Improving Molecular De Novo Drug Design with Transformers - Prototype

Dhaval Soni, ID 1189821, MSc in Computer Science, Lakehead University, Thunder Bay
Gaminee Ram, ID: 1190137, MSc in Computer Science, Lakehead University, Thunder Bay
Hashmath Shaikh, ID: 1190451, MSc in Computer Science, Lakehead University, Thunder Bay
Kavan Raval, ID: 1190125, MSc in Computer Science, Lakehead University, Thunder Bay
Manish Jadhav, ID: 1194246, MSc in Computer Science, Lakehead University, Thunder Bay
Payal Devalia, ID: 1191996, MSc in Computer Science, Lakehead University, Thunder Bay
Venkata Naga Sai Kinnera, ID: 1193728, MSc in Computer Science, Lakehead University, Thunder Bay

Overview

The fields of protein synthesis and structural biology have been revolutionized by the emergence of transformer-based machine learning models, which have offered new opportunities to understand and predict protein structures and functions. Essential roles in biological processes are played by proteins, making it crucial for their structural and functional characteristics to be comprehended. This comprehension not only advances scientific knowledge but also contributes to practical applications, including disease elucidation and therapeutic development. Initially designed for natural language processing, exceptional adaptability in handling sequential data has been demonstrated by transformers, making them indispensable tools for a wide range of protein-related tasks. The generation of 3D protein structures from amino acid sequences is enabled by transformer architectures, facilitating de novo structure prediction. Moreover, contributions to predicting protein-protein interactions are made by transformers through the integration of data from diverse sources to improve predictive accuracy. Overall, significant advancements in research in protein synthesis and structural biology have been brought about by transformer-based models, offering promising avenues for further exploration and discovery.

I. Molecular Structure Generation and Analysis

We use the system to generate novel molecular structures tailored to specific properties we're interested in. By inputting our desired criteria, the system employs sophisticated algorithms and machine learning models to produce a range of candidate structures. Our outputs include detailed representations of these structures in SMILES format, along with predictions regarding their properties, such as binding affinity.

A. Molecular Property Prediction and Similarity Search

In our research endeavors, we predict the binding affinities of ligands with target proteins and conduct similarity searches to find molecules akin to our reference compounds. Leveraging machine learning models, we can accurately predict binding affinities, while the similarity search functionality helps us identify structurally similar compounds. We receive outputs comprising predicted binding affinities for ligands and ranked lists of compounds sharing similarities with our reference molecules.

B. Molecular Visualization and Virtual Screening

As students and researchers, we often rely on the system to visualize intricate molecular structures and screen chemical libraries for potential lead compounds. Using visualization tools integrated into the system, we can generate interactive 3D representations of molecules for educational purposes. Additionally, the virtual screening feature enables us to evaluate compound binding affinities with target proteins, aiding in drug discovery efforts. Our outputs include engaging 3D visualizations and prioritized lists of compounds based on their predicted binding affinities.

C. Data Processing and Model Training

To further our research goals, we engage in data processing tasks such as grouping, merging, filtering, and adjusting molecular data using the system's functionalities. Moreover, we utilize the system to train and refine machine learning models for improved performance in predicting molecular properties. This involves tokenizing SMILES strings and preparing training datasets tailored to our specific research objectives. Our outputs encompass informative visualizations, refined datasets, and finely-tuned models ready for predictive analysis.

D. Exporting and Tracking Generated Molecules

After generating SMILES strings representing molecular structures, we ensure their validity, track their details, and export them for further analysis. The system seamlessly appends validated molecules to a master tracking table, assigning unique identifiers for efficient organization. Once exported to SDF files, we receive confirmation messages indicating the successful completion of the export process.

II. Utilization of the Molecular Generation and Analysis System in Research Tasks

In our research efforts, the Molecular Generation and Analysis System has been utilized to address various tasks crucial for advancing our understanding of molecular structures and properties. Here's an overview of how the system has been employed across different scenarios:

A. Molecular Data Analysis:

In this scenario, the distribution of binding affinities within a dataset of molecular structures was analyzed by our research team. By importing the dataset into the system and leveraging code execution capabilities, the distribution was visualized using histograms and scatter plots. Valuable insights into the binding affinity distribution and its correlation with other molecular properties were provided by the generated visualizations, aiding in the identification of patterns and trends crucial for our research.

