PPE Kit Detection

Submitted in Partial fulfilment of the

requirements of the degree of

Bachelor of Engineering

in

Artificial Intelligence and Machine Learning

by

Aditya Randive *121A9002*

Pranav Pillai *121A9043*

Suyash Utekar *121A9062*

UNDER THE GUIDANCE OF

Dr. Nita Patil

In

Artificial Intelligence and Machine Learning

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING ENGINEERING

SIES GRADUATE SCHOOL OF TECHNOLOGY

NERUL, NAVI MUMBAI – 400706

ACADEMIC YEAR

2024 – 2025

Certificate

This is to certify that the project entitled "**PPE Kit Detection**" is a bonafide work carried out by the following students of Third year in Artificial Intelligence and Machine Learning.

Sr.no	Name	Roll No.
1	Aditya Randive	121A9002
2	Pranav Pillai	121A9043
3	Suyash Utekar	121A9062

The report is submitted in partial fulfillment of the degree course of Bachelor of Engineering in Artificial Intelligence and Machine Learning, of University of Mumbai during the academic year 2024 – 2025.

Dr. Varsha Patil	**Dr. Varsha Patil**	**Dr. K Laskshmisudha**
Internal Guide	**Head of Department**	**Principal**

We have examined this report as per university requirements at SIES Graduate School of Technology, Nerul, Navi Mumbai on _____

Project Report Approval

This project report entitled **TransBioRetro: A Transformer-based Self Correcting Beam Search for Bio Retrosynthesis Prediction** by the following student is approved for the degree of **Bachelor of Engineering** in **Artificial Intelligence and Machine Learning**.

Aditya Randive	121A9002
Pranav Pillai	121A9043
Suyash Utekar	121A9062

Name of External Examiner: _____

Signature: _____

Name of Internal Examiner: _____

Signature: _____

Date:

Place:

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute anc can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Name	Roll No.	Signature
Aditya Randive	121A9002	
Pranav Pillai	121A9043	
Suyash Utekar	121A9062	

Date:

Acknowledgement

We would like to express our thanks to the people who have helped us the most throughout our project. We are grateful to our guide **Dr. Varsha Patil** and coordinator **Prof. Nita Patil** for nonstop support for the project.

A special thanks goes to each other who worked together as a team in completing the project, where we all exchanged our own interesting ideas, thoughts and made it possible to complete our project with all accurate information. We also wish to thank our parents for their personal support and attention who inspired me to go my own way.

We would also like to extend our sincere gratitude to our Head of the Department **Dr. Varsha Patil** and our Principal **Dr. K Lakshmisudha** for their continuous support and encouragement.

We also would like to thank our other faculty members for providing us with all the required resources and references for the project.

Project Team

Aditya Randive

Pranav Pillai

Suyash Utekar

Abstract

As an alternative to traditional organic synthesis, bioretrosynthesis uses enzyme conversion to provide an effective and sustainable synthetic method for target molecules. This strategy is promising in the fields of organic fuels, fine chemicals and pharmaceuticals. The broad implementation is hampered by issues such as scalability, product allocation with unclear assumptions, and insufficient reaction information. To improve the accuracy of retro synthetic predictions, this work introduces a profound architecture of hybrids that combine contrasting learning with autocorrelation jet search methods. Our approach optimizes both enzymatic and organic reaction pathways by integrating carefully selected data records and modern trans models of synthetic reactions in organic synthesis. It provides a comprehensive assessment approach that shows significant benefits compared to basic line techniques, and includes the top N accompanying chemical validity tests and ablation search. By bridging the gap between real-world computer prediction and biochemical applications, our work promotes AI-controlled retrosynthesis.

Keywords: *Bio retro synthesis, Sustainable Synthesis, Deep Learning, Transformer Models, Contrastive Learning, SelfCorrecting Beam Search, Retro synthetic Prediction, Bio Synthetic Datasets, Hybrid Framework, Deep Learning, Metabolic Engineering, Synthetic Biology, Sustainable Chemistry.*

Contents

List of Figures

Abbrevations

AI Artificial Intelligence

ML Machine Learning

RL Reinforcement Learning

MCTS Monte Carlo Tree Search

USPTO United States Patent and Trademark Office

GDP Glycerol-3-phosphate dehydrogenase

NLP Natural Language Processing

SBA Synthetic Biology Applications

CADD Computer-Aided Drug Design

SMILES Simplified Molecular Input Line Entry System

QSAR Quantitative Structure-Activity Relationship

MPO Metabolic Pathway Optimization

Chapter 1

Introduction

1.1 Introduction

One of the main challenges of synthetic biology and organic chemistry is the ability to create sustainable and effective synthetic methods for chemical production. A practical alternative to traditional organic synthesis is biotransmitting, which uses enzymatic processes to create biosynthetic pathways. In contrast to traditional retrositetie methods, which primarily rely on organic conversion, the use of biocatalysts has advantages such as selectivity, high selectivity, softer reaction conditions, and reduced environmental effects. These properties make bioretrosynthesis for use in biofuels, fine chemicals, and drug therapy very relevant. Despite this possibility, bioretrosynthesis has not spread due to severe arithmetic and experimental impairments.

One of the biggest challenges of bioretrosynthesis is the lack of complete knowledge about biochemical responses. Enzyme conversion is less characterized than natural chemical processes, and therefore the biosynthetic pathway is lacking. Furthermore, the major issue is ambiguity of the response. Many biological approaches have specific preparations for pioneering products due to context-dependent reactivity and enzyme tumor disease. This complexity makes it difficult to increase the number of prediction algorithms that can correctly determine the greatest possible sequence of reactions. Scalability remains a major situation. Computer models regularly advise artificial routes. This is conceptually impractical, but unrealistic in the real world context, but unrealistic due to its limitations along with the availability of enzymes, kinetics and financial feasibility.

The accuracy of retrosynthetic prediction has considerably stepped forward in latest deep gaining knowledge of advances, mainly transformer-based totally collection-to-collection fashions, to get

over those obstacles. those fashions have carried out properly in natural chemistry by figuring out complex relationships from big response datasets. however, the incapacity of current fashions to integrate natural and enzymatic reaction pathways leads to inaccurate biosynthetic course predictions. similarly proscribing their use in bioretrosynthesis is the truth that most approaches depend on conventional beam search techniques, which are not excellent at managing reaction ambiguities or incorrect precursor-product assignments.

With the intention to enhance retrosynthetic prediction, we gift a hybrid deep studying gadget on this study that combines contrastive cutting-edge with a self-correcting beam seek method. Our approach cutting-edge cautiously chosen biosynthetic reaction datasets and state-of-the-art transformer fashions to optimize enzymatic and natural synthesis pathways. The contrastive getting to know modern aspect reduces ambiguity and improves prediction robustness by using supporting the model distinguish between valid and incorrect precursor-product pairs. moreover, the self correcting beam search process iteratively improves predictions by way of making sure that erroneous intermediate steps are dynamically corrected. via this aggregate, our approach complements the accuracy and value latest bioretrosynthetic forecasts.

Via addressing great issues such response ambiguity, a loss of biochemical statistics, and scalability problems, this work complements AI-pushed bioretrosynthesis.By using combining contrastive gaining knowledge of and self-correcting beam seek, we recommend a hybrid strategy to beautify the accuracy and utility of retrosynthetic forecasts,which is a step in the direction of greater reliable and comprehensible biosynthetic pathway design. The statistics gained from this look at targets to make contributions to the development of greater efficient andsustainable chemical production methods via bridging the distance among computational retrosynthesis and real-world biochemical programs.

