`PatentAgent`: Intelligent Agent for Automated Pharmaceutical Patent Analysis

Xin Wang[1*], **Yifan Zhang**[1*], **Xiaojing Zhang**[1], **Longhui Yu**[1], **Xinna Lin**[1], **Jindong Jiang**[1], **Bin Ma**[1]
Kaicheng Yu[1]

[1]Westlake University

buhangyunfei@gmail.com, yifanzhang2024@u.northwestern.edu, dexianz07@gmail.com, longhuiyu98@gmail.com,
linxinna@westlake.edu.cn, jiangjinkekao@gmail.com, binma@stu.xjtu.edu.cn, kyu@westlake.edu.cn

Abstract

Pharmaceutical patents play a vital role in biochemical industries, especially in drug discovery, providing researchers with unique early access to data, experimental results, and research insights. With the advancement of machine learning, patent analysis has evolved from manual labor to tasks assisted by automatic tools. However, there still lacks an unified agent that assists every aspect of patent analysis, from patent reading to core chemical identification. Leveraging the capabilities of Large Language Models (LLMs) to understand requests and follow instructions, we introduce the **first** intelligent agent in this domain, `PatentAgent`, poised to advance and potentially revolutionize the landscape of pharmaceutical research. `PatentAgent` comprises three key end-to-end modules — *PA-QA*, *PA-Img2Mol*, and *PA-CoreId* — that respectively perform (1) patent question-answering, (2) image-to-molecular-structure conversion, and (3) core chemical structure identification, addressing the essential needs of scientists and practitioners in pharmaceutical patent analysis. Each module of `PatentAgent` demonstrates significant effectiveness with the updated algorithm and the synergistic design of `PatentAgent` framework. *PA-Img2Mol* outperform existing methods across CLEF, JPO, UOB, and USPTO patent benchmarks with an accuracy gain between 2.46% and 8.37% while *PA-CoreId* realizes accuracy improvement ranging from 7.15% to 7.62% on PatentNetML benchmark. Our code and dataset will be publicly available.

1 Introduction

Patents provide a unique wealth of information for the early stages of chemical research and development, especially in freedom-to-operate analysis (Mucke 2023), prior-art search (Setchi et al. 2021), and landscape analysis (Ohms 2021). They play a critical role in lead and target discovery (Zdrazil et al. 2024; Senger 2017), offering earlier access to innovative insights and biochemical data than publishable journals (Southan et al. 2013). Given the competitive nature of the drug development business (Bregonje 2005), patents are a vital source of information. Aiming for the wealth of information provided by patents, scientists continuously work to extract relevant contents in patents and analyze them. However, *scientists struggle with extracting and organizing information due to the lack of an unified agent.* Specifically, existing approaches have the following shortcomings:

Manual approaches require excessive human effort to extract information. Traditional methods before the existence of computational technologies, such as manual review and keyword search (Hu et al. 2018; Andres and Treanor 2010), are often viewed as the golden standards in patent analysis. However, these methods require scientists to spend considerable time and effort extracting information. These methods rely heavily on human expertise to interpret complex chemical information, making them expensive and time-consuming.

Current computational tools are highly specialized and lack a holistic solution. For instance, methods like text mining (Gadiya et al. 2023; Liu et al. 2011; Yun and Geum 2020) and chemical structure exploration (Morin et al. 2024; Zdrazil et al. 2024) operate independently, with little integration or standardized criteria for data formats, algorithms, or evaluation metrics. This lack of common standards makes it difficult to coordinate efforts across multiple modules, a necessity in comprehensive patent analysis. As a result, researchers, particularly those without a background in computer science, such as medical researchers, struggle to effectively utilize these tools. The need for coding skills and the lack of seamless integration create significant barriers, limiting the accessibility and efficiency of these computational methods.

Existing tools still struggle to identify the core compound successfully. Identifying the core compound structure among hundreds or even thousands of chemicals (Habibi et al. 2016) in a pharmaceutical patent is one of the major tasks for scientists and practitioners (Akhondi et al. 2019). Given the complexity of the task, however, current tools struggle at an accuracy level of random guesses (Tyrchan et al. 2012; Falaguera and Mestres 2021; Zhu et al. 2024), including the most recent one.

