that for all combined association score methods, the scores for indications are significantly ($p < 0.05$) higher than the scores non-indications; 2) that this difference is far more pronounced for the Average Association Score than the other three combined association score methods; and 3) for all scoring methods, there are some non-indication drug-disease associations with either maximal scores (in the case of Average Association Score) or scores as high or higher than the highest indication's score (in the case of the other three scoring methods).

To test the predictive power, the ROC (Receiver Operating Characteristic) curve (true positive rate against the false positive rate) was plotted and the AUC (area under the curve) calculated. The true positive rate and false positive rate were calculated from the matrix of combined association scores produced for a given combined association score method (one score for each drug-disease pairing, as in Figure 5) with the categorisation of each score as an indication or non-indication being known. These results

are displayed in Figure 7 and show the values for the AUC of each method, which measures the probability that a randomly chosen indication category score is higher than a randomly chosen non-indication category score. The key result here is that for each method, the AUC is above well above 0.5

Figure 5. Heatmap to show the Average Association Score (as defined in Table 2) of 192 drugs and 196 diseases, each disease being the indication for one of the drugs. Squares with a pink outline show the scores for drugs with their indications and darker coloured squares represent higher scores on a scale from 0 to 1.

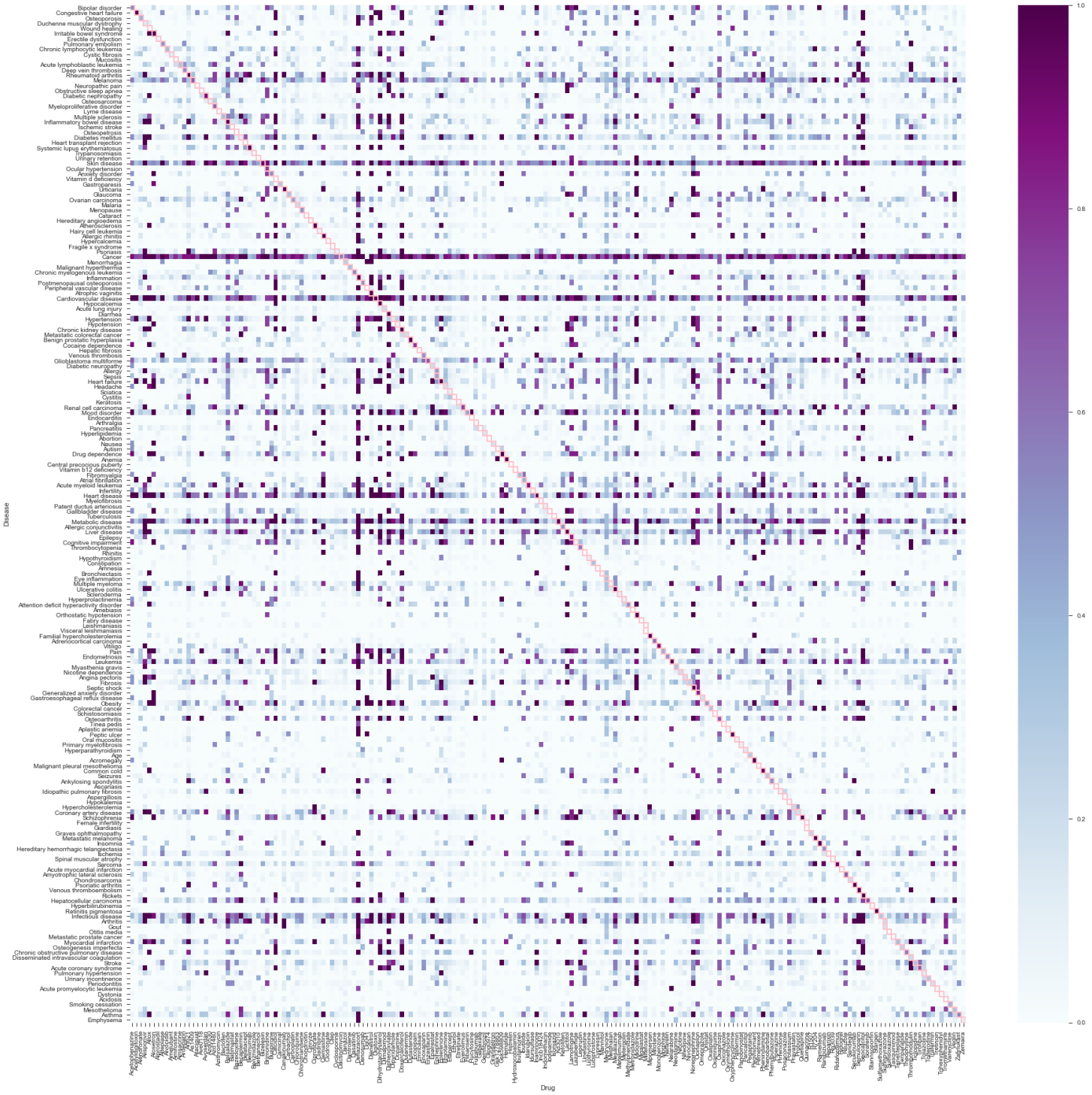


Figure 6. Results of comparing the scores assigned by combined scoring methods (as defined in Table 2), between those drug-disease associations expected by disease indication data and all other drug-disease associations. In each plot, "Indications" includes combined association scores for each of the 192 drugs in the testing dataset and their indicated diseases, whereas "Non-indications" contains scores for each drug with all the other diseases in the sample (there are a total of 196 diseases). **(A)** refers to the results for Target Weighted Association Score, **(B)** the Average Association Score, **(C)** the Evidence Weighted Association Score and **(D**) the Evidence Association Score as described in Table 2. Except for **(B)**, these combined association scores are unbounded (there is no maximum score, see Table 2) and are plotted on a log scale for clarity of presentation.

