About

Problem statement

- The formulation development stage in drug discovery is often hindered by extensive trial-and-error experimentation, leading to high costs and long timelines. Despite the availability of numerous approved drugs with known formulations, there is no systematic approach to leverage this existing knowledge for new or generic compounds.
- Developing a data-driven framework that identifies formulation analogs based on molecular similarity and proximity in the physico-chemical property space, could enable researchers to initiate formulation studies from known starting points.

Solution

- This portal attempts to address this challenge by enabling formulation scientists to search existing formulations using an API name, or SMILES representation of an API of interest to instantly retrieve formulations for similar (w.r.t. structure and/or physicochemical properties) APIs.
- The expectation is that these search results provide a scientifically informed starting point for developing formulations for novel molecules, reducing experimental redundancy and accelerating R&D workflows.

Data sources

- Formulation related data has been retrieved from the DailyMed, an US-FDA's repository. Based on this data, a master list of unique APIs was prepared, the same is discussed in section 3.1. Formulation details. The data was fetched on 21 April 2025.
    - Total unique APIs extracted from 152605 formulations after cleanup were = 2380
- Data related to experimental properties *viz*. melting point, pKa, etc. were collected/ extracted from multiple public sources
    - This data was available in the form of web-pages, databases, search-engine, publications, public datasets, scientific handbooks, etc.

Search

- ‘SMILES’ based search: Users can search using SMILES (canonical/isomeric) that can either be input as text or generated by drawing chemical structures. Search is then based on tanimoto similarity of Morgan ECFPs after canonicalizing the input SMILES
- ‘Name’ based search: Users can search with an API name or by brand name. This is mostly useful for known API and not for novel APIs
- ‘Physicochemical property’ search: User can get APIs of desired physicochemical property