Internship Project Report

Name: Janesh

Organization: Bversity

Period of Internship: 26.11.2024 to 30.11.2024

Mode: Online

Overview:

The internship focused on applying bioinformatics in drug discovery and development, emphasizing the role of artificial intelligence (AI) and machine learning (ML) in bioinformatics and cheminformatics. These technologies were explored to enhance the efficiency and speed of drug and compound production while maintaining safety and efficacy standards.

The primary tool used was KNIME, a free and open-source platform for data analytics, reporting, and integration. KNIME's modular "Building Blocks of Analytics" framework enables the creation of workflows for machine learning, data mining, and data analysis. The program also introduced the basics of the "molecule-to-market" process, which involves identifying a potential drug candidate and progressing through its development to commercialization.

Problem Statement:

A company, XYZ, wants to develop a new drug for a disease since the existing drug in the market is expensive and not affordable to all segments of the population. The company wants the new drug to be developed with similar efficacy to the existing one, but at a lesser price.

Disease Overview:

Neurofibromatosis Type 1 (NF1), or von Recklinghausen disease, is a genetic disorder marked by the growth of benign tumors (neurofibromas) along nerves, affecting about 1 in 3,000 individuals worldwide. It is caused by mutations in the NF1 gene, leading to a loss of neurofibromin, a protein that regulates cell growth and division. This results in abnormal cell proliferation and tumor formation.

Selumetinib, a MEK inhibitor, offers potential treatment for NF1-related complications, particularly in cases with inoperable plexiform neurofibromas. By inhibiting the MEK1/2 enzymes in the MAPK/ERK pathway, which is often overactive in NF1, selumetinib can reduce tumor size and alleviate symptoms, providing hope for affected patients.

Approach:

- To tackle this task, a disease was chosen for analysis.
- The disease chosen in this case was “neurofibromatosis type 1” and the existing drug in the market for it is “Selumetinib” (Koselugo).
- The molecular properties of the drug can be studied in public databases such as PubChem.
- The mechanism of action of the drug is studied to understand the target sites of the drug and its clinical pathway.
- Then based on its action mechanism, a drug library is selected for comparison.
- The drug library in this case is a “MEK Inhibitors” library.
- The drug library consists of several drugs to be compared with our drug of interest. The drug with the most similarity is then used to develop the new drug.
- The software then runs the program to search for similarities between the drug and the drug library. Then, the result is obtained in the form of visual chart(HEAT Map).

KNIME:

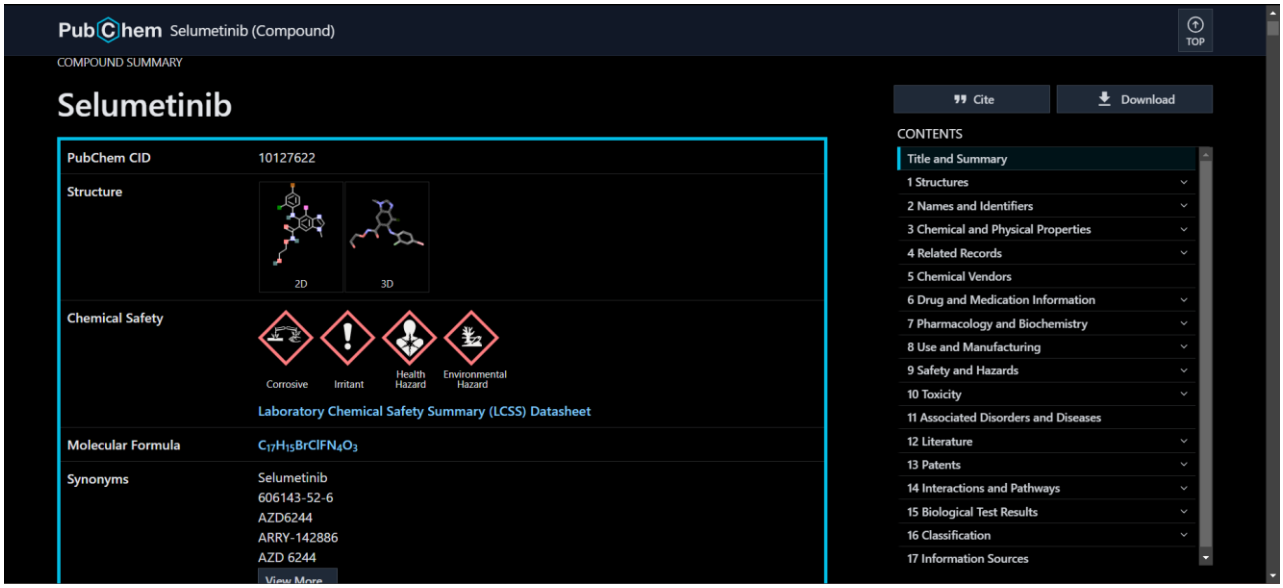
The entire process of the comparison of drug comparison and analysis is carried out using the KNIME software. The process is run through protocols called “workflows”. The workflows are made up numerous connections of nodes. The nodes are small blocks of codes that can be connected to perform certain functions. It is a functionality/small script that performs a certain function.