Those shortcomings make it challenging for AI to assist scientists in analyzing pharmaceutical patents. A unified agent with the capability to precisely interpret and execute scientific requests is essential. To bridge this gap, we aim to create an integrated system that can fully leverage the potential of computational methods to streamline patent analysis (Morin et al. 2024). Fortunately, Large Language
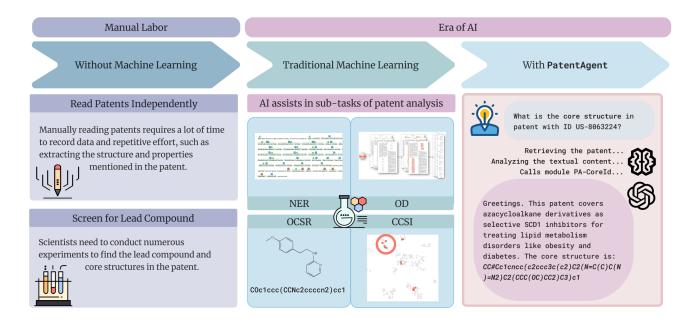
*Equal Contribution

Figure 1: Comparison of patent analysis between manual, traditional machine learning, and `PatentAgent`. The sub-tasks listed in the *Traditional Machine Learning* phase of patent analysis are: Named Entity Recognition (NER), Object Detection (OD), Optical Chemical Structure Recognition (OCSR), and Core Chemical Structure Identification (CCSI).

Models(LLMs) have demonstrated their exceptional ability to understand human instructions (Zeng et al. 2023) and effectively utilize tools (Ruan et al. 2023), making them well-suited to assist scientists in conducting patent analysis. With the integration of LLMs, `PatentAgent` intelligently understands human intentions in natural language even when they are multi-fold and involves several patents.

Besides the LLM orchestrator, `PatentAgent` consists of three main modules, *PA-QA*, *PA-Img2Mol*, and *PA-CoreId*:

1. *PA-QA*: A question-answering chatbot that faithfully responds users' inquiries about patents.
2. *PA-Img2Mol*: An ensemble of deep learning models and transformer-based models that uses Visual Language Model (VLM) as evaluator. It achieves accuracy gain between 8.37% and 2.46% over the best performing models on widely recognized benchmarks of Optical Chemical Structure Recognition (OCSR), including CLEF, JPO, and UOB.
3. *PA-CoreId*: A machine learning classifier to identify the core chemical structure amongst various chemicals. It realizes better performance on the most updated and accepted PatentNetML dataset for Core Chemical Structure Identification (CCSI).

The success of `PatentAgent` and its modules across multiple datasets highlights robustness and its ability to consistently outperform existing methods, underscoring its transformative potential. Specifically, `PatentAgent` not only improves accuracy in critical tasks in chemical structure analysis but also significantly reduces time and effort required to analyze complex pharmaceutical patents. Such effectiveness is further demonstrated in a case study in Sec-

tion 3.3, where `PatentAgent` simplifies the patent analysis process, enabling researchers to gain actionable insights with unprecedented pace and reliability.

2 Method

This section presents our `PatentAgent` framework (Fig 2). Our LLM orchestrator is introduced in Section 2.1, and the three succeeding subsections discuss three major modules used in `PatentAgent`. Section 2.2 explains the question answering chatbot *PA-QA* that to process natural language queries and generate accurate responses based on the knowledge base of the patent file. Section 2.3 describes the module *PA-Img2Mol* and its updated algorithm in converting images of chemical structures to molecular expression with Simplified Molecular Input Line Entry System (SMILES) (Weininger 1988) . Section 2.4 introduces module *PA-CoreId* that identifies the common chemical scaffold across a series of diverse chemical compounds.

2.1 LLM Orchestrator

The LLM orchestrator serves as the central component that connects and coordinates the three major modules: *PA-QA*, *PA-Img2Mol*, and *PA-CoreId*. Upon receiving a user query, the LLM orchestrator interprets the input and determines whether the query involves retrieving textual information, converting images to molecular structures, or identifying core chemical structures through contextual understanding. It then processes the queries to the requested format and routes the processed queries to corresponding module, aggregates the results and returns a comprehensive response to the user.
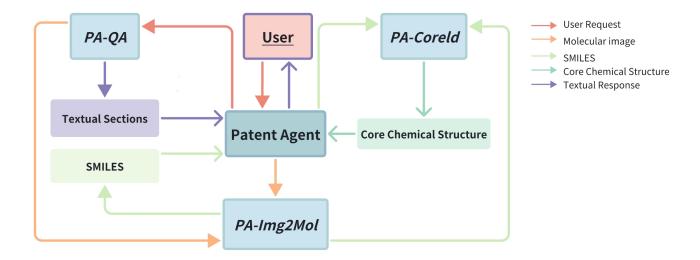
Figure 2: Workflow Illustration for `PatentAgent`. `PatentAgent` consists of three major modules, namely *PA-QA*, *PA-Img2Mol*, and *PA-CoreId*. *PA-QA* processes user requests and outputs patent segmentation (texts or images), *PA-Img2Mol* processes molecular images and output SMILES, and *PA-CoreId* processes SMILES and outputs the core chemical structure. Refer to the color coding legends for better understanding.