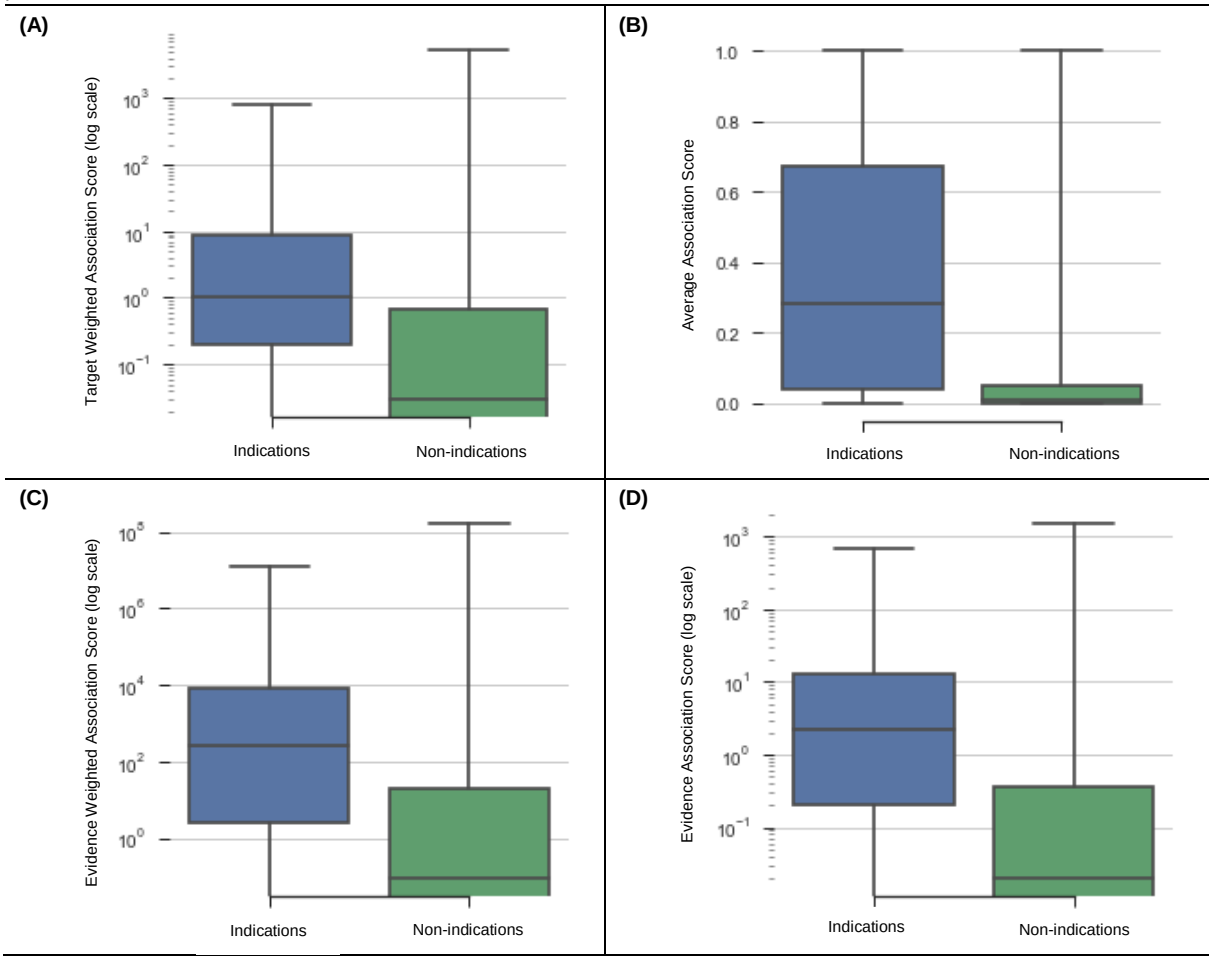
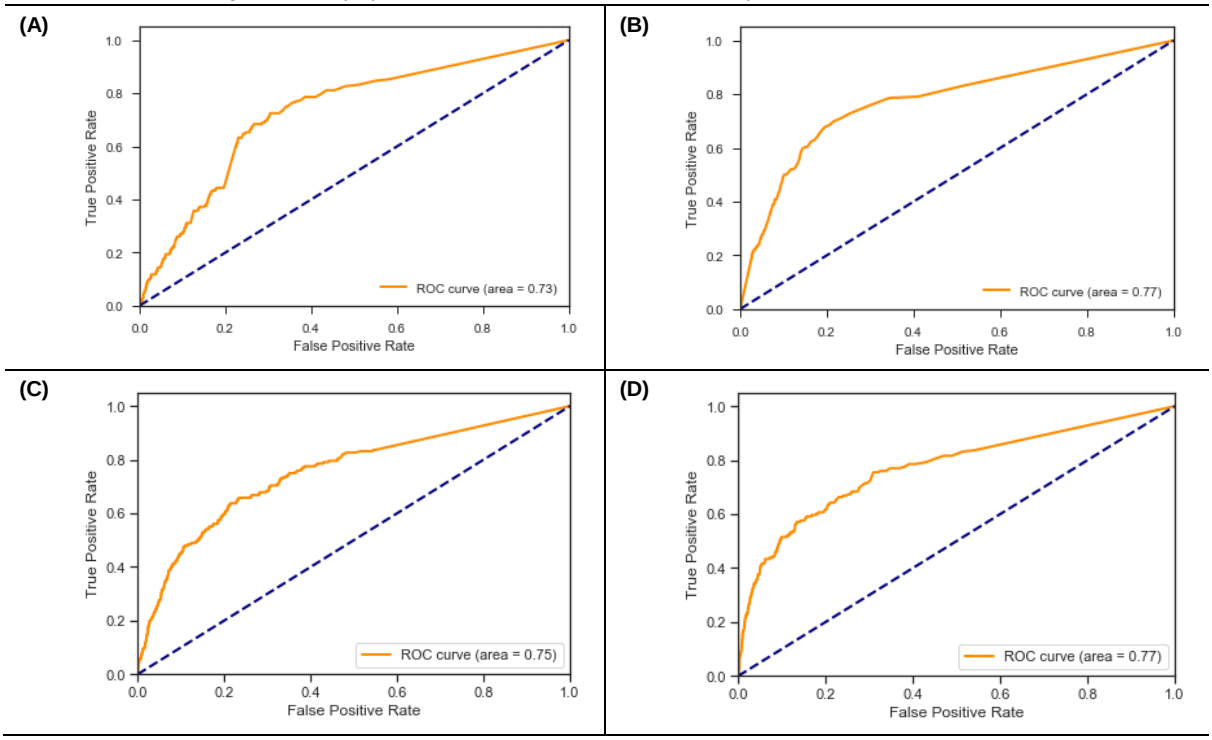


Table 3. Significance of the results shown in Figure 6, calculated with Welch's *t*-test for two samples of unequal variances between the “Indications” and “Non-indications” categories. The “Non-indications” category is much larger that the “Indications” category, because each drug only has one or two indications, making all the other diseases in the testing dataset non-indications for that drug.

Scoring method	T statistic	*p* value
Target Weighted Association Score	3.019008000131943	0.002873589724678486
Average Association Score	11.041333993479755	2.413761857264011e-22
Evidence Weighted Association Score	2.350457662434032	0.01973780261803963
Evidence Association Score	3.7666821089318656	0.00021895815428483382

Figure 7. ROC curves for **(A)** the Target Weighted Association Score, **(B)** the Average Association Score, **(C)** the Evidence Weighted Association Score and **(D)** the Evidence Association Score, each as defined in Table 2. The true positive rates and false positive rates are calculated from matrices of scores for the testing dataset of 192 drugs with 196 diseases (that are each an indication of one drug in the sample). Results above the dotted line in each plot are better than random.