1.2 Need of Project

The need for this project arises from the growing demand for sustainable and efficient synthetic methods in chemical production, particularly in the fields of biofuels, pharmaceuticals, and fine chemicals. Traditional organic synthesis often faces limitations, including low selectivity, harsher reaction conditions, and significant environmental impacts. Bioretrosynthesis, which utilizes enzymatic processes, offers a greener alternative; however, its widespread implementation is hindered by challenges such as incomplete biochemical knowledge, scalability issues, and

ambiguity in reaction pathways.

Consequently, there is a pressing need for advanced predictive models that can accurately determine viable biosynthetic routes. This project introduces a hybrid deep learning framework that combines contrastive learning and self-correcting beam search techniques to enhance the accuracy and practicality of retrosynthetic predictions, aiming to bridge the gap between computational models and real-world biochemical applications.

This advancement will significantly contribute to the development of more efficient and sustainable chemical production methodologies.

1.3 Scope

The scope of this project encompasses the development and implementation of a hybrid deep learning framework designed to advance bioretrosynthesis predictions. Specifically, the project focuses on creating a transformer-based model that integrates contrastive learning and a self-correcting beam search method to improve the accuracy and reliability of retrosynthetic pathway predictions.

It aims to address key challenges such as response ambiguity, the lack of comprehensive biochemical data, and scalability. By utilizing a rich dataset combining biosynthetic and organic reaction data, the project intends to enhance the model's ability to propose feasible synthetic pathways that consider both chemical and enzymatic transformations.

Additionally, it seeks to produce more interpretable and robust predictions that can facilitate real-world applications in various fields, including drug development and environmentally friendly chemical synthesis. Ultimately, the project aims to contribute to the overarching goals of creating sustainable and efficient methods for chemical production through innovative computational strategies.

1.4 Organization of the report

The report is organized into six well-defined chapters, each essential for achieving the project's goals and contributing to scholarly knowledge. Chapter 1 lays the groundwork by introducing the project's core concepts, outlining the problem it addresses, and detailing its objectives. This chapter serves as a pivotal element, giving readers a comprehensive overview of the project's

scope and direction.

Chapter 2 presents a comprehensive Literature Survey, exploring existing research relevant to the project. By situating the current work within the broader academic context, this chapter highlights gaps and opportunities for innovation, enriching the scholarly conversation and promoting collaborative efforts. It also contributes insights to a mini project, enhancing the exchange of knowledge.

In Chapter 3, the focus shifts to the Proposed System, emphasizing its design for simplicity and efficiency. This chapter outlines the system's key features and operational processes, addressing limitations of existing systems and setting the stage for a novel contribution to the field.

In Chapter 4, an in-depth analysis of evolutionary algorithms takes center stage, serving as a cornerstone of the project's methodology. Through detailed examination, this chapter elucidates the rationale behind the chosen approach, preparing for empirical validation in subsequent chapters.

In Chapter 5, the emphasis is on empirical validation and system analysis. Through rigorous experimentation and evaluation, this chapter demonstrates the practical effectiveness of the proposed system. User flow illustrations and usability assessments offer insights into its functionality, reinforcing its relevance in real-world applications.

The final chapter, Chapter 6, discusses future directions and broader implications of the research. By synthesizing the key findings, this chapter outlines potential avenues for further exploration, highlighting the report's contributions to scholarly dialogue and its capacity to inspire future research and development.

In conclusion, this thoughtfully structured report provides an in-depth examination of the project's area, serving as a guide for ongoing innovation and progress. With its organized approach and thorough analysis, the report delivers valuable insights and establishes a foundation for future advancements in the field.

Chapter 2

Literature Survey

2.1 Reinforcement Learning for Bioretrosynthesis

2.1.1 Survey of Existing system

The survey of existing systems in bioretrosynthesis and metabolic engineering reveals a diverse array of tools designed to assist researchers in the design and optimization of biosynthetic pathways. One of the earliest tools, RetroPath, employs a rule-based approach to predict biosynthetic pathways, allowing users to input target compounds and retrieve potential pathways. Its successor, RetroPath 2.0, enhances these capabilities by incorporating more sophisticated algorithms and a larger dataset, improving accuracy in finding relevant pathways for metabolic engineering.

The most advanced among them, RetroPath RL, utilizes the Monte Carlo Tree Search (MCTS) algorithm, offering an open-source Python package that combines biochemical scoring with chemical similarity for pathway prediction. This tool has demonstrated a high success rate in identifying experimentally relevant pathways and provides metabolic engineers with multiple options, although it faces challenges in pathway ranking and handling stereochemistry. Other notable systems include Pathway Tools, which offers a comprehensive database for pathway analysis and visualization, and MetaNetX, which focuses on metabolic network analysis rather than direct pathway design. Overall, while these tools vary in their approaches and features, the development of RetroPath RL marks a significant advancement in the field, leveraging modern algorithms to enhance the efficiency and accuracy of pathway predictions, thereby addressing some limitations of earlier systems. [1].

2.1.2 Research Gaps

Despite the advancements in bioretrosynthesis and metabolic engineering tools, several research gaps remain that hinder the full realization of their potential. One significant gap is the integration of real-time data and dynamic modeling into pathway design, as most systems rely on static databases and predefined reaction rules, limiting their adaptability to new findings and conditions. Additionally, while tools like RetroPath RL have made strides in utilizing advanced algorithms such as Monte Carlo Tree Search, there is still a need for improved methods to incorporate learned values and machine learning techniques that enhance pathway predictions based on historical data. Furthermore, the handling of stereochemistry in retrosynthetic analysis remains a challenge, as many tools do not adequately account for the spatial arrangement of atoms, which is crucial for the synthesis of complex natural products. Another area that requires attention is the ranking and evaluation of suggested pathways; current systems often lack robust criteria for assessing the feasibility and efficiency of proposed routes, which can lead to suboptimal experimental outcomes. Lastly, the accessibility and usability of these tools for non-expert users pose a barrier, as many systems are complex and require a deep understanding of metabolic engineering principles. Addressing these gaps could significantly enhance the effectiveness of bioretrosynthesis tools and facilitate more successful metabolic engineering projects.

2.1.3 Problem Definition

Bioretrosynthesis and metabolic engineering face several challenges in the design and optimization of biosynthetic pathways. Many existing tools depend on static databases and predefined reaction rules, limiting their ability to adapt to new experimental data or discoveries. This rigidity constrains innovation and reduces the accuracy of pathway predictions due to the lack of advanced algorithms and machine learning integration.

A key issue is the handling of stereochemistry, where the spatial arrangement of atoms is often overlooked, particularly in complex natural product synthesis, leading to inaccurate predictions and failed experiments. Additionally, the lack of robust criteria for evaluating and ranking suggested pathways complicates the decision-making process for researchers, often resulting in inefficient routes. Lastly, the complexity of these tools and their steep learning curve make them less accessible to non-experts, hindering wider adoption in the field.

2.1.4 Objectives

The objectives outlined in the proposed research adopt a structured and methodical approach, aimed at addressing key challenges in the field of bioretrosynthesis and metabolic engineering. These objectives are formulated with clarity and precision, adhering to standards of academic rigor and professionalism.

1. The primary objective is to conduct an in-depth survey of existing bioretrosynthesis tools and methodologies, focusing on their current capabilities, limitations, and applicability in metabolic engineering. This review will provide a broad understanding of the strengths and weaknesses of the prevailing systems, helping to identify areas where innovation is needed. By analyzing the contemporary landscape, the survey aims to offer researchers valuable insights that will inform the development of more advanced and adaptable tools.

