Lantern Pharma

Hackathon Challenge #1: NLP extraction to create ML training data

Classify drugs that are BCRP inhibitors or non-inhibitors

Generating data at scale from lab experiments on a question of interest costs significant money and time. Institutional public repositories of data are helpful, but are not available for many of the questions we'd like to model and answer. This challenge is about scraping and organizing data for a question of interest, so that a predictive model can then be built for the study.

The following research paper provides a good example of the approach and how the data is then used successfully in machine learning models:
"A Novel Approach for Predicting P-Glycoprotein (ABCB1) Inhibition Using Molecular Interaction Fields" Fabio Broccatelli, et. al.
DOI: 10.1021/jm101421d
https://pubs.acs.org/doi/abs/10.1021/jm101421d

Free access to full paper:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3069647/

The PGP inhibition determination was not done directly in a lab by the authors, it comes from their large and accurate literature collection which includes 52 cited publications with study dates ranging from 1995 to 2010. See link above which includes a spreadsheet organized with Drug Name, SMILES string, and response (PGP inhibitor or non-inhibitor) along with a second tab that includes references to research papers where each drug response comes from.

This challenge will be to use a similar approach and output format for drugs that are substrates vs. inhibitors of the Breast Cancer Resistance Protein (BCRP/ABCG2). BCRP is expressed in many normal tissues and plays an important role in drug absorption, distribution, and elimination.This information could be useful as a feature in a variety of ML tasks including but not limited to predicting blood-brain-barrier permeability for a given drug.

For more information on the topic, these references are a good starting point:

Role of the Breast Cancer Resistance Protein (BCRP/ABCG2) in Drug Transport
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287283/

Breast Cancer Resistance Protein - an overview | ScienceDirect Topics
https://www.sciencedirect.com/topics/medicine-and-dentistry/breast-cancer-resistance-protein

Transporters BCRP (breast cancer resistance protein)
https://www.solvobiotech.com/transporters/bcrp

Blood–Brain Barrier Transporters: Opportunities for Therapeutic Development in Ischemic Stroke
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287283/

Task : Collecting and organizing the Training data

The PGP reference paper above defined a criteria where drugs with IC50 lower than 15 μM were labeled as inhibitors and drugs with IC50 greater than 100 μM were labeled as non-inhibitors (drugs falling between those thresholds were excluded). While using this approach is ideal, it may be too time consuming to develop for this challenge. For this challenge it is acceptable to just mine the text in articles where drugs are already labeled as inhibitors or non-inhibitors even if the drug IC50 is not specifically listed to extract and apply appropriate labels.

After reading the above references, note that some of the papers include tables with summaries of the drugs, while others include the information in sentence form throughout the paper. It will be up to the data scientist to decide on NLP workflow to extract from sentences vs. a tool to extract from summarized tables in the document. Ideally a combination of these and other methods would be used to obtain as large sample size as possible, but if you choose only one of these methods then text extraction is preferred.

A few synonyms and relationships to be aware of:

Protein Name Representations with same meaning: “Breast Cancer Resistance Protein”, “BCRP”, “ABCG2”

Representations of an inhibitor: “inhibitor”, “transport inhibitor”, “inhibits”, “non transporter”, “unable to transport” (different authors may describe it different but from examples here see that they have the positive descriptors of “inhibit” and negative descriptors of “transport”)

Representations of a non-inhibitor: “non inhibitor”, “substrate”, “transport substrate”, “transporter” (different authors may describe it different but from examples here see that they have the negative descriptors of “inhibit” and positive descriptors of “transport”)

Resources with API’s for searching through and identifying papers on the topic include but are not limited to:
PubMed Central: https://www.ncbi.nlm.nih.gov/pmc/:
Biorxiv: https://www.biorxiv.org/

Note that the first step would be using API search tools to identify a list of documents relevant to the topic, so that only relevant documents go through nlp extraction.

Similar to the PGP reference project, after collecting and organizing the lists of drugs and their BCRP class, each drug will need to have its Canonical SMILES string (which represents its molecular structure) merged into the training set. Note that there may be some drug names that exist in one of these two sources but not the other. If you start by incorporating one source and have missing values, you may want to add in additional source.

Some resources to collect canonical SMILES by drug name using an api are:
PubChem: https://pubchem.ncbi.nlm.nih.gov/
ChemBL: https://www.ebi.ac.uk/chembl/

You may notice when reading some of these papers, that without subject matter expertise it may be difficult to distinguish drug names from gene names, protein names, or other pharmacological language. The base tools you may typically use for named entity recognition/extraction may not include drug names. You can choose any method you prefer for solving this challenge, but here’s an example tool ready in a package:

NLP Named Entity Recognition tool compatible with spaCy or NLTK
“Drug named entity recognition Python library by Fast Data Science”:
https://pypi.org/project/drug-named-entity-recognition/

Checking your answers:
FDA has a searchable database where you can filter to just BCRP which lists 8 BCRP transport inhibitors and 2 transport substrates. As a minimum check on the quality of your generated training dataset, it is recommended that your outputs have collected and properly labeled these well known small samples

https://www.fda.gov/drugs/drug-interactions-labeling/healthcare-professionals-fdas-examples-drugs-interact-cyp-enzymes-and-transporter-systems

It is also recommended that you directly read some of the research papers and manually annotate your own test set with the correct labels so that your code output can be compared for accuracy and improved as needed.

