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 is challenging due to the multi-modal distribution of information (text, tables, figures, structures) and the complexity of chemical representations (e.g., Markush structures). Furthermore, the lack of standardized benchmarks impedes the evaluation and development of extraction methods. In this work, we introduce BioMiner, a multi-modal system designed to automatically extract protein-ligand bioactivity data from thousands to potentially millions of publications. BioMiner employs a modular, agent-based architecture, leveraging a synergistic combination of multi-modal large language models, domain-specific models, and domain tools to navigate this complex extraction task. To address the benchmark gap and support method development, we establish BioVista, a comprehensive benchmark comprising 16,457 bioactivity entries and 8,735 chemical structures curated from 500 publications. On BioVista, BioMiner validates its extraction ability and provides a quantitative baseline, achieving F1 scores of 0.22, 0.45, and 0.53 for bioactivity triplets, chemical structures, and bioactivity measurement with high throughput (14s/paper on 8 V100 GPUs). We further demonstrate BioMiner's practical utility via three applications: (1) extracting 67,953 data from 11,683 papers to build a pre-training database that improves downstream models performance by 3.1%; (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 vast amounts of previously inaccessible bioactivity data and accelerate data-driven drug discovery. All codes and data are available at GitHub.

Main

Protein-ligand bioactivity data represents a cornerstone of modern drug discovery[1]. Quantifying the interaction strength between potential drug molecules (ligands) and their biological targets (proteins), this information provides fundamental insights into structure-activity relationships (SAR) and the core principles of medicinal chemistry[2–7]. Crucially, it serves as the essential fuel for developing and validating predictive computational models – from traditional quantitative structure-activity relationship (QSAR) models to the artificial intelligence (AI) algorithms increasingly employed for virtual screening and hit identification[8,9]. Access to comprehensive and accurate bioactivity data is therefore indispensable for accelerating the discovery and design of novel therapeutics. Over the past several years, there has been a considerable increase in experimentally determined bioactivity data, meticulously curated and made publicly accessible through databases such as ChEMBL[10], BindingDB[11], PDBbind[12], and PubChem[13]. These invaluable resources serve as foundational pillars for drug discovery efforts.

However, the construction of these datasets relies on a labor-intensive process, requiring domain experts to painstakingly identify and transcribe bioactivity data from a vast number of scientific publications. This manual curation process can often take hours per publication, representing a significant bottleneck in scaling up bioactivity databases. With the exponential growth of scientific literature, a substantial portion of valuable bioactivity data remains inaccessible and unstructured, highlighting the urgent need for robust and efficient automated tools capable of extracting bioactivity data at scale.

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)[4,14] and optical chemical structure recognition (OCSR)[15]. 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[16,17]. However, these methods, while demonstrating progresses, are not yet equipped to handle the extraction of protein-ligand bioactivity data.

Specifically, given a literature, the goal of bioactivity extraction is to extract all protein-ligand-bioactivity triplets (protein name, ligand SMILES, bioactivity value). Three unique challenges distinct from prior extraction tasks are presented. **Firstly**, the task is fundamentally multi-modal. Bioactivity data are distributed across different modalities, including text, tables, figures, and crucial chemical structures. This multi-modal challenge requires a sophisticated approach capable of cross-modal reasoning and information fusion, going beyond simple text mining or image mining. **Secondly**, the accurate recognition and conversion of chemical structures, particularly the widely used Markush structures (which represent groups of related chemical compounds)[18], 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. **Finally**, besides these intrinsic data complexities, the filed 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 intelligent system specialized for automatically extracting protein-ligand bioactivity data from literature, and establish a comprehensive and challenging benchmark **BioVista**. We first describe BioMiner, which employs a modular, agent-based architecture to navigate this complex extraction task (Figure 1(a)). Rather than training a single end-to-end extraction model, BioMiner decomposes the extraction process into manageable sub-tasks: document parsing (data preprocessing), chemical structure extraction, bioactivity measurement extraction, and cross-modal data integration (data postprocessing). A synergistic combination of general multi-modal large language models (MLLMs), domain-specific models (DSMs), and domain tools (DTs) is employed in the framework. Particularly, for chemical structure extraction, the system incorporates a MLLM-DSM-DT collaborative strategy specifically designed to resolve and enumerate 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 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[12]. 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 system evaluation (Figure 1(c)). When evaluating BioMiner on the BioVista dataset, the system achieves F1 scores of 0.22 for extracting complete bioactivity triplets, 0.45 for chemical structures (ligand coreference-SMILES), and 0.53 for bioactivity measurements (protein-ligand coreference-bioactivity value). This performance is complemented by high throughput, processing literature at approximately 14 seconds per paper on 8 V100 GPUs. Notably, BioMiner demonstrates effectiveness in processing complex chemical representations, achieving an F1 score of 0.56 for the challenging task of Markush structure enumeration. These results validate BioMiner's bioactivity data extraction ability and providing a quantitative baseline for future research.

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 67,953 data points extracted from 11,683 papers within two days. Models pre-trained on this BioMiner-generated database demonstrate improved performance (3.1% and 1.9% improvement in RMSE metric) on two independent test sets (PDBbind v2016 core set[19] and CSAR-HiQ[20]) 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[21], 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[22] and Enamine[23]) 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

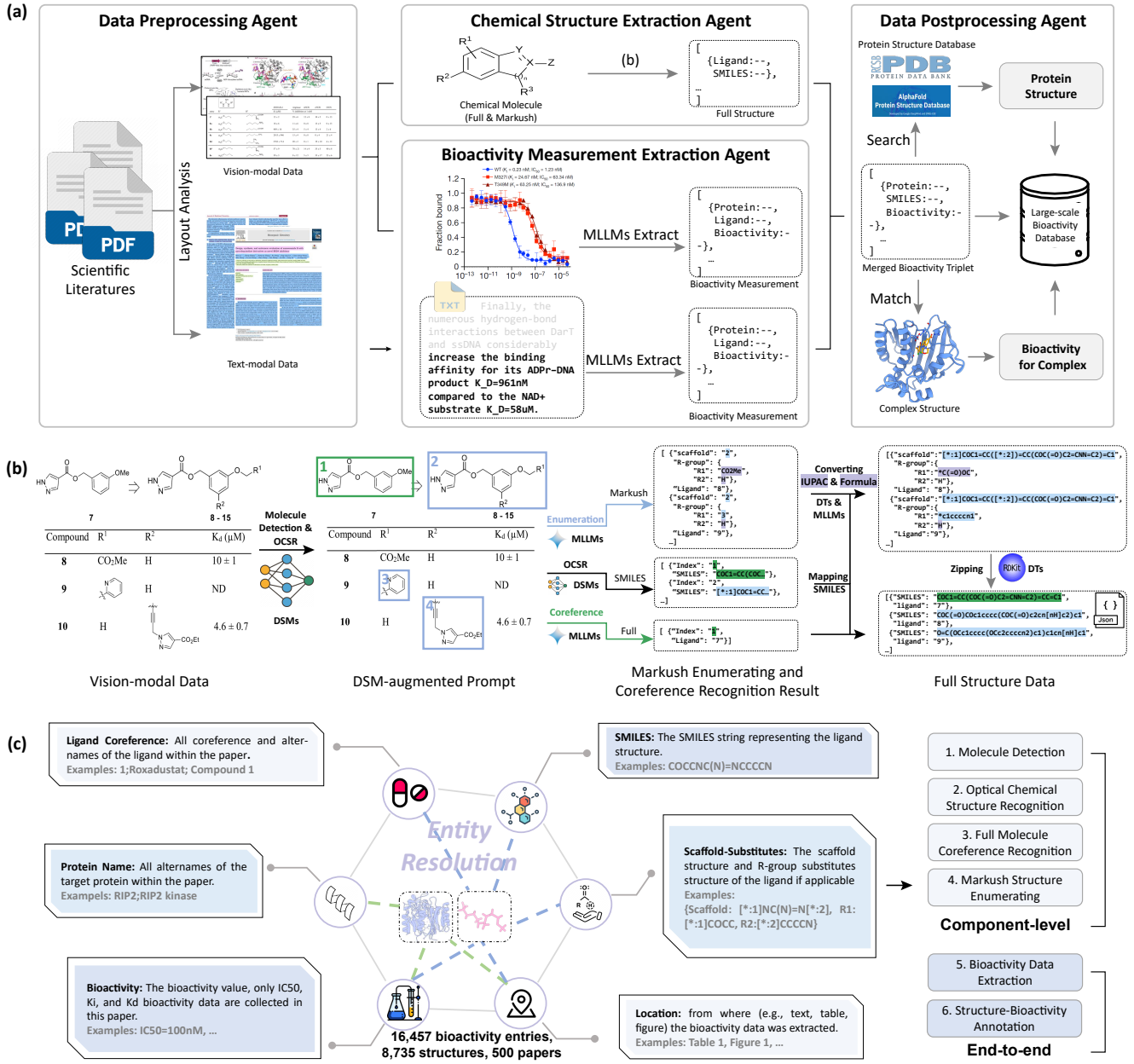


Figure 1. Overview of protein-ligand bioactivity extraction framework BIOMINER and benchmark BIOVISTA. (a) The whole protein-ligand bioactivity extraction framework BIOMINER. (b) The chemical structure extraction agent. In this agent, explicit full structures and Markush structures are both processed. 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.

datasets like PDBbind. Using a HITL annotation workflow similar to bioactivity extraction, we label 242 structures within the famous PoseBusters database[24] 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.

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

The BIOMINER system, depicted in Figure 1(a), is a multi-modal system engineered for the automated extraction of protein-ligand bioactivity data from scientific literature.

Overview of BIOMINER

BIOMINER integrates four specialized agents-data preprocessing, chemical structure extraction, bioactivity measurement extraction, and data postprocessing—working collaboratively to address the complexity of this task. The key design principle of BIOMINER is tuning-free. Instead of training a novel end-to-end bioactivity extraction model, BIOMINER decomposes the complex extraction problem into four fundamental sub-tasks and leverages a synergistic combination of general multi-modal large language models (MLLMs), domain-specific models (DSMs), and domain tools (DTs) to solve them. This design eliminates the need for extensive collection of labeled data and training of task-specific models, which are common challenges in the development of specialized tools.