In more complicated cases that span multiple aspects of patent analysis, such as identifying the core chemical structure directly from a patent file, the LLM orchestrator sequences tasks accordingly. It first retrieves chemical images in the pdf files and convert the chemical images to Simplified Molecular Input Line Entry System (SMILES) (Weininger 1988) with *PA-Img2Mol*, then identifies the core chemical structure with *PA-CoreId*, finally synthesizing the information into a unified response. More intuitive workflow can be viewed in Figure 2.

2.2 *PA-QA*

PA-QA is designed to interpret and respond to natural language queries about patents, extracting precise information from the text of patent documents. After collecting user's request from LLM Orchestrator, *PA-QA* focuses on specific sections of the patent that are most likely to contain relevant answers. When retrieving relevant sections, *PA-QA* utilizes a doc-layout model comprised of object detection and Optical Character Recognition (OCR) to extract specific regions in the patents, such as the abstract, claims, and detailed description.

After getting specific section, *PA-QA* employs a transformer-based retrieval model that performs semantic search across the parsed sections of the patent document. Once the relevant information are retrieved, *PA-QA* synthesizes into a coherent answer. If necessary, the LLM orchestra would rephrase the text to ensure the response is directly aligned with the user's query.

2.3 *PA-Img2Mol*

The task of *PA-Img2Mol* is to convert images to molecular sequences, which is a task of Optical Chemical Structure Recognition (OCSR). Here specifically we wish to generate Simplified Molecular Input Line Entry System (SMILES) (Weininger 1988).

Firstly, We use MolDetect (Ozymandias314 2023)[1] to extract specially images of chemical compounds from patents. An intuitive visual illustration can be viewed in Figure 4. After retrieving all the chemical images, we followe Algorithm 1 to convert chemical images to SMILES.

Considering the different architectures and training data of each OCSR model, it is natural that each model has its own strengths and weaknesses in learning and identifying the relevant features. Therefore, we select three OCSR models, namely DECIMER 2.0 (Rajan, Zielesny, and Steinbeck 2021), MolScribe (Qian et al. 2023b), and SwinOCSR (Xu et al. 2022) for the best performance and use GPT-4o (Achiam et al. 2023) as the VLM evaluator. An intuitive illustration of the framework can be viewed in Figure 3.

In this module, for each chemical image, we use the three models: DECIMER, MolScribe, and SwinOCSR, to generate SMILES. After standardizing the generated SMILES, we use RDKit (RDKit 2015) to convert SMILES back to chemical images. Then we use GPT-4o (Achiam et al. 2023) the VLM to calculate the similarity between the regenerated image and the original image, so that the similarity considers possible re-orientation and scaling. By comparing the results, we are able to find the best SMILES with the highest similarity.

2.4 *PA-CoreId*

PA-CoreId is designed to identify the core chemical structure when given a list of chemicals, most possibly retrieved from the same patent file. The mechanism for identifying

---

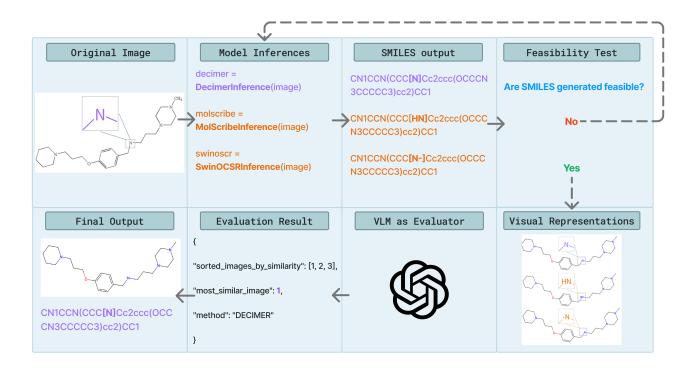[1]Github repo forked from RxnScribe (Qian et al. 2023a)

Figure 3: Complete workflow of converting a molecular image to SMILES: from model inference through VLM evaluation to final output.

Algorithm 1: The OCSR Algorithm with Model Ensemble and VLM as Evaluator

---

Given three Image-to-SMILES models M_1, M_2, and M_3 and an input image I, this algorithm outputs the final SMILES: $\text{SMILES}_{\text{final}}$

for *each model M* **do**

$\quad \text{SMILES}_M = f_M(I)$ // Image to SMILES

$\quad \text{SMILES}_{\text{std}} = g(\text{SMILES}_M)$ // Standardizes

$\quad I' = \text{RDKit}(\text{SMILES})$ // Convert back to image

$\quad \text{similarity}_M = \text{VLM}(I, I')$

end

$\text{SMILES}_{\text{final}} = \text{SMILES}_{M*}$, when $\text{similarity}_{M*} = \max\{\text{similarity}_{M_1}, \text{similarity}_{M_2}, \text{similarity}_{M_3}\}$

---

the core structure is to find the common structure in all the chemicals, with the understanding that the maximal common substructure is highly likely the core structure (Zhu et al. 2024).