Principle:

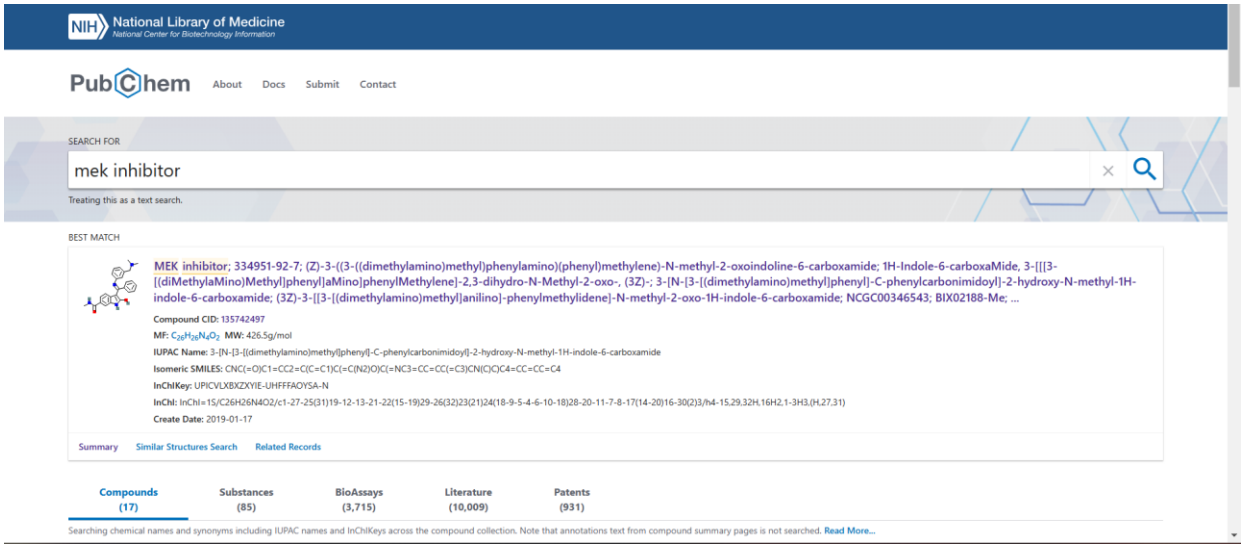
The main logic behind this process is inputting the drug of interest and drug library in two different slots and using machine learning process to find out which drug from the drug library matches the drug of interest.

Data Collection

Selumetinib Structure:
The 2D and 3D structures of selumetinib were obtained from PubChem in SDF format, providing comprehensive details on its molecular configuration and atomic connectivity.