The extraction framework begins with the data preprocessing agent, which utilizes the DSM toolkit MinerU[25] to parse PDF documents. Through layout analysis and reading order determination, this agent parses both textual and visual content with high fidelity. Subsequently, the chemical structure extraction and bioactivity measurement extraction agents operate in parallel on the parsed data. The bioactivity measurement extraction agent employs the MLLM Gemini-2.0-flash[26], guided by carefully crafted prompts, to identify quantitative bioactivity measurements (protein-ligand coreference-bioactivity value) from text, tables, and figures. Meanwhile, the chemical structure extraction agent, detailed below, employs a MLLM-DSM-DT collaborative strategy to resolve complex chemical structures (ligand corefernence-SMILES), including Markush structures. Finally, the postprocessing agent integrates the extracted bioactivity measurements from different modalities with their corresponding ligand SMILES strings to obtain the full bioactivity triplets (protein-SMILES-bioactivity value). This agent can optionally enrich the extracted data by querying external databases (e.g., UniProtKB[27], AlphaFoldDB[28], PDB[29]) for protein structural information or annotating reported PDB complex structures with the corresponding extracted bioactivity data, thereby creating highly actionable datasets. This tuning-free pipeline—enabled by the strategic orchestration of MLLMs' reasoning capabilities, DSMs' domain knowledge, and DTs' precise rules—delivers robust performance without the need for new model training, distinguishing BIOMINER as a scalable and adaptable solution in a dynamic research field.

Chemical Structure Extraction

The chemical structure extraction agent, illustrated in Figure 1(b), is a cornerstone of BIOMINER 's ability to handle the complex chemical structures of bioactivity data. It operates in three stages to extract and resolve ligand structures, including Markush structures, which are prevalent in medicinal chemistry literature and represent a significant challenge. First, DSMs (MolMiner[30], supplemented by YOLO[31] for tables) detect 2D molecular structure depictions within the document. MolParser[32] then performs OCSR to generate SMILES strings for these depictions. Second, using images augmented with DSM-detected structure bounding boxes and indices (Figure 1(b)), MLLM Gemini performs coreference recognition and Markush enumeration. For explicit full structure, Gemini directly recognizes its coreference, linking chemical depictions to their textual references. Critically, for Markush structures, Gemini identifies Markush scaffold structures and enumerates associated R-group substituents. Third, BIOMINER uses RDKit[33] 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

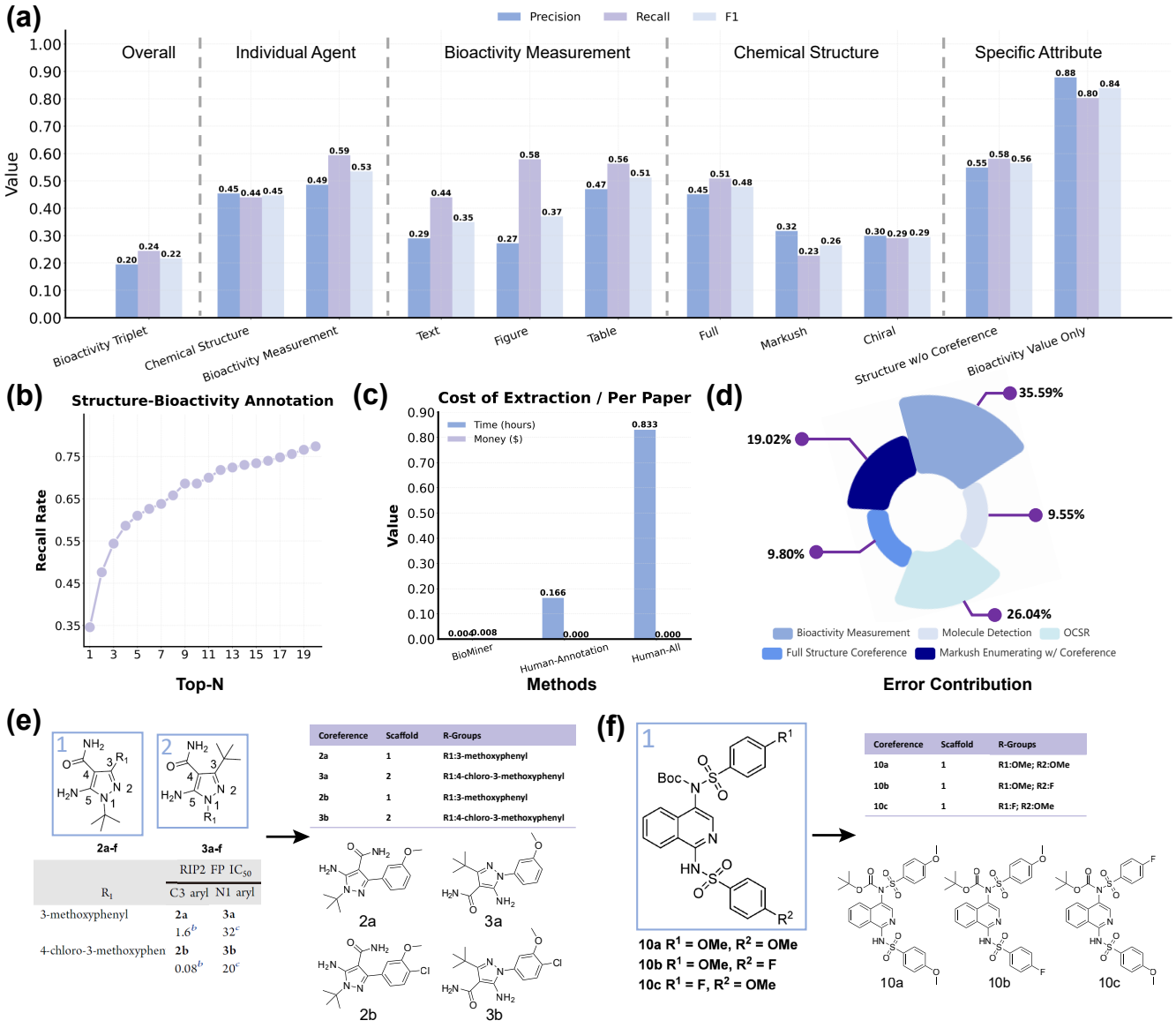


Figure 2. Benchmarking extraction performance of BioMiner on BioVista. (a) Performance of bioactivity triplet and individual attributes extraction. (b) Performance of structure-bioactivity annotation. Given a PDB structure and corresponding paper, BioMiner first extracts all bioactivity from the paper and then ranks these data based on ligand similarity between extracted data and PDB structure. Top-N recall rate represents the ratio of target bioactivtiy data in Top-N extracted data. (c) Comparison on consuming time and money between BioMiner and human. (d) Detailed error source analysis of bioactivity triplet extraction. (e,f) Two examples of Markush structures successfully processed by BioMiner.

IUPAC names and a Gemini-assisted, manually curated mapping table for abbreviations and chemical formulas. This hybrid MLLM-DSM-DT approach, which obviates the need for task-specific model training, enables the successful resolution of both explicit full and Markush structures, representing a key innovation of BioMiner. Implementation details of this agent are provided in the Method and Supporting Information section.

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 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.

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. In BioMiner, we choose Gemini-2.0-flash as our employed MLLM based on this table results. The best results are bolded, and the second best results are underlined.

Tasks	Metric	**Gemini-2.0-flash**	Gemini-2.0-pro	GPT-4o	GPT-4o-mini	Claude-3.7-sonnet	Claude-3.5-sonnet	Grok-3	Qwen-2.5
Component-level evaluation for chemical structure extraction									
Full Structure Coreference	Recall	**0.786**	0.763	0.749	0.552	0.749	0.755	0.529	0.748
	Precision	**0.783**	0.760	0.745	0.550	0.757	0.754	0.542	0.745
	F1	**0.784**	0.761	0.747	0.551	0.753	0.754	0.535	0.746
Markush Enumerating w/ Coreference	Recall	**0.579**	0.386	0.430	0.039	0.465	0.477	0.190	0.209
	Precision	**0.545**	0.377	0.426	0.041	0.475	0.472	0.190	0.212
	F1	**0.561**	0.381	0.428	0.040	0.470	0.474	0.190	0.210
Markush Enumerating w/o Coreference	Recall	**0.628**	0.425	0.500	0.050	0.503	0.516	0.241	0.252
	Precision	**0.594**	0.448	0.495	0.058	0.514	0.511	0.257	0.281
	F1	**0.611**	0.436	0.497	0.054	0.508	0.513	0.249	0.266
Different attributes evaluation for end-to-end extraction and annotation									
Structure w/ Coreference	Recall	**0.440**	0.374	0.406	0.277	0.436	0.434	0.275	0.373
	Precision	0.454	0.395	0.419	0.305	**0.462**	0.459	0.326	0.403
	F1	0.447	0.384	0.412	0.290	**0.449**	0.446	0.298	0.387
Bioactivity Measurement	Recall	**0.594**	0.584	0.418	0.250	0.480	0.472	0.386	0.552
	Precision	**0.486**	0.465	0.364	0.219	0.396	0.410	0.285	0.431
	F1	**0.535**	0.518	0.389	0.233	0.434	0.439	0.328	0.484
Bioactivity Triplet	Recall	**0.244**	0.220	0.171	0.088	0.208	0.208	0.109	0.199
	Precision	**0.195**	0.172	0.149	0.077	0.178	0.185	0.081	0.149
	F1	**0.217**	0.193	0.159	0.082	0.192	0.196	0.093	0.170
Bioactivity-Structure Annotation	Top-1	0.346	0.288	0.318	0.258	0.330	**0.348**	0.264	0.322
	Top-3	**0.544**	0.482	0.530	0.454	0.518	0.538	0.434	0.494
	Top-10	0.686	0.634	0.688	0.612	0.670	**0.706**	0.614	0.680
Price comparison for different MLLMs									
1M Tokens Cost $	Input	**0.100**	1.250	2.500	0.150	3.300	3.300	0.830	-
	Output	**0.400**	5.000	10.000	0.600	16.500	16.500	4.160	-

Construction of BioVista

BioVista is derived from 500 recent publications referenced in PDBbind v2020[12], ensuring its relevance to real-world scenarios. 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 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 robustness, we implement a closed-loop bootstrapping procedure. Initially, BioMiner is used to perform automatic extractions on the curated publications, facilitating the identification and correction of potential inconsistencies or errors in the initial annotations (e.g., missing alternames, incorrect transcription). Feedback from these automatic extraction results are then incorporated into the refinement of BioMiner 's design, improving prompt engineering strategies and optimizing individual agent performance. Subsequently, the enhanced version of BioMiner is re-applied to comprehensively validate and confirm the accuracy of the final BioVista dataset. 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).

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=0.20, Recall=0.24, and F1=0.22 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=0.45, with performance varying by complexity (Explicit Full F1=0.48 vs. Markush F1=0.26). Bioactivity measurement extraction (protein-ligand coreference-bioactivity value) achieved F1=0.53, with table-based extraction being the most effective (Table F1=0.51 vs. Figure F1=0.37, Text F1=0.35). Extraction of isolated attributes is even more proficient, with solely bioactivity value F1=0.84 and SMILES F1=0.56. 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 0.69 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.

