Lead optimization is the phase in drug discovery where we want to optimise properties of the molecule, so that it will meet all the requirements for being a drug. First of all, it cannot be toxic for people. So one of the properties which is checked, is activity on hERG which may lead to cardiotoxicity. In order to speed up lead optimization, machine learning models predicting properties can be applied and this is exactly your task: build ML model predicting hERG. The assignment is a toy problem that reflects small part of challenges that you might expect when working in Data Science department at Ryvu Therapeutics..

Data descriptions:

Dataset coming from paper https://jcheminf.biomedcentral.com/articles/10.1186/s13321-021-00541-z

- The training data used in this study can be found at https://github.com/Abdulk084/CardioTox/blob/master/data/train_validation_cardio_tox_data.tar.xz.
- Test set-I: The positively biased test set can be found at https://github.com/Abdulk084/CardioTox/blob/master/data/external_test_set_pos.csv.
- Test set-II: The negatively biased test set can be found at https://github.com/Abdulk084/CardioTox/blob/master/data/external_test_set_neg.csv.
- Test set-III: Relatively larger negatively biased test set can be found at https://github.com/Abdulk084/CardioTox/blob/master/data/external_test_set_new.csv

You first task is to build model in active learning fashion in Jupyter Notebook. You should:

- Prepare the data for modelling
- Choose data representation
- Build starting model
- Decide upon active learning strategy and apply it to improve your model
- Decide when to stop labelling new data in lab as labelling compounds via lab experiments is always additional cost

Your second task is to prepare model for production by serving the model as API and package it into Docker. Show us your MLOps skills like. Please include git history or provide link to the repo with the solution to the recruitment task:

- Organising the code
- Building the REST API
- Wrapping served model into Docker
- Testing
- Logging
- Load testing of deployed model inference

Please send us the link to the repo with the solution or zipped folder.

Please plan your time appropriately in order to be able to build end-to-end ML solution. The key thing is to have working prototype.

Good luck!