To identify the core compound, we trained a machine learning model which ensembles XGBoost (Chen and Guestrin 2016) and Random Forest (Breiman 2001). Bayesian Optimization was used to identify the optimal parameters for each model. Boruta (Kursa, Jankowski, and Rudnicki 2010) is used for feature selection. We trained the model with mini dataset[2] of PatentNetML (Zhu et al. 2024)

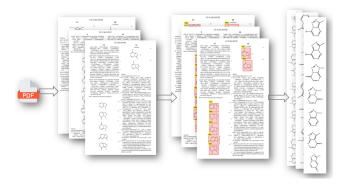[2]The mini train consists of only 112 data points

Figure 4: This figure shows the process of extracting related images of chemical structure in pdf patent file.

with 3080ti for one day.

In preparing the features for the model, considering the focus of our application in pharmaceutical patents, we selected Extended-Connectivity FingerPrints 4 (ECFP4) (Rogers and Hahn 2010), adapted from Morgan algorithm (Morgan 1965), to capture the structural information of the chemicals, as ECFP4 is the algorithm that captures the most information. Then we convert the ECFP4 fingerprint and physicochemical properties into a graph network, using cutoff values from 0.4 to 0.9. We also create an adjacency matrix where each pair of vertices is labeled 1 or 0 depending on whether they are adjacent.
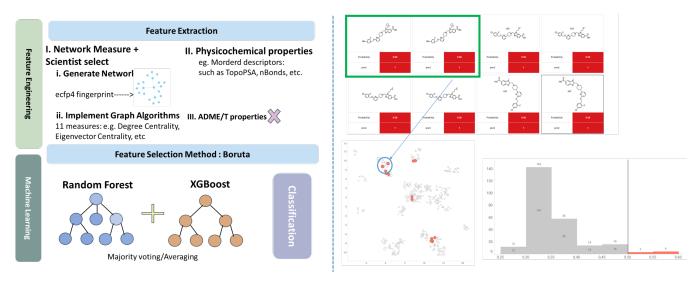
Figure 5: The left image illustrates the framework for achieving specific lead identification (leadid), covering the process from feature selection to model training. The right image shows an example of leadid classification within a patent: (a) in the top left visualizes the lead id, (b) in the bottom left projects the molecules from the patent and marks the position of the lead id and (d) in the top right displays a probability distribution graph of the molecules in the patent.

3 Experiments

We design experiments to evaluate `PatentAgent` to answer the following research questions:

RQ1: How effectively does *PA-Img2Mol* convert patent images to molecular structures compared to existing methods?

RQ2: How accurately does *PA-CoreId* identify core chemical structures within pharmaceutical patents, comparing to other state-of-the-art methods?

RQ3: How does `PatentAgent` perform in a real-world application, such as analyzing a complex pharmaceutical patent, and what insights can it provide to researchers?

3.1 RQ1: *PA-Img2Mol* converts patent images to molecular structure more effectively than existing methods

To answer **RQ1**, we compared `PatentAgent` with other openly available tools for Optical Chemical Structure Recognition (OCSR):

1. **OSRA**: Filippov and Nicklaus (2009) propose the first open-source tool to identify chemical structure from images using traditional computational techniques.
2. **Imago**: Smolov, Zentsev, and Rybalkin (2011) suggest a novel algorithm for processing images of chemicals by separating out the graphical and symbolic layers.
3. **MolVec**: National Center for Advancing Translational Sciences (NCATS (2017)) publishes a public repository vectorizing chemical images into chemical objects.
4. **Img2Mol**: Clevert et al. (2021) train and publicize a deep convolutional neural network that consists of an encoder that converts images into latent space and a decoder that translates latnent space information into SMILES.

5. **SwinOCSR**: Xu et al. (2022) propose a Swin transformer based approach to train an end-to-end model.
6. **MolScribe**: Qian et al. (2023b) present a generational model that predicts atom, bond, and layer to construct molecular structure.
7. **DECIMER**: Rajan, Zielesny, and Steinbeck (2021) brings an image transformer that trains on Pub-Chem (Kim et al. 2023) dataset and a dataset generated with Markush structure.

Experiments comparing `PatentAgent` and the above models run on these datasets with real chemical structure images without distortion [3]:

1. **CLEF**: 992 images and chemical molecule file pairs published by the Conference and Labs of the Evaluation Forum (CLEF) in 2012.
2. **JPO**: 450 images and chemical molecule file pairs published by Japanese Patent Office (JPO).
3. **UOB**: 5470 image and chemical molecule file pairs published by the University of Birmingham (UOB), United Kingdom.
4. **USPTO**: 5719 image and chemical molecule file pairs published by the US PaTent Office (USPTO).
5. **ACS**: 331 image and chemical molecule file pairs published by American Chemistry Society (ACS) and manually annotated by MolScribe (Qian et al. 2023b).

