BioMiner: A Multi-modal System for Automated Mining of Protein-Ligand Bioactivity Data from Literature

Jiaxian Yan[1*], Jintao Zhu[2*], Yuhang Yang[1], Qi Liu[1,✉], Kai Zhang[1], Zaixi Zhang[3], Xukai Liu[1], Boyan Zhang[1], Kaiyuan Gao[4], Jinchuan Xiao[5], and Enhong Chen[1]

[1]State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
[2]Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University
[3]Princeton University
[4]Huazhong University of Science and Technology
[5]Infinite Intelligence Pharma
[*]Jiaxian Yan and Jintao Zhu contribute equally to this work.
[✉]Qi Liu is the corresponding author. qiliuql@ustc.edu.cn.

ABSTRACT

Protein-ligand bioactivity data published in literature are essential for drug discovery, yet manual curation struggles to keep pace with rapidly growing literature. Automated bioactivity extraction remains challenging because it requires not only interpreting biochemical semantics distributed across text, tables, and figures, but also reconstructing chemically exact ligand structures (e.g. Markush structures). To address this bottleneck, we introduce BioMiner, a multi-modal extraction framework that explicitly separates bioactivity semantic interpretation from ligand structure construction. Within BioMiner, bioactivity semantics are inferred through direct reasoning, while chemical structures are resolved via a chemical-structure-grounded visual semantic reasoning (CSG-VSR) paradigm, in which multi-modal large language models operate on chemically grounded visual representations to infer inter-structure relationships, and exact molecular construction is delegated to domain chemistry tools. For rigorous evaluation and method development, we further establish BioVista, a comprehensive benchmark comprising 16,457 bioactivity entries curated from 500 publications. On BioVista, BioMiner validates its extraction ability and provides a quantitative baseline, achieving F1 scores of 0.33 for complex bioactivity triplets. BioMiner's practical utility is demonstrated via three applications: (1) extracting 82,262 data from 11,683 papers to build a pre-training database that improves downstream models performance by 3.9%; (2) enabling a human-in-the-loop workflow that doubles the number of high-quality NLRP3 bioactivity data, helping 38.6% improvement over 28 QSAR models and identification of 16 hit candidates with novel scaffolds; and (3) accelerating the annotation of the protein-ligand structures in PoseBusters benchmark with reported bioactivity, achieving a 5-fold speed increase and 10% accuracy improvement over manual methods. BioMiner and BioVista provide a scalable extraction methodology and a rigorous benchmark, paving the way to unlock bioactivity data that previously required extensive human effort. All codes and data are available at GitHub.

Main

Protein-ligand bioactivity data represents a cornerstone of modern drug discovery[1,2], underpinning structure-activity relationships (SAR) analysis[3–11], quantitative structure-activity relationship (QSAR) modeling, and AI-driven virtual screening[12,13]. Despite the availability of large public resources such as ChEMBL[14], BindingDB[15], and PDBbind[16], the continual expansion of these databases still relies predominantly on manual expert curation. This reliance has emerged as a fundamental scalability bottleneck, increasingly unable to keep pace with the exponential growth of the scientific literature.

Automated data extraction tools have been developed in various scientific domains, yet a critical gap exists in specialized tools tailored for the extraction of protein-ligand bioactivity data. Early tools, constrained by the limitations of prior natural language processing (NLP) and computer vision (CV) techniques, primarily focused on tasks of extracting fundamental chemical information, such as named entity recognition (NER)[17–19] and optical chemical structure recognition (OCSR)[20]. Recently, the development of large language models (LLMs) advances some more promising text-mining tools. For example, GPT-based tools have been proposed for extracting metal-organic framework (MOF) synthesis conditions and enzyme-substrate interactions from literature[21–23].

However, these methods, while demonstrating progresses, are not yet equipped to handle the extraction of protein-ligand bioactivity data (details in Table S1), which requires jointly resolving biochemical semantics and chemically exact

structural representations. This intrinsic coupling fundamentally distinguishes bioactivity extraction from conventional scientific information extraction and gives rise to three core challenges. **First**, bioactivity data is inherently multi-modal, distributed across different modalities, including text, tables, figures, and crucial chemical structures. This multi-modal challenge requires robust cross-modal reasoning rather than unimodal extraction. **Second**, the accurate recognition and conversion of chemical structures, particularly the widely used Markush structures (which represent groups of related chemical compounds)[24], remains a significant obstacle. While existing OCSR methods have ability to recognize some Markush scaffold or R-group substitute structures, the crucial step of enumerating the specific, individual compounds they represent – a necessity for precise bioactivity data extraction – has been largely unexplored. **Third**, besides these intrinsic data complexities, the field lacks standardized, large-scale benchmarks. This critical gap severely hinders the rigorous evaluation, comparative analysis, and systematic advancement of automated methodologies, impeding the development and validation of robust, generalizable solutions capable of reliably unlocking the wealth of bioactivity data embedded within the scientific literature.

To overcome these challenges, we propose **BioMiner**, a multi-modal agentic system designed to automatically extract protein-ligand bioactivity data from literature. Rather than relying on one-shot end-to-end extraction, BioMiner decomposes the extraction task into document parsing, chemical structure extraction, bioactivity measurement extraction, and cross-modal integration, with each subtask handled by specialized agents. As illustrated in Figure 1, the two core subtasks—bioactivity measurement extraction and chemical structure extraction—are explicitly separated and addressed using fundamentally different strategies, where bioactivity measurements are directly extracted through semantic reasoning, whereas chemical structures are resolved via a chemical-structure-grounded visual semantic reasoning (CSG-VSR) mechanism. This design is motivated by their intrinsic difference: bioactivity measurement extraction is dominated by biochemical semantic reasoning, while chemical structure extraction additionally requires the exact construction of chemically valid symbolic representations. We therefore decouple chemical semantic reasoning from strict symbolic construction and integrate MLLM reasoning with domain grounding models and chemistry tools for chemical symbolic processing. Such strategy enables the resolution of complex chemical representations, particularly enumerating complex Markush structures into specific full molecular structures—a pivotal step previously unaddressed at scale for automated bioactivity extraction (Figure 1(b)).

To provide standardized resources for evaluation and support future research, we further introduce benchmark **BioVista**, to our knowledge, the largest benchmark dedicated to protein-ligand bioactivity extraction. Meticulously curated by domain experts, BioVista comprises 16,457 bioactivity entries and 8,735 unique chemical structures extracted from 500 recent publications indexed in PDBbind v2020[16]. The data originates from diverse modalities within the source papers: text (15.8%), figures (11.6%), and tables (72.5%) (details in Figure S1), with a substantial fraction of chemical structures (48.7%) derived from challenging Markush representations. BioVista supports six distinct evaluation tasks, ranging from end-to-end bioactivity extraction to component-level Markush enumeration, allowing for comprehensive evaluation (Figure 1(c)). To enable rigorous and unbiased evaluation, BioVista is constructed with a strictly held-out test set and a separate validation set used exclusively for model development and ablation, preventing any form of test-time tuning or leakage. When evaluated on BioVista, BioMiner achieves an F1 score of 0.33 for extracting complete bioactivity triplets. In contrast, a one-shot end-to-end extraction baseline that directly processes full paper text and images attains an F1 score of only 0.00037 (Figure 2(a)), highlighting the intrinsic difficulty of the task and the necessity of principled task decomposition.

We further perform three applications to showcase BioMiner's practical utility and broad applicability. **First**, we use BioMiner to construct a bioactivity pre-training database containing 82,262 data points extracted from 11,683 papers within two days. Models pre-trained on this BioMiner-generated database demonstrate improved performance (3.9% and 3.6% improvement in RMSE metric) on two independent test sets (PDBbind v2016 core set[25] and CSAR-HiQ[26]) compared to models trained only on existing curated data (PDBbind v2016 refined set, 3,767 data points). **Second**, we implement a human-in-the-loop (HITL) bioactivity extraction workflow, where human experts collect data by reviewing the extraction result of BioMiner rather than *de novo* identification and transcription, to correct errors of fully automated extraction and ensure data quality. With HITL workflow, 1,592 bioactivity data of NLRP3[27], a high-priority target for anti-inflammatory therapies, are collected from 85 papers in 26 hours, doubling the NLRP3 data available in ChEMBL (820 data). This expanded dataset yields improved QSAR models (38.6% improvement in EF1% metric), based on which we screen molecule libraries (ChemDiv[28] and Enamine[29]) and identify 16 hit candidates with novel scaffolds. **Third**, besides bioactivity extraction, we utilize BioMiner to label complex structures with reported bioactivity data, which is important in structure-based drug design for establishing datasets like PDBbind. Using a HITL annotation workflow similar to bioactivity extraction, we label 242 structures within the famous PoseBusters database[30] at an average rate of 2 minutes/structure, achieving 97% annotation accuracy—a 5-fold speedup and a 10% accuracy improvement over purely manual annotation. These experiments, taken together, demonstrate BioMiner's efficiency, applicability, and potential to accelerate drug discovery through both fully automated and HITL workflows. In summary, BioMiner and BioVista provide a new system and evaluation standard for automated bioactivity data extraction. These contributions offer a pathway to unlock vast amounts of previously inaccessible bioactivity data, accelerating data-driven drug discovery and establishing a foundation for future progress in automated scientific literature mining.