2. Another critical objective involves identifying the limitations in current tools, specifically in areas such as static database reliance, stereochemistry handling, and pathway evaluation. By systematically assessing these gaps, the research will highlight computational inefficiencies, scalability challenges, and usability issues. This evaluation serves as a foundation for future innovations, pinpointing areas where new algorithms, real-time data integration, and machine learning techniques can significantly enhance pathway predictions and metabolic engineering outcomes.

3. The final objective focuses on developing a next-generation bioretrosynthesis tool that integrates advanced algorithms like Monte Carlo Tree Search, machine learning, and real-time dynamic modeling. This tool will enhance pathway prediction accuracy, improve stereochemistry handling in retrosynthesis, and introduce strong criteria for pathway evaluation. It will prioritize user accessibility by offering a modular, intuitive platform for both experts and non-experts, creating a flexible and responsive solution for evolving biosynthetic pathway design needs.

In conclusion, these objectives reflect a commitment to advancing research in bioretrosynthesis through methodical analysis, critical evaluation, and innovative tool development. By adhering to these goals, the research aims to contribute meaningful advancements in the design and optimization of biosynthetic pathways, facilitating broader application in metabolic engineering.

2.2 Transformer Model for Hybrid Retrosynthesis Planning

2.2.1 Survey of Existing system

Existing systems for retrosynthesis planning primarily focus on predicting synthetic pathways for chemical compounds using various computational approaches. These systems often rely on extensive databases such as MetaCyc, KEGG, Rhea, and BRENDA for biological reactions, along with the USPTO database for chemical reactions, which provide essential data for model training and prediction. Traditional retrosynthesis tools typically employ either template-based or template-free methods; the former uses predefined reaction templates to guide synthesis predictions, while the latter relies on machine learning techniques to generate pathways without such constraints. Most systems operate on a two-tiered approach, where a single-step prediction model generates potential precursors for a target molecule, which are then iteratively fed back into the model to explore further synthesis steps until a termination condition is met.

To efficiently rank precursor candidates, scoring methods are implemented, helping to identify the most reliable pathways in a timely manner. However, many current models predominantly rely on experimentally validated reactions, which can limit their ability to explore novel pathways. Additionally, the inherent complexity of biological reactions can complicate predictions, as variations in reaction conditions, enzyme activity, and substrate specificity play significant roles. Performance metrics often evaluate these systems based on their success rates in predicting viable synthetic pathways; for instance, BioNavi has demonstrated a high success rate in generating pathways for natural products and drugs, outperforming many existing models. Other models also leverage heuristic approaches and optimization algorithms to enhance their efficiency.

Overall, while significant strides have been made in retrosynthesis planning, challenges related to data limitations, pathway diversity, and the complexity of biological reactions remain. These challenges prompt the development of advanced models like BioNavi that integrate deep learning and reaction templates to enhance prediction accuracy and pathway exploration while facilitating the discovery of innovative synthetic routes that were previously unexplored. By addressing these limitations, future advancements can lead to more comprehensive tools that can predict a broader range of synthetic pathways and support more efficient laboratory synthesis, ultimately contributing to advancements in pharmaceuticals and biotechnology. [2].

2.2.2 Research Gaps

Research gaps in retrosynthesis planning for models like BioNavi include limited reaction data, particularly negative data for low-yield reactions, and a lack of comprehensive stereochemical information essential for accurate biological synthesis predictions. Selecting appropriate reaction conditions is also challenging, as current models often neglect this aspect. Additionally, improving pathway diversity and integrating reaction yield predictions are crucial for enhancing accuracy. Finally, scalability and efficiency must be addressed to broaden applicability across various chemical classes.

2.2.3 Problem Definition

This paper addresses challenges in retrosynthesis planning for high-value chemicals. Current models, such as BioNavi, rely on experimentally validated reactions, limiting the exploration of novel pathways due to insufficient datasets, especially for low-yield reactions. Incomplete stereochemical information and the need for better reaction condition selection also hinder predictions. Enhancing model scalability and efficiency is essential for broader applicability across chemical classes, ultimately leading to more effective tools for retrosynthesis planning.

2.2.4 Objectives

The primary objective of this research is to develop a more robust and comprehensive model for retrosynthesis planning that integrates both chemical and biological synthesis pathways. The specific objectives include:

1. Incorporating negative reaction data and complete stereochemical information into the dataset.
2. Optimizing reaction condition selection for better synthetic pathway predictions.
3. Enhancing model scalability and efficiency for broader chemical class applicability.
4. Generating diverse and feasible synthetic pathways.
5. Promoting efficient production of high-value-added chemicals to advance synthetic chemistry.

2.3 Retro synthetic planning with Dual Value Network

2.3.1 Survey of Existing system

Existing studies have investigated various methods for retrosynthetic planning, including traditional heuristic approaches and modern machine learning techniques. The Planning with Dual Value Networks (PDVN) algorithm introduces an innovative solution to enhance retrosynthesis by integrating reinforcement learning with a structured evaluation of synthesis pathways. PDVN employs two value networks: the Synthesizability Network, which assesses whether a molecule can be synthesized from available building blocks, and the Cost Network, which estimates synthesis costs based on reaction complexity and reagent availability. The algorithm also features a two-branch policy network, combining a fixed single-step model for valid reactions with a learnable single-step model that adapts through feedback. Utilizing a tree-shaped Markov Decision Process (MDP), PDVN alternates between planning and updating phases to refine its predictions. Performance improvements include increased success rates—such as the Retro* planner's success rate rising from 85.79 percent to 98. percent — and reduced synthesis route lengths. PDVN has shown effectiveness across diverse datasets, demonstrating its robustness in tackling various retrosynthetic challenges. [3].

2.3.2 Research Gaps

The paper highlights three key limitations of the Planning with Dual Value Networks (PDVN) algorithm in retrosynthetic planning:

A.Dependence on Quality of Training Data: PDVN's performance relies heavily on the quality and diversity of the training data. Inaccuracies or insufficient representation can negatively impact predictions.

B.Complexity of Reaction Space: The vast and intricate retrosynthetic space presents challenges for navigating numerous possible reactions and pathways, especially for complex target molecules.

C.Generalization to Novel Reactions: PDVN may struggle to generalize to new or unconventional reactions not well-represented in the training data, limiting its ability to propose valid synthesis routes.

2.3.3 Problem Definition

Retrosynthesis involves deconstructing a target molecule into simpler precursors, aiming to identify a series of chemical reactions that synthesize the target from available starting materials. Traditional methods often rely on single-step reaction predictors, optimizing accuracy for individual reactions but failing to consider the overall synthesis route. Multi-step planners may generate complete routes but do not effectively leverage single-step predictors, leading to lower success rates and longer routes.

There is a need for a more integrated approach that combines the strengths of single-step predictors with multi-step planning. The existing systems often fail to optimize for both the synthesizability and cost of the entire synthesis route, which is crucial for practical applications in drug discovery and materials design.

2.3.4 Objectives

The objectives of the research paper on Planning with Dual Value Networks (PDVN) in retrosynthetic planning can be summarized as follows:

A. Develop an Integrated Algorithm: To create a novel algorithm, PDVN, that integrates single-step reaction prediction with multi-step retrosynthetic planning, thereby optimizing the overall synthesis route.

B. Enhance Performance of Multi-Step Planners: To significantly improve the performance metrics of existing multi-step retrosynthetic planners, particularly in terms of success rates and route quality, by leveraging the capabilities of the PDVN algorithm.