DISCUSSION AND CONCLUSIONS

I have developed and tested the functionality of "Open Targets Platform: Multi-Target Search", a bioinformatics tool for drug efficacy prediction that takes a list of drug targets as input and ranks EFO disease terms on their combined association to the target list, based on association scores from Open Targets Platform.

I have shown that for a large testing dataset of 192 approved drugs and their respective indications (a unique sample of diseases, that are also EFO terms), my tool generally assigns combined association scores for drugs and their indications that are higher than for those drugs and other diseases. This result is statistically significant for all combined association score methods tested, as can be seen from Table 3 and is highly significant for the Average Association Score (p ~ 2.41e-22), which is a simple average of the association scores for a given disease across each of a drug's targets from Open Targets Platform target searches (see Table 2). Comparison of the true positive rate and false positive rate in the form of a ROC curve reveals that for each method, the AUC is at least 0.7 (refer to Figure 7), with the Average Association Score and Evidence Association Score having the highest at 0.77. This suggests that regardless of which method is used by "Open Targets Platform: Multi-Target Search" to calculate combined association scores, the tool has some predictive power in its suggestion of disease candidates for a target list.

However, the many false positives displayed in the heatmap presented as Figure 5 (darker shaded cells outside the diagonal line of indication diseases) suggest that the predictive power of the tool could be relatively low, with many non-indication diseases scoring highly. This is also reflected in Figure 6, which shows that for three of the combined association score methods, some of the "Non-indication" category scores are actually much higher than any of the scores of actual indications (note the log scale) and for the Average Association Score, some non-indication scores are maximal.

<u>Limitations of the testing dataset, methods and results</u>

The results presented in this report may not represent the true level of predictive power that "Open Targets Platform: Multi-Target Search" tool has for several reasons. Firstly, it's possible that the testing dataset (the sample of 192 drugs, their targets and disease indications) was suboptimal for the purpose of evaluating the tool. Some of the drugs in the testing set have indications with little or nothing to do with their molecular profile (targets) in the human body. For instance, *Isoniazid* has the indication "Tuberculosis", but this drug works as an antibacterial agent against the pathogen that causes the disease (information gathered from DrugBank

(Wishart et al., 2018)). As such, it's not surprising that the tool has not assigned a high combined association score between *Isoniazid* and Tuberculosis. Instead, the diseases the tool has assigned a high combined association score with *Isoniazid*, are diseases associated with the list of targets *Isoniazid* has in human cells, because these were the targets sampled from the data files provided by UCB (refer to Figure 3). In addition, some of the disease indications in the testing dataset are generic high-level disease categories. *Dactinomycin* for example has the indication "Cancer", for which my tool has assigned a high association with most of the sampled drugs (see Figure 5). This is not just unsurprising but expected, since many of the other drugs have indications that are a specific cancer. It does however mean that a lot of the high scores present in the "Non-indication" category in Figure 6 could be working to wrongly suggest that the tool has a lower predictive power than it truly does. This is because these scores are being treated as false positives when they are actually true positives. Similar examples of high level categories that have a high association to many of the sampled drugs include *Digitoxin*, with the indication of "Cardiovascular disease" and *Bizelesin*, with "Skin disease". Each of these (and others) can be picked out by visible horizontal lines in Figure 5 and may help to explain some of the high scores that can be seen in the "Non-indication" category of each box plot in Figure 6.

I have also noticed on examination of the results (displayed in Figure 5) that my data extraction Python script has mistakenly assigned the drug *Pegaptanib* with the indication "Age", when in the source data file, its indications are "Age-related macular degeneration" and "Neovascular age-related macular degeneration". This suggests that the data extraction Python script I wrote has worked imperfectly. I don't however, consider this realisation to be of detriment to the conclusion, supported by the results, that my tool generally assigns combined association scores for drugs and their indications that are higher than for other diseases. This is because, if anything, any incorrect indication is likely to have generated multiple false positives (high combined association scores to other drugs) and one possible false negative (a low score with the drug that has this indication) both of which would only serve to weaken the implied predictive power of the tool.

Limitations of the tool

Some of the low/zero combined association scores of the indication category diseases in Figure 5 (the white cells within the highlighted diagonal line of indication diseases) could be false negatives, resulting from the limitations of the target-disease evidence data collected from Open Targets Platform. It could be that

the target-disease associations suggested by Open Targets Platform are not always supported by the kinds of evidence that would be useful for suggesting a possible therapeutic use for drugs that have those targets. For example, evidence of association between targets and diseases from RNA expression data (taken from the EMBL-EBI Expression Atlas, see Table 1) could come from experiments that are specific to a cell type that is different from the cell type in which those targets interact with a given drug. In addition, since the tool relies on Open Targets Platform data, which in turn comes from external sources (see Table 1) it relies on these resources being reliable and comprehensive. It may simply be the case that many of the aforementioned false negatives simply lack evidence of association in the source databases and literature at this time. As described in the methods section of this report, the testing dataset came from data files provided by UCB (see Figure 3) and therefore could contain drug-target and drug-disease associations that are not covered by the target-disease associations in Open Targets Platform.

Conclusion

The results of this study indicate that the specific predictive power of the Multi-Target search may be limited in its ability to prioritise specific disease candidates over others for a new drug with an experimentally determined molecular profile. It could nonetheless be extremely useful in its primary aim of providing an initial computational screen for the process of drug development, as evidenced by the significantly higher combined association scores for disease indications compared with non-indications already discussed. For example, the search results displayed in Figure 2B show that for an example search of 20 targets, many of the top results are diseases are kinds of cancer (or neoplasm). This indicates that a drug with these targets could be effective in treating cancer and would therefore guide someone researching the hypothetical drug towards prioritising clinical experimentation for cancer treatment over other kinds of disease. In follow-up to an initial search of a novel drug's target list (such as in Figure 2A), a screening procedure making use of "Open Targets Platform: Multi-Target Search" could also include a second search, filtered by a selection of more specific disease terms (like in Figure 2C). These diseases could be those of particular interest to the researcher, or some of those prioritised in the initial search results, or indeed any diseases suggested by other computational screening tools and clinical experimentation.

Further research ideas

In further validation of a multi-target search method to get diseases associated with a drug, it seems that during the term of this project, the developers of Open Targets Platform have added a "Batch search" feature

to their website (available at https://www.targetvalidation.org/batch-search). Like the tool developed in this project, this allows the user to search with more than one target and returns a priority list of associated diseases that also shows which of the targets were associated with each disease, in a similar fashion to my tool as shown by Figure 2B (an example of batch search is provided as Figure 9 in Appendix I). In light of this, further research could involve carrying out a comparative analysis of Open Targets Platform's batch search feature with the tool developed for this project. It may also be useful to separate out EFO disease terms in the search results of my tool into general therapeutic area labels and specific diseases, as is done with batch search (see Figure 9), allowing users to more easily see which specific diseases their experimental drug could be useful in the treatment of. This would require further investigation of how to use the Open Targets Platform API to do this and modification of my code accordingly.

