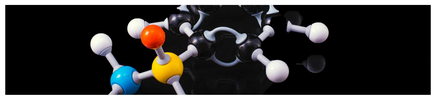
Bristol-Myers Squibb – Molecular Translation - Exploratory Data Analysis

Quick Exploratory Data Analysis for Bristol-Myers Squibb – Molecular Translation challenge

In this competition, you are provided with images of chemicals, with the objective of predicting the corresponding International Chemical Identifier (InChI) text string of the image. The images provided (both in the training data as well as the test data) may be rotated to different angles, be at various resolutions, and have different noise levels.



Overview

In [ ]:

train_labels.csv - ground truth InChi labels for the training images

In [ ]:

train/ - the training images, arranged in a 3-level folder structure by image_id

In [ ]:

In [ ]:

In [ ]:

test/ - the test images, arranged in the same folder structure as train/

Data Visualization

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

Competition Metric

Submissions are evaluated on the mean Levenshtein distance between the InChi strings you submit and the ground truth InChi values.

The Levenshtein distance between two strings **a**,**b** (of length **|a|** and **|b|** respectively) is given by **lev(a,b)** where

$$\mathrm{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0 \\ |b| & \text{if } |a| = 0 \\ \mathrm{lev}(\mathrm{tail}(a),\mathrm{tail}(b)) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \mathrm{lev}(\mathrm{tail}(a), b) \\ \mathrm{lev}(a, \mathrm{tail}(b)) \\ \mathrm{lev}(\mathrm{tail}(a), \mathrm{tail}(b)) \end{cases} & \text{otherwise.} \end{cases}$$

where the **tail** of some string **x** is a string of all but the first character of **x**, and **x[n]** is the *n*th character of the string **x**, starting with character 0.

Note that the first element in the minimum corresponds to deletion (from **a** to **b**), the second to insertion and the third to replacement.

The ima
from
Underst
the
Levensh
Distanc
Equatio
Review