B. Binding Affinity Prediction:

The prediction of the binding affinity of novel ligands with a target protein was conducted by our pharmaceutical company. Leveraging machine learning models trained on known ligand-protein interactions, the system was employed to predict the binding affinities of new ligands. The system's output of predicted binding affinities enabled us to prioritize ligands with the highest predicted affinities for further experimental validation, streamlining our drug discovery process.

TABLE I: Example Table for Modified Scores

ID	**Generation**	**SCORE**
AAZD	2	-5.2
AAZE	2	-5.5
AAZF	2	-5.4
AAZG	2	-6.8
AAZH	2	-5.6

C. Structural Similarity Analysis:

As medicinal chemists, the identification of structurally similar compounds to a lead molecule with known biological activity was aimed for. By inputting the lead molecule's SMILES string or structure into the system and executing code for similarity search, a ranked list of molecules ordered by their structural similarity to the lead compound was received. This assisted in identifying potential analogs for further investigation, enhancing our understanding of structure-activity relationships.

D. Data Preprocessing and Visualization:

In our exploratory analysis, the preprocessing and visualization of a dataset of molecular structures were necessary. Leveraging Python libraries such as pandas, RDKit, and matplotlib within the system, data preprocessing, calculation of molecular descriptors, and generation of visualizations were conducted. The system's output of descriptive statistics, molecular visualizations, and distribution plots provided us with valuable insights into the dataset's characteristics, guiding subsequent analysis decisions.

E. Chemical Library Screening:

Our drug discovery team aimed to screen a chemical library for potential drug candidates targeting a specific protein. By importing the chemical library into the system and executing code for evaluating binding affinities, rankings of compounds based on their predicted binding affinities were received. This facilitated the identification of lead compounds for further experimental validation in our drug discovery campaigns, accelerating our research efforts.

III. Application of the Molecular Generation and Analysis System in Research Tasks

In our research endeavors, we've found the Molecular Generation and Analysis System to be indispensable in tackling various tasks crucial for advancing our understanding of molecular structures and properties. Here's how we've utilized the system across different scenarios:

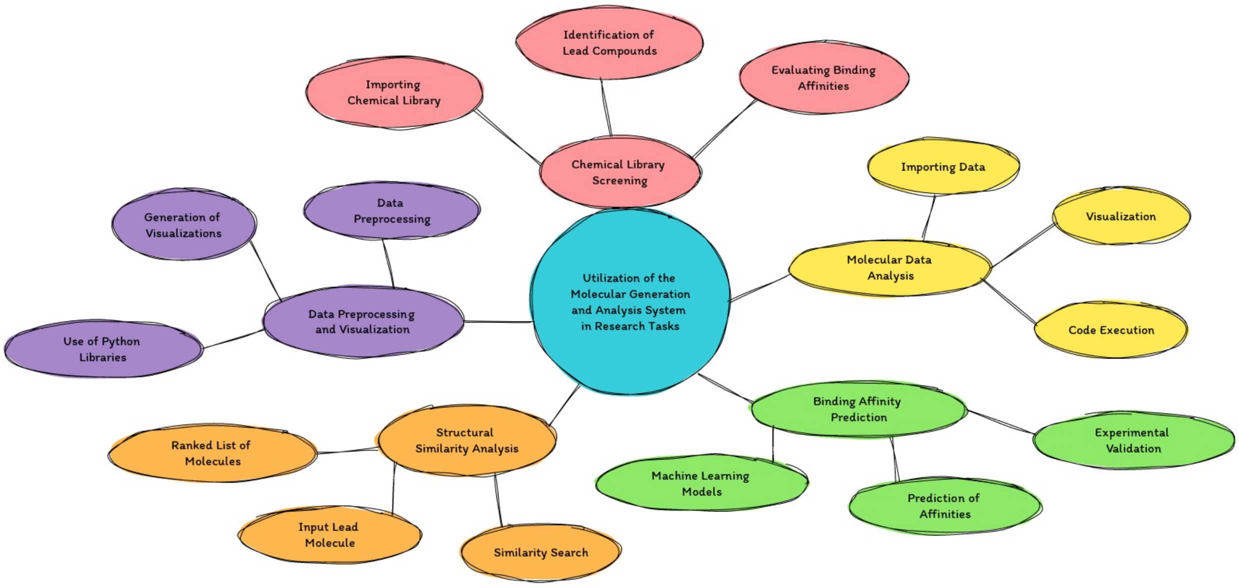
Identification of Lead Compounds
Importing Chemical Library
Evaluating Binding Affinities
Chemical Library Screening
Importing Data
Data Preprocessing
Generation of Visualizations
Visualization
Molecular Data Analysis
Utilization of the Molecular Generation and Analysis System in Research Tasks
Data Preprocessing and Visualization
Use of Python Libraries
Code Execution
Binding Affinity Prediction
Structural Similarity Analysis
Experimental Validation
Ranked List of Molecules
Machine Learning Models
Prediction of Affinities
Input Lead Molecule
Similarity Search