In calculating the overall scores for each target-disease association from individual sorted evidence scores, Koscielny et al., 2017 use a harmonic sum function (described by Figure 10 in Appendix I). As an alternative method for calculating the combined association score of multiple targets with a disease, to those I have already tested in this project (see Table 1), the harmonic sum of all the evidence scores across all targets could be taken, or indeed the harmonic sum of the overall association scores from each target. These would constitute two additional combined association score methods to be tested in the same manner as the other four.

To expand upon the results presented in this report, it could also be useful to carry out the same steps described in Figure 4 and generate an equivalent set of results shown by figures 5, 6 and 7 but with the `datatype` parameter of the core Python module's functions set to a specific evidence type (see Table 1). Filtering by data from known drugs only (from the ChEMBL data source), or one of the other evidence types, could offer a version of the tool with greater predictive power, but this is as yet untested.

The next stage of research, were this project to be continued, could be to further validate the utility of the developed tool with target data from an experimental drug(s). This could be a new drug that has been shown to be clinically effective for a disease(s), but for which the association is novel and therefore not represented in literature/databases already. If results of a search with the list of targets that form this drug's molecular profile yielded the expected disease as a high/top scoring association, this would provide further evidence that the tool could be useful for computational screening of newer compounds.

REFERENCES

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., … Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research*, *41*(D1), 991–995. https://doi.org/10.1093/nar/gks1193

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Guney, E., Menche, J., Vidal, M., & Barábasi, A.-L. (2016). Network-based in silico drug efficacy screening. *Nature Communications*, *7*(May 2015), 10331. https://doi.org/10.1038/ncomms10331

Katsila, T., Spyroulias, G. A., Patrinos, G. P., & Matsoukas, M. (2016). Computational approaches in target identification and drug discovery. *CSBJ*, *14*, 177–184. https://doi.org/10.1016/j.csbj.2016.04.004

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., … Bryant, S. H. (2016). PubChem substance and compound databases. *Nucleic Acids Research*, *44*(D1), D1202–D1213. https://doi.org/10.1093/nar/gkv951

Koscielny, G., An, P., Carvalho-Silva, D., Cham, J. A., Fumis, L., Gasparyan, R., … Dunham, I. (2017). Open Targets: A platform for therapeutic target identification and Validation. *Nucleic Acids Research*, *45*(D1), D985–D994. https://doi.org/10.1093/nar/gkw1055

Lee, B. K. B., Tiong, K. H., Chang, J. K., Liew, C. S., Abdul Rahman, Z. A., Tan, A. C., … Cheong, S. C. (2017). DeSigN: connecting gene expression with therapeutics for drug repurposing and development. *BMC Genomics*, *18*(S1), 934. https://doi.org/10.1186/s12864-016-3260-7

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., … Parkinson, H. (2010). Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, *26*(8), 1112–1118. https://doi.org/10.1093/bioinformatics/btq099

Martínez, V., Navarro, C., Cano, C., Fajardo, W., & Blanco, A. (2015). DrugNet: Network-based drug-disease prioritization by integrating heterogeneous data. *Artificial Intelligence in Medicine*. https://doi.org/10.1016/j.artmed.2014.11.003

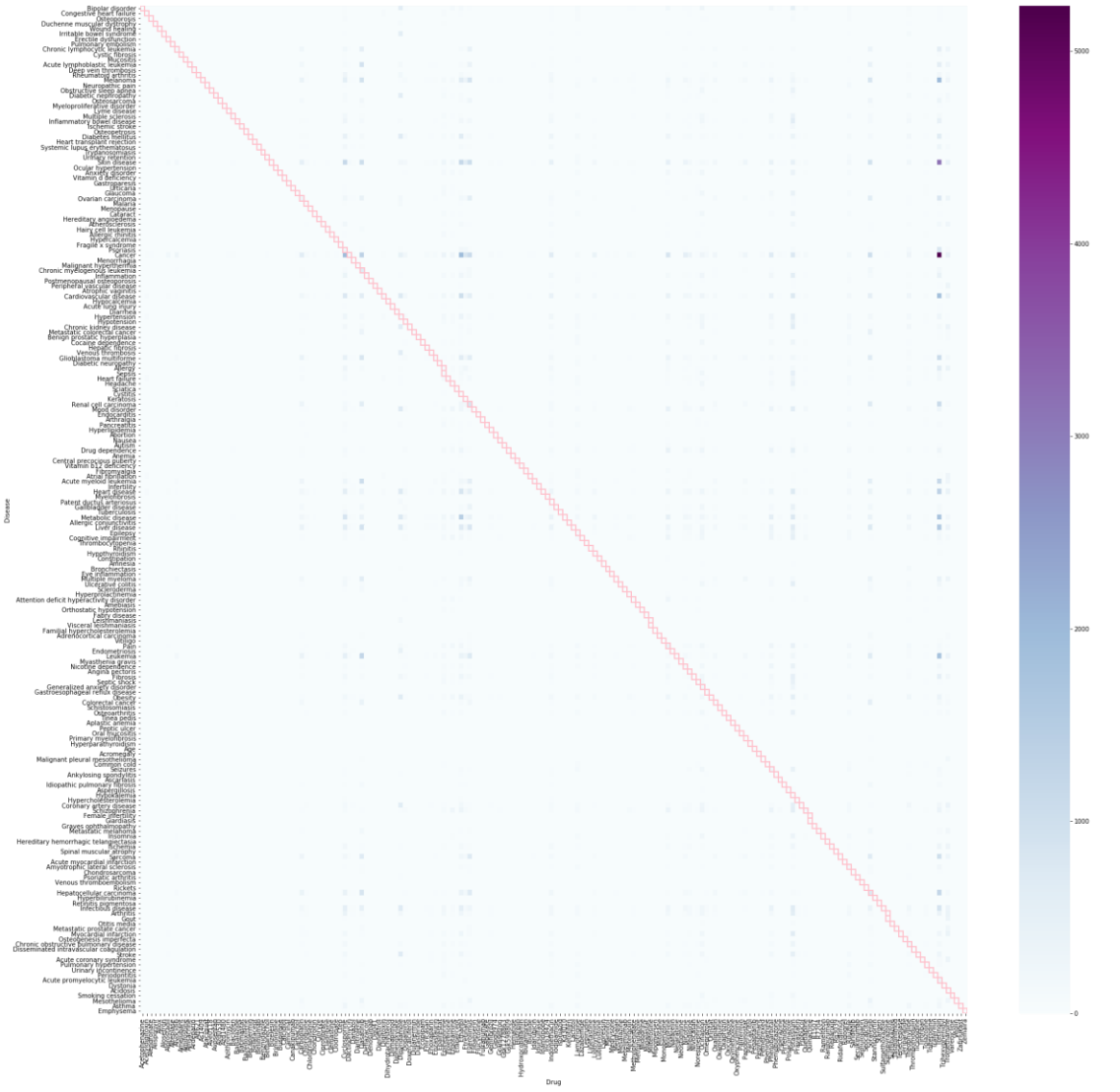
Szklarczyk, D., Santos, A., Von Mering, C., Jensen, L. J., Bork, P., & Kuhn, M. (2016). STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Research*, *44*(D1), D380–D384. https://doi.org/10.1093/nar/gkv1277