C. Utilize Dual Value Networks: To construct and utilize two separate value networks within the PDVN framework that predict the synthesizability of molecules, ensuring that proposed routes are feasible and the synthesis cost of molecules, allowing for the identification of economically viable routes.

D. Maintain Single-Step Accuracy: To design a two-branch policy network structure that retains high accuracy in single-step predictions while optimizing for complete synthesis routes, thus balancing the need for precision in individual reactions with the overall synthesis strategy.

2.4 RL Optimization of Reaction Routes with Hybrid Organic and Synthetic Biology Data

2.4.1 Survey of Existing system

The proposed model features a sophisticated reinforcement learning (RL) framework designed to optimize biochemical synthesis routes by leveraging a hybrid dataset that combines organic chemistry and synthetic biology data. Central to this framework is the RL agent, which makes decisions based on a state space of possible molecular structures and an action space of available reactions. The hybrid dataset integrates information from the Reaxys® Database, detailing organic reactions, and the Kyoto Encyclopedia of Genes and Genomes (KEGG), which provides insights into metabolic pathways and enzymatic reactions, enabling exploration of both traditional and biocatalytic synthesis routes.

The learning process balances exploration of new pathways with exploitation of known successful routes, using a reward system to reinforce efficient synthesis characterized by high atom economy and fewer reaction steps. The agent updates its policy through algorithms like Q-learning.

The model evaluates routes based on criteria such as atom economy, reaction steps, and the cost of building blocks. It demonstrated effective convergence to optimal solutions within 20 iterations, showcasing its ability to identify near-optimal synthesis pathways, thus enhancing the efficiency and cost-effectiveness of biochemical synthesis. [4].

2.4.2 Research Gaps

The research paper outlines several limitations of the proposed reinforcement learning (RL) model for optimizing biochemical synthesis routes. Key issues include a heavy reliance on dataset quality, with a smaller biological dataset (about 30,000 molecules) limiting generalization. Uncertainty in one-step predictions can propagate through multi-step planning, especially in complex reactions, affecting accuracy. The complexity of reaction networks can hinder effective convergence, leading to suboptimal synthesis routes. Its reliance on historical data restricts applicability to emerging synthetic methods, and its ability to generalize across diverse chemical spaces may be limited. Finally, the current model does not incorporate data on engineered enzymes, which could enhance its effectiveness.

2.4.3 Problem Definition

The research paper tackles the challenge of optimizing biochemical synthesis routes by integrating organic chemistry and synthetic biology data. It focuses on designing efficient, cost-effective pathways for complex molecules, especially drug candidates that are difficult to synthesize with traditional methods. This involves merging established organic synthesis routes with alternative enzymatic reactions from synthetic biology. Key aspects include retrosynthesis planning, which requires deconstructing target molecules into simpler precursors while considering criteria like atom economy and reaction costs. The complexity of biochemical reactions, characterized by multiple pathways, adds to the difficulty, while uncertainty in reaction predictions can lead to failures. The model aims to enhance decision-making in selecting reaction steps through reinforcement learning, allowing for better exploration of potential pathways.

2.4.4 Objectives

The proposed model in the research paper aims to optimize biochemical synthesis routes using reinforcement learning (RL) and integrates data from organic chemistry and synthetic biology. The key objectives are:

1. Develop methods to suggest near-optimal pathways for complex molecules, streamlining the synthesis process.

2. Employ RL techniques to learn from past reactions, enabling informed decisions that maximize efficiency and minimize costs.

3. Evaluate proposed pathways based on metrics like atom economy, reaction steps, and cost, ensuring feasibility and economic viability.

4. Demonstrate that integrating chemical and biological reactions leads to significant cost savings and efficiency improvements over conventional methods.

5. Ensure the model is scalable and applicable across diverse chemical spaces and synthetic challenges.

2.5 Probabilistic Hypergraphic Enumeration(PHE) for Novel Biochemical Pathways

2.5.1 Survey of Existing system

The proposed model in the research paper focuses on efficiently enumerating branched novel biochemical pathways using a probabilistic technique. It employs a hypergraphic reaction network to represent biochemical reactions, enabling a more complex depiction than simple graph-based networks by capturing multiple substrates and products in a single reaction. The model incorporates four types of edges, representing various reaction types, which enhances its flexibility in accounting for complex biochemical interactions. Through a series of expansion steps, new compounds are generated based on predefined reaction rules, allowing for greater exploration of potential pathways compared to traditional models. Additionally, probabilistic methods are utilized to navigate the vast space of biochemical pathways, improving the identification and enumeration of biologically relevant pathways. The model's effectiveness is validated through the generation of "Golden Set Pathways," which are tested using reinforcement learning techniques for bioretrosynthesis, demonstrating its capability to find relevant pathways that vary based on reaction complexity. [5].

2.5.2 Research Gaps

The proposed model in the research paper has several limitations. Firstly, it relies on a specific set of reaction rules, which may not encompass all biochemical reactions, potentially limiting the diversity of generated pathways. Additionally, while the hypergraphic approach captures complex reactions, it may not fully represent the intricacies of biochemical systems, including enzyme kinetics and regulatory mechanisms. The model also faces challenges with computational resources due to the combinatorial explosion of pathways in hypergraphic networks, making it less applicable to larger networks. Furthermore, assumptions made during network growth may oversimplify biochemical processes, leading to inaccurate predictions. Validation through "Golden Set Pathways" might not reflect practical applications in biochemistry, and the model's accuracy is significantly dependent on the quality of the underlying reaction data, as incomplete or biased data can result in misleading predictions.

2.5.3 Problem Definition

The problem definition in the research paper revolves around the challenge of efficiently enumerating novel biochemical pathways that are both biologically relevant and feasible for synthetic applications. Here are the key aspects of the problem as outlined in the paper:

1. Biochemical pathways are complex, involving numerous reactions in various combinations, making it challenging to systematically explore potential pathways for applications like bioremediation and drug development.

2. Traditional pathway enumeration methods often use simple graph representations, which may not accurately capture reactions with multiple substrates and products, leading to incomplete pathway predictions.

3. Generated pathways must be both novel and biologically relevant, requiring models that reflect biochemical principles and constraints to ensure viability in real biological systems.

2.5.4 Objectives

The objectives of the proposed model in the research paper focus on addressing the challenges associated with enumerating novel biochemical pathways. Here are the key objectives outlined in the paper:

1. Develop algorithms to efficiently enumerate a wide range of biochemical pathways while minimizing computational demands.

2. Leverage hypergraphic representations to accurately capture the complexity of multiple substrates and products, enhancing pathway prediction.

3. Employ probabilistic methods to navigate the combinatorial complexity of biochemical reactions while maintaining prediction accuracy.

4. Validate the model's predictions against established "Golden Set Pathways" to assess performance and reliability in identifying novel, relevant pathways.

2.6 Molecular Transformer with Hyper-Graph Exploration for Retrosynthesis

2.6.1 Survey of Existing system

The research paper presents a novel approach to retrosynthesis by integrating the Molecular Transformer architecture with a hyper-graph exploration strategy. The Molecular Transformer, a machine learning model that treats chemical reactions as a language, effectively predicts reaction outcomes and enhances the prediction of retrosynthetic pathways. A key innovation is the introduction of a single-step retrosynthetic model that predicts both reactants and reagents, addressing a more complex aspect of retrosynthesis. The model employs a hyper-graph exploration strategy, dynamically constructing the hyper-graph during predictions, filtering and expanding nodes based on a Bayesian-like probability score, and identifying optimal pathways through a beam search algorithm. To evaluate performance, four new metrics are introduced: Coverage, which measures the ability to find viable disconnection sites; Class Diversity, which ensures a variety of disconnection strategies; Round-Trip Accuracy, which assesses the regeneration of target molecules; and Jensen–Shannon Divergence (JSD), quantifying similarity between disconnection classes to minimize bias. While the framework performs exceptionally well across various disconnections, it has limitations, such as bias towards certain reaction classes and challenges in handling stereochemical reactions due to the quality of the training dataset. [6].