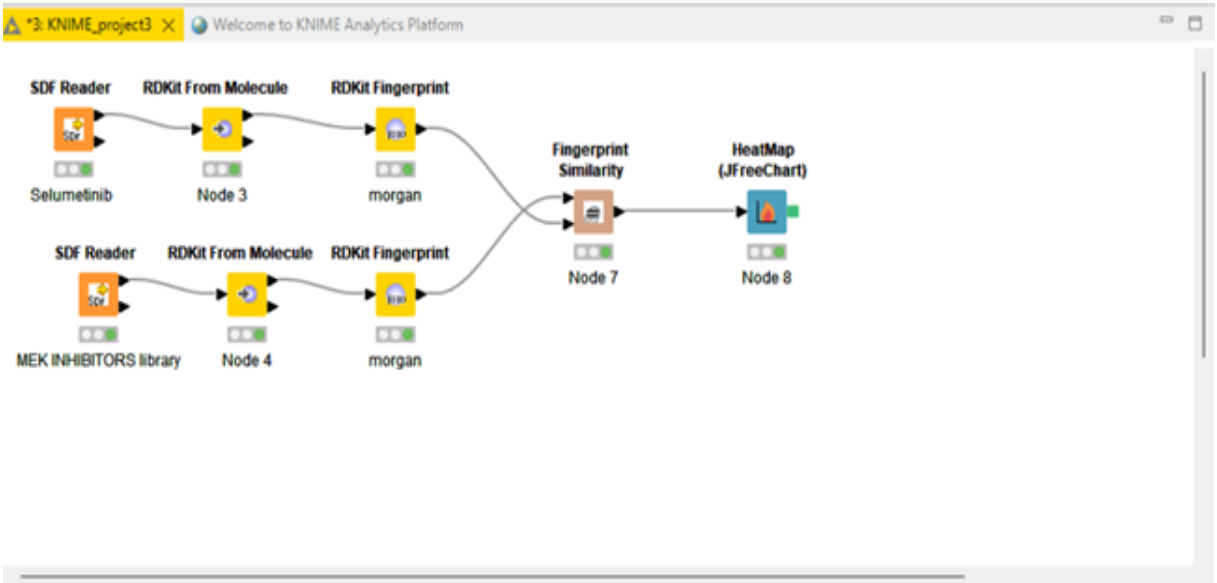
MEK Inhibitors:
A dataset of known MEK (Mitogen-Activated Protein Kinase/ERK Kinase) inhibitors was compiled from literature and PubChem, focusing on compounds that have demonstrated the ability to inhibit MEK activity, a critical component in various signaling pathways associated with cancer and other diseases.



Workflow: (Using 2D Conformation)

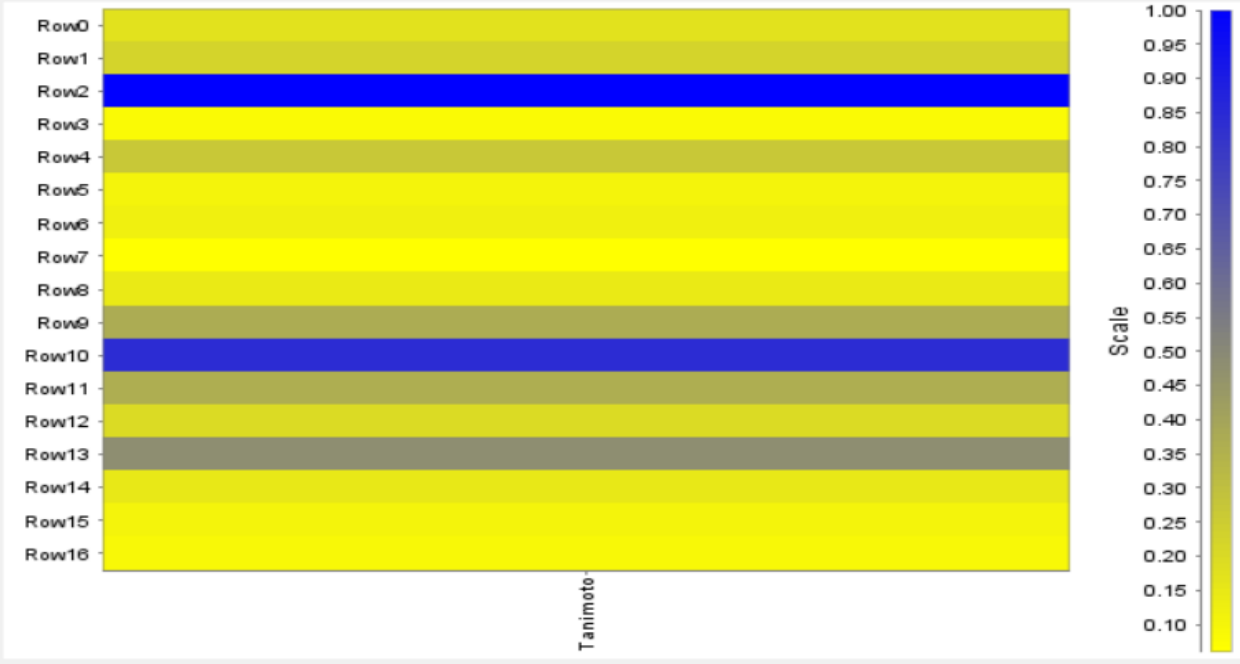
- **Download SDF Files**: The SDF files containing the 2D molecular structures of the target drug and the drug library are obtained from a public database.
- **SDF Reader Configuration**: These files are loaded into the workflow using the "SDF Reader" node and executed.
- **Molecular Fingerprint Conversion**: To enable comparison, the molecular structures are converted into fingerprints. Molecular fingerprints are binary representations that capture the unique features of a molecule, facilitating identification and comparison.
    - The nodes used for this process are "RDKit from Molecule" and "RDKit Fingerprint."
- **Fingerprint Configuration**: In the "RDKit Fingerprint" node, the fingerprints are configured using two types:
    - **MACCS (Array Format)**
    - **Morgan (Spatial Arrangement)**
- **Fingerprint Similarity**: The fingerprints are compared using the "Fingerprint Similarity" node, which calculates the similarity between the two inputs.
- **Result Visualization**: The results are visualized using a "Heatmap" node, which represents the similarity data using the Tanimoto coefficient. Shades on the heatmap indicate similarity levels, with values closer to "1" denoting higher similarity.