For end-to-end bioactivity extraction and structure-bioactivity annotation, BioMiner offers transformative efficiency gains as indicated in Figure 2(c). It processes literature at an average rate of 0.004 hours/paper (on 8 V100 GPUs, about 14s/paper) with a very cheap cost ($0.008/paper), orders of magnitude faster than human structure annotation and bioactivity extraction (0.166-0.833 hours/paper). These results demonstrate BioMiner's potential for cost-effective, large-scale data mining.

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 S2), OCSR (Table S3), and the second stage full structure coreference recognition, Markush enumeration (Table 1).

For first stage molecule detection, BioMiner employs MolMiner supplemented by YOLO for table molecules, achieving mAP=0.857, AP_{75}=0.892, and AP-Small=0.185 (Table S2). Compared with directly using YOLO or MolMiner, BioMiner's strategy achieves better performance for detecting small molecules (best AP-Small), usually R-group within tables, and meeting high Intersection over Union (IoU) requirement (best AP_{75}). As for OCSR, compared to other DSMs, the employed MolParser exhibits significantly better ability to recognize Markush structure (Accuracy=0.733 for Markush structures), and thus leading to best OCSR performance (Accuracy=0.703 for all structures) (Table S3). When it comes to second stage tasks, BioMiner achieves an F1 score of 0.784 and 0.561 on full structure coreference recognition and Markush enumeration (with coreference) using Gemini-2.0-flash (Table 1), respectively. If we only consider the structure enumeration result and ignore coreference, F1 score of Markush enumeration rises to 0.611. This result indicates the ability of BioMiner to process the challenging Markush structures. Different MLLMs, including Gemini, GPT[35], Claude[36], Qwen[37], and Grok[38], have been tested on these two tasks, and Gemini-2.0-flash 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 (35.59%), followed by OCSR inaccuracies (26.04%), and the Markush structure enumeration (19.02%). Molecule detection (9.55%) and explicit full structures coreference recognition (9.80%) contribute less to the overall error. This breakdown provides critical insights for directing future research toward optimizing the system components.

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

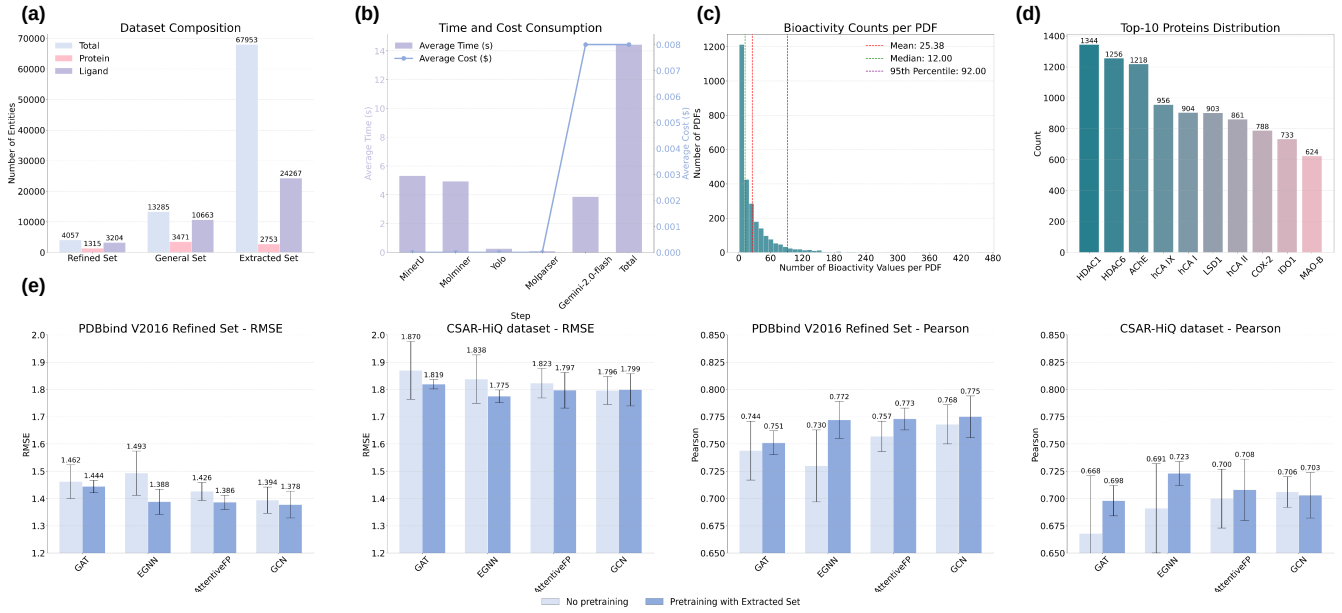


Figure 3. Using BioMiner collecting large-scale bioactivity data for deep learning model training. (a) Statistical comparison between manually curated PDBbind v2016 dataset and our extracted dataset; (b) Consuming time and money for BioMiner; (c) Number of extracted bioactivity data in each paper; (d)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

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 two days (about 14s/paper on 8 V100 GPUs), a speed that would be practically impossible with manual curation (Figure 3(b)). From these paper, 195,046 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. 67,953 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 25.38 bioactivity values per paper, with a median of 12 and a 95th percentile of 92, 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. We pre-train several graph neural network (GNN) architectures (GAT[39], EGNN[40], AttentiveFP[41], and GCN[42]) 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. As demonstrated in Figure 3(e), 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 3.1% and 1.9% in the RMSE metric are acheived in the PDBbind core set and CSAR-HiQ set, respectively. 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,

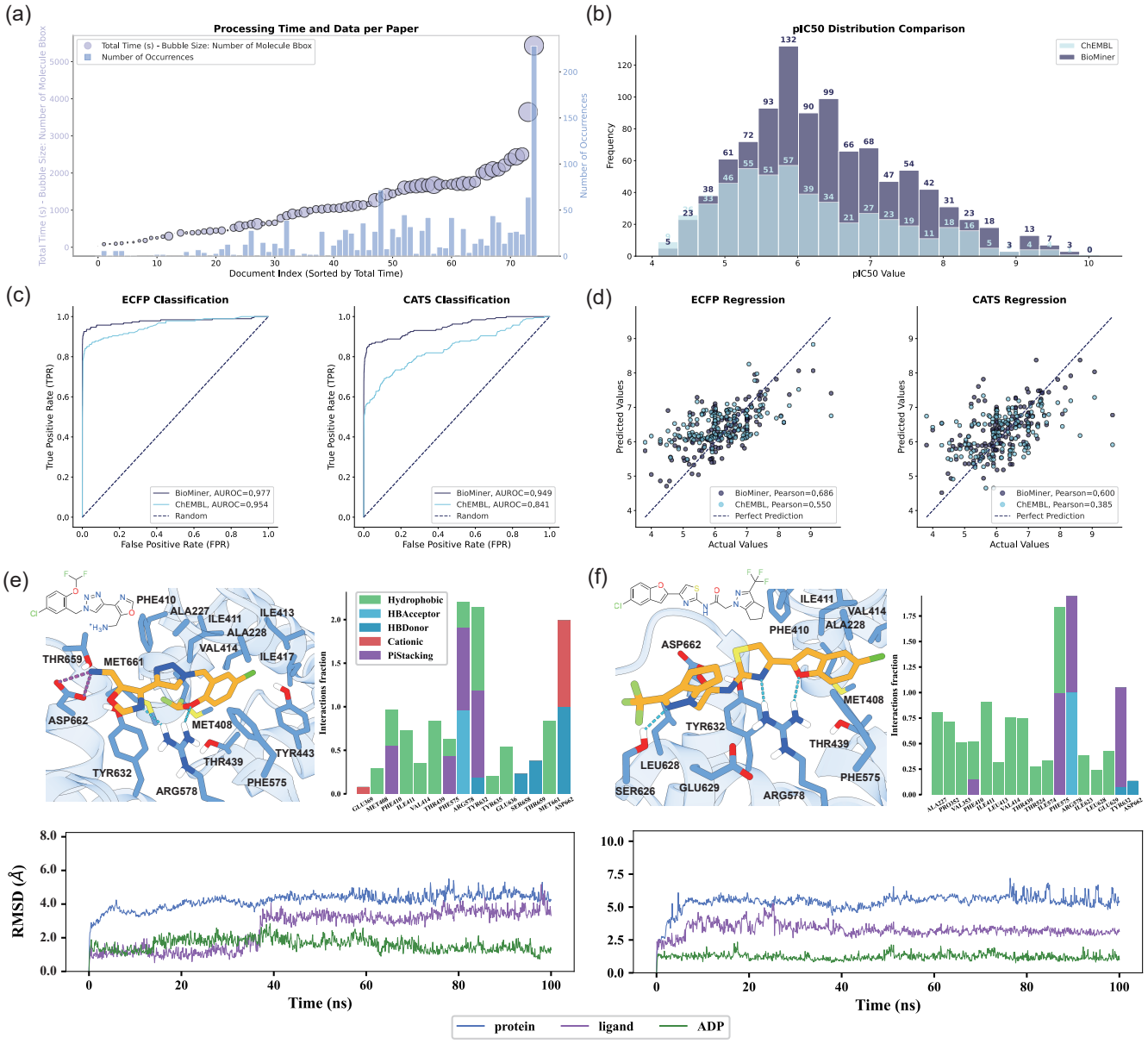


Figure 4. Using BioMiner for human-in-the-loop NLRP3 bioactivity data collection from 85 papers and inhibitor screening. (a) Consuming time (about 18.4 minutes in average) and number of collected data for each paper. The final consuming time is highly related with the number of bioactivity data and chemical 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.

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 S2, S3, and S4, 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[21] 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.