2.6.2 Research Gaps

The proposed model in the research paper has several limitations. Firstly, it exhibits bias towards certain reaction classes, such as reduction and oxidation, which can lead to illogical disconnection strategies. Additionally, it struggles with managing stereochemical aspects due to a lack of diverse examples in the training dataset. The model's performance is also heavily dependent on the quality and diversity of the training data; if coverage is insufficient, predictions may be limited. While designed for single-step retrosynthesis, it may not effectively handle the complexities of multi-step pathways. Moreover, its reliability decreases when encountering novel reactions not represented in the training set. Lastly, the evaluation metrics introduced may not fully capture all qualitative factors that influence the practical applicability of the predicted pathways.

2.6.3 Problem Definition

The research paper addresses the challenge of automating retrosynthetic analysis in organic chemistry, highlighting several key issues. Retrosynthesis is complex due to the multitude of possible reactions and factors like reaction conditions and reagent availability. Existing models have primarily focused on single-step predictions, often neglecting multi-step synthesis and the prediction of necessary reagents, solvents, and catalysts. Furthermore, biases in training datasets can lead to suboptimal disconnection strategies, and the quality of data significantly impacts performance, especially in stereochemical reactions. There is also a lack of robust evaluation metrics to comprehensively assess model performance. Lastly, scalability remains a concern, as effective models must provide reliable predictions for complex synthetic problems in real-world applications.

2.6.4 Objectives

The objective of the research paper is to develop and present an advanced model for retrosynthetic analysis that addresses the limitations of existing approaches in organic chemistry. The specific goals of the research include:

1. Develop a model for autonomous prediction of retrosynthetic pathways to streamline the process for chemists.

2. Introduce a model that predicts reactants, reagents, solvents, and catalysts for target molecule synthesis, enhancing practical applicability.

3. Implement a strategy for efficient identification of optimal synthetic pathways to avoid selectivity traps and improve route quality.

4. Establish metrics like coverage, class diversity, round-trip accuracy, and Jensen–Shannon divergence for a comprehensive assessment of model performance.

2.7 Computer-Assisted Retrosynthesis Based on Molecular Similarity

2.7.1 Survey of Existing system

The proposed model in the research paper introduces a similarity-based approach to retrosynthesis that operates as follows.

This data-driven model relies on a dataset of 40,000 known reactions from patent literature, eliminating the need for predefined chemical rules or heuristics. Central to its functionality is a molecular similarity metric that quantifies similarity scores between products on a scale of 0 to 1, allowing for the identification and ranking of one-step retrosynthetic disconnections. The model retrieves up to 100 relevant reaction precedents based on these scores to ensure computational efficiency while providing a robust selection of options.

From each precedent, it extracts highly localized transforms that focus on the atoms directly involved in the reaction, such as leaving groups, distinguishing it from traditional methods that may require broader contextual information. The approach is deterministic and non-parametric, functioning directly on the available data without parameter tuning or training, effectively mimicking the "average retrosynthetic strategy" implicit in the reaction corpus.

Performance metrics indicate strong efficacy, with reactants appearing in the top 10 proposed precursors for 74.1 percent of test reactions and achieving a perfect recommendation rate of 52.9 percent when the reaction class is known. Additionally, the model can extend its one-step strategy to multistep pathway planning, making it suitable for complex synthetic routes in medicinal chemistry. [7].

2.7.2 Research Gaps

The proposed model has several limitations: it may generate synthetically unviable suggestions, relies on potentially low-quality patent data, and lacks contextual information about reagents and conditions. While it mimics average strategies, it struggles with novel chemistries and is dependent on the diversity of its training corpus. By limiting retrieved precedents to 100 for efficiency, it may miss relevant reactions, and the lack of explicit chemical knowledge can hinder accuracy.

2.7.3 Problem Definition

The research paper addresses the challenge of automated retrosynthesis, aiming to develop a method that automatically suggests viable synthetic pathways for target molecules. It emphasizes a data-driven approach that leverages a large corpus of known reactions instead of predefined rules. The model utilizes molecular similarity to identify relevant reaction precedents, ensuring that proposed reactions not only fit the target molecule but also have a high likelihood of success in synthesis. The paper critiques existing methods for their reliance on heuristic scoring and highlights the need for a more objective evaluation of synthetic accessibility. Additionally, the model seeks to generalize across various reaction classes, allowing for diverse transformation origins in retrosynthetic pathways.

2.7.4 Objectives

The objectives of the model presented in the research paper are as follows:

1. To develop a system that can automatically generate retrosynthetic disconnections for target molecules, mimicking the analytical process of a trained synthetic chemist.

2. To leverage molecular similarity as a key metric for identifying relevant reaction precedents from a large corpus of known reactions, thereby facilitating the generation of synthetic routes based on analogy.

3. To implement a purely data-driven methodology that relies on historical reaction data, avoiding the need for predefined rules or heuristics, and allowing for a more objective evaluation of synthetic pathways.

4. To ensure that the suggested disconnections lead to chemically valid precursor molecules by using highly local transforms that focus on the atoms directly involved in the reactions.

5. To provide an open-source framework that allows other researchers to adapt the model to different datasets and enhance its applicability in various contexts, promoting collaboration and further development in the field of automated retrosynthesis.

2.8 Using Machine Learning to Predict Suitable Conditions for Organic Reactions

2.8.1 Survey of Existing system

The landscape of machine learning applications in organic synthesis has seen significant advancements over the past decade. Quantitative Structure-Activity Relationship (QSAR) models are traditional computational techniques used to predict chemical reaction outcomes based on the molecular structures of compounds. While they have been effective in some scenarios, they face significant limitations when applied to complex organic reactions due to the high dimensionality of chemical data and the inherent noise within it. These factors reduce the overall predictive power and generalizability of QSAR models, making them less suitable for intricate reaction mechanisms.

Machine learning techniques like Random Forests and Support Vector Machines (SVM) have been employed to predict reaction conditions from historical reaction data. Although these models provide reasonable predictions, they are often hindered by limited interpretability and scalability issues, which become pronounced when applied to large chemical datasets. Neural networks, particularly deep learning models, have addressed some of these challenges by offering greater accuracy in modeling reaction outcomes and predicting reaction conditions. However, their success heavily depends on the availability of large, high-quality training datasets, making them less accessible for areas with limited data. [8].

2.8.2 Research Gaps

Despite advancements in computational models for predicting chemical reactions, several gaps remain in the current literature. One major challenge is limited generalization, where models trained on specific datasets struggle to accurately predict outcomes for new or unseen reactions.Data scarcity is another issue, as high-quality datasets covering diverse reaction conditions are limited, which restricts the development of robust models. Additionally, many existing models fail to account for the integration of multivariate factors such as temperature, pressure, and solvent, leading to less accurate predictions. The lack of real-time adaptability further limits the application of these models in dynamic experimental settings.