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., … Wilson, M. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, *46*(D1), D1074–D1082. https://doi.org/10.1093/nar/gkx1037

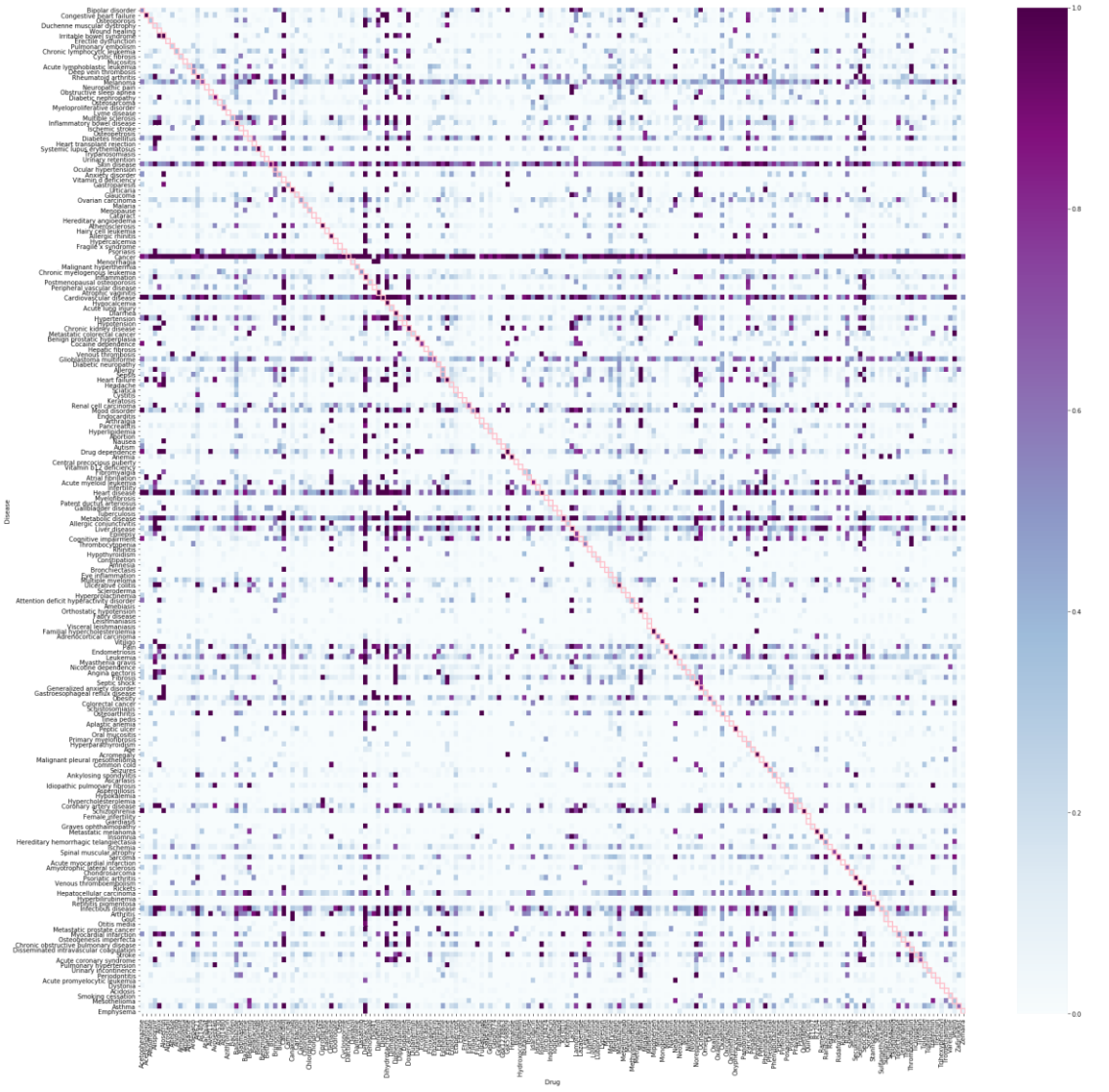
APPENDIX I: SUPPLEMENTARY RESULTS AND FIGURES

Figure 8. Heatmaps to show the combined association scores of 196 diseases, each of which is an indication for one of 192 drugs in the testing dataset. Squares with a pink outline show the scores for drugs and the diseases they treat and darker coloured squares represent higher scores. **(A)** and **(B)** refer to the results for Target Weighted Association Score, **(C)** and **(D)** for Evidence Weighted Association Score and **(E)** and **(F)** for Evidence Association Score, each as described in Table 2. In **(B)**, **(D)** and **(F)** the scores are normalised for each drug, meaning that the disease with the highest score (for each drug) is set to 1 and all lower scores are represented as a proportion of the highest score (a value between 1 and 0). This is for clarity of presentation.