Fig. 1: Utilization of Molecular Generation

A. Data Grouping and Calculation

In this scenario, we needed to group molecular data by ligand and calculate the minimum binding affinity for each group. The system allowed us to easily group the data by ligand, perform the necessary calculations, and extract generation and ID information from the ligand column. This resulted in a new dataset containing ligand-specific minimum binding affinities, empowering us for downstream analysis and visualization.

B. Data Visualization

Here, our aim was to visualize the distribution of binding affinity scores across different generations. Leveraging the system's capabilities, we generated a swarm plot illustrating binding affinity scores grouped by generation. This visualization provided valuable insights into how binding affinity scores varied across generations, helping us identify trends and patterns in the data.

C. Data Merging and Updating

This scenario involved merging and updating two datasets based on shared identifiers. By leveraging the system, we merged a master dataset with a new dataset based on ID and generation columns, updating the binding affinity scores seamlessly. The system facilitated the generation of a consolidated dataset with updated binding affinity scores, ensuring the inclusion of the latest data for downstream analysis and reporting.

D. Molecular Weight Calculation

In this case, a chemist needed to calculate the molecular weight of compounds based on their SMILES representation. Leveraging the system's features, we computed the molecular weights of compounds using RDKit's MolWt function applied to SMILES strings. The system efficiently added the computed molecular weights as a new column in the dataset, providing crucial information for molecular analysis and property prediction.

E. Top SMILES Extraction for Training

A machine learning researcher required the extraction of top SMILES strings for training a predictive model. Leveraging the system, we extracted a set of unique SMILES strings with the lowest binding affinity scores for model training purposes. This ensured that the model was trained on diverse molecular structures with low binding affinities, ultimately enhancing its predictive performance.

IV. Expanding the Utility of the Molecular Generation and Analysis System

In our ongoing exploration of the Molecular Generation and Analysis System, we've identified several key areas where its capabilities can be extended to further support our research endeavors. Here's a glimpse into the new scenarios we've envisioned:

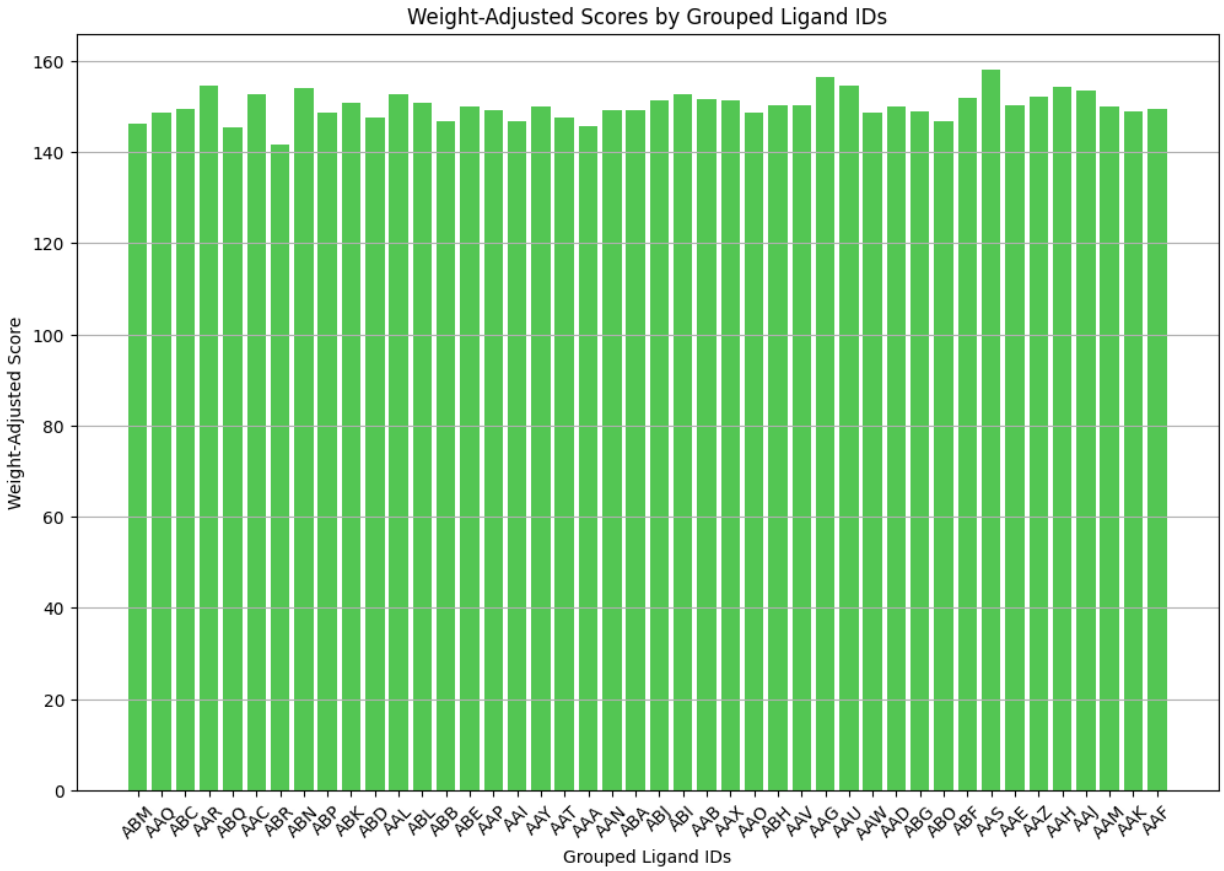

Fig. 2: Weight-Adjusted Scores by Grouped Ligand IDs

A. Data Filtering and Adjustment

Chemists often require precise control over their datasets, necessitating the ability to filter out compounds based on specific criteria such as molecular weight thresholds. Additionally, the adjustment of scores based on factors like molecular similarity and weight can provide deeper insights into molecular properties. With this in mind, we aim to leverage the system to create modified datasets tailored to our research needs, shedding light on the intricate relationship between molecular attributes and corresponding scores.

B. Data Visualization for Adjusted Scores

Visual representations play a crucial role in elucidating complex relationships within datasets. Analysts seeking to understand the interplay between molecular weight and adjusted scores can benefit greatly from intuitive visualizations. By harnessing the system's visualization capabilities, we intend to generate scatter plots that vividly illustrate how adjusted scores fluctuate in response to changes in molecular weight. These visual aids will serve as invaluable tools for interpretation and decision-making in our research pursuits.

C. Model Training Preparation

As we delve deeper into the realm of predictive modeling, the preparation of training data assumes paramount importance. Tokenizing SMILES strings and extracting top-performing candidates based on adjusted scores are critical steps in this process. Leveraging the system's functionality, we plan to streamline these tasks, ensuring that our data is primed for use in deep learning algorithms. By doing so, we aim to accelerate the development of predictive models tailored to the intricacies of chemical informatics applications.