2.8.3 Problem Definition

The primary problem addressed in this paper is the challenge of accurately predicting optimal conditions for organic reactions through machine learning techniques. Existing methods face significant hurdles due to high dimensionality, as numerous factors influence reaction outcomes, complicating the identification of suitable conditions. Additionally, the **limited diversity of available datasets** restricts the ability of these models to generalize across the wide range of organic reactions encountered in practical applications. Moreover, there is a pressing need for model interpretability, as predictive models must not only demonstrate high performance but also offer insights into their underlying decision-making processes to foster trust and usability among chemists.

2.8.4 Objectives

The objectives of the model presented in the research paper are as follows:

1. Develop a Comprehensive Machine Learning Framework: Create a model that effectively predicts suitable conditions for a wide range of organic reactions while integrating multiple reaction factors.

2. Utilize Diverse Datasets: Leverage various existing datasets and possibly incorporate synthetic data to improve model training and generalization.

3. Enhance Model Interpretability: Implement techniques to enhance the interpretability of the model's predictions, making it easier for chemists to understand the rationale behind suggested conditions.

4. Real-Time Adaptation: Design a system capable of adapting to new data in real-time, facilitating dynamic learning and improving the predictive accuracy over time.

5. Validate with Experimental Data: Test the developed framework against experimental results to ensure practical applicability and reliability of the predictions.

2.9 Artificial Intelligence for Retrosynthesis Prediction

2.9.1 Survey of Existing system

Rule-based systems were among the earliest approaches to retrosynthesis prediction, with notable examples such as CASES and Synthia (formerly known as Chematica). These systems rely on manually encoded reaction rules derived from existing chemical knowledge, which can effectively suggest synthetic pathways for known reactions. However, their scalability and generalization capabilities are inherently limited, as they can only operate within the confines of the encoded rules and the prior knowledge of reactions, making them less adaptable to novel or complex organic reactions.

In contrast, template-based models utilize reaction templates extracted from extensive databases, such as Reaxys and the USPTO database, to enhance their predictive capabilities. While these models offer more flexibility compared to rule-based systems, their performance is constrained by the availability and comprehensiveness of reaction templates. As a result, template-based approaches struggle to accurately predict outcomes for novel reactions that do not fit existing templates, which limits their overall utility in diverse synthetic scenarios. [9].

2.9.2 Research Gaps

Despite advancements in AI-based retrosynthesis prediction, several challenges and gaps persist. One significant issue is the limited generalization to novel reactions, as current systems often struggle to accurately predict outcomes for compounds or pathways not represented in their training data. Additionally, the availability and quality of data present further obstacles, as many AI models rely on large datasets that may be lacking, especially in niche chemical areas or for proprietary molecules; moreover, poor data quality or imbalanced datasets can result in biased or inaccurate predictions. Efficiency in navigating the vast chemical reaction space is another challenge, as finding optimal retrosynthetic pathways can be computationally expensive, particularly for complex or multi-step reactions. Finally, while AI for chemical retrosynthesis has progressed, the integration of biological reactions into retrosynthetic planning remains a critical challenge, especially in the context of bioretrosynthesis, where both chemical and biological steps need to be effectively optimized.

2.9.3 Problem Definition

The problem addressed by this research is the development of a robust AI-based retrosynthesis prediction system that can generalize to novel chemical reactions, handle large and diverse reaction datasets, and efficiently predict multi-step reaction pathways. Traditional systems rely heavily on expert-curated rules or templates and are often limited by the quality and size of available reaction datasets. Moreover, current AI systems lack the ability to seamlessly integrate biological reactions into retrosynthetic planning.

2.9.4 Objectives

The objectives of the model presented in the research paper are as follows:

1. Develop a Reinforcement Learning (RL)-based framework for retrosynthesis prediction: The framework should improve the efficiency and accuracy of retrosynthesis by learning from a combination of chemical reaction data and biological reaction data.

2. Enhance generalization for novel reactions: Build a model that can predict novel reactions that are not present in the training dataset by leveraging graph neural networks or transformer-based architectures, which generalize better across chemical reaction spaces.

3. Integrate biological and chemical pathways: Develop a hybrid AI system that combines chemical and biological reaction steps, enabling the prediction of bioretrosynthesis pathways, which are crucial for sustainable bio-based chemical production.

4. Improve data handling and efficiency: Design strategies to handle noisy or incomplete reaction data while ensuring that the system remains computationally efficient, even in large reaction spaces.

5. Multi-task learning for reaction prediction: Incorporate multi-task learning to predict reaction outcomes, retrosynthesis, and optimization of reaction conditions simultaneously, enabling a more holistic retrosynthesis planning system.

2.10 Learning Retrosynthetic Planning through Simulated Experience

2.10.1 Survey of Existing system

Traditional retrosynthetic analysis primarily depends on human expertise and heuristic rules for planning synthetic routes, which can be time-consuming and subjective. Existing software tools such as Synthia and Reaxys utilize rule-based approaches to assist chemists in retrosynthesis; however, these systems often struggle with adaptability, particularly when confronted with novel compounds or complex retrosynthesis pathways that fall outside their predefined rules. As a result, while these tools can offer valuable insights, their reliance on established knowledge limits their effectiveness in dynamic chemical environments where innovation is crucial.
In contrast, recent advancements have integrated machine learning approaches, particularly deep learning, to enhance the prediction of feasible synthetic routes by leveraging extensive datasets of chemical reactions. Models like Seq2Seq have been effectively employed to generate reaction sequences, thereby improving the automation of retrosynthetic planning and reducing the dependency on human intervention. Furthermore, techniques such as reinforcement learning (RL) have emerged as valuable tools in various fields, including chemical synthesis, by simulating experiences to enhance decision-making processes. [10].

2.10.2 Research Gaps

Current systems in retrosynthesis prediction face several critical challenges that hinder their effectiveness. One major issue is the **limited exploration of novel reactions**, as these systems often rely heavily on historical data, which restricts their ability to generate innovative synthetic routes. Additionally, there is a **lack of robustness** in many machine learning models, as they frequently fail to account for the complexity and variability inherent in real-world chemical reactions, leading to less reliable predictions. Moreover, existing approaches often struggle to effectively **integrate domain knowledge**, such as fundamental chemical principles, with data-driven methods, which limits their applicability in practical scenarios. Although **simulated experience application** has been explored through reinforcement learning, its potential to enhance the learning process in retrosynthesis remains largely underutilized, indicating a significant area for improvement in developing more effective predictive models.

2.10.3 Problem Definition

The primary problem addressed in this paper is the challenge of efficiently generating robust retrosynthetic pathways that not only utilize historical reaction data but also incorporate simulated experiences to explore novel synthetic routes. The limitations of current models in terms of adaptability and exploration hinder their application in complex retrosynthetic problems.

2.10.4 Objectives

The objectives of the model presented in the research paper are as follows:

1. To create an RL-based approach that leverages simulated experiences for learning retrosynthetic planning.
2. To improve the generation of synthetic routes by incorporating exploration mechanisms that encourage the discovery of novel reactions.
3. To combine machine learning techniques with chemical knowledge to enhance the robustness and reliability of the retrosynthetic planning process.
4. To assess the performance of the proposed system in terms of accuracy and novelty of the generated synthetic routes compared to existing methods.
5. To provide a framework that can be applied in real-world scenarios for efficient retrosynthetic planning in various fields, including pharmaceuticals and materials science.

Chapter 3

Proposed System

3.1 Present Report on Investigation

This segment describes the method for growing a dependable bio-retrosynthesis version, including the framework for assessment, inference approach, version design, and data education. "Fig.1", Our method carries a transformer-based totally sequence-to-sequence model that has been refined the use of a carefully decided on dataset of chemical and biosynthetic reactions. using contrastive mastering and self-correcting beam seek, we enhance retrosynthetic prediction accuracy and dependability. moreover, model training is optimized the use of parameter-efficient nice-tuning techniques like QLoRA [**?**], which assure computational efficiency while preserving proper predictive overall performance. to assess the efficacy of our method, the assessment framework uses a diffusion of indicators, consisting of chemical validity assessments and top-N accuracy.