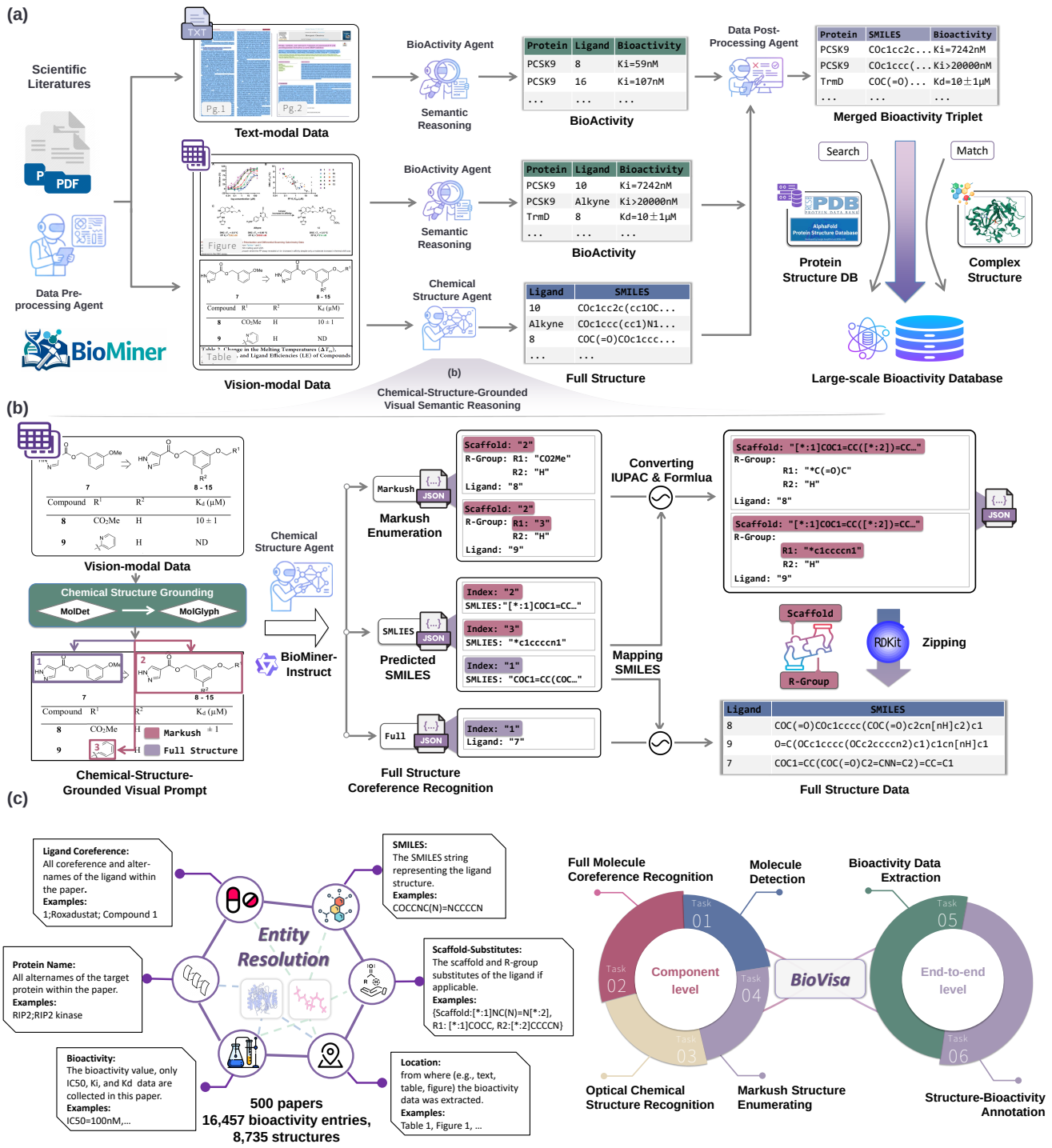

Figure 1. Overview of protein-ligand bioactivity extraction framework BIOMINER and benchmark BIOVISTA. (a) The whole protein-ligand bioactivity extraction framework BIOMINER. (b) The chemical structure extraction agent. In this agent, explicit full structures and Markush structures are both processed. For clarification, explicit full structures are plotted in green boxes and Markush scaffolds and R-group substitues are plotted in blue boxes. (c) The new benchmark BIOVISTA, containing 16,457 bioactivity data and 8,735 structures as ground-truth labels from 500 publications. Based on these data, six evaluation tasks are designed for comprehensive evaluation, ranging from end-to-end tasks to component-level tasks.

Results

To elucidate the performance and utility of our proposed framework, BioMiner, we first briefly outline its architecture, with a focus on the core chemical structure extraction agent that underpins its ability to handle complex structure data (including Markush structures). Next, we introduce BioVista, to the best of our knowledge, the largest benchmark designed to rigorously evaluate bioactivity extraction performance. We provide a detailed analysis of BioMiner's extraction performance on diverse tasks of BioVista. Finally, we demonstrate BioMiner 's practical value through three real-world applications, highlighting its versatility and impact in benefiting bioactivity data extraction and drug design.

Framework BioMiner

BioMiner, depicted in Figure 1(a), is a multi-modal agentic system specially designed for protein-ligand bioactivity extraction.

Overview of BioMiner

Rather than introducing BioMiner as a collection of agents or models, we first clarify the design principle that governs the entire system. The central challenge in automated protein–ligand bioactivity extraction lies in two fundamentally different requirements: (1) semantic reasoning over heterogeneous, multi-modal bioactivity evidence, and (2) the exact construction of chemically valid ligand structures, particularly Markush structures. End-to-end extraction approaches entangle these requirements, rendering the task structurally brittle, particularly in the presence of complex chemical representations such as Markush structures.

Instead of performing an one-shot end-to-end prediction, BioMiner explicitly separates bioactivity semantic interpretation from ligand structure construction. Specifically, BioMiner separates the task into four stages—document parsing, bioactivity measurement interpretation, chemical structure resolution, and cross-modal integration—each aligned with a distinct type of uncertainty and handled by specialized components. Following document parsing using MinerU[31], bioactivity measurement extraction and chemical structure resolution proceed in parallel. Bioactivity measurements are extracted via semantic reasoning across text, tables, and figures using BioMiner-Instruct, a domain-specialized multi-modal language model fine-tuned from Qwen3-VL-32B[32]. In contrast, ligand structures are resolved through a chemical-structure-grounded visual semantic reasoning strategy, described in detail below, which anchors MLLM reasoning to detected chemical depictions and enforces chemical validity through domain-specific models and tools. Finally, a post-processing agent integrates semantic bioactivity tuples with their resolved ligand structures to produce complete protein–ligand–bioactivity triplets.

Chemical-Structure-Grounded Visual Semantic Reasoning

Chemical structure resolution constitutes the primary structural bottleneck in automated bioactivity extraction, particularly for Markush representations that encode combinatorial chemical spaces rather than explicit molecules. To address this challenge, BioMiner introduces a chemical-structure-grounded visual semantic reasoning (CSG-VSR) mechanism, which anchors high-level MLLM reasoning to chemically grounded visual and symbolic representations.

CSG-VSR operates in three stages. First, domain-specific models detect and parse 2D chemical depictions from figures and tables. MolDetv2[33] identifies molecular structure regions, followed by optical chemical structure recognition using MolGlyph (details in Methods section) to generate SMILES representations. This stage provides a chemically grounded visual substrate that constrains subsequent reasoning. Second, an MLLM performs visual-semantic reasoning over augmented images containing indexed chemical depictions. For explicit full structures, the model resolves coreference between textual mentions and detected depictions. For Markush structures, the model identifies the Markush scaffold and semantically enumerates associated R-group definitions, which may be expressed visually, textually, or symbolically. Crucially, the MLLM is responsible only for relational and semantic reasoning (e.g., scaffold–substituent associations), not for enforcing chemical validity. Third, chemical symbolic construction is carried out deterministically using domain tools to systematically zip the recognized Markush scaffold SMILES with the enumerated R-group substituent SMILES, generating the final list of specific, full chemical structures represented by the Markush definition. Notably, R-group substituents are often described textually (e.g., IUPAC names, abbreviations, chemical formulas) rather than visually. Before the RDKit zipping step, these 1D R-group substituents are converted into SMILES, employing OPSIN[34] for IUPAC names and a Gemini-assisted, manually curated mapping table for abbreviations and chemical formulas.

By decoupling semantic reasoning from chemical symbolic construction and grounding both stages in domain-specific representations, CSG-VSR enables scalable and reliable resolution of complex Markush structures without task-specific model training. This capability is essential for automated bioactivity extraction at scale and underpins the performance gains observed in subsequent evaluations.

Benchmark BioVista and Performance Evaluation

To enable rigorous evaluation and facilitate the development of future extraction methods, we introduce BioVista, a comprehensive and challenging benchmark for protein-ligand bioactivity data extraction. To our knowledge, this is the