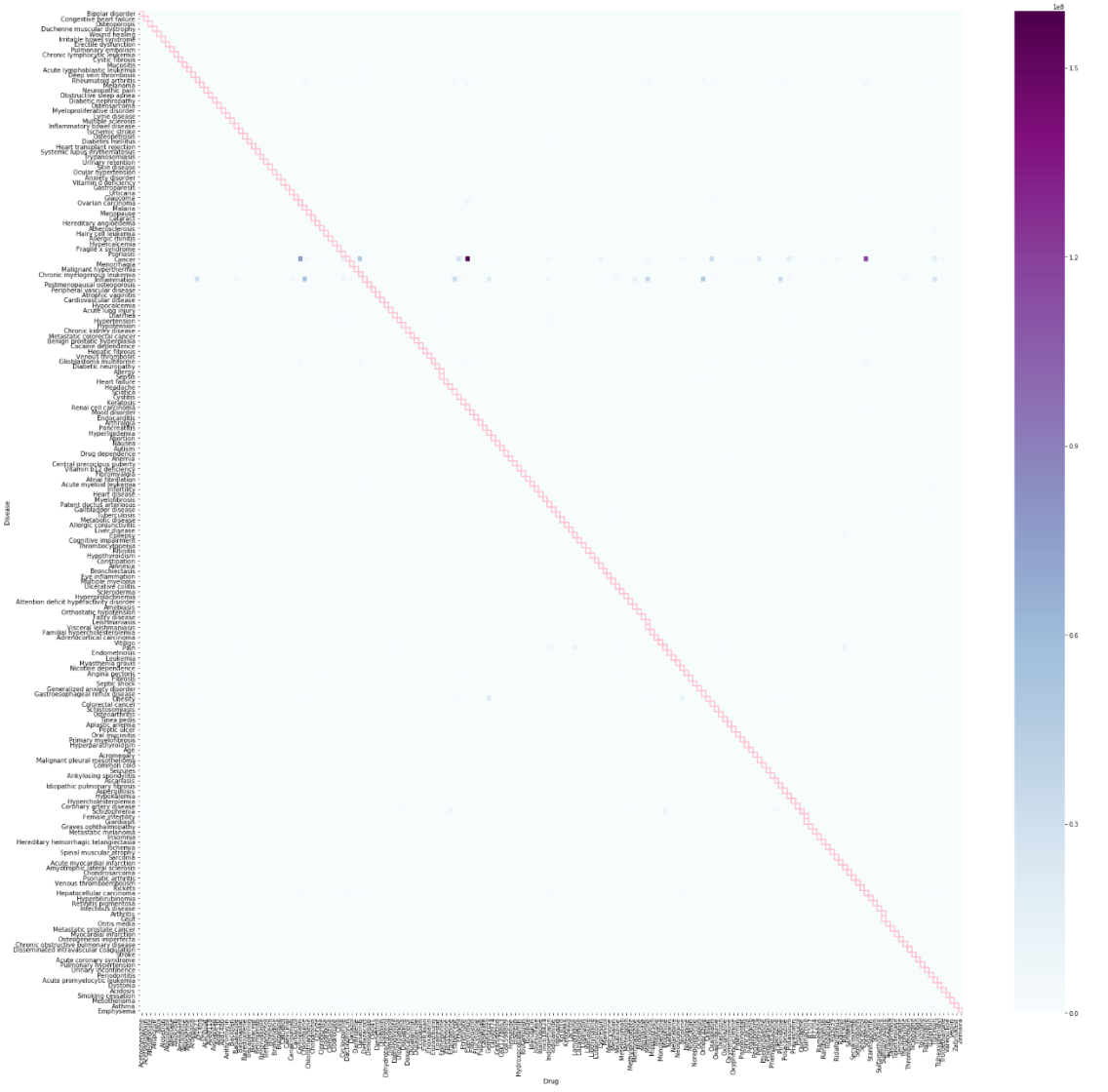
(A)



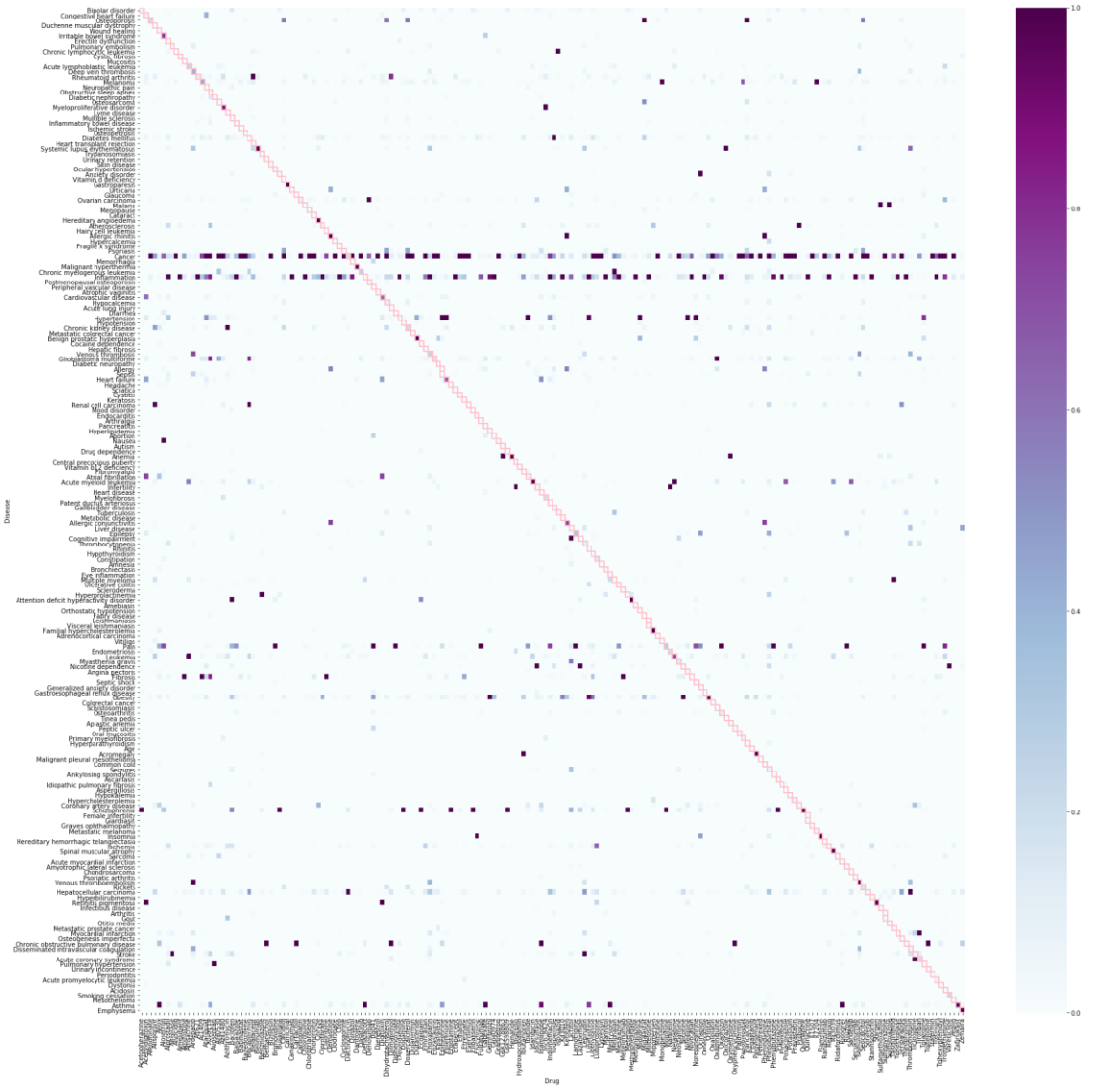
(B)



