Task 1: Gene pathway images

We have created a website that shows relevant details about a potential therapeutic target that the user has searched for by gene symbol or name (called the Target Detail page). We don't currently have pathway images on the page, so users also do a Google image search to find easily interpretable expert-created schematics of signaling pathways the target is involved in (ex: Fig 1).
You are tasked with adding a Signaling Pathway Images section to the Target Detail page which provides images that help investment team members quickly understand the key biology of a target they should keep in mind when evaluating the biological hypothesis of a company we are considering investing in.

1. **Based on your expertise, what are the important things to show in a pathway for drug discovery related questions?**

   a. For quickly understanding the minimum biology needed to evaluate a company's hypothesis a Signaling Pathway image must contain, at the very least, a set of molecular species connected via arrows, showing the barebones structure of the pathway.

   b. Within this basic pathway structure, it is generally important to distinguish certain kinds of molecular identities, e.g. activators/repressors, enzyme-catalyzed reactions, activated/non-activated forms, etc. This is because the molecular identity of a target can determine its efficacy. For example, it is generally a bad idea to target enzymes in many pathways, since even small concentrations of enzymes are enough to catalyze chemical reactions at the same rate.

   c. Based on my experience, an important step in understanding the biology of a signaling cascade is teasing out the spatial context in which it exists. For example, Fig 1 shows a schematic representation of OPN-mediated MMP9 activation via the MAPK and the IKK/IkB/NF-κB pathways:

      i. OPN, present in the extracellular space, binds to the αvβ3 integrin molecule embedded in the cell membrane.

      ii. This binding induces a signaling cascade in the cytoplasm, ultimately forming an activated NF-κB/p65/p50 complex that translocates into the nucleus and induces expression of uPA and pro-MMP9.

      iii. uPA binds to pro-MMP9 and converts it into MMP9, which is a well-known ECM remodeler and is commonly exported out of the cell for ECM degradation.

      Information about the precise spatial localization of any druggable target is invaluable for decision making regarding therapeutic modality and efficacy. Therefore, it is important to show this information in signaling pathways for answering target selection/prioritization related questions.

   d. It is important to have as much information as possible about potential cross-pathway activation. A diversity of pathways involving the target may directly invalidate the biological hypothesis or invoke legitimate safety concerns.

   e. Modern pathway modeling is not just qualitative, but also quantitative. Although generally information about reaction kinetics is not presented in schematics of signaling pathways found online, it is advisable to extract such information from biochemistry/biophysics research papers and create custom pathway models to perform numerical simulations. Not only do such simulations provide insight into pathway dynamics but also act as cost effective computational methods for target selection and drug screening.

2. **What data sources, libraries, algorithms, or models would you consider using for either finding or creating pathway images that include the features you mention in Question 1?**

   a. **Data sources**: KEGG (https://www.genome.jp/kegg/), Reactome (https://reactome.org/), BioGRID (https://thebiogrid.org/), PathBank (https://www.pathbank.org/), MetaCyc (https://www.metacyc.org/), SABIO-RK (http://sabio.h-its.org/), NDEx (https://www.ndexbio.org/index.html#/), PantherDB (https://www.pantherdb.org/), SignaLink (http://signalink.org/), Biocarta (https://maayanlab.cloud/Harmonizome/dataset/Biocarta+Pathways), Biology/Biochemistry journal websites (e.g. https://www.sciencedirect.com/search?qs=MAPK%20pathway), PubMed (https://pubmed.ncbi.nlm.nih.gov/), ResearchGate (https://www.researchgate.net/), Google scholar (https://scholar.google.com/), Google images (https://images.google.com/)

   b. **Algorithms/Libraries**: Web scraping (BeautifulSoup/Scrapy, urllib/requests), optical character recognition (pytesseract), image processing (PIL), pathway building, modeling and visualization (PySB/Tellurium, NetworkX/Escher), cross-pathway analysis (PETAL)

   c. **Models**: CNNs, Vision transformers, Diffusion or other flow matching models for a more advanced implementation

3. **How would you evaluate the quality of the results of your implementation?**

We could evaluate the quality of our results based on at least the following two metrics:

   a. **Accuracy**: It is of paramount importance to ensure accuracy of the pathway because an incorrect pathway could lead to dangerous consequences for our investment strategy. Ideally, we would like to use/create a dataset consisting of "ground truth" for a few pathways that we can compare with the results from our signaling pathway image search algorithm for testing purposes.

b. **Usability**: Since the point of the Target Detail page is to provide quick and easy access to high quality pathway information to the investment team, it is essential for us to get direct feedback from the users on whether or not they find the results useful and reliable.

Additionally, we may also consider evaluating our results based on diversity of pathways observed and the amount of details that can be extracted from the results (e.g. spatial context, molecular identities, etc.).

4. Coding task: Use Google's API to find pathway images.
You have decided to use Google's API to retrieve potential target pathway images, and to write a ranking algorithm to identify the best image(s) to embed in the Target Detail page. For simplicity, just use the metadata returned by the API for the coding task. Then, write up what additional data sources, algorithms, or models you would use to improve upon the baseline performance of your implementation.

See the jupyter notebook, named Coding Task – Find and Rank Pathway Images, for an answer to this question.