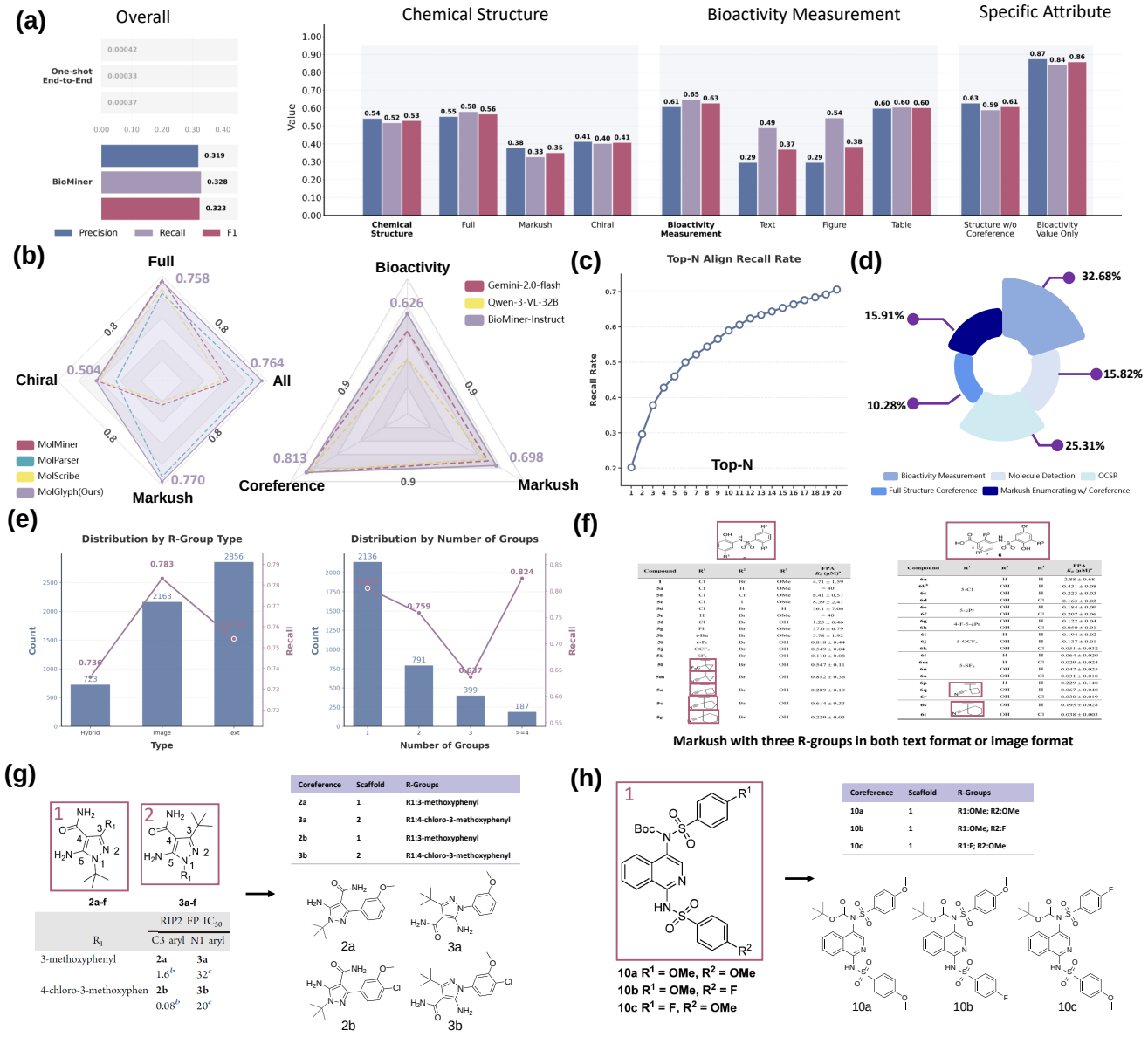

Figure 2. Benchmarking extraction performance of BIOMINER on BIOVISTA. (a) Performance of bioactivity triplet and individual attributes extraction. (b) Performance of OCSR models and MLLM models. (c) Performance of structure-bioactivity annotation task. (d) Detailed error source analysis of bioactivity triplet extraction. (e) Quantitative analysis of Markush enumeration recall under varying structural complexity and information modalities. (f) Failed example of Markush structures with three R-groups. (g,h) Two examples of Markush structures successfully processed by BIOMINER.

largest benchmark dedicated to the bioactivity extraction task. In this subsection, we first present the construction process of BioVista, and then detail BioMiner's performance.

Construction of BioVista

BioVista is derived from 500 recent publications referenced in PDBbind v2020[16], ensuring its relevance to real-world scenarios. While sourced from PDBbind, our benchmark represent a diverse and challenging distribution of scientific literature. Analysis shows that these 500 publications cover 102 distinct journals, exhibiting a "long-tail" distribution unlike standard databases (e.g., ChEMBL) which are heavily concentrated in a few top-tier journals (Figure S2). This ensures the model is evaluated on a wide variety of layout styles and textual reporting habits. Unlike PDBbind data curation, BioVista includes all bioactivity data reported within these publications, not just the reported bioactivity to the PDB structure. Domain experts manually (details about experts in Table S3) curate all bioactivity data points, annotating attributes beyond the core triplet (protein, ligand SMILES, bioactivity value) to enable fine-grained analysis (Figure 1(c)). These attributes (e.g., location, ligand coreference, scaffold-substitutes) include information such as the source modality to analyze the ability to process different modal data. Notably, all alternative names (alternames) for proteins and ligands are carefully collected, ensuring completeness and preventing ambiguity or overlap.

To ensure high data quality and completeness, we employed a model-assisted verification strategy aimed at identifying potential omissions (false negatives) in the initial human annotations. Specifically, the BioMiner was utilized to scan the curated publications, and its extractions were cross-referenced with human annotations. Any discrepancies or newly identified data points were rigorously reviewed by domain experts. Only data points with explicit location verified by humans were added to the ground truth, ensuring the dataset remains objective and free from model-induced hallucinations. After annotation, we follow[35] to conduct an Inter-Annotator Agreement study on a random 10% subset of 50 papers to analysis the quality of BioVista. Such analysis yields a F1 coefficient of 0.899, indicating high consistency and reliability. The finalized BioVista dataset comprises 16,457 bioactivity data points source from text (15.8%), figures (11.6%), and tables (72.5%) (Figure S1), along with 8,735 chemical structures (of which 48.7% are derived from Markush structures). To strictly prevent "tuning on the test set" and ensure unbiased evaluation, the BioVista dataset (500 papers) was randomly partitioned into a Validation Set (50 papers, 10%) and a held-out Test Set (450 papers, 90%). All prompt engineering, hyperparameter tuning, and rule refinement described in this study were conducted exclusively based on the analysis of the Validation Set. The Test Set remained unseen during the development phase and was used solely for the final performance evaluation.

BioVista defines two end-to-end tasks and four component-level tasks to thoroughly assess the extraction performance of BioMiner. The two end-to-end tasks evaluate overall extraction capability, specifically: (1) extracting all bioactivity data reported in a publication, and (2) annotating PDB structures with bioactivity information presented in associated papers. These tasks directly reflect the practical utility of bioactivity extraction methods. Additionally, four specialized component-level tasks—molecule detection, optical chemical structure recognition (OCSR), full-structure coreference resolution, and Markush structure enumeration—provide deeper insights into the performance of critical chemical structure extraction processes, facilitating method development and optimization. Detailed descriptions of these evaluation tasks can be found in Supporting Information Note 1. Altogether, this suite provides a robust foundation for evaluating and advancing bioactivity extraction systems. Comprehensive evaluation across the six tasks defined within the BioVista benchmark rigorously assesses BioMiner's capabilities and establishes a performance baseline for further works. Performance metrics, error analysis, cost-efficiency analysis, and case study are all included in the following part.

End-to-end Evaluation

End-to-end bioactivity extraction performance is shown in Figure 2(a), where BioMiner achieves Precision=0.319, Recall=0.328, and F1=0.323 for the complex bioactivity triplet. In addition to the evaluation of whole bioactivity triplet, which integrates multi-modal attributes, analysis of individual attributes reveals stronger performance. Chemical structure extraction (ligand coreference-SMILES) achieves F1=0.528, with performance varying by complexity (Explicit Full F1=0.565 vs. Markush F1=0.349). Bioactivity measurement extraction (protein-ligand coreference-bioactivity value) achieved F1=0.626, with table-based extraction being the most effective (Table F1=0.600 vs. Figure F1=0.382, Text F1=0.368). Extraction of isolated attributes is even more proficient, with solely bioactivity value F1=0.857 and SMILES F1=0.606. These scores underscore the significant challenge of bioactivity extraction and indicate component strengths of BioMiner despite integration challenges.

Performance on the other end-to-end task structure-bioactivity annotation is presented in Figure 2(b). Given a complex structure and the associated paper, BioMiner extracts all bioactivity data from the paper, and ranks them based on ligand similarity between extracted data and given structure. When taking Top-10 bioactivity data into consideration, BioMiner achieves a recall rate of 0.598 for the reported bioactivity. This indicates that while the Top-ranked match may not always be correct, the reported value of the given structure is often among the top candidates, supporting efficient structure-bioactivity annotation or human validation.

Table 1. Performance of BioMiner with different MLLMs on BioVista. This table focuses on presenting MLLMs-related results and comparing the performance of different MLLMs. Our BioMiner-Instruct outperform other models significantly based on this table results. The best results are bolded, and the second best results are underlined.

Tasks	Metric	Gemini-2.0-flash	GPT-4o-mini	GPT-4.1-mini	Grok-4-fast	Claude-haiku-4-5	Qwen3-VL-32B	Qwen3-VL-235B	BioMiner-Instruct
Component-level evaluation for chemical structure extraction									
Full Structure Coreference	Recall	0.786	0.552	0.781	0.559	0.722	0.766	0.785	**0.815**
	Precision	0.783	0.550	0.777	0.558	0.719	0.762	0.782	**0.812**
	F1	0.784	0.551	0.779	0.558	0.720	0.764	0.783	**0.813**
Markush Enum. w/ Coreference	Recall	0.579	0.039	0.219	0.168	0.309	0.375	0.611	**0.709**
	Precision	0.545	0.041	0.220	0.162	0.289	0.369	0.576	**0.687**
	F1	0.561	0.040	0.219	0.165	0.299	0.372	0.593	**0.698**
Markush Enum. w/o Coreference	Recall	0.628	0.050	0.255	0.223	0.380	0.422	0.655	**0.743**
	Precision	0.594	0.058	0.287	0.212	0.363	0.422	0.618	**0.719**
	F1	0.611	0.054	0.270	0.217	0.371	0.422	0.636	**0.731**
Different attributes evaluation for end-to-end extraction and annotation									
Structure w/ Coreference	Recall	0.482	0.310	0.413	0.413	0.269	0.406	0.462	**0.516**
	Precision	0.520	0.360	0.482	0.468	0.312	0.489	0.498	**0.541**
	F1	0.500	0.333	0.445	0.439	0.289	0.444	0.479	**0.528**
Bioactivity Measurement	Recall	0.583	0.251	0.363	0.513	0.400	0.628	0.611	**0.647**
	Precision	0.522	0.236	0.422	0.368	0.255	0.501	0.423	**0.606**
	F1	0.551	0.243	0.390	0.429	0.311	0.557	0.500	**0.626**
Bioactivity Triplet	Recall	0.284	0.089	0.145	0.225	0.112	0.253	0.271	**0.328**
	Precision	0.260	0.082	0.171	0.174	0.073	0.202	0.199	**0.318**
	F1	0.272	0.085	0.157	0.196	0.088	0.224	0.230	**0.323**
Bioactivity-Structure Annotation	Top-1	0.198	0.218	0.192	0.198	0.214	0.198	**0.238**	0.202
	Top-3	0.362	**0.458**	0.400	0.372	0.400	0.376	0.444	0.378
	Top-10	0.568	**0.652**	0.598	0.570	0.626	0.586	**0.652**	0.598

Component-level Evaluation

Chemical structure extraction, which is highly related to final extraction performance and the core section of BioMiner, is particularly evaluated on four specially designed component-level tasks. As introduced above, BioMiner resolves ligand structures in three stages. These four component-level tasks respectively evaluate the first stage molecule detection (Table S6), OCSR (Table S7), and the second stage full structure coreference recognition, Markush enumeration (Table 1).

For first stage molecule detection, BioMiner employs MolDetv2, achieving mAP=0.747, AP_{50}=0.922, and AP-Small=0.050 (Table S6). As for OCSR, compared to other DSMs, the employed MolGlyph exhibits significantly better ability to recognize Markush structure (Accuracy=0.770 for Markush structures), and thus leading to best OCSR performance (Accuracy=0.764 for all structures) (Table S7). When it comes to second stage tasks, BioMiner achieves an F1 score of 0.814 and 0.698 on full structure coreference recognition and Markush enumeration (with coreference) using BioMiner-Instruct (Table 1), respectively. If we only consider the structure enumeration result and ignore coreference, F1 score of Markush enumeration rises to 0.731. This result indicates the ability of BioMiner to process the challenging Markush structures. Different MLLMs, including Gemini, GPT[36], Claude[37], Qwen[38], and Grok[39], have been tested on these two tasks, and BioMiner-Instruct achieves the best performance in both price and accuracy. These component-level evaluations provide a detailed analysis of extraction results and effectively benefit the design of chemical structure extraction agent.