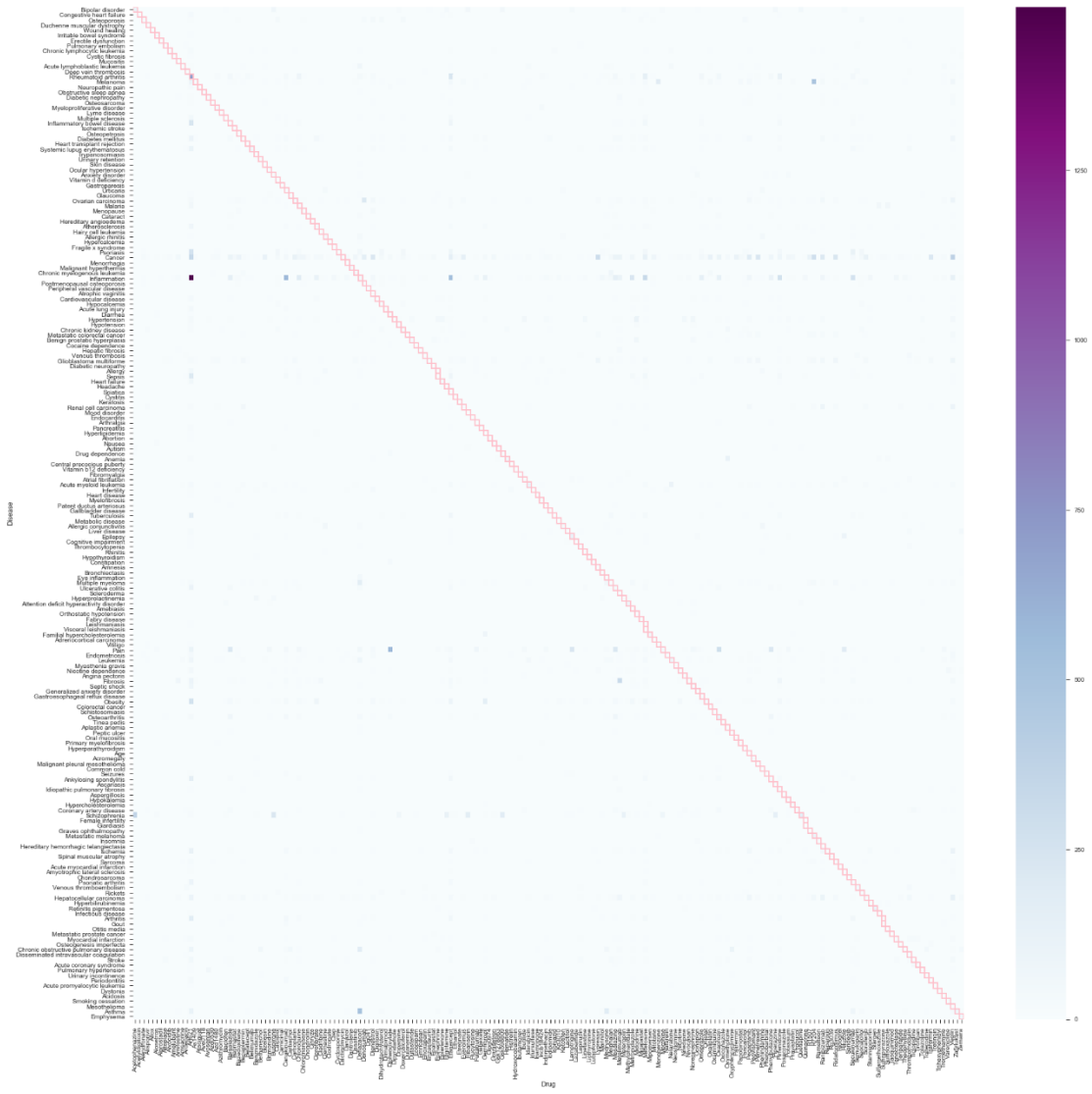
(C)



(D)



(E)



(F)