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 S5 and Table S6, various QSAR models with different algorithms (e.g., Random Forest[43], SVM[44], etc.), different tasks (regression, classification), and different molecular representations (ECFP[45], CATS[46]) 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 S8 and see Supplementary Information for details), sixteen compounds are manually selected as potential hit candidates, exhibiting both high structural diversity and novelty (Table S7). Further MM/PBSA binding free energy calculation reveals six compounds exhibit comparable or better binding free energies than both MCC950[47] and NP3-562[48], suggesting their potential as highly potent NLRP3 inhibitors (Table S7). 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

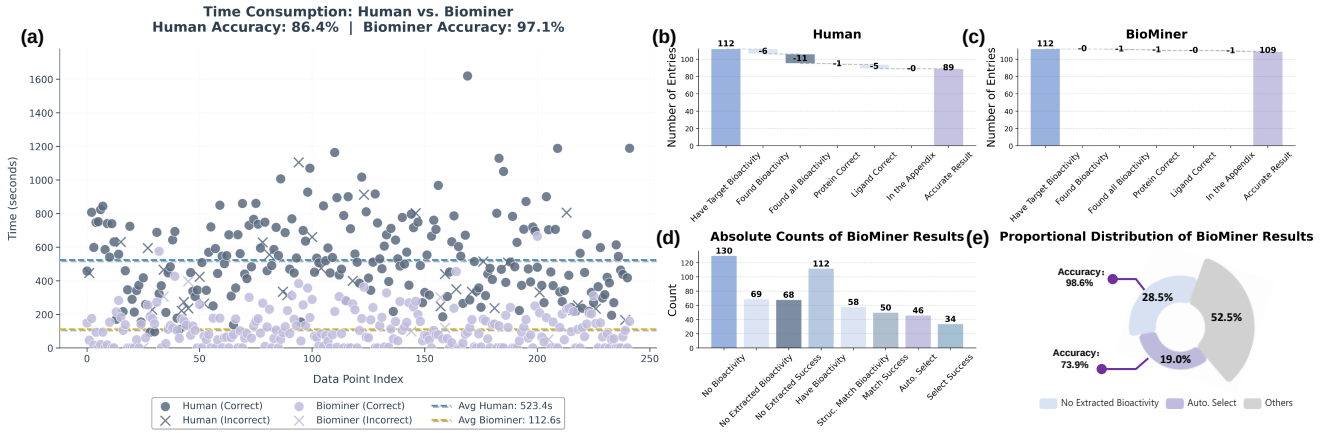


Figure 5. Annotating protein-ligand complex structures in PoseBusters with corresponding experimental bioactivity values. (a) Comparison of consuming time and annotation accuracy between human and BIOMINER; (b) Waterfall plot of error analysis of human extraction; (c) Waterfall plot of error analysis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

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[24], 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 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 69 of 242 complexes, BioMiner finds no bioactivity data in the publications. Among these complexes, 68 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, 58 complexes have at least one extracted bioactivity measurement associated with a ligand structure perfectly matching the PDB ligand (chemical similarity = 1.0). For these 58 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 leveraging the synergistic combination of multi-modal large language models (MLLMs), domain-specific models (DSMs), and domain tools (DTs) to automate the extraction of protein-ligand-bioactivity triplets from scientific papers. To rigorously evaluate such systems and guide future development, we established BioVista, the largest bioactivity extraction benchmark featuring 16,457 bioactivity entries and 8,735 chemical structures across six distinct evaluation tasks.

On BioVista, BioMiner demonstrated considerable extraction capabilities, achieving F1 scores of 0.22, 0.45, and 0.53 for bioactivity triplets, chemical structures, and bioactivity measurements, respectively, alongside high operational efficiency (14s/paper on 8 V100 GPUs). The practical significance of BioMiner was further underscored by three applications. Its scalability enabled the rapid generation of a large-scale bioactivity database (67,953 data points from 11,683 papers in 2 days), leading to improvements (e.g., 3.1% RMSE reduction) in downstream deep learning models pre-trained on this data (Figure 3). Moreover, human-in-the-loop (HITL) workflows showcased the powerful synergy between BioMiner and human expertise. This approach doubled the high-quality NLRP3 bioactivity data compared to ChEMBL in just 26 hours, resulting in markedly improved QSAR models (38.6% EF1% gain) and the identification of 16 novel hit candidates (Figure 4). Similarly, HITL annotation of the PoseBusters benchmark yielded a 5-fold speedup and enhanced accuracy (97.1% vs. 86.4% manual baseline) (Figure 5), highlighting BioMiner's value in curating crucial structural datasets.

While BioMiner marked significant progress, the modest end-to-end F1 score (0.22 for triplets) underscored the inherent complexity of integrating information across multiple modalities, interpreting complex chemical representations like Markush structures, and accurately linking entities scattered throughout publications. Our error analysis (Figure 2(d)) pinpointed specific bottlenecks demanding further refinement: reliably extracting and associating bioactivity measurements from diverse formats (35.6% error contribution), improving the robustness of OCSR (26.0%), and enhancing the handling of Markush structure enumeration (19.0%). Additional limitations included an intrinsic dependency on the rapidly evolving capabilities of underlying MLLMs. Future work should therefore prioritized advancing cross-modal reasoning mechanisms, strengthening component integration, and potentially exploring targeted fine-tuning or domain adaptation strategies to optimize performance on critical sub-tasks while maintaining adaptability.

Looking forward, BioMiner's modular architecture and its proven ability to handle complex chemical information suggested significant potential for generalization. The system's core capabilities in multi-modal integration, entity recognition, relationship extraction, and chemical structure processing could be readily adapted related extraction tasks in cheminformatics and bioinformatics, such as mining RNA-ligand interactions or molecule ADMET properties. Tailoring specific agents, such as the bioactivity measurement extractor or the chemical structure extractor for different molecular types or assay readouts, while leveraging the established framework for document parsing, multi-modal integration, and coreference resolution, would be key to extending BioMiner's utility.

In conclusion, BioMiner represented a significant advance in automated bioactivity data extraction, complemented by BioVista as an essential standard for rigorous evaluation and comparison. Together, they offered a scalable pathway to unlock vast amounts of previously inaccessible bioactivity data trapped in literature, thereby accelerating data-driven drug discovery and laying a robust foundation for future progress in automated scientific information extraction.

Methods

Preliminary

For notation, we represent a literature document as D. Each protein-ligand bioactivity data point b in the document is represented as a tuple $b = (p, l, v)$ with three attributes: the protein target name p, the ligand SMILES string l, and the bioactivity value v. Denoting the set of bioactivity data points within the literature as $B = \{b_1, b_2, ..., b_n\}$, where n is the number of bioactivity data points. The target of BIOMINER can be formally written as:

$$B = \text{BIOMINER}(D). \tag{1}$$

To achieve this, BIOMINER comprises four specialized agents: the data preprocessing agent A^{pre}, the chemical structure extraction agent A^{struc}, the bioactivity measurement extraction agent A^{measr}, and the data postprocessing agent A^{post}. In the following, we detail the functionality of each agent. *Additional information, including ablation study, MLLMs' prompt, technical tricks, and model implementations, can be found in the Supporting Information section.*

Data Preprocessing Agent

The data preprocessing agent A^{pre} initiates the pipeline by parsing the multi-modal content of document D, aiming to preserve essential information while minimizing noise for subsequent extraction tasks. To accomplish this, agent A^{pre} first employs the MinerU toolkit[25], which integrates a suite of models for document parsing, including layout analysis and reading order determination. This operation is formalized as:

$$D^{text}, D^{layout} = MinerU(D), \tag{2}$$

where D^{text} represents the extracted text containing all information for extracting bioactivity measurement in text modality, and D^{layout} denotes the layout analysis output, comprising categorized elements (e.g., text block, figure body, figure caption) and their corresponding regions within each page of D.

The agent further processes D to prepare visual data for downstream analysis. Pages of the document are converted to images D^{vision}_{page}. Then, based on D^{layout}, figures and tables are segmented from the document. To ensure completeness, only when captions are detected, the tables or figures are segmented out; otherwise, the entire page is treated as a single image. This process yields the visual data subset D^{vision}_{seg}, which encapsulates figures, tables, and their associated metadata. Here, D^{vision}_{page} supports molecule detection (not yet implemented in MinerU), and D^{vision}_{seg} facilitates Markush structure enumeration and full structure coreference recognition by isolating irrelevant text surrounding. The output of A^{pre} is thus a dual-modality representation of the document, defined as:

$$D^{text}, D^{vision} = A^{pre}(D), \tag{3}$$

where $D^{vision} = (D^{vision}_{page}, D^{vision}_{seg})$. This structured parsing into textual D^{text} and visual D^{vision} components provides a robust foundation for the extraction of bioactivity measurements and chemical structures, as described in the following subsections.

Chemical Structure Extraction Agent

The chemical structure extraction agent A^{struc} is tasked with detecting and recognizing chemical structures within a document, producing a set of ligand coreferences and their corresponding SMILES strings. This is formalized as:

$$(c_i, l_i) = A^{struc}(D^{vision}), i \in \{0, 1, ..., K^{struc} - 1\}, \tag{4}$$

where c_i is the ligand coreference (e.g., textual name), l_i is the SMILES string, and K^{struc} is the number of structures.

Chemical structures in scientific literature generally appear as either explicit full structures or Markush structures[18]. Explicit full structure is standalone molecule requiring only coreference resolution to link them to textual mentions. Markush structures, however, represent families of related compounds, typically with variable substitutions (R-groups), variable attachment positions, or repeating substructures (Figure S6). This work focuses on Markush structures with variable substitutions, where the Markush scaffold is representable as a SMILES string, as other types (e.g., variable positions or repeats) are incompatible with SMILES notation. To tackle this challenging structure extraction task, A^{struc} employs a novel MLLM-DSM-DT hybrid strategy that splits the extraction into three stages, as shown in Figure 1(b).

Molecule Detection and OCSR. The process begins with DSMs detecting 2D molecular graphs in D^{vision} and converting them to 1D SMILES. Both explicit full structures and Markush structures (including both scaffolds and R-group substituents) are detected and processed. The distinction between full structures and Markush structures is made based on the presence of attachment points in the recognized SMILES strings. This operation is expressed as:

$$l^{full}_i, l^{part}_i, x^{full}_j, x^{part}_j = DSM(D^{vision}), i \in \{0, 1, ..., K^{full} - 1\} \; and \; j \in \{0, 1, ..., K^{part} - 1\}, \tag{5}$$

where l_i^{full}, x_i^{full}, and K^{full} denote ligand SMILES strings, bounding boxes, and numbers for full structures (no attachment points), and l_j^{part}, x_j^{part}, K^{part} for Markush components (with attachment points). MolMiner[30] detects molecules across pages D_{page}^{vision}, supplemented by YOLO[31] specially for segmented tables D_{seg}^{vision}, addressing MolMiner's tendency to merge closely spaced structures. MolParser[32], a recently proposed model can handle different plot styles of Markush attachment points based on its novel end-to-end training paradigm (Table S2), then performs OCSR.

Markush Enumerating and Coreference Recognition. With detected structures, we first create DSM-augmented images, $D^{aug} = (D^{aug,full}, D^{aug,part})$, by plotting the bounding boxes onto the original images and adding unique numerical identifiers (artificial indices) to each box (top-left corner), as shown in Figure 1(b). Here, the creation of DSM-augmented images leverages D_{seg}^{vision} to isolate irrelevant text surrounding. These augmented images are processed separately for explicit full structures and Markush structures by MLLM Gemini-2.0-flash[26].