Error Analysis and Cases Study

Based on the component-level evaluation results, we plot the error contribution analysis for the end-to-end triplet extraction in Figure 2(d). Errors in bioactivity measurement extraction contribute the largest share (32.68%), followed by OCSR inaccuracies (25.31%), and the Markush structure enumeration (15.91%). Molecule detection (15.82%) and explicit full structures coreference recognition (10.28%) contribute less to the overall error. This breakdown provides critical insights for directing future research toward optimizing the system components.

We observe that, in addition to bioactivity extraction, errors arising from OCSR and Markush enumeration constitute the most substantial portions of the overall error budget. Accordingly, we conduct a more detailed analysis of these two components in Figure 2(b) and Figure 2(e). For OCSR, the results indicate that, beyond Markush structures, chirality recognition remains a major challenge. Even for our best-performing OCSR model, the accuracy on chiral structures is limited to 0.504, highlighting the intrinsic difficulty of precise stereochemical interpretation from heterogeneous visual inputs. For Markush enumeration, we find that the recall is lowest (0.736) when R-groups are jointly specified in both textual and graphical forms, suggesting that

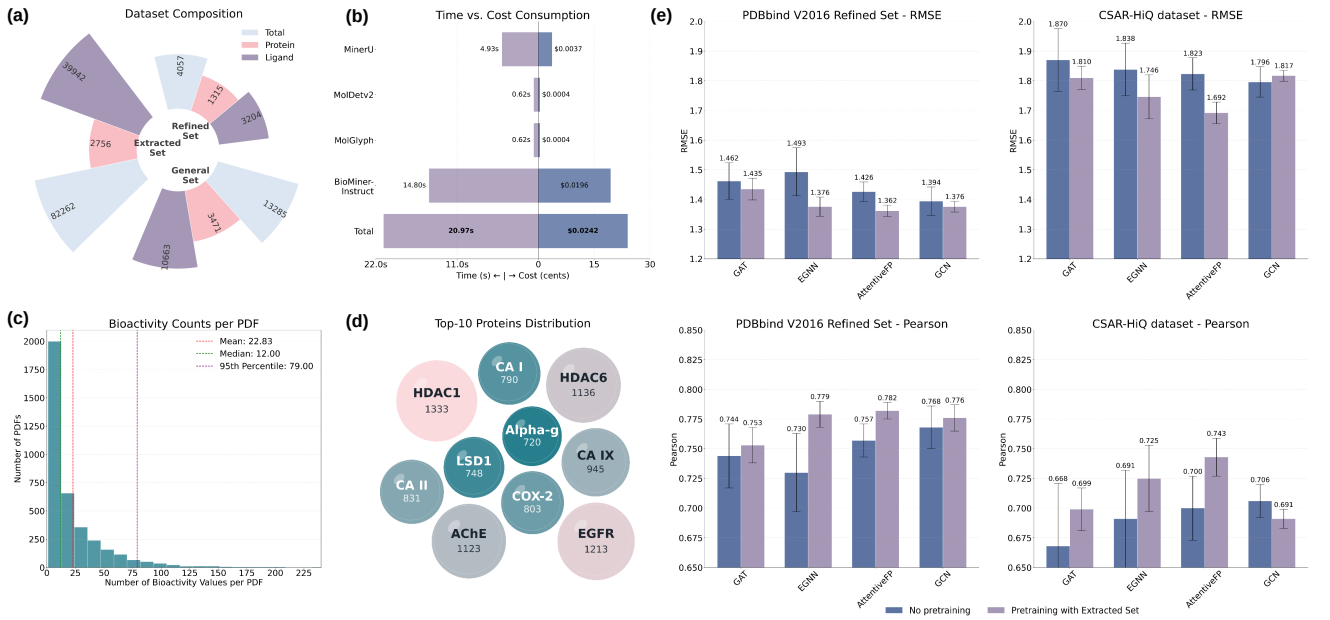

Figure 3. Using BioMiner collecting large-scale bioactivity data for deep learning model training. (a) Statistical comparison between manually curated PDBbind v2016 dataset and our extracted dataset; (b) Consuming time and money for BioMiner; (c) Number of extracted bioactivity data in each paper; (d) Protein names distribution within extracted bioactivity data. (e) Performance comparison of models with and without pre-training on BioMiner extracted set.

cross-modal alignment between textual R-group descriptions and visual depictions is particularly error-prone. Moreover, we observe a pronounced decline in recall when the number of R-groups increases to three, reflecting the combinatorial complexity introduced by multi-substituent enumeration. These analyses provide concrete guidance for future improvements, indicating that targeted advances in stereochemical OCSR and cross-modal Markush enumeration are likely to yield the greatest gains in end-to-end extraction performance.

Qualitative evidence of BioMiner's advanced capabilities is shown in Figure 2(e, f), which displays examples of complex Markush structures successfully processed by the system. This includes correct scaffold identification, R-group enumeration, 1D R-group processing (IUPAC and chemical formula), and accurate generation of the final enumerated SMILES strings, indicating its ability to process challenging Markush structures.

In summary, the benchmarking results on BioVista validate BioMiner's capacity for automated, multi-modal bioactivity data extraction. While quantifying the performance across various dimensions and pinpointing key challenges for future work, the results demonstrate certain bioactivity extraction ability and substantial efficiency advantages over manual curation. This positions BioMiner as a valuable tool for accelerating data acquisition and unlocking previously inaccessible data in the vast body of drug discovery literature.

Automated Collection of Large-scale Bioactivity Data for Model Training

A key objective in developing BioMiner is to overcome the limitations of existing, manually curated bioactivity databases, and to unlock the wealth of previously inaccessible data hidden within scientific literature. To demonstrate BioMiner's capacity for large-scale data acquisition, we apply it to construct a bioactivity database. Deep learning-based bioactivity prediction models are further pre-trained on this extracted database to demonstrate its utility.

To construct a large-scale dataset, we target the European Journal of Medicinal Chemistry (EJMC), a high-impact journal known for its density of protein-ligand bioactivity data, making it an ideal source. 11,683 articles published in EJMC since 2010 are collected, excluding earlier articles due to potential low resolution. Employing BioMiner, we process the 11,683 papers within just three days (cost about 21 second and 0.024$ per paper, as shown in Figure 3(b)), a speed that would be practically impossible with manual curation (details in Table S4). From these paper, 22,6076 bioactivity triplets are extracted. To support the training of bioactivity prediction models, we further enrich these data with protein structure information. As discussed above, using the extracted protein name, BioMiner searches external structure databases (including AlphaFoldDB and PDB) for protein structure information. 82,262 data points are successfully enriched with protein structure information, providing a large-scale extracted database significantly expands the available data compared to PDBbind's refined and general sets (Figure 3(a)). After excluding paper contains no bioactivity data, analysis of the extracted dataset reveals a mean of 22.83

bioactivity values per paper, with a median of 12 and a 95th percentile of 79, highlighting the density of information contained within individual publications (Figure 3(c)). Further analysis of the extracted dataset shows the Top-10 protein distributions, indicating the most researched proteins and their potential application in bioactivity prediction (Figure 3(d)).

Critically, we evaluate the impact of this automatically extracted database on the performance of downstream deep learning models. While the end-to-end extraction precision is approximately 0.32, indicating the presence of noise (e.g., incomplete experimental conditions), deep learning models have shown remarkable robustness to massive label noise when trained on sufficiently large datasets[40]. We hypothesized that the scale of our extracted data allows the model to "average out" the noise and capture generalized interaction features. We pre-train several graph neural network (GNN) architectures (GAT[41], EGNN[42], AttentiveFP[43], and GCN[44]) on the database constructed by BioMiner and compared their performance against models trained solely on the PDBbind v2016 refined set (3,767 data points), across two independent test sets: PDBbind v2016 core set and CSAR-HiQ set. Models pre-trained on the BioMiner-derived database consistently achieved performance improvements, as measured by both Root Mean Squared Error (RMSE) and Pearson correlation coefficient. Average improvements of 3.9% and 3.4% in the RMSE metric are acheived in the PDBbind core set and CSAR-HiQ set, respectively.

In addition, to rigorously distinguish the value of the mined bioactivity signals from the benefit of simply scaling up structural data (unsupervised representation learning), we conducted additional comparisons against two baselines: (1) an unsupervised pre-training strategy that masks atoms/residues without using bioactivity labels, and (2) a negative control using shuffled bioactivity labels. As shown in Figure S7, the downstream performance follows a clear hierarchy: Shuffled « No Pre-train < Unsupervised < Ours (BioMiner). Our method (3.9% improvement) outperforms the unsupervised strategy (1.4% improvement), confirming that the mined bioactivity values—despite their inherent noise—provide unique, critical supervision signals regarding binding physics that cannot be learned from structural data alone.

These results underscore the value of automatically extracted bioactivity data for enhancing the accuracy and predictive power of computational models in drug discovery. BioMiner can serve as an efficient and cost-effective method for mining scientific literature, which is crucial for data-driven AI-powered drug design. Details about model training can be found in Supporting Information.

Human-in-the-loop Curation of NLRP3 Bioactivity Data and Enhanced Inhibitor Screening

While fully automated extraction using BioMiner demonstrates high throughput, potential data errors will be introduced. In practical scenario, correcting potential errors and ensuring high data quality is crucial for sensitive downstream applications like quantitative structure-activity relationship (QSAR) model building and drug screening. To address this, in this subsection, we introduce a human-in-the-loop (HITL) workflow to enable high-quality and efficient data collection aided by BioMiner and validate its effectiveness in a practical task.