The results of the experiments are presented in Table 1. The results show that *PA-Img2Mol* achieves better performance in four of the five benchmarks. The lower performances of all models the other models on JPO are likely due to the noises in the images, including unsegmented labels, numbers, and English or Japanese characters (Rajan et al. 2023, 2020). The same happens in our experiments too.

---

[3] These datasets are all obtained from validation datasets at OCSR Benchmark (Rajan et al. 2020)

	CLEF	JPO	UOB	USPTO	ACS
OSRA	84.6	55.3	78.5	87.4	55.3
MolVec	82.8	67.8	80.6	**88.4**	47.4
Imago	59.0	40.0	58.0	87.0	-
Img2Mol	18.3	16.4	68.7	26.3	23.0
SwinOCSR	30.0	13.8	44.9	27.9	27.5
MolScribe	88.9	76.2	87.9	79.0	71.9
DECIMER	62.7	55.2	88.2	41.1	46.5
`PatentAgent`	**89.28**	**78.68**	**96.57**	82.63	**72.34**

Table 1: This table presents the experiment results of converting chemical structure images to SMILES. *PA-Img2Mol* recognizes optical chemical structures more accurately than all the other strong baseline models on four of the five benchmarks and competitively on USPTO.

(Top)	1	5	10	5%	10%
CSA	6.14	27.00	-	-	-
MI	5.41	18.00	-	-	-
FOG	6.37	26.00	-	-	-
PatentNetML	7.14	35.71	50.00	28.57	50.00
`PatentAgent`	14.29	35.71	50.00	35.71	50.00

Table 2: *PA-CoreId* predicts the core chemical in patents more accurately than the recent machine-learning and cheminformatics approaches. It reaches SoTA in all accuracy measures.

When using only *PA-Img2Mol* module, the accuracy for JPO benchmark is only 58%. However, when running the benchmark using the whole agent, the accuracy boosts to 78.68%, which is the highest of existing tools, which further validates the effectiveness of our `PatentAgent` framework.

Moreover, for the USPTO benchmark, following the remarks of Clevert et al. (2021); Rajan et al. (2023) all three models surpassing us, OSRA, MolVec, and Imago, are overfitted to available benchmarks. While our *PA-Img2Mol* contains no training process and utilizes none of the overfitted models, it still reaches competitive accuracy on USPTO benchmark.

3.2 RQ2: *PA-CoreId* identifies core chemical structure more accurately comparing to other SoTA methods

To answer **RQ2**, we compared our model with PatentNetML (Zhu et al. 2024) and other traditional Cheminformatics Methods, including Cluster Seed Analysis(CSA) (Hattori, Wakabayashi, and Tamaki 2008), Molecular Idol(MI) (Lee et al. 2019), and Frequency of Group Analysis(FOG) (Tyrchan et al. 2012) on PatentNetML dataset (Zhu et al. 2024).

We present the results from the experiments in Table 2 with all the accuracy measures, including Top 1, Top 5, Top 10, Top 1%, Top 5%, and Top 10%. We got the highest percentage in all these accuracy measures with the updated feature selection algorithm.

3.3 RQ3: `PatentAgent` improves the efficiency when dealing with complex patent — a case study

We select the patent with ID US-8063224 (LaChance et al. 2011) for this case study. This patent covers azacycloalkane derivatives acting as selective inhibitors intended for the treatment and prevention of diseases related to abnormal lipid metabolism, such as cardiovascular disease, obesity, diabetes, and metabolic syndrome.

To demonstrate `PatentAgent`'s ability under real-world circumstances, we want to test if `PatentAgent` successfully complete the requests of the users and achieves high accuracy when identifying core chemical structure.

When asked to identify the core chemical compound from this patent, after scanning the file, extracting chemical compound images, converting these images to SMILES, and identifying the top ten candidates with the highest probabilities, the core chemical compound ranks sixth. This lower ranking compared to using *PA-CoreId* alone is due to discrepancies in retrieving all chemical images, as many chemicals in the patent are not listed as images. Nevertheless, `PatentAgent` effectively reduces the effort needed to identify the core chemical compound from fifty compounds, saving time for researchers in testing and linking tools.

4 Discussion

Our experiments show that `PatentAgent` is a robust tool capable of addressing key aspects of patent analysis within a single, integrated framework. The effective operation across core modules—patent question answering, image-to-molecular-structure conversion, and chemical identification—demonstrates its functionality and ability to streamline traditionally fragmented workflows.