For explicit full structures, Gemini is prompted to identify the coreference corresponding to each numbered bounding box. The output is a set:

$$(c_i^{full}, \hat{c}_i^{full}) = Gemini(D^{aug,full}, T^{full}), i \in \{0, 1, ..., K^{full} - 1\}, \tag{6}$$

where c_i^{full} is the recognized coreference for a full structure box, $\hat{c}_i^{full}$ is the artificial index, and T^{full} represents the prompt for this full structure process step. For Markush structures, Gemini is prompted to identify Markush scaffolds, determine if they represent variable substitutions, and enumerate all R-group substituents for each scaffold in. The output is:

$$(c_i^{part}, \hat{c}_i^{part,scaf}, \hat{C}_i^{part,R}) = Gemini(D^{aug,part}, T^{part}), i \in \{0, 1, ..., \hat{K}^{full} - 1\}, \tag{7}$$

where c_i^{part} is the coreference for the full structure derived by combining Markush scaffold and R-group substitutes, $\hat{c}_i^{part,scaf}$ is the artificial index of this full structure's scaffold box, $\hat{C}_i^{part,R}$ is the set of R-group substitutes, T^{part} is the prompt for this markush structure process step, and $\hat{K}^{full}$ is the number of combination of Markush scaffolds and R-group substitutes. Each element in $\hat{C}_i^{part,R}$ is the artificial index of a 2D R-group depiction or the 1D textual representation of R-group (IUPAC name, abbreviation, or chemical formula).

Substitutes Processing and Full Structure Conversion. For explicit full structure, we can easily obtain the ligand coreference-SMILES pair (c_i^{full}, l_i^{full}) by replacing artificial index $\hat{c}_i^{full}$ in coreference recognition result $(c_i^{full}, \hat{c}_i^{full})$ as its SMILES. Then for Markush structures $(c_i^{part}, \hat{c}_i^{part,scaf}, \hat{C}_i^{part,R})$, substituents in $\hat{C}_i^{part,R}$ are converted to SMILES: IUPAC names via OPSIN[34], and abbreviations or formulas via a manually curated mapping table assisted by Gemini. RDKit[33] then combines scaffold and substituent SMILES to yield (c_i^{part}, l_i^{part}). Both explicit structure and Markush structure sets are merged into the final output (c_i, l_i).

Bioactivity Measurement Extraction Agent

The agent A^{measr} is design to extract bioactivity measurement data (p, c, v) from text and vision data using MLLMs:

$$(p_i, c_i, v_i) = A^{measr}(D^{text}, D_{page}^{vision}), i \in \{0, 1, ..., K^{measr} - 1\}, \tag{8}$$

where K^{measr} is the number of bioactivity measurement. Gemini is utilized to processes text and visual data separately:

$$(p_i^{text}, c_i^{text}, v_i^{text}) = Gemini(D^{text}, T^{text}), i \in \{0, 1, ..., K^{measr,text} - 1\}, \tag{9}$$

$$(p_i^{vision}, c_i^{vision}, v_i^{vision}) = Gemini(D_{page}^{vision}, T^{vision}), i \in \{0, 1, ..., K^{measr,vision} - 1\}, \tag{10}$$

where T^{text} and T^{vision} are prompts designed for text and vision modality respectively. As both outputs share a consistent format, they are merged into (p_i, c_i, v_i).

Data Postprocessing Agent

The data postprocessing agent A^{post} integrates bioactiviy measurements (p_i, c_i, v_i) and full chemical structures (c_i, l_i) via ligand coreference, yielding triplets (p_i, l_i, v_i). To enhance utility for downstream tasks (e.g., bioactivity prediction), beyond the literature itself, A^{post} enriches these with protein and complex structure data from external databases. To this end, beyond this triplet, A^{post} further enriches the bioactivity data with protein structure information or even protein-ligand complex structure, providing actionable information for downstream analysis.

Protein Structure Retrieval. A^{post} enriches the protein structure information by searching protein structure database based on extracted protein name p_i. It leverages Gemini's knowledge to filter invalid entries, then queries PDB[12] and UniProtKB[27] APIs to retrieve experimental co-crystal or AlphaFold2-predicted structures[49].

Structure-Bioactivity Annotation. If the document has a corresponding PDB complex structure, A^{post} attempts to identify the complex structure's bioactivity data point within the extracted set. It compares the SMILES string of the ligand in the PDB entry to the SMILES strings of all extracted ligands. Exact matches (ligand similarity = 1.0) trigger protein name comparison using Gemini, which selects out the corresponding bioactivity data based on PDB titles and extracted protein names p_i. The final output of A^{post} is an enriched set of bioactivity data triplets, including protein structure information and, when available, links to PDB complex structures.

Overall,* BIOMINER *synergistically combines MLLMs' reasoning, DSMs' domain expertise, and DTs' precise rules to deliver a scalable, zero-shot solution for bioactivity data extraction, requiring no task-specific training.

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. 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).
3. Mastropietro, A., Pasculli, G. & Bajorath, J. Learning characteristics of graph neural networks predicting protein-ligand affinities. *Nat. Mac. Intell.* **5**, 1427–1436 (2023).
4. Lan, T. *et al.* Generating mutants of monotone affinity towards stronger protein complexes through adversarial learning. *Nat. Mac. Intell.* **6**, 315–325 (2024).
5. Zhang, Z., Shen, W., Liu, Q. & Zitnik, M. Efficient generation of protein pockets with pocketgen. *Nat. Mac. Intell.* **6**, 1382–1395.
6. Feng, B. *et al.* A bioactivity foundation model using pairwise meta-learning. *Nat. Mac. Intell.* **6**, 962–974 (2024).
7. Zheng, S., Li, Y., Chen, S., Xu, J. & Yang, Y. Predicting drug–protein interaction using quasi-visual question answering system. *Nat. Mach. Intell.* **2**, 134 – 140 (2019).
8. Gentile, F. *et al.* Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. *Nat. Protoc.* **17**, 672 – 697 (2022).
9. 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).
10. 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).
11. Liu, T. *et al.* Bindingdb in 2024: a fair knowledgebase of protein-small molecule binding data. *Nucleic Acids Res.* **53**, D1633–D1644 (2025).
12. Liu, Z. *et al.* Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics* **31 3**, 405–12 (2015).
13. Kim, S. *et al.* Pubchem substance and compound databases. *Nucleic Acids Res.* **44**, D1202 – D1213 (2015).
14. 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).
15. 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).
16. 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).
17. 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).
18. Simmons, E. S. Markush structure searching over the years. *World Pat. Inf.* **25**, 195–202 (2003).
19. Su, M. *et al.* Comparative assessment of scoring functions: The casf-2016 update. *J. Chem. Inf. Model.* **59 2**, 895–913 (2018).
20. Dunbar, J. B. *et al.* Csar data set release 2012: Ligands, affinities, complexes, and docking decoys. *J. Chem. Inf. Model.* **53**, 1842 – 1852 (2013).
21. 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).
22. Chemdiv, available: https://www.chemdiv.com/ (ChemDiv, 2023).
23. Enamine, available: https://enamine.net/ (Enamine, 2023).
24. 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).
25. Wang, B. *et al.* Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839* (2024).

26. Reid, M. *et al.* Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

27. Boutet, E. *et al.* Uniprotkb/swiss-prot, the manually annotated section of the uniprot knowledgebase: How to use the entry view. *Methods molecular biology* **1374**, 23–54 (2016).

28. Váradi, M. *et al.* Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439 – D444 (2021).

29. Velankar, S. *et al.* Pdbe: Protein data bank in europe. *Nucleic Acids Res.* **39**, D402 – D410 (2010).

30. Xu, Y. *et al.* Molminer: You only look once for chemical structure recognition. *J. Chem. Inf. Model.* **62**, 5321 – 5328 (2022).

31. Khanam, R. & Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725* (2024).

32. Fang, X. *et al.* Molparser: End-to-end visual recognition of molecule structures in the wild. *arXiv preprint arXiv:2411.11098* (2024).

33. Landrum, G. *et al.* rdkit/rdkit: 2021_03_4 (q1 2021) release (2021).

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

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

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

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

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

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

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

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

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

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

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

46. 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).

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

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

49. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* **630**, 493 – 500 (2024).

50. Rajan, K., Brinkhaus, H. O., Agea, M. I., Zielesny, A. & Steinbeck, C. Decimer. ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nat. Commun.* **14**, 5045 (2023).

51. Qian, Y. *et al.* Molscribe: robust molecular structure recognition with image-to-graph generation. *J. Chem. Inf. Model.* **63**, 1925–1934 (2023).

52. Chen, Y. *et al.* Molnextr: a generalized deep learning model for molecular image recognition. *J. Cheminformatics* **16** (2024).

53. O'Boyle, N. M. *et al.* Open babel: An open chemical toolbox. *J. Cheminformatics* **3**, 33 – 33 (2011).

54. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. S. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *ICLR'23* (2023).

55. Stärk, H., Ganea, O.-E., Pattanaik, L., Barzilay, R. & Jaakkola, T. Equibind: Geometric deep learning for drug binding structure prediction. In *ICML'22* (2022).

56. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *ICLR '15* (2015).

57. Shi, C. *et al.* Discovery of nlrp3 inhibitors using machine learning: Identification of a hit compound to treat nlrp3 activation-driven diseases. *Eur. J. Medicinal Chem.* **260**, 115784 (2023).

58. Friesner, R. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Medicinal Chem.* **47**, 1739–1749 (2004).

59. Valdés-Tresanco, M. S., Valdés-Tresanco, M. E., Valiente, P. A. & Moreno, E. gmx_mmpbsa: a new tool to perform end-state free energy calculations with gromacs. *J. Chem. Theory Comput.* **17**, 6281–6291 (2021).

60. Burley, S. K. *et al.* Updated resources for exploring experimentally-determined pdb structures and computed structure models at the rcsb protein data bank. *Nucleic Acids Res.* **53**, D564–D574 (2025).

61. Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Mol. Des.* **27**, 221–234 (2013).

62. Søndergaard, C. R., Olsson, M. H., Rostkowski, M. & Jensen, J. H. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p k a values. *J. Chem. Theory Comput.* **7**, 2284–2295 (2011).

63. Webb, B. & Sali, A. Comparative protein structure modeling using modeller. *Curr. Protoc. Bioinforma.* **54**, 5–6 (2016).

64. Schrödinger release 2021-4: Glide; schrödinger, llc: New york, ny (Schrödinger, LLC, 2021).

65. Huang, J. *et al.* Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).

66. Vanommeslaeghe, K. *et al.* Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *J. Comput. Chem.* **31**, 671–690 (2010).

67. Jo, S., Kim, T., Iyer, V. G. & Im, W. Charmm-gui: a web-based graphical user interface for charmm. *J. Comput. Chem.* **29**, 1859–1865 (2008).