Bioactivity Data Collection HITL Workflow

In the HITL workflow, human experts extract bioactivity data by reviewing and validating the chemical structure and bioactivity measurement extraction results of BioMiner, which consists the final bioactivity triplet. For chemical structures, as shown in Figure S3, S4 and S5, human expert verifies the first stage molecule detection, OCSR, and the second stage full structure coreference recognition, Markush enumeration, sequentially. For bioactivity measurement, the expert directly confirms the correctness of the extracted protein-ligand coreference-bioactivity value tuples, the output of the bioactivity measurement extraction agent. Critically, human intervention is targeted only at these verification steps. Other components within BioMiner, such as 1D R-group processing and Markush scaffold-R-group zipping, are rule-based steps that do not require manual review. Overall, this HITL workflow omits the *de novo* identification and transcription of chemical structure and bioactivity, and experts primarily focus on verifying BioMiner's outputs, significantly accelerating curation compared to fully manual extraction.

NLRP3 Inflammasome Bioactivity Data Collection

To showcase this workflow's practical utility, we apply it to curate NLRP3[27] inflammasome bioactivity data. NLRP3 is a high-priority target for anti-inflammatory therapies, yet publicly available bioactivity data, such as in ChEMBL, remains relatively sparse. Our objective was to expand the size of high-quality NLRP3 bioactivity dataset and utilize this enriched data to develop more accurate models for inhibitor discovery.

We identify 85 relevant scientific publications reporting NLRP3 bioactivity data through literature searches and then collect 1,592 data from these publications with this HITL workflow. As shown in Figure 4(a), the time required per paper varied, correlating with the number of chemical structures and bioactivity data points present. On average, it takes approximately 18.4 minutes for each paper and consumes 26 hours for all 85 papers. This effort effectively doubles the amount of NLRP3 bioactivity data previously available in ChEMBL (Figure 4(b)), providing a substantially larger and chemically diverse dataset. The expert oversight inherent in the HITL process also allows for meticulous handling of ambiguities and edge cases, further ensuring data quality.

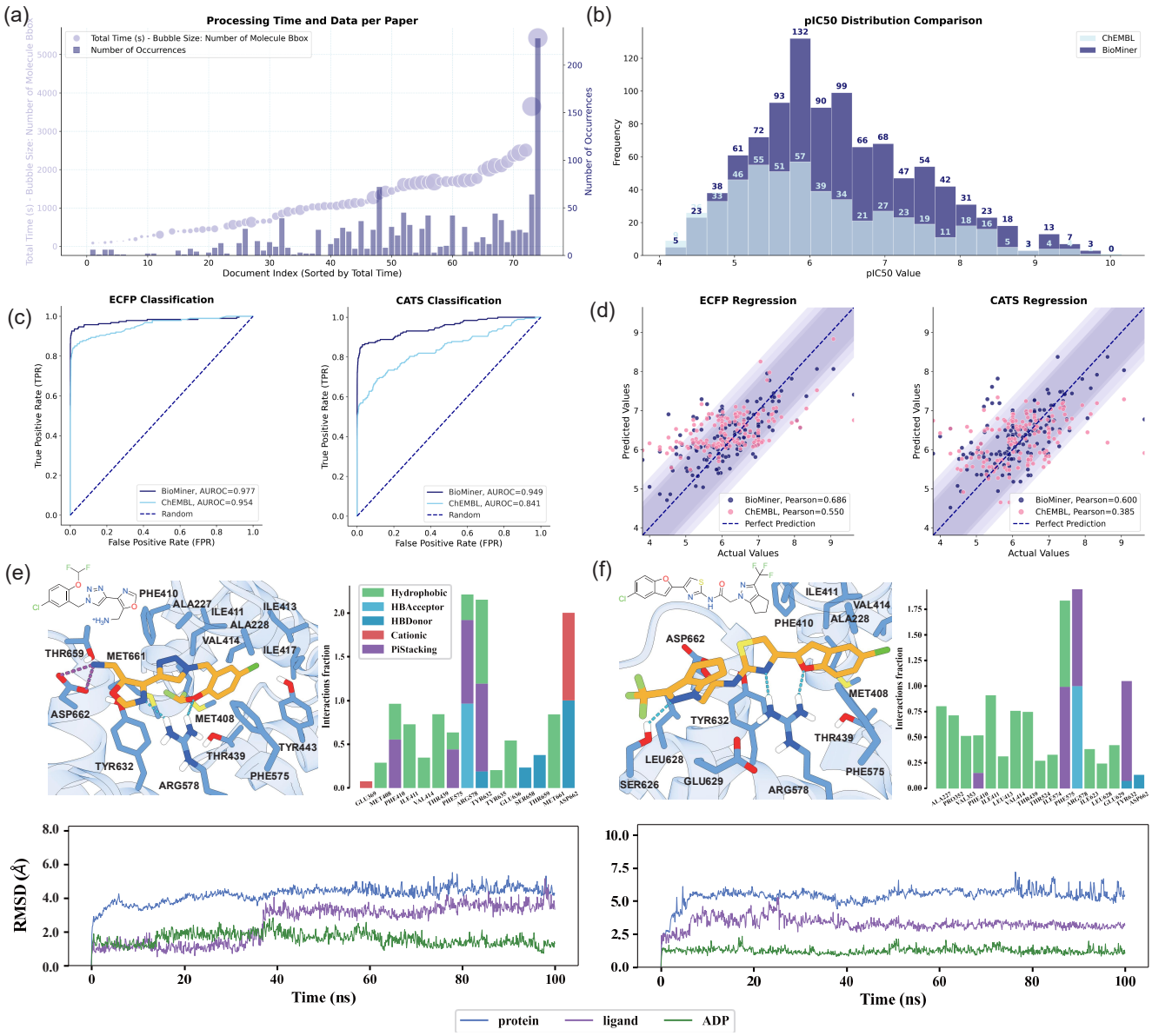

Figure 4. Using BioMiner for human-in-the-loop NLRP3 bioactivity data collection from 85 papers and inhibitor screening. (a) Consuming time (about 18.4 minutes in average) and number of collected data for each paper. The final consuming time is highly related with the number of bioactivity data and chemical structures; (b) Comparison of pIC50 distribution between BioMiner collected bioactivity data and ChEMBL data; Classification (c) and regression (d) performance comparison between QSAR models trained on BioMiner data and ChEMBL data; The Glide-XP docked binding pose within the NP3-253-binding pocket (PDBID: 9GU4), stacked bar plot of the interaction fraction during MD simulation and RMSD curve of protein, ligand and ADP for Z6739936901 (e) and Z5232931194 (f), respectively.

QSAR Models Training

To assess the impact of this data expansion, we train quantitative structure-activity relationship (QSAR) models for predicting NLRP3 inhibition using both the original ChEMBL NLRP3 dataset and our expanded, BioMiner-curated NLRP3 dataset, respectively. As shown in Table S8 and Table S9, various QSAR models with different algorithms (e.g., Random Forest[45], SVM[46], etc.), different tasks (regression, classification), and different molecular representations (ECFP[47], CATS[48]) are evaluated comprehensively. Models trained on the BioMiner-curated dataset consistently outperformed those trained on the standard ChEMBL data across all settings evaluated. On average, considering the EF1% metric across 28 distinct QSAR model configurations, the BioMiner-curated dataset yielded a 38.6% performance improvement. Figure 4(c) and (d) visualize the performance of the top-performing models under various settings. Notably, for classification using ECFP fingerprints, the AUROC improved from 0.954 (ChEMBL) to 0.977 (BioMiner-curated). Similarly, for CATS-based regression, the Pearson correlation coefficient increased substantially from 0.385 to 0.600. This marked enhancement in predictive performance significantly increases the reliability of bioactivity prediction for novel chemical entities, thereby improving prospects for effective inhibitor screening.

NLRP3 Inhibitor Screening

Leveraging the superior QSAR models trained on the BioMiner-curated data, we initiate a virtual screening against chemical libraries (i.e., ChemDiv and Enamine). Through our established rational screening pipeline (Figure S10 and see Supplementary Information for details), sixteen compounds are manually selected as potential hit candidates, exhibiting both high structural diversity and novelty (Table S10). Further MM/PBSA binding free energy calculation reveals six compounds exhibit comparable or better binding free energies than both MCC950[49] and NP3-562[50], suggesting their potential as highly potent NLRP3 inhibitors (Table S10). Docking simulation demonstrates two promising candidates, Z6739936901 and Z5232931194, showing considerable binding mode between the NBD, HD1, WHD, and HD2 domains (Figure 4(e,f)). MD simulation of 100 ns is performed to investigate the binding stability of identified compounds. The RMSD of protein and ligand shows that Z6739936901 and Z5232931194 bind stably with the binding pocket in the last 60 ns. Z6739936901 maintains persistent hydrogen bonding, cation-π and π-π interactions with ARG578 and TYR632. Its solvent-exposed charged amino group forms strong salt bridges and hydrogen bonds with ASP662, anchoring the oxazole moiety rigidly. Meanwhile, the phenyl fragment remains tightly bound within a large hydrophobic subpocket. Z5232931194 frequently forms hydrogen bonds with ARG578, PHE575, and TYR632, along with cation-π and π-π interactions. The bromine-substituted benzofuran moiety remains tightly bound within the hydrophobic subpocket. Notably, the ADP bound in the substrate pocket remains stable throughout MD simulations, with low RMSD values (<1.5 Å) confirming its rigid positioning with minimal conformational changes. This observation suggests that both Z6739936901 and Z5232931194 effectively stabilize the inactive conformation of the NBD, HD1, WHD, and HD2 domains. Collectively, our BioMiner-curated data-augmented QSAR models identifies several structurally novel hit candidates exhibiting high potential as potent NLRP3 inhibitors. These candidates represent promising chemotypes for future experimental validation, potentially unlocking new avenues in NLRP3-targeted drug discovery.

In summary, this study demonstrates the effectiveness of a BioMiner-powered HITL workflow for rapidly constructing large, high-quality, target-specific bioactivity datasets. The efficiency gain (approx. 18.4 min/paper) over manual curation, combined with the significant improvement in downstream QSAR model performance, underscores the potential of this approach to accelerate data-driven drug discovery efforts.

Enhancing Structure-Bioactivity Annotation for PoseBusters

Besides bioactivity data extraction, annotating experimentally determined protein-ligand complex structures with their reported bioactivity measurements in the literature is also important in structure-based drug design field for establishing datasets like PDBbind. In the experiments on the BioVista, BioMiner has demonstrated its capability to perform this task. The reported bioactivity measurement of a given structure is often found among the top candidates of the extracted bioactivity data (Top-10 Recall=0.686). In this subsection, we further demonstrate BioMiner's utility in a practical annotation scenario, proposing a HITL annotation workflow and presenting a detailed analysis for fully automated data annotation beyond Top-N Recall.