Workflow :



Result:

2D Conformer result:

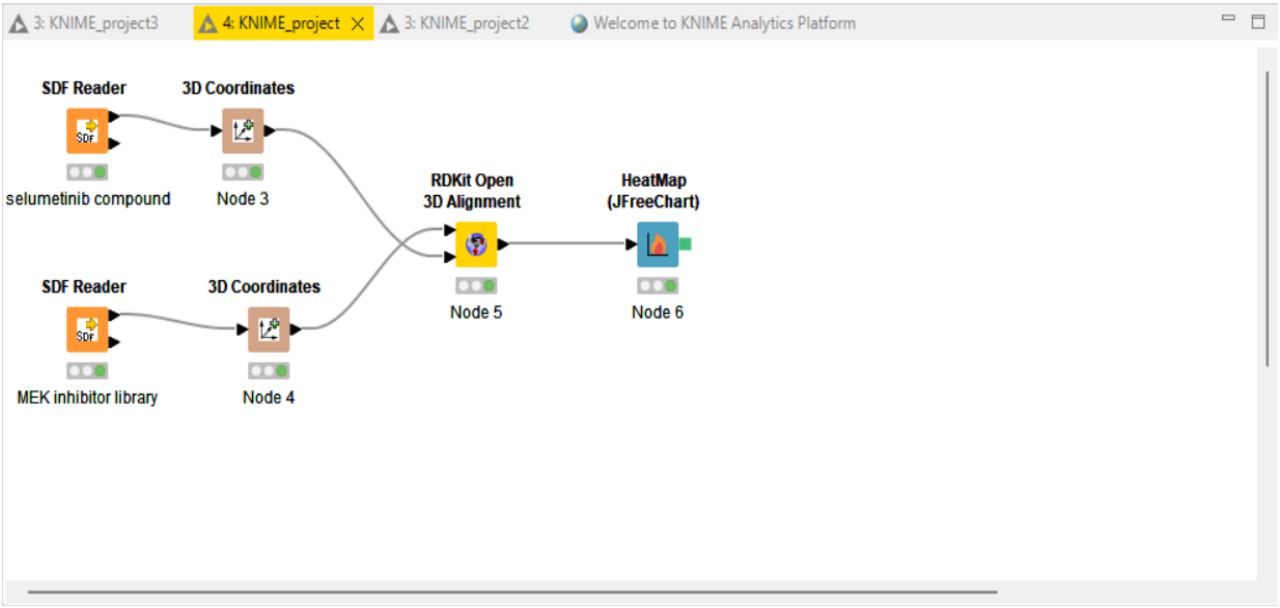


From this result we can interpret that the dark blue represents more similarity and yellow represents less similarity since the dark blue's Tanimoto value is closer to "1". From observing the above chart, we can conclude that the drug from the drug library which has the most similarity to our drug of interest is the drug on "row 2".

Workflow: (Using 3D Conformation)