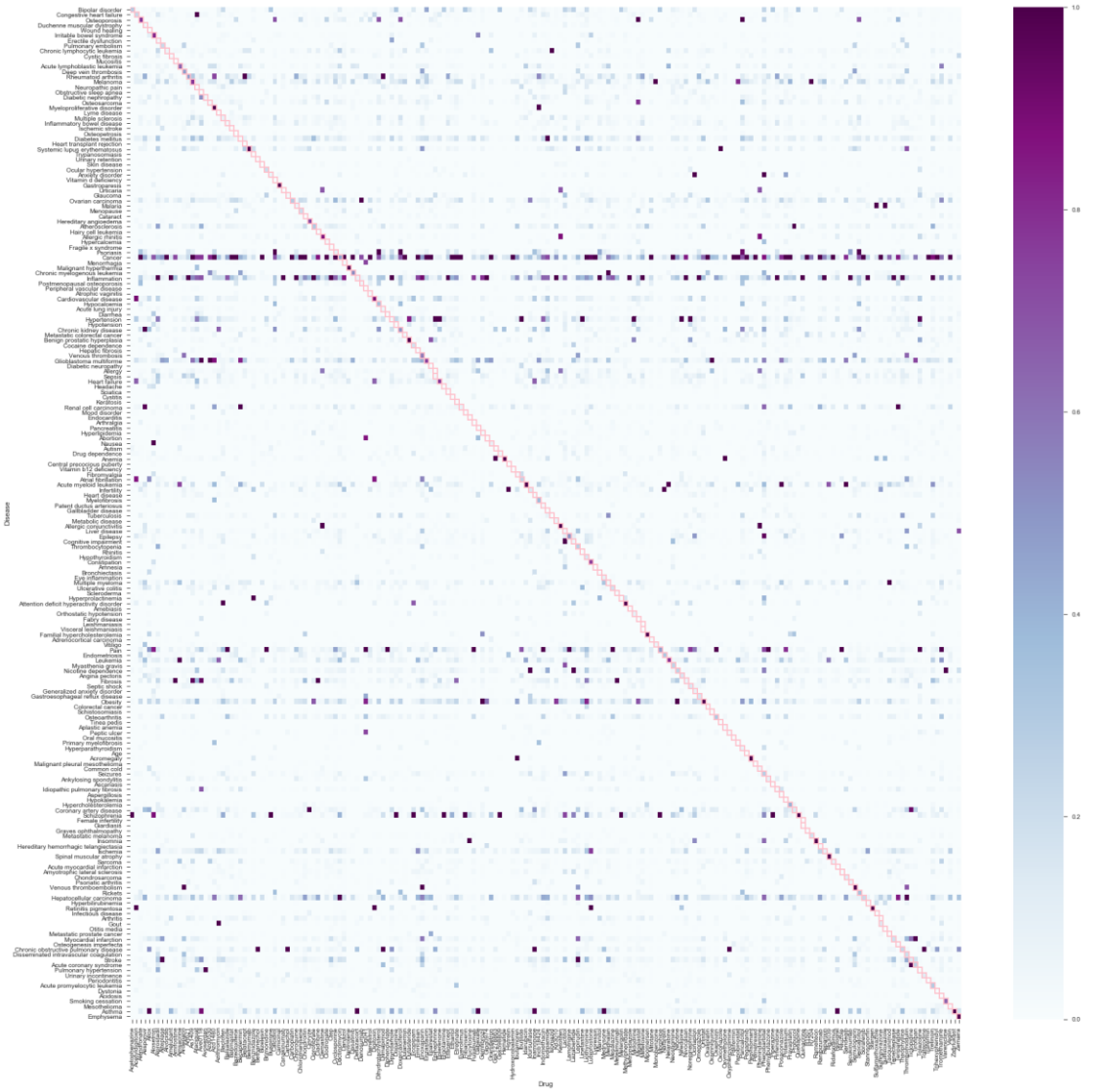


Figure 9. Screenshot of Open Targets Platform's batch search results for the same 20 targets listed in Figure 2. Unlike the tool developed for this project, batch search separates the EFO disease terms into therapeutic areas and specific diseases. You can see some of the same labels appearing here as in Figure 2B and some more specific ones such as "Gallbladder Adenocarninoma" (one of the results in Figure 2B is "Adenocarcinoma").

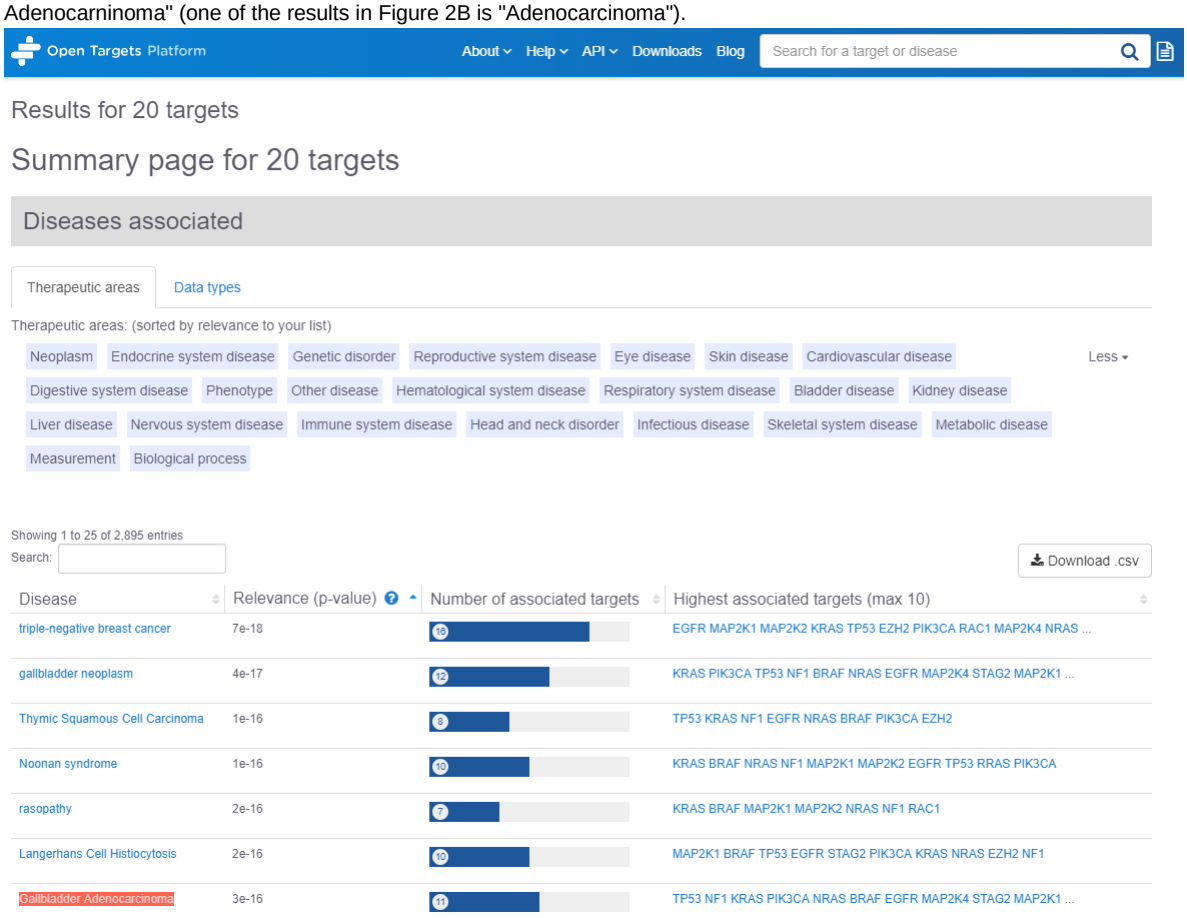


Figure 10. The harmonic sum function used by Koscielny et al., 2017 to compute an overall score for a target-disease association, based on scores from individual pieces of evidence supporting the association. S_1, S_2,...,S_i are the individual sorted evidence scores in descending order.

$$S_{1..i} = S_1 + \frac{S_2}{2^2} + \frac{S_3}{3^2} + \frac{S_4}{4^2} \ldots + \frac{S_i}{i^2}$$

APPENDIX II: HOW TO USE OPEN TARGETS PLATFORM: MULTI-TARGET SEARCH

Code for the project and web tool can be found at:
https://github.com/edwardchalstrey1/msc_project

After cloning the GitHub repository locally, the web tool can be launched with a working installation of Python 3. The only prerequisite Python package is that of Open Targets Platform, which can be installed with the command: `pip install opentargets`

A local test server can be opened from the root directory of the repository with the command:
`python -m http.server --bind localhost --cgi 8000`

APPENDIX III: ALTERNATIVE PROJECT IDEAS

This project took place over the course of 2 years on a part time basis, with a significant amount of the work being done in the summer of 2017. Before the hypothesis described in the introduction of this report was chosen as the aim for this project and the methods described were developed, alternative ideas for a computational tool to aid in the early stages of drug development were discussed and explored (another option was to base the project around finding applications for novel drugs using existing tools, rather than developing a new one). These ideas also worked around the starting point of utilizing freely available online data resources in novel ways with experimental data. Resources such as STITCH (Szklarczyk et al., 2016) and PubChem (Kim et al., 2016) were considered as options for the retrieval of useful biological interaction data (before Open Targets Platform was settled on) and before I focussed on drug target lists as the data input of choice for my tool, I considered making use of drug perturbation data from the Gene Expression Omnibus (Barrett et al., 2013).

Since these alternative ideas for the project never got developed into hypotheses and tested, I have not written about them in the main section of this report. Please see Appendix II for a link to the main GitHub repo for this project. The previous work mentioned in this section was developed in a separate repo that can be found here:
https://github.com/edwardchalstrey1/birkbeck_notes/tree/master/project