Structure-Bioactivity Annotation HITL Workflow

In the annotation HITL workflow, given a PDB complex structure and associated publication, BioMiner first extracts all potential protein-ligand-bioactivity triplets and ranks extracted data based on ligand similarity between extracted SEMILS and PDB ligand SMILES. Then, a human expert rapidly verifies the bioactivity data in the ranked list one by one. Compared with purely human annotation, such a HITL streamlines the expert's task to validation rather than *de novo* annotation.

PoseBusters Structures-Bioactivity Annotation

The famous PoseBusters benchmark[30], which contains 308 high-quality complex structures, is selected for HITL annotation. After careful preparation of the input, 242 PDB structures and associated publications are collected. To highlight our workflow's ability, the annotation results between a human expert using the HITL workflow and the other human expert without it are

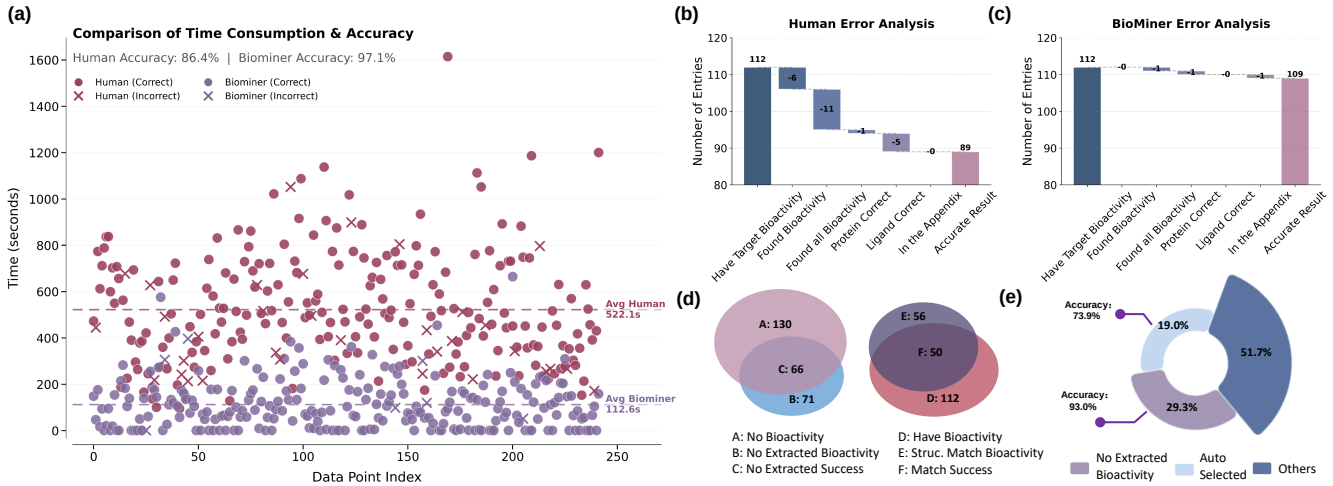

Figure 5. Annotating protein-ligand complex structures in PoseBusters with corresponding experimental bioactivity values. (a) Comparison of consuming time and annotation accuracy between human and BIOMINER; (b) Waterfall plot of error analysis of human extraction; (c) Waterfall plot of error analysis of BIOMINER extraction; (d) Number analysis of different data of BIOMINER results; (e) Proportional analysis of different data of BIOMINER results.

compared. As shown in Figure 5(a), this BIOMINER-assisted HITL workflow demonstrates significant efficiency gains, requiring an average of only about 2 minutes per publication for verification and annotation. The result represents an estimated 5-fold acceleration compared to purely manual curation efforts. Crucially, the accuracy of associating the correct bioactivity value with the specific complex structure improves substantially to 97.1%, compared to an estimated 86.4% accuracy baseline for manual annotation alone. An analysis of annotation errors (Figure 5(b,c)) reveals that manual annotation is prone to overlooking the correct bioactivity data (failed to find or cannot find all correct data), a category of error effectively mitigated by BIOMINER's systematic extraction and presentation of candidates. The improved accuracy of the HITL approach can thus be attributed to BIOMINER's comprehensive extraction combined with expert validation of a focused, ranked list, leading to a more robust and efficient annotation process.

Fully Automated Annotation

Based on the HITL annotation result, we further analyze the performance of the fully automated annotation beyond the Top-N Recall (Figure 5(d,e)). For 71 of 242 complexes, BIOMINER finds no bioactivity data in the publications. Among these complexes, 66 complexes are correct (consistent with manual verification confirming absence). For the remaining 173 complexes where bioactivity data are potentially present, BIOMINER successfully extracts candidate bioactivity measurements. Within this group, 56 complexes have at least one extracted bioactivity measurement associated with a ligand structure perfectly matching the PDB ligand (chemical similarity = 1.0). For these 56 matches, the data postprocessing agent (utilizing the Gemini MLLM) selects the most probable bioactivity value based on contextual consistency with the protein name and PDB structure title. This automated selection process identifies a candidate value in 46 cases. Manual validation confirms that 34 of these 46 automatically selected annotations are correct, yielding an accuracy of 73.9% for this specific subset. Combining the 69 cases correctly identified as having no data with the 46 cases where a candidate value is proposed, BIOMINER demonstrates reliable automated handling for 115 out of 242 structures (47.5%). This suggests that nearly half of the annotation task for this dataset could potentially be automated with high confidence (Accuracy=0.887), significantly reducing manual effort, while the remaining cases benefit from the efficient HITL verification.

The ability to rapidly generate and verify high-fidelity structure-activity annotations for large structural datasets like PoseBusters is highly valuable for the development of structure-based drug design algorithms. By significantly accelerating the creation and curation of these critical datasets with enhanced accuracy, BIOMINER directly facilitates progress in structure-guided drug discovery and the application of machine learning in structural biology.

Discussion

In this work, we introduced BIOMINER, a multi-modal system for automated extraction of protein–ligand bioactivity data from scientific literature, built upon a principled integration of multi-modal large language models (MLLMs), domain-specific models, and chemistry-aware tools. To support rigorous evaluation and foster systematic progress in this area, we further

established BioVista, the largest benchmark to date dedicated to protein–ligand bioactivity extraction, comprising 16,457 bioactivity entries and 8,735 chemical structures across six carefully designed evaluation tasks.

Evaluated on BioVista, BioMiner demonstrated meaningful extraction capability under this highly challenging setting, achieving F1 scores of 0.32 for complete bioactivity triplets, 0.59 for chemical structure extraction, and 0.61 for bioactivity measurement extraction. Beyond benchmark performance, the practical utility of BioMiner was substantiated through three real-world applications. First, its scalability enabled the rapid construction of a large-scale bioactivity database (82,262 entries from 11,683 papers within two days), which yielded consistent improvements in downstream binding affinity prediction models, including a 3.8% reduction in RMSE after pre-training (Figure 3). Second, human-in-the-loop (HITL) workflows highlighted the complementary strengths of automated extraction and expert validation: for the therapeutically important target NLRP3, HITL-assisted extraction doubled the volume of high-quality bioactivity data relative to ChEMBL within 26 hours, leading to substantially improved QSAR performance (38.6% gain in EF1%) and the identification of 16 novel hit candidates (Figure 4). Third, BioMiner proved effective for accelerating structure-centric annotation tasks, achieving a 5-fold speedup and higher accuracy (97.1% versus 86.4%) in labeling bioactivity-linked structures within the PoseBusters benchmark (Figure 5).

Despite these advances, the modest end-to-end F1 score for complete triplet extraction underscores the intrinsic difficulty of the task. Protein–ligand bioactivity extraction requires precise cross-modal integration, reliable chemical structure reconstruction (particularly for Markush representations), and accurate association of entities and measurements scattered across heterogeneous document elements. Our error analysis (Figure 2(d)) identifies several dominant bottlenecks, including bioactivity measurement extraction (32.68% of errors), optical chemical structure recognition (25.31%), and Markush structure enumeration (19.0%). These findings indicate that performance limitations primarily stem from fundamental task complexity rather than superficial system design choices.

Looking forward, the demonstrated ability of BioMiner to handle complex chemical information suggests strong potential for generalization. The system's core capability in chemical structure processing can be naturally extended to related extraction tasks in cheminformatics and bioinformatics, such as mining molecular ADMET properties. In such structure-centric scenarios, the chemical structure extractor agent in BioMiner serves as a general-purpose component. Adapting the framework to new tasks typically requires only task-specific prompt design for data extraction, which can be guided by domain expertise or, in some cases, directly supported by large language models.

In summary, BioMiner represents a step toward scalable, structure-aware automation of bioactivity data extraction, while BioVista establishes a much-needed standard for systematic evaluation in this domain. Together, they provide a foundation for unlocking large volumes of previously inaccessible bioactivity data embedded in scientific literature, enabling more comprehensive data-driven modeling and accelerating future advances in drug discovery and automated scientific knowledge extraction.

Methods

Preliminary

For notation, a scientific article is represented as a multi-modal document D (PDF), containing text, tables, and figures. The goal is to extract a set of protein–ligand bioactivity records $\mathscr{B} = \{(p_i, \ell_i, v_i)\}_{i=1}^{N}$, where p_i is a protein target mention, ℓ_i is a chemically valid ligand representation (SMILES), and v_i is a quantitative bioactivity value with its associated assay context.

A key challenge is that bioactivity mining from literature *structurally couples*: (i) semantic reasoning over heterogeneous modalities (to interpret bioactivity measurements and entity relations) and (ii) *exact* symbolic construction of ligand structures (especially for Markush representations). End-to-end generation is brittle.

BIOMINER addresses this by decoupling:

$$\underbrace{\mathscr{M}(D) \to \{(p,c,v)\}}_{\text{bioactivity measurement: semantic reasoning}} \quad \text{and} \quad \underbrace{\mathscr{S}(D) \to \{(c,\ell)\}}_{\text{chemical structure: grounded + constrained construction}}, \tag{1}$$

and then integrating them via ligand coreference c:

$$\mathscr{B} = \mathscr{J}\left(\{(p,c,v)\}, \{(c,\ell)\}\right). \tag{2}$$

Here c denotes a ligand coreference string/identifier (e.g., compound name, label, or alias) used as the join key across modalities.