Supplementary Contents

Supplementary Notes

- Supplementary Note 1 – Details of Six Evaluation Tasks of BIOVISTA.
- Supplementary Note 2 – Implementation Details of Employed Models and Hyperparameter Settings.
- Supplementary Note 3 – Implementation Details of Protein-ligand Bioactivity Prediction Model Training.
- Supplementary Note 4 – Details of Human-in-the-loop NLRP3 Bioactivity Data Collection.
- Supplementary Note 5 – Details of NLRP3 QSAR Model Training and Inhibitor Screening.
- Supplementary Note 6 – Details of PoseBusters Structure-bioactivity Annotation.
- Supplementary Note 7 – Prompts of Four Agents in BIOMINER.
- Supplementary Note 8 – Technical Tricks of BIOMINER.
- Supplementary Note 9 – Ablation Study of BIOMINER.

Supplementary Figures

- Supplementary Figure S1– Construction Process and Overall Statistics of Benchmark BIOVISTA.
- Supplementary Figure S2 – MolMiner Platform for Reviewing Molecule Detection and OCSR Result in Human-in-the-loop Workflow.
- Supplementary Figure S3 – Success Case Study of Chemical Structure Extraction and Bioactivity Data Extraction.
- Supplementary Figure S4 – Failed Case Study of Chemical Structure Extraction and Bioactivity Data Extraction.
- Supplementary Figure S5 – Protein Target Distribution of Extracted Data.
- Supplementary Figure S6 – Different Types of Markush Structures.
- Supplementary Figure S7 – Different Attachment Points of Markush Structures.
- Supplementary Figure S8 – The workflow of NLRP3 inhibitor screening.

Supplementary Tables

- Supplementary Table S1 – Evaluation Metrics and Data Statistics of BIOVISTA's 6 Evaluation Tasks.
- Supplementary Table S2 – Benchmarking Different Molecule Detection Models on BIOVISTA.
- Supplementary Table S3 – Benchmarking Different OCSR Models on BIOVISTA.
- Supplementary Table S4 – Ablations Study Results of BIOMINER.
- Supplementary Table S5 – QSAR Performance of Different Classification Methods
- Supplementary Table S6 – QSAR Performance of Different Regression Methods
- Supplementary Table S7 – MMPBSA Result of NLRP3 Inhibitors.

Supplementary Notes

Details of Six Evaluation Tasks of BioVista

BioVista consists of six evaluation tasks, including two end-to-end tasks and four component-level tasks. This section details the information on these six evaluation tasks.

End-to-end Tasks

Two end-to-end evaluation tasks, protein-ligand bioactivity extraction and structure-bioactivity annotation, are designed to measure the overall extraction ability of protein-ligand bioactivity extraction methods.

- Protein-ligand Bioactivity Extraction: Given a paper, extracting all protein-ligand bioactivity triplet data. This task contains 500 papers with 16,457 bioactivity data as testing set. Recall, Precision, and F1 score are calculated as metrics.
- Structure-bioactivity Annotation: Given a PDB structure and corresponding paper, annotating the structure with reported bioactivity data in the paper. This task contains 500 PDB structures and 500 corresponding papers as testing set. Top-N Recall is calculated as metrics.

Component-level Tasks

Four component-level tasks, molecule detection, OCSR, explicit full structure corefernece recognition, and Markush enumeration, are specially designed to provide a deep analysis of the core chemical structure extraction agent of BioMiner.

- Molecule Detection: Given a paper, detecting boundary boxes of all 2D molecule depictions. This task contains 11,212 boundary boxes from 500 papers as ground-truth labels. Average Precision (AP) score is calculated as metrics.
- Optical Chemical Structure Recognition: Given a 2D molecule depiction, converting the molecule image into 1D molecule SMILES. This task contains 8,861 molecule depictions and corresponding SMILES as ground-truth labels. Accuracy score is calculated as metrics.
- Explicit Full Structure Coreference Recognition: Given a DSM-augmented image with explicit full structures, recognizing structures' coreference within the image. This task contains 962 images with 5,105 explicit full structures from 500 papers as ground-truth labels. Recall, Precision, and F1 score are calculated as metrics.
- Markush Enumeration: Given a DSM-augmented image containing Markush structures, enumerating all Markush full structures and coreferences within the image. This task contains 355 images with 3,513 Markush scaffold-R group-coreference pairs from 500 papers as ground-truth labels. Recall, Precision, and F1 score are calculated as metrics.

Implementation Details of Employed Models and Hyperparameter Settings

BioMiner integrates Multi-modal Large Language Models (MLLMs), Domain-Specific Models (DSMs), and Domain Tools (DTs) in a tuning-free pipeline. Below are the specifics of the components employed.

General Multi-modal Large Language Models

The core MLLM utilized across all agents in BioMiner for tasks such as bioactivity measurement extraction, chemical structure coreference recognition, and Markush structure enemueration was Google's **Gemini-2.0-flash**. This model was selected based on internal evaluations balancing performance, cost, and API availability at the time of development. As presented in Table 1 of the main text, Gemini-2.0-flash offered best performance compared to other leading MLLMs tested, including OpenAI's GPT-4o and Anthropic's Claude-3.7-sonnet, while being significantly more cost-effective for large-scale processing. All interactions with MLLMs were performed via their respective official APIs using carefully engineered prompts designed for each specific sub-task within the extraction workflow.

Domain-specific Models

Specialized models were employed for tasks requiring high precision within specific domains (document parsing, molecule detection, OCSR).

MinerU[25]: Utilized for initial PDF document preprocessing. MinerU performs layout analysis, Optical Character Recognition (OCR), and text block ordering to accurately extract textual content, images, tables, and their spatial relationships. We used the publicly available implementation (GitHub) with its default settings, leveraging its underlying PDF-Extract-Kit models and rule-based refinements for high-fidelity document parsing.

MolParser[32]: Employed as the primary OCSR engine within the chemical structure extraction agent. MolParser is an end-to-end image-to-SMILES model trained via curriculum learning on diverse synthetic datasets, enabling it to recognize a wide range of chemical structures, especially Markush structures. We accessed MolParser's capabilities through its provided web API.

MolMiner[30]: Used primarily for the initial detection of molecular structure depictions within document images. MolMiner integrates deep learning models (e.g., YOLOv5) for object detection. We utilized the MolMiner software downloaded from its repository (GitHub) for this detection step. Detected image snippets were then passed to MolParser for OCSR.
YOLO[31]: While MolMiner was the primary detector, experiments were also conducted using standalone YOLO models for molecule detection as part of our component evaluation, confirming the effectiveness of deep learning-based detection. The YOLO model is available in GitHub.
DECIMER[50]: An open-source platform for the identification, segmentation and recognition of chemical structure depictions in the scientific literature. In this work, DECIMER is employed as baseline for molecule detecton and OCSR. The codes are downloaded from GitHub.
MolScribe[51]: An image-to-graph OCSR model that explicitly predicts atoms and bonds, along with their geometric layouts, to construct the molecular structure. In this work, MolScribe is employed as baseline for OCSR. The codes are downloaded from GitHub.
MolNexTR[52]: Another image-to-graph model fusing ConvNeXt and Vision Transformer architectures In this work, MolNexTR is employed as baseline for OCSR. The codes are downloaded from GitHub.

Domain Tools

Established cheminformatics tools were integrated for specific, rule-based chemical operations.
OPSIN[34]: Used within the chemical structure extraction agent to convert textual representations of R-groups (IUPAC names) into SMILES strings. This conversion is crucial for enumerating Markush structures when substituents are described textually rather than graphically. We utilized the web service available at API.
RDKit[33]: An open source toolkit for cheminformatics, including a collection of 2D and 3D molecular operations. In BioMiner, RDKit is employed for zipping Markush scaffold SMILES and R-group SMILES into full structure SMILES. The RDKit can be installed by its tutorial.

Implementation Details of Protein-ligand Bioactivity Prediction Model Training

This section details the training settings of protein-ligand bioactivity prediction model for the large-scale data collection experiment (Figure 3).
Overall Model. In this work, we implement a simple bioactivity prediction model to demonstrate the effectiveness of BioMiner-extracted dataset. Formally, a data entry from the extracted set can be denoted as $d = (p, l, v, \mathscr{P})$, where p is the protein name, l is the ligand SMILES, v is the bioactivity value, and $\mathscr{P}$ is the enriched protein 3D structure. We generate ligand 3D structure $\mathscr{L}$ based on ligand SMILES by RDKit[33]. 3D protein and 3D ligand are represented as molecular $\mathscr{G}^{\mathscr{P}}$ and $\mathscr{G}^{\mathscr{L}}$. Our model makes prediction based on protein graph $\mathscr{G}^{\mathscr{P}}$ and ligand graph $\mathscr{G}^{\mathscr{L}}$:

$$o^{\mathscr{P}} = GNN^{\mathscr{P}}(\mathscr{G}^{\mathscr{P}}), \tag{11}$$

$$o^{\mathscr{L}} = GNN^{\mathscr{L}}(\mathscr{G}^{\mathscr{L}}), \tag{12}$$

$$\hat{v} = MLP(o^{\mathscr{P}} || o^{\mathscr{L}}), \tag{13}$$

where $GNN^{\mathscr{P}}$ and $GNN^{\mathscr{L}}$ are two individual graph neural network (GNN) encoders for protein and ligand respectively. $o^{\mathscr{P}}$ and $o^{\mathscr{L}}$ are graph representation after encoding. $\hat{v}$ is the bioactivty predicted by a multi-layer perception (MLP) based on graph representation of ligand and protein. Four types of popular GNNs were used, including graph convolutional networks (GCN)[42], graph attention network (GAT)[39], AttentiveFP[41], and equivariant graph neural network (EGNN)[40].
Ligand Graph. The input ligand L is first represented as a ligand graph $\mathscr{G}^{\mathscr{L}} = (\mathscr{V}^{\mathscr{L}}, \mathscr{E}^{\mathscr{L}})$, where $\mathscr{V}^{\mathscr{L}}$ is the node set and node i represents the i-th atom in the ligand. Each node $v_i^{\mathscr{L}}$ is also associated with an atom coordinate $x_i^{\mathscr{L}}$ retrieved from the individual ligand structure $\mathscr{P}$ and an atom feature vector $h_i^{\mathscr{L}}$. The edge set $\mathscr{E}^{\mathscr{L}}$ is constructed according to the spatial distances among atoms. More formally, the edge set is defined to be:

$$\mathscr{E}^{\mathscr{L}} = \left\{ (i,j) : |x_i^{\mathscr{L}} - x_j^{\mathscr{L}}|^2 < cut^{\mathscr{L}}, \forall i,j \in \mathscr{V}^{\mathscr{L}} \right\}, \tag{14}$$

where $cut^{\mathscr{L}}$ is a distance threshold, and each edge $(i,j) \in \mathscr{E}^{\mathscr{L}}$ is associated with an edge feature vector $e_{ij}^{\mathscr{L}}$. The node and edge features are obtained by OpenBabel[53].
Protein Graph. For protein Graph $\mathscr{G}^{\mathscr{P}} = (\mathscr{V}^{\mathscr{P}}, \mathscr{E}^{\mathscr{P}})$, $\mathscr{V}^{\mathscr{P}}$ is the node set and the node i represents the i-th residue in the protein. Each node $v_i^{\mathscr{P}}$ is also associated with an C_α coordinate of the i-th residue $x_i^{\mathscr{P}}$ retrieved from the individual protein structure and a residue feature vector $h_i^{\mathscr{P}}$. The edge set $\mathscr{E}^{\mathscr{P}}$ is constructed according to the spatial distances among atoms. More formally, the edge set is defined to be:

$$\mathscr{E}^{\mathscr{P}} = \left\{ (i,j) : |x_i^{\mathscr{P}} - x_j^{\mathscr{P}}|^2 < cut^{\mathscr{P}}, \forall i,j \in \mathscr{V}^{\mathscr{P}} \right\}, \tag{15}$$

where $cut^{\mathscr{P}}$ is a distance threshold, and each edge $(i, j) \in \mathscr{E}^{\mathscr{P}}$ is associated with an edge feature vector $e_{ij}^{\mathscr{P}}$. The edge features are obtained following[54]. As for the node features, they are obtained following[55].