Integration of Multiple Aspects. One of the most significant finding is `PatentAgent`'s ability to unify diverse tasks into a unified system, effectively bridging gaps that usually require multiple tools. This integration ensures consistency and enhances efficiency in patent analysis, suggesting that `PatentAgent` could serve as an all-in-one solution.

Accuracy Boosts. Our experiments, particularly those conducted on the JPO dataset, indicate that `PatentAgent` improves accuracy in identifying and analyzing chemical structures within patents, leading to more reliable insights for subsequent research.

Efficiency in Real-Life Applications. `PatentAgent` simplifies tasks traditionally handled by separate modules into one streamlined step, significantly reducing time and effort. This makes patent analysis more accessible and manageable, broadening its use among researchers.

These findings collectively reinforce the initial claim that `PatentAgent` has the potential to transform the drug discovery routine by accelerating the process from patent analysis to actionable insights, contributing to faster and more effective drug development.

Figure 6: User Interface of `PatentAgent`.

5 Related Works

This section provides an overview of existing approaches in patent analysis in time and method development order, along with gaps that our approach aims to address.

5.1 Text Mining and Natural Language Processing

With advancements in computational linguistics, NLP techniques have been applied to patent analysis. SCRIPDB provides a database that automatically retrieves relevant patent information, including syntheses, chemicals, and reactions (Heifets and Jurisica 2012). Chemical NER methods like tmChem (Leaman, Wei, and Lu 2015) and ChemSpot (Rocktäschel, Weidlich, and Leser 2012) efficiently identify chemical names in patent text. Feng and Fuhai (2012) combines text mining with informetric methods for morphological analysis of patent technology. Chikkamath et al. (2020) introduces novelty detection and evaluates various language models. Kronemeyer, Draeger, and Moehrle (2021) presents a process model identifying frugal patents.

Recent works leverage AI and language models for patent innovation and validation. Inspired by TRIZ (Ilevbare, Probert, and Phaal 2013), Trapp and Warschat (2024) identifies contradictions in patents using LLMs to drive technological progress. PARIS and LE-PARIS generate effective responses to Office Actions (OAs) (Chu et al. 2024a), while PRO combines legal knowledge graphs with LLMs through retrieval-augmented generation (RAG) for more faithful OA responses (Chu et al. 2024b).

5.2 Chemical Structure Analysis

Chemical structures are the vital treasure of information in biochemical patents, especially in drug discoveries. There are plenty of works related to chemical structures in patents, and the first one is mining chemicals in patents.

SureChEMBLcss (Falaguera and Mestres 2021) contains authentic information on core chemicals in patents in the SureChEMBL datasets. PatentNetML (Zhu et al. 2024) predicts key compound using machine learning networks,

a model MOMP (Turutov and Radinsky 2024) is proposed to optimize molecules with patentability constraint. CASTER (Huang et al. 2020) predicts drug-drug interactions (DDIs) given chemical structures of drugs.

5.3 Integrated AI Systems

The increasing complexity and volume of patent data, with the development of large language models (LLMs), have led to increasing efforts in integrating AI into a whole system consisting of multiple analytical modules. MolTailor (Guo et al. 2024) combines molecular representation model as a knowledge base with language model as an agent to achieve molecule-text multi-task regression (MT-MTR). DECIMER.ai (Rajan et al. 2023) is an open platform for automatic chemical structure segmentation, classification, and translation in scientific publications, which can also be generalized into patents.

6 Conclusion

In conclusion, in this paper we propose `PatentAgent` as a novel approach in utilizing Large Language Models (LLMs) to link vital tools in pharmaceutical patent analysis so that researchers and practitioners can interactively and efficiently analyze patents from both text and chemical compounds. Moreover, *PA-Img2Mol* and *PA-CoreId* with updated algorithm achieve better performances than current tools. The case study also presents as a qualitative validation of the real-life application of `PatentAgent` even under unusual and complex circumstances.

While our results are encouraging, they represent just the beginning of what `PatentAgent` could achieve. With further development and refinement, `PatentAgent` has the potential to not only streamline patent analysis but also fundamentally change the drug discovery routine by making the process faster, more accurate, and more accessible to researchers worldwide. As the field of automatic patent analysis develops, we wish to see more evaluation datasets and benchmarks for assessing intelligent automatic patent analysis agent like `PatentAgent`.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akhondi, S. A.; Rey, H.; Schwörer, M.; Maier, M.; Toomey, J.; Nau, H.; Ilchmann, G.; Sheehan, M.; Irmer, M.; Bobach, C.; et al. 2019. Automatic identification of relevant chemical compounds from patents. *Database*, 2019: baz001.

Andres, C. J.; and Treanor, R. L. 2010. Patents in drug discovery: Case studies, examples, and simple steps medicinal chemists can take to protect Hard-Won intellectual property. In *Annual reports in medicinal chemistry*, volume 45, 449–463. Elsevier.