System overview

BIOMINER is implemented as four specialized agents: data preprocessing A^{pre}, chemical structure extraction A^{str}, bioactivity measurement extraction A^{mea}, and postprocessing/integration A^{post}. At deployment, BIOMINER runs inference-only: it orchestrates fixed-weight models and deterministic chemistry tools without document-specific training.

$$(D^{txt}, D^{vis}) = A^{pre}(D), \quad \mathscr{S} = \{(c,\ell)\} = A^{str}(D^{vis}), \quad \mathscr{M} = \{(p,c,v)\} = A^{mea}(D^{txt}, D^{vis}), \quad \mathscr{B} = A^{post}(\mathscr{S}, \mathscr{M}). \tag{3}$$

Backbone models and tools. BIOMINER uses a locally deployable multi-modal language model BIOMINER-INSTRUCT (fine-tuned from Qwen3-VL-32B) for cross-modal reasoning, a domain-specific OCSR model MOLGLYPH for chemical structure recognition in scientific figures, and domain tools including RDKit for chemical graph operations and OPSIN for IUPAC-to-SMILES conversion. (Implementation details and prompts are provided in Supporting Information.)

Agent 1: data preprocessing (A^{pre})

The preprocessing agent parses D into aligned textual and visual representations while preserving layout and reading order. We apply MinerU to obtain text and layout elements:

$$(D^{txt}, D^{lay}) = \texttt{MinerU}(D), \tag{4}$$

where D^{txt} contains extracted text segments and D^{lay} stores page-level layout elements with their bounding boxes and categories (e.g., paragraph, table, figure, caption).

For visual grounding, each page is rasterized into an image and we additionally extract figure/table crops when possible:

$$D^{vis} = \left(D^{vis}_{page}, D^{vis}_{seg}\right). \tag{5}$$

D^{vis}_{page} contains full-page images, which are robust to imperfect caption detection and support molecule detection at page scale. D^{vis}_{seg} contains segmented figure/table regions (with captions when detected) to reduce surrounding noise for downstream visual reasoning.

Agent 2: chemical structure extraction (A^{str})

A^{str} produces a set of ligand coreferences and their chemically valid SMILES:

$$\mathscr{S} = \{(c_k, \ell_k)\}_{k=1}^{K} = A^{str}(D^{vis}). \tag{6}$$

Chemical-Structure-Grounded Visual-Semantic Reasoning (CSG-VSR)

Chemical structures in literature appear as: (i) explicit full structures, and (ii) Markush structures that define families of compounds via variable substituents (R-groups). Our scope focuses on Markush instances where the scaffold and attachment points are representable in SMILES; other Markush types (e.g., variable positions/repeats not representable in SMILES) are handled as out-of-scope and reported separately.

CSG-VSR resolves chemical structures through a three-stage grounded pipeline:

$$\text{(detect + recognize)} \to \text{(grounded relation inference)} \to \text{(tool-constrained symbolic construction)}. \tag{7}$$

Stage I: molecule detection and OCSR. We detect 2D chemical depictions in D^{vis}_{page} and D^{vis}_{seg} and recognize each depiction into a preliminary SMILES string. Let $\{(x_m, \tilde{\ell}_m)\}_{m=1}^{M}$ denote detected bounding boxes x_m with OCSR outputs $\tilde{\ell}_m$. Depictions with attachment points are tagged as Markush components; otherwise they are treated as explicit full structures.

Stage II: grounded relation inference (coreference + Markush enumeration). To reliably refer to visual structures, we create an augmented image by overlaying bounding boxes and unique indices on each detected depiction. Given an augmented image I^{aug}, BIOMINER-INSTRUCT infers:

- **Coreference for explicit structures:** mapping from visual indices to ligand identifiers c (e.g., "compound 12", "AZD1234", "R1").
- **Markush enumeration:** for each Markush scaffold, identify the scaffold index and enumerate substituent sets (R-groups), where substituents can be (a) visual indices of depicted fragments or (b) textual 1D mentions (IUPAC, abbreviations, formulas).

Stage III: tool-constrained symbolic construction. We convert R-group substituents into SMILES: IUPAC names via OPSIN, and abbreviations/formulas via a curated mapping table (assisted by BIOMINER-INSTRUCT during construction; the finalized table is fixed for deployment). Finally, RDKit composes scaffold and substituent SMILES by attachment-point "zipping" to generate chemically valid full structures. This stage enforces chemical validity deterministically and isolates the MLLM from error-prone string construction.

The final output $\mathscr{S}$ merges: (i) explicit structures (c, ℓ) and (ii) Markush-enumerated structures (c, ℓ) obtained by RDKit construction.

Agent 3: bioactivity measurement extraction (A^{mea})

A^{mea} extracts bioactivity measurements as semantic tuples (p, c, v) from both text and visual modalities:

$$\mathscr{M} = \{(p_t, c_t, v_t)\}_{t=1}^{T} = A^{mea}(D^{txt}, D^{vis}). \tag{8}$$

We run BIOMINER-INSTRUCT over (a) D^{txt} and (b) page-level images D^{vis}_{page} to capture measurements reported in paragraphs, tables, and figure panels. Outputs are normalized into a unified schema, including value types (e.g., IC_{50}, K_i), numeric values, and units when available. Text- and vision-derived tuples are then merged with deduplication based on entity/value consistency rules (Supporting Information).

Agent 4: postprocessing and integration (A^{post})

The postprocessing agent joins $\mathscr{M}$ and $\mathscr{S}$ via ligand coreference to produce final triplets:

$$\mathscr{B} = A^{post}(\mathscr{S}, \mathscr{M}) = \{(p, \ell, v)\}. \tag{9}$$

Optional enrichment. For downstream usability, A^{post} optionally enriches records with external identifiers and structures by querying public databases (e.g., UniProtKB, PDB, AlphaFoldDB), and can link extracted bioactivities to reported PDB complexes by SMILES matching and target consistency checks.

Summary. Overall, BIOMINER operationalizes a structure-grounded paradigm: BIOMINER-INSTRUCT performs cross-modal *relation inference* (what refers to what, which R-groups belong where, which measurements belong to which targets), while domain-specific models provide perception (OCSR/detection) and domain tools guarantee *exact* chemical symbolic construction and validity.

Model training for BIOMINER

BIOMINER is an inference-only agentic system at deployment time; however, we develop two open-weight models to enable fully reproducible local execution: (i) an optical chemical structure recognition (OCSR) model MOLGLYPH for robust recognition of complex chemical depictions in scientific figures, and (ii) a bioactivity-oriented multimodal instruction model BIOMINER-INSTRUCT for grounded cross-modal reasoning (coreference, Markush enumeration, and image-based bioactivity parsing). Importantly, the training of these models is *document-independent* and strictly isolated from BIOVISTA to avoid data leakage.

Training OCSR MOLGLYPH

Architecture. MOLGLYPH follows an encoder–decoder OCSR design instantiated from MolScribe, using a Swin-Transformer visual encoder and an Transformer decoder with 12 layers, 16 attention heads, and 512 hidden dimension.

Tokenizer and data format. We implement the tokenizer and sequence format following the MolParser paper specification. MolParser does not provide an open-source implementation; therefore, all tokenization, rendering, and training pipelines are independently implemented in this work, while remaining compatible with the released MolParser pretraining data.

Training Data. We use three complementary data sources: (1) **MolParser pretraining set** (8M images), which contains ∼2M synthesized Markush-like samples and provides broad coverage of depiction styles; (2) **MolParser real-world set** (91k images) as a curated evaluation-oriented distribution of real depictions; (3) **Literature-specific OCSR set** (170k images) constructed in this work to mitigate domain shift from patents to scientific articles. The MolParser corpus is predominantly patent-derived, which differs from the visual style and composition of biomedical literature (e.g., tighter layouts, multi-panel figures, smaller line widths, and heavier co-occurrence with biological annotations). To bridge this gap, we collect chemical depiction panels from (i) PDBbind-cited publications prior to 2019 and (ii) high-bioactivity-density medicinal chemistry journals (JMC, JNP, BMC, BMCL). To prevent evaluation leakage, we ensure that journals/papers included in BioVista are excluded from this OCSR training pool; additionally, to reduce redundancy, we sample 3–5 figure panels per paper and cap the number of samples per article source when necessary. Among the three data sources, the literature-specific set is weakly labeled by MolParser (closed API) to obtain paired (image, SMILES) supervision. We treat these labels as *pseudo-ground-truth* and rely on large-scale diversity + curriculum learning (below) to reduce sensitivity to occasional labeling noise.

Curriculum learning and augmentation. MolGlyph is trained with a curriculum strategy: we first pretrain on the large-scale MolParser pretraining set, and then progressively introduce stronger augmentations and harder samples. After convergence, we perform joint training on a mixed corpus consisting of: the MolParser real-world 91k set, a balanced subset of ∼200k synthesized Markush samples, and the 170k literature-specific set. This final mixture emphasizes the scientific-figure distribution while retaining sufficient Markush coverage.

Optimization and hyperparameters. The model is trained end-to-end using teacher-forced cross-entropy loss with label smoothing of 0.1. We use the AdamW optimizer with separate learning rates of 4×10^{-4} for both encoder and decoder. A linear warm-up schedule is applied for the first 2% of training steps, followed by cosine decay. Training is conducted for 10 epochs over the combined corpus. To improve stability on noisy pseudo-labeled data, we filter invalid structure graphs during training and apply SMILES syntax constraints during decoding. It takes 7 days to train MolGlyph on 4 A800 80G GPUs.

***Training MLLM* BioMiner-Instruct**

Base model. We initialize BioMiner-Instruct from Qwen3-VL-32B-Instruct and perform multi-task supervised instruction tuning (SFT) on a curated multi-modal dataset covering three tasks used by BioMiner: (1) **Markush enumeration** (scaffold identification and R-group set enumeration), (2) **Full-structure coreference recognition** (mapping depicted structures/fragments to textual identifiers), (3) **Image bioactivity extraction** (interpreting dense quantitative bioactivity values).

Instruct-tunning Data. All candidate papers are drawn from PDBbind-cited publications prior to 2019 and the same four high-bioactivity journals (JMC, JNP, BMC, BMCL), with explicit exclusion of the 500 BioVista papers. This isolation is enforced at the paper level (no shared PDF sources). We first run a proprietary MLLM (Gemini-2.0-flash) to generate candidate input–output pairs for all three tasks. Then, we filter and mine difficult examples to avoid over-representing trivial instances. For Markush enumeration, we discard invalid outputs and cases enumerating fewer than 8 structures. For image bioactivity extraction we prioritize dense figures. And for coreference recognition, since BioMiner adopts a divide-and-conquer strategy with a controlled number of molecules per image (typically $\leq$4), we select data randomly. Followed by human expert checking to ensure data quality, the final multi-task SFT set contains approximately 754 Markush enumeration instances, 2,215 coreference instances, and 4,375 image bioactivity instances.