Datasets. For pre-training, the dataset is the BIOMINER-extracted dataset, comprising 67,953 protein-ligand-bioactivity data points extracted from 11,683 EJMC articles. For fine-tuning, the PDBbind v2016 refined set is employed, with 3,390 data for training and 377 data for validation. The final testing is conducted on PDBbind v2016 Core Set (N=285) and CSAR-HiQ Set (N=135).

Training Procedure. The models were trained using Adam[56] with an initial learning rate of 10^{-3} and an L_2 regularization factor of 10^{-6}. The learning rate was scaled down by 0.6 if no drop in training loss was observed for 10 consecutive epochs. The dropout value was set to 0.1 and batch size was set to 256 and 64 for pre-training and fine-tuning respectively. The mean squared error (MSE) between predicted and true log-transformed bioactivity values is employed as training loss for both pre-training and finetuning. For pre-training, the number of training epochs was set to 100 with an early stop of 20 epochs, while for fine-tuning, the number of training epochs was set to 1000 with an early stopping rule of 70 epochs if no improvement in the validation performance was observed.

Details of Human-in-the-loop NLRP3 Bioactivity Data Collection

This section elaborates on the Human-in-the-Loop (HITL) workflow used to curate NLRP3 bioactivity data (Figure 4).

Literature Identification. Relevant publications reporting NLRP3 inhibitor bioactivity were identified through keyword searches (e.g., "NLRP3 inhibitors") within journals covered by ChEMBL, supplemented by reviewing recent review articles on NLRP3. This yielded 85 primary research articles.

HITL Workflow. In the HITL workflow, human experts focus on review the intermediate results of chemical structure extraction agent and bioactivity measurement extraction agent. Firstly, molecule detection and OCSR results are carefully checked to ensure the accuracy of following full structure recognition. As shown in Figure S2, the check of this step is conducted in the MolMiner platform. Based on the corrected detected and recognized structures, human experts then check the results of full structure coreference recognition and Markush structure enumerating. This step is conducted by directly modifying the intermediate results (Figure S4 and S3). The final check is the bioactivity measurement, and the check process is similar to the check of full structure coreference recognition. Due to the clear and structured intermediate data, t he whole process will not cost too much effort, leading to the collection of 1,592 high-quality NLRP3 bioactivity data points from the 85 papers in approximately 26 person-hours.

Details of NLRP3 QSAR Model Training and Inhibitor Screening

To demonstrate the effectiveness of BIOMINER-collected data, we train QSAR models based on ChEMBL data and BIOMINER data respectively and compare their performance (Table S5 and S6).

Dataset. For training QSAR models, we only keep IC50 bioactivity data about IL-1β. After this filter, 503 ChEMBL data and 994 BIOMINER data remain. We then apply a data split based on the scaffold of the merge set of ChEMBL data and BIOMINER data with train:test ratio of 4:1. After remove duplicate data between ChEMBL data and BIOMINER data, we obtain a test set with 213 data for testing models trained on 436 ChEMBL data or 770 BIOMINER data.

Ensemble QSAR Models Training and Testing. We trained 2D QSAR models based on two molecule features, Extended-connectivity fingerprints (ECFP)[45] and Chemically Advanced Template Search (CATS)[46]. Both regression and classification models are trained. For regression, $-log10(IC50)$ is set as the training target, and the models we employed are Liner Regression, Ridge Regression, Lasso Regression, ElasticNet Regression, Support Vector, K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and Multi-layer Perceptron. For classification, 10uM is set the threshold to determine positive and negative samples. As a result, BIOMINER set contains 443 positive data and 327 negative data, and ChEMBL set contains 191 positive data and 245 negative data. Particularly, we follow[57] to randomly select ligand SMILES has not appeared in training set from the whole ChEMBL dataset as negative data to augment the positive and negative ratios of the training and testing sets for model training and testing to 1:4 and 1:128. Classification models we employed including Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, XGBoost, Multi-layer Perceptron, and K-Nearest Neighbors. Due to the limited training data, we create ensemble QSAR model by 5-fold cross-validation. The final prediction of the ensemble QSAR model is the average of 5 models yielded by 5-fold cross-validation.

Inhibitor Screening Based on the test results Table S5 and S6, models trained on BIOMINER data significantly better than trained on ChEMBL. We finally choose SVM model with ECFP feature and Random Forest model with CATS feature for further inhibitor screening. Two commercially available compound libraries were screened: ChemDiv (approx. 1.67 million compounds) and Enamine REAL Space (approx. 4.41 million compounds). Two ensemble models were used individually. The SVM ECFP model identified 794 and 1,743 positive data from ChemDiv and Enamine respectively. And Random Forest CATS model screened 510 and 637 respectively. The predicted active molecules were further filtered based on their ECFP similarity to the reported active molecules in the BIOMINER collection, using a similarity cutoff of 0.4, thereby yielding molecules

with novel structures. Glide-XP[58] was employed for subsequent docking simulations to model the interactions between the molecules and NLRP3, with the NLRP3 structure derived from the crystal structure 9GU4. Next, the molecules obtained after PAINS and drug-likeness filtering were clustered based on ECFP fingerprints. The binding modes of the molecules were manually inspected, and up to three representative molecules were selected from each cluster. Finally, molecular dynamics (MD) simulations were performed for 16 molecules, and the binding free energies were calculated using the end-state MM/PBSA method[59]. Each molecule was independently run three times.

System Preparation We collected 7 NLRP3 holo structures from PDB[60]: 7ALV, 7PZC, 8ETR, 8RI2, 8WSM, 8ZEM and 9GU4. All structures were prepared using the Protein Preparation wizard[61], including bond order assignment, hydrogen addition, repair of missing side chain atoms, generation of protonation states using PROPKA[62], and optimization of hydrogen bonding networks at pH = 7, followed by a restrained minimization which only allows hydrogen atoms movable. The final structures were obtained by modeling the missing loops with MODELLER software[63]. Notably, all crystallization additives were removed, while ADP was retained in the final structure. Molecular structures were prepared using Schrödinger's LigPrep module, initialized from SMILES representations. Preparation parameters included: (1) tautomer generation via Epik (pH 7.0 ± 1.0) and (2) stereoisomer enumeration (maximum 16 per compound). Default protonation states were maintained using both Epik and PROPKA.

Glide-XP Docking To perform ligand docking simulation, we used Glide[58] in the XP (extra precision) mode from Schrödinger suite[64] (2021-3 release). The docking grid box was centered on the ligand from the crystal structure 9GU4. The grid outer box size was defined as 16 Å+ 0.8 × diameter, where diameter represents the maximum interatomic distance within the ligand, whereas the inner box size was fixed to 15 Å. Halogen atoms were treated as hydrogen bond acceptors, while aromatic hydrogens were considered as potential hydrogen bond donors. All other docking parameters were set to default. For subsequent analysis of the docking results, only the top-scoring pose from Glide was retained.

Molecular Dynamics Simulations We used the GROMACS2021 software to perform MD simulations for docked structures of identified hits and collected PDB structures. TIP3P water model and CHARMM36m forcefield[65] were employed, and ligand/ADP were parameterized using CGenFF forcefield[66]. 0.15 M sodium and chloride ions were added to neutralize the system's charge. Energy minimization and equilibration followed the default protocol of CHARMM-GUI[67]. Subsequent MD production runs of 100 ns were independently conducted three times at a temperature of 310.15 K and 1 bar using V-rescale thermostat and isotropic Parrinello-Rahman barostat, respectively. The final MD trajectories were analyzed and visualized using PyMOL software. The last 10 ns frames were evenly taken to compute MM/PBSA by gmx_MMPBSA[59].

Details of PoseBusters Structure-bioactivity Annotation

This section details the workflow used to for the annotation of the PoseBusters benchmark[24] with corresponding bioactivity data (Figure 5)

Dataset. The PoseBusters benchmark dataset (version 2, comprising 308 data points) served as the target for annotation. Of the 308 complex structures in PoseBusters, 252 were associated with corresponding literature references. From these, 242 data points were utilized for analysis. The first 10 entries were excluded to allow familiarization with the extraction framework, enabling a comparison of extraction accuracy and efficiency between purely human annotators and BioMiner.

Automated Matching Strategy. For each publication associated with a PoseBusters PDB entry, BioMiner extracted all potential protein-ligand-bioactivity triplets. The SMILES string of the ligand in the PDB entry (sourced from the Protein Data Bank) was compared to SMILES strings extracted by BioMiner from the publication. Ligand similarity was quantified using the Tanimoto coefficient, calculated based on ECFP4 fingerprints generated with RDKit. Exact matches (Tanimoto coefficient = 1.0) were prioritized. Extracted bioactivity data points were then ranked by ligand similarity score. For candidates with a similarity score of 1.0, BioMiner selected the best match by comparing the protein name associated with the extracted bioactivity data to the structure title of the PDB entry, leveraging the MLLM Gemini.

HITL Process. An expert reviewer was provided with the PDB ID, the PDB ligand structure, and the associated publication. The BioMiner system presented a ranked list of extracted bioactivity triplets for the publication, with the top-ranked match highlighted. Source context, including text snippets and table excerpts, was readily accessible to support verification. The expert's task was to confirm the accuracy of the top-ranked match. If incorrect, the expert could select the appropriate bioactivity value from the ranked list. For structures lacking extracted data, the expert could quickly skip to the next entry. The average review time, approximately 2 minutes per publication, was determined by timing expert interactions across a representative subset of PoseBusters entries during the annotation process.