By combining two fundamental statistics resources, the examine made use of an extensive response dataset for bio-retrosynthesis. We began by using integrating a carefully selected biosynthetic response dataset (BioChem) that covered validated bio-alterations and was sourced from well-known enzyme databases like KEGG [11], MetaCyc [12], and MetaNetX. second, we delivered USPTO NPL natural response facts, which changed into selected particularly for reactions that resembled natural products. complete insurance of each enzymatic and organic reaction information is ensured via this blended method.

To correctly distinguish among the source (supposed product) and precursor (predicted substrate), each reaction turned into encoded as a response SMILES string utilizing a custom delimiter for records illustration. Every reaction whilst necessary, stereochemical records become preserved

all through the canonicalization of SMILES. To assure that all chemical symbols and unique characters are handled efficaciously, we installed region a tokenizer that is area-tailored.

3.2 Architecture of Proposed System

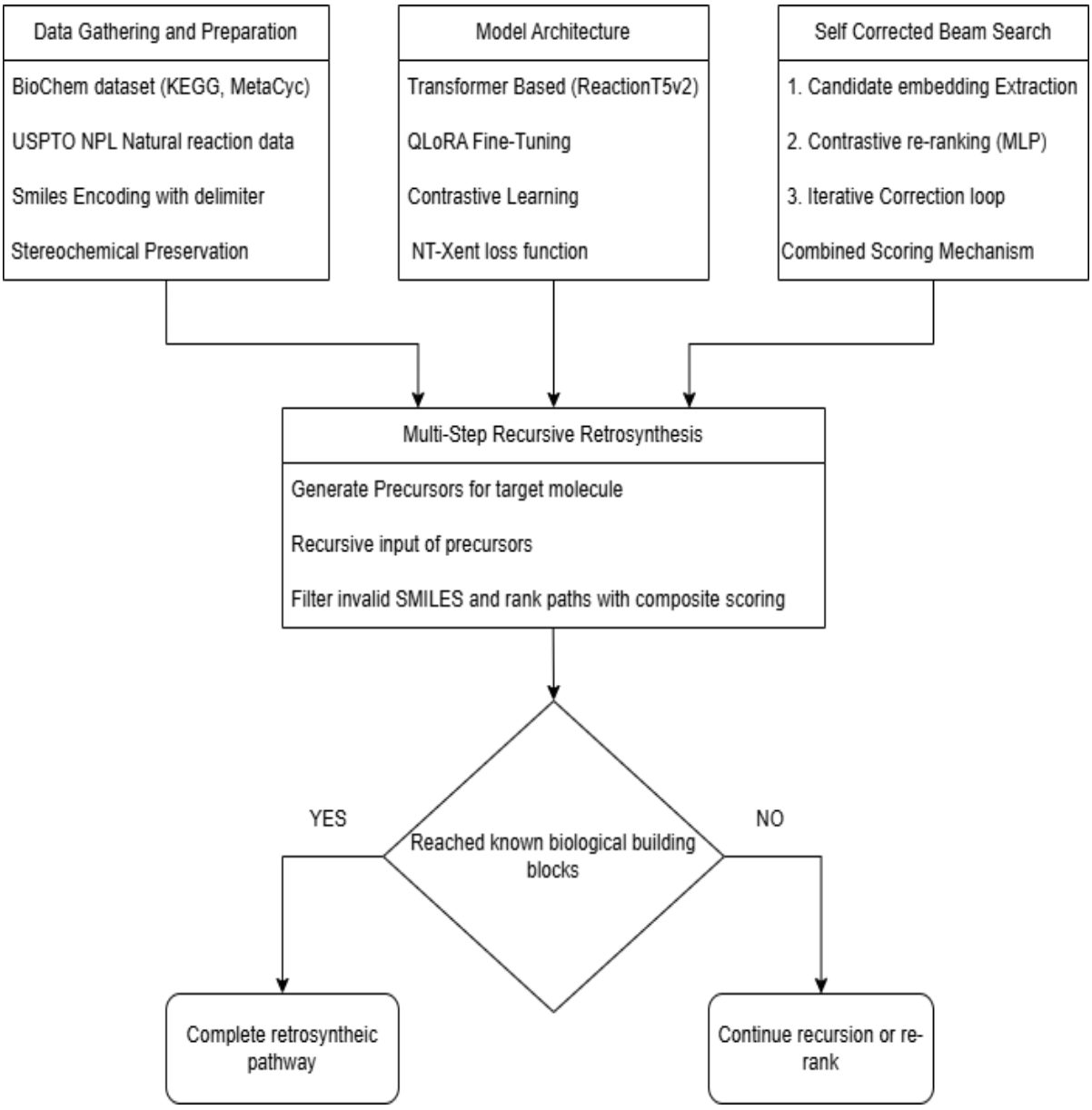

Figure 3.1: Flowchart of Proposed System

This flowchart Fig. 3.1 gives a clear and structured overview of how the components of your proposed system interact to achieve bioretrosynthesis planning.

Our methodology is built on a transformer-based totally structure, in particular a chain-to-collection version that has been pre-educated on full-size datasets of chemical reactions. Our

precise bio-retrosynthesis dataset turned into used to refine this base model if you want to determine the right mapping among product SMILES (input) and precursor SMILES (output). At some point of the training section, we delivered a contrastive getting to know factor to triumph over the inherent uncertainties in retrosynthetic predictions. bad pairs have been created by way of cautiously manipulating both the enter or output SMILES, while fine pairs have been described as accurate precursor-product relationships. functional organization substitutions and small chemical mistakes were brought during this perturbation method. To assure that the learnt representations of accurate precursor-product pairs remained towards one another than mismatched pairings, the NT-Xent loss characteristic changed into applied.

Implementation of Self-Corrected Beam Search:

- **Candidate Embedding Extraction:** the usage of encoderdecoder hidden states, the machine takes intermediate representations for each beam candidate out of the exceptional-tuned transformer.
- **Contrastive Re-ranking:** A contrastive re-scoring module based on learnt MLP assesses how nicely candidate representations in shape the right transformation patterns that have been obtained from training times. thru this process, the initial beam ratings are changed to definitely weight candidates according to retrosynthetic common sense and chemical plausibility.
- **Iterative Correction:** candidates are constantly reranked using a mixture of log-probability and contrastive scores as the beam seek runs in a self-correcting loop. This iterative method improves the chemical validity of final outputs and makes it simpler to repair faults from previous beam expansions.

Multi-Step Recursive Retrosynthesis Strategy:

To predict complete synthetic routes, we implemented a recursive retrosynthesis pipeline. The model generates precursors for a target molecule, and each precursor is recursively input until all branches reach known biological building blocks. Invalid or chemically implausible SMILES are filtered at each step. A priority queue ranks partial paths using a composite score of model likelihood and contrastive plausibility. Recursion depth is capped to ensure efficiency. This approach enables the discovery of full retrosynthetic pathways, enhancing

practical synthesis planning and enabling real-world applications in bio-manufacturing, green chemistry, and metabolic engineering.