Sample Workflows:
Simple workflows for relevant text extraction are acceptable, for ideas on more sophisticated approaches, these sources include description of other methods and pseudo code for inspiration:
"Extracting Drug Names and Associated Attributes From Discharge Summaries: Text Mining Study"
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8135022/

"Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning"
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7489056/

Submission Requirements:

Your work will be done in Code Ocean which uses cloud computing resources, and your submission will be completed by sharing your capsule code and captured outputs within Code Ocean.

Start by creating an account on the Code Ocean Server setup specifically for this challenge:
You will be emailed an invitation to setup your user account

Once you have created your account, you will have access to the capsule titled "Lantern Hackathon Challenge #1 Description & Sample Submission" (and linked below) The capsule includes the challenge description in README.md as well as a sample submission containing the file names and formats your output should contain.
https://hackathon2023.radr-ai.com/capsule/5391547/tree

Note the capsule includes a pre-built virtual environment including python packages most likely to be needed for this challenge. To save environment build time, you can go to the "Capsule" tab in the upper left corner and click duplicate to start your own new capsule. You can rename the capsule to whatever you choose, and if needed you can click on the environment section to add any additional packages you need for your

approach. Note that adding additional packages means the environment will be re-built which takes some time, therefore it is recommended you do this as a first step with a simple quick running script. After the first run, the environment will not have to be rebuilt at each run as long as you don't add more packages.

When you are ready to submit your entry, please go to the "Share" button in the upper right corner of both the capsule and your captured data asset (the output) with:
Rick Fontenot
Peter Carr
Maximilian Carr

Output files required for submission (see sample files in Code Ocean capsule)

training_data.tsv → columns should include:
"Name" : string containing the name of the drug
"SMILES" : string containing the canonical SMILES obtained by using ChemBL, PubChem, or other by searching the drug name
"Activity" : the class label as you've determined through mining the text on the given drug. The values should be "BCRP_substrate" or "BCRP_inhibitor"
"Ref1": An numeric identifier for the reference source, i.e. which paper did this drug come from. A single reference may have a single drug or many drugs. Note that some drugs may be identified in multiple reference papers, in this case you should not duplicate the drug in multiple rows. Each drug should have a single row and additional reference columns such as Ref2, Ref3, etc can be added. This will be used in conjunction with a lookup table in the next file described.
"ChemBL_ID" : The reference ID for the source of your SMILES string if the sample's information is from ChemBL. Note that only one source is required for obtaining a SMILES string. Some drugs may only be in one source or the other, in that case the PubChem_CID can be left blank
"PubChem_CID" : The reference CID for the source of your SMILES string if the sample's information is from PubChem. Note that only one source is required for obtaining a SMILES string. Some drugs may only be in one source or the other, in that case the ChemBL_ID can be left blank

Name	SMILES	Activity	Ref 1	Ref 2	Ref 3	Ref 4	Ref 5	Ref 6	Ref 7	Ref 8	Ref 9	ChemBL_ID	PubChem_CID
Mitoxantrone	C1=CC(=C2C(=C1NCCNCCO)C(=O)C3=C(C=CC(=C3C2=O)O)O)NCCNCCO	BCRP_substrate	1										4212
Bisantrene	C1CN=C(N1)NN=CC2=C3C=CC=CC3=C(C4=CC=CC=C42)C=NNC5=NCCN5	BCRP_substrate	1										5351322

Aza-anthrapyrazole (BBR3390)	C1=CC=C2C=C3C(=CC2=C1)C=CC4=NNN=C43	BCRP_substrate	1										87489584
Topotecan	CCC1(C2=C(COC1=O)C(=O)N3CC4=CC5=C(C=CC(=C5CN(C)C)O)N=C4C3=C2)O	BCRP_substrate	1									CHEMBL84	
SN-38	CCC1=C2CN3C(=CC4=C(C3=O)COC(=O)C4(CC)O)C2=NC5=C1C=C(C=C5)O	BCRP_substrate	1										104842

references.tsv → this acts as a lookup table for the drugs in your training set columns “Ref1” etc. Each reference article used should have one row and columns should include:
“Code”: The numeric identifier used in the “Ref1” etc. columns of your training data
“Source”: information on how to locate the document that you extracted drug information from. Any format here is acceptable and can be mixed in different formats for each row if needed. Ideally the format would include either 1) The title, authors, and date 2) a doi link or reference or 3) A PubMed PMID number