V. Model Training and Generation

In our pursuit of exploring chemical space and facilitating molecular design, we recognize the importance of incorporating deep learning-based sequence generation into our toolkit. With this in mind, we've introduced a new scenario where researchers can now utilize the system to train and deploy the LSTMChem model for generating SMILES sequences. By initializing and building the LSTMChem model, training it using provided configurations, and saving the trained model weights, researchers can harness the power of deep learning for generating novel molecular structures. Additionally, the system allows for the generation of SMILES sequences based on given starting tokens, enabling seamless exploration of chemical space and molecular design.

VI. Model Refinement and Training

In our continuous quest to enhance the performance of our molecular generation models, a new scenario has been introduced, focusing on the iterative refinement and training of the LSTMChem model. Here's how it operates:

After a set of SMILES strings is generated using the LSTMChem model, the model is further refined by incorporating these generated SMILES into the training data, and then retraining the model. The generated SMILES strings are processed by researchers, ensuring their validity as representations of valid molecules, and a subset is selected for inclusion in the training data. The system then updates the training data with the new SMILES and proceeds to retrain the LSTMChem model using the refined dataset. A refined set of training data is produced by incorporating generated SMILES strings, updating the LSTMChem model, and saving the model's architecture and weights. Additionally, visualizations such as word clouds of SMILES and scatter plots of scores versus generation are generated for analysis and insights into the training process.

This scenario significantly enhances the system's capabilities by allowing users to iteratively refine and train the LSTMChem model using both original and generated data, thereby improving the model's performance and molecular generation capabilities.

VII. Enhancing LSTMChem Model Performance through Iterative Refinement and Fine-tuning

A. *Model Refinement and Training*

In this scenario, the LSTMChem model undergoes further refinement by integrating newly generated SMILES strings into the training data, followed by model retraining. The system processes the generated SMILES, validates them for molecular validity, and selects a subset for training. Subsequently, the training data is updated, and the LSTMChem model is retrained using the refined dataset. This iterative refinement process enhances the model's capabilities by incorporating both original and generated data, thereby improving its performance in molecular generation tasks.

B. *Fine-tuning LSTMChem Model*

After refining the training data with newly generated SMILES, the LSTMChem model undergoes fine-tuning to further enhance its performance. Leveraging the LSTMChemFinetuner class, the modeler object and a data loader are provided for fine-tuning. The finetuner adjusts model parameters using the specified optimizer and loss function, returning the training history. Finally, the fine-tuned model weights are saved for future use. This iterative fine-tuning process refines the LSTMChem model using both original and generated data, leading to continuous improvement in its performance and the quality of generated molecules.

VIII. Workflow Overview for Molecular Structure Generation and Analysis

In this concluding phase of our workflow, which focuses on the generation, validation, and analysis of molecular structures using the LSTMChem model, we meticulously execute several crucial steps to achieve our objectives. Each step is vital for the workflow's overall success, contributing to the generation of new molecular insights.

A. *Generation of SMILES Strings:*

Initially, we employ the LSTMChemGenerator to produce a specific number of SMILES strings. These strings represent potential molecular structures generated by our model, based on learned patterns from the training dataset. This step is crucial for identifying new molecules that could have significant applications in various scientific fields.

B. *Validation of Generated SMILES:*

Following the generation process, the SMILES strings undergo a validation phase. In this phase, we convert the strings into RDKit Mol objects to ensure they represent valid molecular structures. This step is essential for filtering out any erroneous or impractical molecular representations, ensuring that only feasible molecules are carried forward in the analysis process.

C. *Evaluation Metrics:*

We then assess the generated structures through three key metrics: validity, uniqueness, and originality. Validity measures the proportion of SMILES strings that successfully convert to valid molecular structures, indicating the efficiency of our generation process. Uniqueness is calculated as the percentage of unique structures among the validated molecules, essential for identifying diverse molecular entities. Originality determines the fraction of unique and validated molecules not present in the training dataset, highlighting the model's ability to generate novel structures beyond its training scope.