Bregonje, M. 2005. Patents: A unique source for scientific technical information in chemistry related industry? *World Patent Information*, 27: 309–315.

Breiman, L. 2001. Random forests. *Machine learning*, 45: 5–32.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Chikkamath, R.; Endres, M.; Bayyapu, L.; and Hewel, C. 2020. An empirical study on patent novelty detection: A novel approach using machine learning and natural language processing. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 1–7. IEEE.

Chu, J.-M.; Lo, H.-C.; Hsiang, J.; and Cho, C.-C. 2024a. From PARIS to LE-PARIS: Toward Patent Response Automation with Recommender Systems and Collaborative Large Language Models. *ArXiv*, abs/2402.00421.

Chu, J.-M.; Lo, H.-C.; Hsiang, J.; and Cho, C.-C. 2024b. Patent Response System Optimised for Faithfulness: Procedural Knowledge Embodiment with Knowledge Graph and Retrieval Augmented Generation. In *KNOWLLM*.

Clevert, D.-A.; Le, T.; Winter, R.; and Montanari, F. 2021. Img2Mol–accurate SMILES recognition from molecular graphical depictions. *Chemical science*, 12(42): 14174–14181.

Falaguera, M. J.; and Mestres, J. 2021. Identification of the core chemical structure in SureChEMBL patents. *Journal of Chemical Information and Modeling*, 61(5): 2241–2247.

Feng, X.; and Fuhai, L. 2012. Patent text mining and informetric-based patent technology morphological analysis: an empirical study. *Technology Analysis & Strategic Management*, 24(5): 467–479.

Filippov, I. V.; and Nicklaus, M. C. 2009. Optical structure recognition software to recover chemical information: OSRA, an open source solution.

Gadiya, Y.; Gribbon, P.; Hofmann-Apitius, M.; and Zaliani, A. 2023. Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery perspective. *Artificial Intelligence in the Life Sciences*, 3: 100069.

Guo, H.; Zhao, S.; Wang, H.; Du, Y.; and Qin, B. 2024. Moltailor: Tailoring chemical molecular representation to specific tasks via text prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18144–18152.

Habibi, M.; Wiegandt, D. L.; Schmedding, F.; and Leser, U. 2016. Recognizing chemicals in patents: a comparative analysis. *Journal of cheminformatics*, 8: 1–15.

Hattori, K.; Wakabayashi, H.; and Tamaki, K. 2008. Predicting key example compounds in competitors' patent applications using structural information alone. *Journal of chemical information and modeling*, 48(1): 135–142.

Heifets, A.; and Jurisica, I. 2012. SCRIPDB: a portal for easy access to syntheses, chemicals and reactions in patents. *Nucleic acids research*, 40(D1): D428–D433.

Hu, J.; Li, S.; Yao, Y.; Yu, L.; Yang, G.; and Hu, J. 2018. Patent keyword extraction algorithm based on distributed representation for patent classification. *Entropy*, 20(2): 104.

Huang, K.; Xiao, C.; Hoang, T.; Glass, L.; and Sun, J. 2020. Caster: Predicting drug interactions with chemical substructure representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 702–709.

Ilevbare, I. M.; Probert, D.; and Phaal, R. 2013. A review of TRIZ, and its benefits and challenges in practice. *Technovation*, 33(2-3): 30–37.

Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. 2023. PubChem 2023 update. *Nucleic acids research*, 51(D1): D1373–D1380.

Kronemeyer, L. L.; Draeger, R.; and Moehrle, M. G. 2021. Stimulating R&D by finding frugal patents: A process model and a comparison between different evaluation methods. *IEEE Transactions on Engineering Management*, 70(2): 615–630.

Kursa, M. B.; Jankowski, A.; and Rudnicki, W. R. 2010. Boruta–a system for feature selection. *Fundamenta Informaticae*, 101(4): 271–285.

LaChance, N.; Li, C. S.; Leclerc, J.-P.; and Ramtohul, Y. K. 2011. Azacycloalkane derivatives as inhibitors of stearoyl-coenzyme a delta-9 desaturase.

Leaman, R.; Wei, C.-H.; and Lu, Z. 2015. tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of cheminformatics*, 7: 1–10.

Lee, S. D.; Priest, C.; Bjursell, M.; Gao, J.; Arneson, D. V.; Ahn, I. S.; Diamante, G.; van Veen, J. E.; Massa, M. G.; Calkin, A. C.; et al. 2019. IDOL regulates systemic energy balance through control of neuronal VLDLR expression. *Nature metabolism*, 1(11): 1089–1100.