Fine-tuning strategy. BioMiner is obtained by supervised instruction tuning of Qwen3-VL-32B-Instruct using parameter-efficient low-rank adaptation (LoRA)[51]. We adopt 4-bit quantization during training (QLoRA)[52] to reduce GPU memory usage while preserving model capacity. The instruction dataset jointly covers three tasks: Markush enumeration, full-structure coreference recognition, and image-based bioactivity extraction. These three tasks are mixed uniformly during training without task-specific weighting.

Optimization and hyperparameters. Training is performed using the AdamW optimizer with a peak learning rate of 5×10^{-5}. A cosine learning-rate scheduler with a warm-up ratio of 0.1 is applied. It takes one day to train BioMiner-Instruct on 4 A800 80G GPUs.

Acknowledgements

We sincerely thank Xi Fang from DP Technology for their support of MolParser. We also extend our gratitude to Ruikang Li and Qingchuan Li from USTC, as well as Qian Yan, Qian Xie, Yiwen Zhang, Tongtong Yan, Huangdong Liang, and Li Wang

from iFLYTEK, for their invaluable support in the construction of the benchmark BIOVISTA.

Author contributions statement

Conceptualization: J.X.Y., X.K.L., J.T.Z., Z.X.Z., and Q.L.; Data Curation: J.X.Y., K.Z., Y.H.Y., and B.Y.Z.; Formal Analysis: J.X.Y., Y.H.Y., K.Z., Z.X.Z., J.T.Z., and X.K.L.; Investigation: J.X.Y., and J.T.Z.; Methodology: J.X.Y., X.K.L., Z.X.Z., and J.T.Z.; Software: J.X.Y., Y.H.Y., J.T.Z., and J.C.X.; Validation: J.X.Y., J.T.Z., K.Z., Z.X.Z., and Q.L.; Visualization: J.X.Y., and J.T.Z.; Writing – Original Draft: J.X.Y., and J.T.Z.; Writing – Review & Editing: J.X.Y., J.T.Z., Z.X.Z., K.Y.G., and Q.L.; Supervision: Z.X.Z., K.Z., and Q.L.; Project administration: K.Z., and Q.L.; Resources: K.Z., K.Y.G., and Q.L.; Funding Acquisition: Q.L.; All authors reviewed the manuscript.

Data availability

The proposed benchmark BIOVISTA is now available in https://github.com/jiaxianyan/BioMiner. The PDBbind dataset is available in http://www.pdbbind.org.cn/. The ChEMBL dataset is available in https://www.ebi.ac.uk/chembl/. The molecule library ChemDiv is available in https://www.chemdiv.com/. The molecule library Enamine is available in https://enamine.net/. The PoseBuster dataset is available in https://github.com/maabuu/posebusters.

Code availability

All codes of proposed BIOMINER are released in https://github.com/jiaxianyan/BioMiner.

References

1. Gaulton, A. *et al.* Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100 – D1107 (2011).
2. Zaixin Zhang *et al.* Structure-based drug design with geometric deep learning: A comprehensive survey. *ACM Comput. Surv.* **58**, 1 – 35 (2025).
3. Theisen, R., Wang, T., Ravikumar, B., Rahman, R. & Cichońska, A. Leveraging multiple data types for improved compound-kinase bioactivity prediction. *Nat. Commun.* **15** (2024).
4. Lai, H. *et al.* Interformer: an interaction-aware model for protein-ligand docking and affinity prediction. *Nat. Commun.* **15** (2024).
5. Shah, P. M. *et al.* Deepdtagen: a multitask deep learning framework for drug-target affinity prediction and target-aware drugs generation. *Nat. Commun.* **16** (2025).
6. Lu, Z. *et al.* Dtiam: a unified framework for predicting drug-target interactions, binding affinities and drug mechanisms. *Nat. Commun.* **16** (2025).
7. Ye, Q. *et al.* A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. *Nat. Commun.* **12** (2021).
8. Koh, H. Y., Nguyen, A. T. N., Pan, S., May, L. T. & Webb, G. I. Physicochemical graph neural network for learning protein-ligand interaction fingerprints from sequence data. *Nat. Mac. Intell.* **5**, 673–687 (2024).
9. Mastropietro, A., Pasculli, G. & Bajorath, J. Learning characteristics of graph neural networks predicting protein-ligand affinities. *Nat. Mac. Intell.* **5**, 1427–1436 (2023).
10. Zhang, Z., Shen, W., Liu, Q. & Zitnik, M. Efficient generation of protein pockets with pocketgen. *Nat. Mac. Intell.* **6**, 1382–1395.
11. Feng, B. *et al.* A bioactivity foundation model using pairwise meta-learning. *Nat. Mac. Intell.* **6**, 962–974 (2024).
12. Gentile, F. *et al.* Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. *Nat. Protoc.* **17**, 672 – 697 (2022).
13. Cao, D. *et al.* Generic protein-ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling. *Nat. Mac. Intell.* **6**, 688–700 (2024).
14. Zdrazil, B. *et al.* The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **52**, D1180–D1192 (2024).
15. Liu, T. *et al.* Bindingdb in 2024: a fair knowledgebase of protein-small molecule binding data. *Nucleic Acids Res.* **53**, D1633–D1644 (2025).
16. Liu, Z. *et al.* Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics* **31 3**, 405–12 (2015).
17. Lan, T. *et al.* Generating mutants of monotone affinity towards stronger protein complexes through adversarial learning. *Nat. Mac. Intell.* **6**, 315–325 (2024).
18. Song, B., Li, F., Liu, Y. & Zeng, X. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Briefings Bioinforma.* **22**, bbab282 (2021).
19. John Dagdelen *et al.* Structured information extraction from scientific text with large language models. *Nat. Commun.* **15** (2024).
20. Morin, L., Weber, V., Meijer, G. I., Yu, F. & Staar, P. W. J. Patcid: an open-access dataset of chemical structures in patent documents. *Nat. Commun.* **15** (2024).
21. Zheng, Z., Zhang, O., Borgs, C., Chayes, J. T. & Yaghi, O. M. Chatgpt chemistry assistant for text mining and the prediction of mof synthesis. *J. Am. Chem. Soc.* **145**, 18048–18062 (2023).
22. Smith, N., Yuan, X., Melissinos, C. & Moghe, G. D. Funcfetch: an llm-assisted workflow enables mining thousands of enzyme–substrate interactions from published manuscripts. *Bioinformatics* **41** (2024).
23. Yeon Seok Kang & Kim, J. Chatmof: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nat. Commun.* **15** (2024).
24. Simmons, E. S. Markush structure searching over the years. *World Pat. Inf.* **25**, 195–202 (2003).

25. Su, M. *et al.* Comparative assessment of scoring functions: The casf-2016 update. *J. Chem. Inf. Model.* **59 2**, 895–913 (2018).

26. Dunbar, J. B. *et al.* Csar data set release 2012: Ligands, affinities, complexes, and docking decoys. *J. Chem. Inf. Model.* **53**, 1842 – 1852 (2013).

27. Swanson, K. V., Deng, M. & Ting, J. P.-Y. The nlrp3 inflammasome: molecular activation and regulation to therapeutics. *Nat. Rev. Immunol.* **19**, 477 – 489 (2019).

28. Chemdiv, available: https://www.chemdiv.com/ (ChemDiv, 2023).

29. Enamine, available: https://enamine.net/ (Enamine, 2023).

30. Buttenschoen, M., Morris, G. M. & Deane, C. M. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci.* **15**, 3130 – 3139 (2023).

31. Wang, B. *et al.* Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839* (2024).

32. Shuai Bai *et al.* Qwen3-vl technical report (2025).

33. Uni Parser Team. Moldetv2. https://huggingface.co/UniParser/MolDetv2 (2025). Accessed: 20251116.

34. Lowe, D. M., Corbett, P. T., Murray-Rust, P. & Glen, R. C. Chemical name to structure: Opsin, an open source solution. *J. Chem. Inf. Model.* **51 3**, 739–53 (2011).

35. Lombard, M., Snyder-Duch, J. & Bracken, C. C. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Hum. communication research* **28**, 587–604 (2002).

36. Hurst, O. A. *et al.* Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

37. Claude, available: https://www.anthropic.com/news/introducing-claude (Anthropic, 2023).

38. Bai, J. *et al.* Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966* (2023).

39. Grok, available: https://grok.x.ai/ (X team, 2023).

40. David Rolnick, Veit, A., Belongie, S. J. & Shavit, N. Deep learning is robust to massive label noise. *ArXiv* **abs/1705.10694** (2017).

41. Velickovic, P. *et al.* Graph attention networks. *ICLR'18* (2018).

42. Satorras, V. G., Hoogeboom, E. & Welling, M. E(n) equivariant graph neural networks. *ICML'21* (2021).

43. Xiong, Z. *et al.* Pushing the boundaries of molecular representation for drug discovery with graph attention mechanism. *J. Medicinal Chem.* (2020).

44. Kipf, T. & Welling, M. Semi-supervised classification with graph convolutional networks. *ICLR'17* (2017).

45. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

46. Hearst, M. A. Support vector machines. *IEEE Intell. Syst. & Their Appl.* **13**, 18–28 (1998).

47. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50 5**, 742–54 (2010).

48. Reutlinger, M. *et al.* Chemically advanced template search (cats) for scaffold-hopping and prospective target prediction for 'orphan' molecules. *Mol. Informatics* **32**, 133 – 138 (2013).

49. Coll, R. C. *et al.* Mcc950 directly targets the nlrp3 atp-hydrolysis motif for inflammasome inhibition. *Nat. Chem. Biol.* **15**, 556–559 (2019).

50. Velcicky, J. *et al.* Discovery of potent, orally bioavailable, tricyclic nlrp3 inhibitors. *J. Medicinal Chem.* (2024).

51. J. Edward Hu *et al.* Lora: Low-rank adaptation of large language models. *ArXiv* **abs/2106.09685** (2021).

52. Tim Dettmers, Pagnoni, A., Holtzman, A. & Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *ArXiv* **abs/2305.14314** (2023).