D. Tracking and Analysis:

To facilitate further analysis and tracking, new molecules are appended to a master tracking table. Each molecule is assigned a unique identifier, and we catalog details such as the molecule's source and specific properties. This structured tracking approach aids in the systematic study and comparison of molecular structures over time.

E. Exportation to SDF Files:

The validated and cataloged molecules are then prepared for exportation to SDF files. This step involves batching the export process if the number of molecules surpasses a predetermined threshold, ensuring that data management remains efficient. The SDF files serve as a versatile format for storing molecular structures, facilitating their use in various chemical informatics applications and further analyses.

F. Completion Confirmation:

Upon the successful completion of the export process, we print a simple 'ok' message. This confirmation signifies that the molecules are now ready for subsequent analysis, marking the end of this workflow phase. The streamlined process from generation to exportation ensures that we have a coherent and efficient methodology for producing, validating, and analyzing molecular structures.

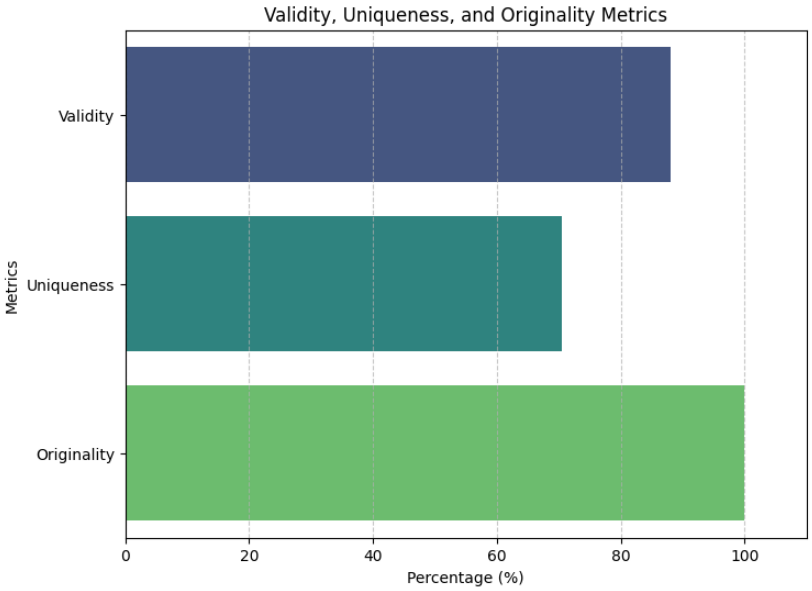

Fig. 3: Evaluation Matrix

This methodology underscores our commitment to advancing scientific research through the generation, validation, and detailed analysis of molecular structures. By following this workflow, we contribute valuable insights and resources for further exploration and discovery in the field.

IX. Front-End Development

A. Key Functionalities

- **SMILES to CSV Conversion:** A function (`smi_to_csv`) is implemented to convert SMILES strings into a structured CSV format. Each SMILES string is stored in a separate row to ensure organized data handling, with leading/trailing whitespace and newline characters removed for data integrity.
- **Molecule Visualization from SMILES:** The RDKit library is utilized to generate 2D depictions of molecules from SMILES strings. Users are enabled to input SMILES representing chemical compounds for visualization, enhancing the interpretability of chemical data by providing visual representations of molecular structures.
- **Integration into Front-End Applications:** The functionalities are designed for integration into front-end applications, such as web interfaces such as Django. Users can upload files containing SMILES data for conversion to CSV, streamlining data preprocessing. Interactive features allow users to input SMILES strings directly and visualize corresponding molecular structures in real-time.