Instructional Procedure:

A multi-task gaining knowledge of approach is used in the education manner, combining: cross-entropy-based collectionto-collection loss for the creation of precursor SMILES Optimizing latent representation distances among correct and incorrect pairings the use of contrastive loss. To maximize computing assets, we used parameter-efficient pleasant-tuning approaches (QLoRA). by using the usage of grid seek optimization on a cut up validation dataset, important hyper-parameters inclusive of learning fee, batch size, and beam width were identified. further to SMILES-based statistics augmentation thru random atom ordering perturbations, the education procedure blanketed dropout and weight decay for regularization.

GPU infrastructure compliant with CUDA 12.6 was used for all research. We hold a public repository with complete supply code and hyperparameter units to guarantee reproducibility. comprehensive documentation on statistics splitting distinct environment necessities are furnished in requirements.txt. Scripts for evaluation and experimental logs.

Chapter 4

Design

4.1 Design details

4.1.1 Dataset

After extensively researching various datasets, models, and tools from research papers, we identified suitable resources for our retrosynthesis process [13]. The biological reactions were sourced from databases such as MetaCyc [11], KEGG [14], Rhea [12], and BRENDA [15], while chemical reactions were obtained from the USPTO dataset. To ensure consistency, any reactions with an unequal number of carbon atoms were excluded. Next, RXNMapper [16] was utilized to map atoms from reactants to products, and RDChiral [17] was employed to extract reaction templates. Reactions that shared the same template (referred to as reference reactions) were grouped together.

In most retrosynthesis scenarios, the model only receives a single target molecule as input. For reactions with multiple products, these were split into individual reactions, keeping all reactants intact but focusing on one product at a time (e.g., a reaction like A + B $\rightarrow$ C + D would be divided into A + B $\rightarrow$ C and A + B $\rightarrow$ D). To minimize complexity, only reactants containing carbon atoms matching those in the product were retained. Reactants and product pairs were standardized using canonical SMILES from RDKit and deduplicated.

Additionally, a label indicating the reaction type was appended to the SMILES string using a vertical bar "|", distinguishing biological (e.g., "Cc1cnc2ccccc2c1|$\langle$B$\rangle$ $\rightarrow$ product") from chemical reactions (e.g., "Cc1ccc(F)cc1CN|$\langle$C$\rangle$ $\rightarrow$ product").

Chapter 5

Conclusion

5.1 Future Scope

The future scope of this project offers significant potential in metabolic engineering and synthetic biology. By integrating advanced machine learning models and expanding datasets, the accuracy and efficiency of retrosynthesis predictions can be enhanced. Future efforts may include real-time data feedback mechanisms for adaptive learning and hybrid approaches combining biological and chemical synthesis methods for more sustainable production processes. The proposed system could also apply to complex natural products, driving innovation in pharmaceuticals, agriculture, and green chemistry. Additionally, improvements in user interfaces will enhance accessibility for researchers and practitioners. Overall, this project lays a strong foundation for advancements in biotechnological applications and synthetic route design.

5.2 Conclusion

This work presents a sophisticated bioretrosynthesis prediction model combining contrastive learning, a self-correcting beam search, and a transformer-based sequence-to-sequence architecture. It uses a curated dataset integrating USPTO NPL and BioChem reactions, encoded via custom SMILES strings with stereochemical fidelity. The model, fine-tuned with QLoRA, employs NT-Xent loss and re-ranks candidate embeddings using contrastive scoring and MLPs. A self-correcting beam search iteratively improves predictions. Multi-task learning with contrastive and cross-entropy losses enhances performance. Evaluations using top-N accuracy, diversity, RDKit validation, and ablation studies show improved accuracy and robustness, advancing AI-driven retrosynthetic analysis. All experiments are done using CUDA 12.6-compatible GPUs.

Bibliography

[1] Mathilde Koch, Thomas Duigou, and Jean-Loup Faulon. Reinforcement learning for bioretrosynthesis. *ACS Synthetic Biology*, 9(1):157–168, 2020. PMID: 31841626.

[2] Tao Zeng, Zhehao Jin, Shuangjia Zheng, Tao Yu, and Ruibo Wu. Developing bionavi for hybrid retrosynthesis planning. *JACS Au*, 4(7):2492–2502, 2024.

[3] Guoqing Liu, Di Xue, Shufang Xie, Yingce Xia, Austin Tripp, Krzysztof Maziarz, Marwin Segler, Tao Qin, Zongzhang Zhang, and Tie-Yan Liu. Retrosynthetic planning with dual value networks. 2023.

[4] Chonghuan Zhang and Alexei A. Lapkin. Reinforcement learning optimization of reaction routes on the basis of large, hybrid organic chemistry–synthetic biological, reaction network data. *React. Chem. Eng.*, 8:2491–2504, 2023.

[5] Zhiqing Xu and Radhakrishnan Mahadevan. Efficient enumeration of branched novel biochemical pathways using a probabilistic technique. *Industrial & Engineering Chemistry Research*, 61(25):8645–8657, 2022.

[6] Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H. Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, and Teodoro Laino. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.*, 11:3316–3325, 2020.

[7] Connor W. Coley, Luke Rogers, William H. Green, and Klavs F. Jensen. Computer-assisted retrosynthesis based on molecular similarity. *ACS Central Science*, 3(12):1237–1245, 2017. PMID: 29296663.

[8] Hanyu Gao, Thomas J. Struble, Connor W. Coley, Yuran Wang, William H. Green, and Klavs F. Jensen. Using machine learning to predict suitable conditions for organic reactions. *ACS Central Science*, 4(11):1465–1476, 2018. PMID: 30555898.

[9] Yinjie Jiang, Yemin Yu, Ming Kong, Yu Mei, Luotian Yuan, Zhengxing Huang, Kun Kuang, Zhihua Wang, Huaxiu Yao, James Zou, Connor W. Coley, and Ying Wei. Artificial intelligence for retrosynthesis prediction. *Engineering*, 25:32–50, 2023.

[10] Bishop KJM Schreck JS, Coley CW. Learning retrosynthetic planning through simulated experience. *ACS Cent Sci*, 5(6):970-981, 2019 Jun 26. PMID: 31263756; PMCID: PMC6598174.

[11] Ron Caspi, Richard Billington, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Peter E Midford, Wai Kit Ong, Suzanne Paley, Pallavi Subhraveti, and Peter D Karp. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Research*, 48(D1):D445–D453, 10 2019.

[12] Parit Bansal, Anne Morgat, Kristian B Axelsen, Venkatesh Muthukrishnan, Elisabeth Coudert, Lucila Aimo, Nevila Hyka-Nouspikel, Elisabeth Gasteiger, Arnaud Kerhornou, Teresa Batista Neto, Monica Pozzato, Marie-Claude Blatter, Alex Ignatchenko, Nicole Redaschi, and Alan Bridge. Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Research*, 50(D1):D693–D700, 11 2021.

[13] Jingxin Dong, Mingyi Zhao, Yuansheng Liu, Yansen Su, and Xiangxiang Zeng. Deep learning in retrosynthesis planning: datasets, models and tools. *Briefings in Bioinformatics*, 23(1):bbab391, 09 2021.

[14] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, 42(D1):D199–D205, 11 2013.

[15] Antje Chang, Lisa Jeske, Sandra Ulbrich, Julia Hofmann, Julia Koblitz, Ida Schomburg, Meina Neumann-Schaal, Dieter Jahn, and Dietmar Schomburg. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Research*, 49(D1):D498–D508, 11 2020.

[16] Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7(15):eabe4166, 2021.

[17] Connor W. Coley, William H. Green, and Klavs F. Jensen. Rdchiral: An rdkit wrapper for handling stereochemistry in retrosynthetic template extraction and application. *Journal of Chemical Information and Modeling*, 59(6):2529–2537, 2019.