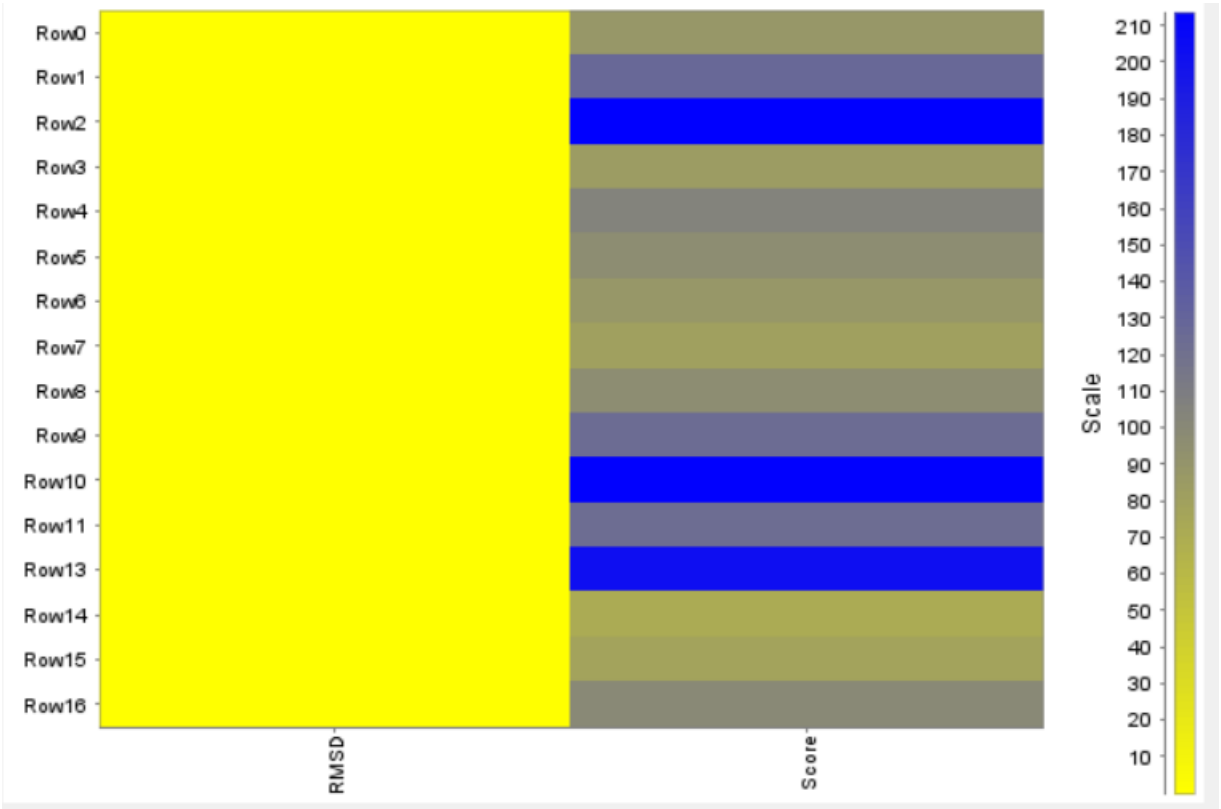
1. **Download SDF Files**: SDF files containing the 3D molecular structures of the target drug and the drug library are retrieved from a public database.
2. **SDF Reader Configuration**: The files are imported into the workflow using the "SDF Reader" node and executed.
3. **3D Structure Formatting**: The molecules' 3D structures are formatted using the "3D Coordinates" node, enabling accurate mapping and simplifying comparison.
4. **3D Alignment**: The formatted 3D coordinate nodes are connected to the "RDKit Open 3D Alignment" node, which performs a structural comparison between the drug of interest and the drug library.
5. **Result Visualization**: The alignment results are visualized using the "Heatmap" node, where the data is represented using RMSD (Root Mean Square Deviation) values. Lower RMSD values indicate higher structural similarity, while higher values indicate greater differences.

Workflow:



Result:

3D Conformer result:



Similarly, from observing and analyzing the RMSD values from the chart, we can conclude that the drug on “row 2” has the most similarity with our drug of interest. The lesser the RMSD value, the higher the similarity and vice versa for score value higher the value ,the higher the similarity.

Significance

- **The Importance of Both 2D and 3D Analysis**:
  For MEK inhibitors, 2D analysis provides a quick and computationally efficient approach for initial screening by focusing on the chemical features of the compounds. In contrast, 3D analysis offers a deeper understanding of molecular alignment and spatial interactions, which is critical for evaluating how these inhibitors bind to their target. Integrating both methods improves prediction accuracy and facilitates the identification of candidates with higher potential for effective MEK inhibition.
- **Potential for Drug Discovery**:
  The MEK inhibitors identified as structurally similar to selumetinib in this study could be further explored as alternative therapeutic options. These compounds may offer comparable or enhanced efficacy while potentially reducing side effects, contributing to the development of safer and more effective treatments for conditions involving MEK signaling pathways.

Conclusion:

By executing the workflows and analyses, we identified the drug best suited to replace the existing market drug. This model serves as a valuable tool for discovering potential new drugs and designing compounds with comparable efficacy and mechanisms of action to those currently available. The integration of bioinformatics, machine learning, and artificial intelligence plays a crucial role in this process, ensuring enhanced efficiency, precision, and innovation in drug development.