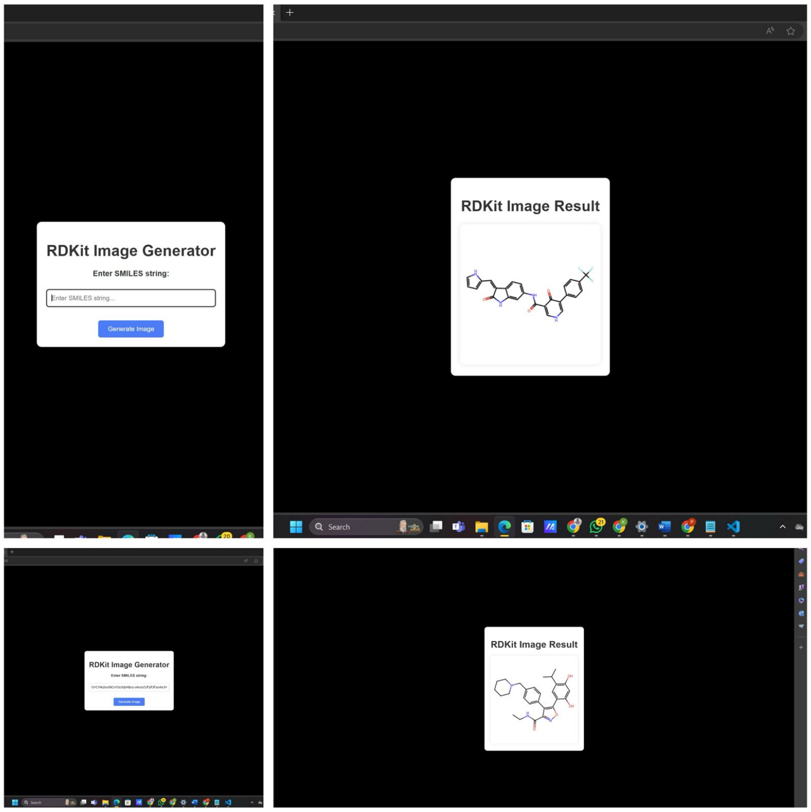

Fig. 4: Website Generated Visualizations using SMILE Strings

B. *Benefit and Impact*

Researchers, educators, and practitioners in the field of cheminformatics are empowered with essential tools for data processing and visualization. Enhanced accessibility and usability are achieved through seamless integration into user-friendly front-end applications, facilitating streamlined analysis and decision-making in various domains, including drug discovery, materials science, and environmental chemistry.

References

[1] Monteiro, Nelson RC, Tiago O. Pereira, Ana Catarina D. Machado, José L. Oliveira, Maryam Abbasi, and Joel P. Arrais. ”FSM-DDTR: End-to-end feedback strategy for multi-objective De Novo drug design using transformers.” Computers in Biology and Medicine 164 (2023): 107285.

[2] Mao, Jiashun, Jianmin Wang, Amir Zeb, Kwang-Hwi Cho, Haiyan Jin, Jongwan Kim, Onju Lee, Yunyun Wang, and Kyoung Tai No. ”Transformer-based molecular generative model for antiviral drug design.” Journal of Chemical Information and Modeling (2023).

[3] Feng, Tao, Pengcheng Xu, Tianfan Fu, Siddhartha Laghuvarapu, and Jimeng Sun. ”Molecular De Novo Design through Transformer-based Reinforcement Learning.” arXiv preprint arXiv:2310.05365 (2023).

[4] Zheng, Yangkun, Fengqing Lu, Jiajun Zou, Haoyu Hua, Xiaoli Lu, and Xiaoping Min. ”De Novo Design of Target-Specific Ligands Using BERT-Pretrained Transformer.” In Chinese Conference on Pattern Recognition and Computer Vision (PRCV), pp. 311-322. Singapore: Springer Nature Singapore, 2023.

[5] Wang, Mingyang, Zhe Wang, Huiyong Sun, Jike Wang, Chao Shen, Gaoqi Weng, Xin Chai, Honglin Li, Dongsheng Cao, and Tingjun Hou. ”Deep learning approaches for de novo drug design: An overview.” Current Opinion in Structural Biology 72 (2022): 135-144.

[6] Wang, Dan, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z. Jane Wang, and Rabab Ward. ”Multi-view 3d reconstruction with transformers.” In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5722-5731. 2021.