Liu, S.-H.; Liao, H.-L.; Pi, S.-M.; and Hu, J.-W. 2011. Development of a Patent Retrieval and Analysis Platform–A hybrid approach. *Expert systems with applications*, 38(6): 7864–7868.

Morgan, H. L. 1965. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2): 107–113.

Morin, L.; Weber, V.; Meijer, G. I.; Yu, F.; and Staar, P. W. 2024. PatCID: an open-access dataset of chemical structures in patent documents. *Nature Communications*, 15(1): 6532.

Mucke, H. 2023. How to conduct a freedom-to-operate analysis for a drug repurposing project. *DrugRxiv*.

NCATS. 2017. molvec. https://github.com/ncats/molvec.

Ohms, J. 2021. Current methodologies for chemical compound searching in patents: A case study. *World Patent Information*, 66: 102055.

Ozymandias314. 2023. MolDetect. https://github.com/Ozymandias314/MolDetect.

Qian, Y.; Guo, J.; Tu, Z.; Coley, C. W.; and Barzilay, R. 2023a. RxnScribe: A Sequence Generation Model for Reaction Diagram Parsing. *Journal of Chemical Information and Modeling*.

Qian, Y.; Guo, J.; Tu, Z.; Li, Z.; Coley, C. W.; and Barzilay, R. 2023b. MolScribe: robust molecular structure recognition with image-to-graph generation. *Journal of Chemical Information and Modeling*, 63(7): 1925–1934.

Rajan, K.; Brinkhaus, H. O.; Agea, M. I.; Zielesny, A.; and Steinbeck, C. 2023. DECIMER. ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nature communications*, 14(1): 5045.

Rajan, K.; Brinkhaus, H. O.; Zielesny, A.; and Steinbeck, C. 2020. A review of optical chemical structure recognition tools. *Journal of Cheminformatics*, 1–13.

Rajan, K.; Zielesny, A.; and Steinbeck, C. 2021. DECIMER 1.0: deep learning for chemical image recognition using transformers. *Journal of Cheminformatics*, 13: 1–16.

RDKit. 2015. RDKit: Open-source cheminformatics. https://www.rdkit.org.

Rocktäschel, T.; Weidlich, M.; and Leser, U. 2012. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12): 1633–1640.

Rogers, D.; and Hahn, M. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5): 742–754.

Ruan, J.; Chen, Y.; Zhang, B.; Xu, Z.; Bao, T.; Du, G.; Shi, S.; Mao, H.; Zeng, X.; and Zhao, R. 2023. Tptu: Task planning and tool usage of large language model-based ai agents. *arXiv preprint arXiv:2308.03427*.

Senger, S. 2017. Assessment of the significance of patent-derived information for the early identification of compound–target interaction hypotheses. *Journal of Cheminformatics*, 9: 1–8.

Setchi, R.; Spasić, I.; Morgan, J.; Harrison, C.; and Corken, R. 2021. Artificial intelligence for patent prior art searching. *World Patent Information*, 64: 102021.

Smolov, V.; Zentsev, F.; and Rybalkin, M. 2011. Imago: Open-Source Toolkit for 2D Chemical Structure Image Recognition. In *TREC*.

Southan, C.; Varkonyi, P.; Boppana, K.; Jagarlapudi, S. A.; and Muresan, S. 2013. Tracking 20 years of compound-to-target output from literature and patents. *PLoS One*, 8(10): e77142.

Trapp, S.; and Warschat, J. 2024. LLM-based Extraction of Contradictions from Patents. *ArXiv*, abs/2403.14258.

Turutov, S.; and Radinsky, K. 2024. Molecular Optimization Model with Patentability Constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38-1, 257–264.

Tyrchan, C.; Boström, J.; Giordanetto, F.; Winter, J.; and Muresan, S. 2012. Exploiting structural information in patent specifications for key compound prediction. *Journal of chemical information and modeling*, 52(6): 1480–1489.

Weininger, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1): 31–36.

Xu, Z.; Li, J.; Yang, Z.; Li, S.; and Li, H. 2022. SwinOCSR: end-to-end optical chemical structure recognition using a Swin Transformer. *Journal of Cheminformatics*, 14(1): 41.

Yun, J.; and Geum, Y. 2020. Automated classification of patents: A topic modeling approach. *Computers & Industrial Engineering*, 147: 106636.

Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; et al. 2024. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, 52(D1): D1180–D1192.

Zeng, Z.; Yu, J.; Gao, T.; Meng, Y.; Goyal, T.; and Chen, D. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Zhu, T.-F.; Qian, R.; Wei, X.; Lu, A.-P.; and Cao, D.-S. 2024. PatentNetML: A Novel Framework for Predicting Key Compounds in Patents Using Network Science and Machine Learning. *Journal of Medicinal Chemistry*, 67-2: 1347–1359.