Code	Source
1	Mao Q, Unadkat JD. Role of the breast cancer resistance protein (BCRP/ABCG2) in drug transport--an update. AAPS J. 2015 Jan;17(1):65-82. doi: 10.1208/s12248-014-9668-6. Epub 2014 Sep 19. PMID: 25236865; PMCID: PMC4287283.
2	TBD
3	TBD

metrics_summary.tsv → this contains a summary on the data you collected and will allow judges to sort entries by the criteria explained in the scoring section below. The columns should contain:
“team_name” : any unique name you’d like whether it’s an individual’s personal name or a name as agreed upon by your group
“team_contact” : a single e-mail address for the team lead in the event Lantern needs to contact you for follow up discussion or questions
“num_samples” : you should calculate this as the total number of unique drugs extracted from documents into your training set. Please be sure to properly de-duplicate

rows in your training set, if review by judges finds that your num_samples is inflated due to duplicate drugs your score will be penalized
“num_references” : you should calculate this as the total number of unique referenced documents that you extracted drug information from. Note this is not the number of documents searched, only the documents that were used on at least one drug included in your training set
“num_inhibitor” : the number of drugs in your training set that are labeled as BCRP_inhibitor
“num_substrate” : the number of drugs in your training set that are labeled as BCRP_substrate
“percent_minority” : should be calculated as the minimum of (num_inhibitor, num_substrate) divided by num_samples to represent how imbalanced the training dataset is.
“num_smiles” : the number of drugs in your dataset that have a canonical smiles string merged on into the column “SMILES”. If you decide to include strings such as “missing” or “not available” to represent missing values, please do not count these in the num_smiles total
“percent_smiles” : calculate as num_smiles divided by num_samples to represent what percentage of drugs in your training set you were able to join the SMILES string information for

team_name	team_contact	num_samples	num_references	num_inhibitor	num_substrate	percent_minority	num_smiles	percent_smiles
lantern_sample	rick@lanternpharma.com	51	1	25	26	0.49	51	1.0

methods_summary.md → A markdown file with a brief writeup on the methods you used, you can also include this information in the README.md of your capsule. Suggested topics to cover include:

1) Resources used in your work, such as PubMed, Biorxiv, ChemBL, PubChem, or others you uniquely added to your workflow
2) Method of document searching and collection, i.e. did you use specific api’s, did you search on the web to save documents and upload them as an input file etc.
3) After document collection, briefly explain your workflow and approach for extracting and labeling the drugs in your training set
4) How did you check your results for validity? For instance, did you manually annotate a smaller list of drugs by reading papers and use that to compare to the outputs of your scripts to check for correctness? If so, how many samples did you manually annotate for spot checking? If you created a more automated method of checking for validity, please briefly explain.

5) Bonus points for discussing how you could adapt your code to work for a protein other than BCRP, or to change topics from transporter inhibition to drug responses for a particular cancer type.

Scoring and determining the winner

Stage 1 Initial Scoring:

1. Teams will be ranked by the num_samples mined and extracted to their training set. The top team will receive 10 points, second ranked will receive 9 points…10th ranked will get 1 point etc. Entries lower than 10th do not earn points on this criteria
2. Teams will be ranked by the num_references mined and extracted to their training set. The top team will receive 10 points, second ranked will receive 9 points…10th ranked will get 1 point etc. Entries lower than 10th do not earn points on this criteria
3. All Teams with percent_minority > 0.2 receive 5 points. While some class imbalance is expected, the search for documents and the method of extracting drugs should focus on both class labels. If a team simply extracts only inhibitors because the document search or extraction method is easier, they will miss out on these points
4. All Teams with percent_smiles > 0.8 receive 5 points. While some drug names may not have information in ChemBL or PubChem, the percentage should be small. If your entry is not above this 80% threshold it is recommended that you attempt searching multiple sources. The named entity recognition tool will also give you synonyms for alternate names of the same drug and this information can be utilized in the search. These points are to encourage sufficient effort after document extraction because the SMILES string is what future ML modeling features will be engineered from.

Teams will be initially ranked by the sum of their scores on the 4 sections above, and the top 5 teams advance to scoring round 2.

Final Scoring:

Judges will review the methods writeup and code for the top 5 teams identified in stage 1 scoring and rank them qualitatively. The top ranked team earns 5 additional points, the second ranked team receives 4 additional points etc. These points will be added to the stage 1 scores to determine the overall ranking and winner.

This qualitative review will be based on the judges opinions on the following criteria:

1) Are the methods well explained and understandable by people outside of the team that created it?
2) Is the method scalable to increase the size of the training set by adding more documents in an automated way?
3) Are the approach and code adaptable for use on other topics or attributes of drugs?
4) Is the code properly organized with comments so that a new team member could follow and modify as needed?
5) Was the output sufficiently validated? (both the method used to validate for correctness as well as the amount of samples checked for correctness)