Fully Automated Performance Analysis. Figure 5 presents the extraction error and data type analyses, detailed below.

Error Analysis: Waterfall plots (Figure 5(b) and (c)) evaluate the following criteria:

- Found Bioactivity: Bioactivity data identified for the complex (correct or incorrect).
- Found all Bioactivity: All correct bioactivity data extracted, accounting for multiple experimental settings per structure.

- Protein Correct: Target protein correctly identified in extracted bioactivity data.
- Ligand Correct: Target ligand correctly identified in extracted bioactivity data.
- In the Appendix: Successfully extracted correct bioactivity located in the publication's appendix file.

Different Data: Figure 5(e) and (f) quantify the following data categories:

- No Extracted Bioactivity (69): Cases where BioMiner found no bioactivity data.
- No Extracted Success (68): Cases where BioMiner found no bioactivity, with manual checks confirming absence in 68/69 instances (one case had data only in the supporting information file).
- Struc. Match Bioactivity (58): Cases where at least one extracted bioactivity had a SMILES string exactly matching the PDB ligand (similarity = 1.0).
- Match Success (50): Subset of Structure-Matched Bioactivity where the correct bioactivity was among the matched data (ligand similarity=1.0).
- Auto. Select (46): Subset of the Struc. Match Bioactivity cases where BioMiner assume the correct bioactivty exists in the matched (ligand similarity=1.0) data and select the best match one based on protein name and structure tile of the PDB entry.
- Select Success (34): Subset of Auto. Selection where manual validation confirmed the selected bioactivity as correct.

Technical Tricks of BioMiner

Beyond the core architecture, two specific strategies are implemented in BioMiner to enhance its robustness and effectiveness.

Divide-and-conquer Strategy.

This strategy directly addresses limitations in MLLM processing capacity when handling dense visual information, particularly for chemical structure extraction (coreference recognition and Markush enumeration, see Figure 1(b)). Scientific publications often present multiple chemical structures on a single page, ranging from a few to several dozen. Our empirical analyses revealed a correlation: as the number of detected structure bounding boxes provided as input to the MLLM increases, the likelihood of errors in both identifying explicit full structure coreferences and correctly enumerating Markush structures also rises. This is likely due to challenges in context management and attention allocation within the MLLM when faced with numerous distinct visual elements requiring detailed analysis. To mitigate this, we employ a divide-and-conquer strategy tailored to the specific sub-task:

- **For explicit full structure coreference recognition:** Pages with a high density of detected structures undergo hierarchical partitioning. First, the page image is segmented into logical units (e.g., figures, tables) based on the layout analysis. If an individual segment (like a single figure panel) still contains numerous explicit structures, the MLLM processes the bounding boxes within that segment iteratively in smaller, more manageable subsets (typically 4 boxes). This adaptive partitioning reduces the computational load and attentional demands on the MLLM during each inference step, thereby improving the accuracy of linking structure depictions to their textual references.
- **For Markush structure enumeration:** For pages containing Markush structures, segmentation is also performed based on layout analysis, isolating figures or tables. However, no further sub-partitioning is performed, as it risks separating a Markush scaffold from its associated R-group definitions, rendering enumeration impossible. The MLLM processes Markush structures on a figure-by-figure basis, assuming that components of a single Markush definition are typically contained within the same figure panel identified by the layout analysis. This ensures semantic coherence is maintained during processing.

By adapting the input granularity based on the task and layout context, this strategy enables BioMiner to handle pages with high structural complexity effectively, enhancing the reliability of chemical structure extraction.

Prompt Updating Strategy.

The performance of MLLMs is highly sensitive to the quality and specificity of the instructional prompts provided. Given the nuanced requirements of extracting protein-ligand bioactivity data – involving cross-modal reasoning, specific formatting for chemical structures (SMILES), and precise identification of bioactivity values and units – crafting optimal prompts is crucial for BioMiner's success. We adopted an iterative prompt refinement methodology. This involved:

- **Initial Prompt Design:** Developing baseline prompts based on a detailed description of each sub-task (e.g., structure coreference, bioactivity measurement extraction, Markush enumeration).

- **Output Generation & Evaluation:** Applying these prompts within BIOMINER to process a development subset of documents (drawn from data similar to BIOVISTA). Evaluating the generated outputs against ground truth using both quantitative metrics and qualitative error analysis.
- **Analysis and Refinement:** Identifying patterns in errors, ambiguities in MLLM interpretation, or incomplete outputs. Modifying the prompts based on this analysis – adjustments included clarifying instructions, adding few-shot examples, refining output format specifications, or explicitly instructing the model on how to handle edge cases.
- **Iteration:** Repeating steps 2 and 3 until performance metrics on the development set stabilized or reached a desired threshold, indicating that the prompts were effectively guiding the MLLM towards accurate and consistent task execution.

Ablation Study of BIOMINER

To obtain a deep understanding of BIOMINER'design, we present a comprehensive ablation study. The ablation study results are shown in Table S4, and the details of each experimental setting are described below. The full BIOMINER system serves as the baseline for comparison.

- **w/o MinerU:** This ablation uses PyPDF2 rather than MinerU for pdf parsing. Without the performing of layout analysis, we can not isolate surrounding text of vision data (figure, tables) and can not use the divide-and-conquer strategy when extracting chemical structures.
- **w/o YOLO for Tables:** This ablation removes the employ of YOLO model for detecting molecules within segmented tables.
- **w/o 1D R-Group Processing:** This ablation removes the convertion of Markush 1D R-Groups (IUPAC, chemical formula, and abbreviation) to SMILES.
- **w/o Divide-and-conquer:** This ablation removes the divide-and-conquer strategy used in chemical structure extraction.
- **w/o Prompt Updating:** This ablation uses the initial prompt rather than updated prompt.

Prompt of Four Agents in BioMiner

Prompt for extracting bioactivity data from text data

System Message:
You are a scientific paper-reading assistant.

User Message:
Task: Extract all protein-ligand bioactivity data from the provided scientific literature text.

Instructions:

1. The following context is the text of a scientific literature. Extract protein-ligand bioactivity data from the context.

2. Protein-ligand bioactivity data consists of three components:
Protein: The name of the protein.
Ligand: The ligand coreference in the paper.
bioactivity: The bioactivity measurement, which includes:
(1) Type: One of "IC50", "Kd", or "Ki".
(2) Value: The numerical value of the bioactivity measurement.
(3) Unit: The unit of the bioactivity measurement (e.g., nM, μM, mM).

3. Output the extracted data in the following JSON format strictly:"json“‘["protein": "protein name", "ligand": "ligand coreference" , "bioactivity": "type": "ki/kd/ic50", "value": "bioactivity value", "unit": "bioactivity unit"]“‘"

4. Only consider protein-ligand bioactivity measurements of type "IC50", "Kd", or "Ki". Ignore other types of bioactivity measurements.

Text Context of the Paper: [Text context]

Expected Answer: "json“‘["protein": "", "ligand": "" , "bioactivity": "type": "", "value": "", "unit": "", ...]“‘"

Prompt for extracting bioactivity from figure and tables

System Message:
You are a scientific paper-reading assistant

User Message:
Task: Extract all protein-ligand bioactivity data from the provided image.

Instructions:

1. The provided image is from a scientific literature. Extract protein-ligand bioactivity data from the image.

2. Protein-ligand bioactivity data consists of three components:
Protein: The name of the protein.
Ligand: The ligand coreference in the paper.
bioactivity: The bioactivity measurement, which includes:
(1) Type: One of "IC50", "Kd", or "Ki".
(2) Value: The numerical value of the bioactivity measurement.
(3) Unit: The unit of the bioactivity measurement (e.g., nM, μM, mM).

3. Output the extracted data in the following JSON format strictly:"json“‘["protein": "protein name", "ligand": "ligand coreference" , "bioactivity": "type": "ki/kd/ic50", "value": "bioactivity value", "unit": "bioactivity unit"]“‘"

4. Only consider protein-ligand bioactivity measurements of type "IC50", "Kd", or "Ki". Ignore other types of bioactivity measurements.

[Image of a PDF page]

Expected Answer: "json“‘["protein": "", "ligand": "" , "bioactivity": "type": "", "value": "", "unit": "", ...]“‘"

Prompt for extracting explicit full molecule coreference

System Message:
You are a scientific paper-reading assistant

User Message:
Task: Recognizing the identifier for every molecule within the plotted red box in provided picture.

Instructions:

1. This picture is a page of a scientific paper.

2. In the figure, there are some 2D molecule structures. We plot red boxes for detected molecules. Each red box has an box index (red number) on the left top corner.

3. For each 2D molecule structure, their usually exist identifier as the coreference for this moleucle in the paper.

4. We already have employed optical molecule structure recognition model to recognize the SMILES of the detected molecules.

5. At this time, we want align the molecule SMILES and identifier in the paper by the box index.

6. Therefore, you need to recognize the identifier for each detected molecule (in the red box).

7. Please outputing the results in this json format: "json“‘["index": "box index of the detected molecule", "identifier": "coreference of the molecule in the paper"]“‘"

8. The box index of detected molecules are: <index></index>. Therefore, in the result, the box index must be in them.

[Image of a PDF page]

Expected Answer: "json“‘["index": "", "identifier": ""]“‘"

Prompt for enumerating Markush structures

System Message:
You are a scientific paper-reading assistant

User Message:
Task: Extracting R-group for each molecular scaffold.

Instructions:

1. Identify Scaffolds and R-groups:

Locate the red boxes in the image and note the artificial index (red number) in the top-left corner. Classify each structure in the red box as either a molecular scaffold or an R-group. For each scaffold, identify all associated R-groups. A complete molecule consists of one scaffold and one or more R-groups.

2. R-group Naming:

Name R-groups based on their position (e.g., R1, R2, R3, X, Y) as indicated in the image or table. If an R-group is not in a red box, provide its chemical expression.

3. Output Format:
Use the following JSON format for each scaffold and its R-groups: "json“‘["scaffold": "artificial index of scaffold","R-group": "name of R-group1": "artificial index of R-group1 or chemical expression",...,"identifier": "compound identifier from the table or image"]“‘"

4. Additional Notes:

Ensure that the scaffold and R-group indices match those in the image. Map the compound identifiers from the table to the corresponding structures in the image.

5. For scaffolds and R-groups that are not text-format, their artifical index must in the following indexs: <index></index>

[Image of a PDF page]

Expected Answer: "json“‘["scaffold": "","R-group": "name of R-group1": "",...,"identifier": ""]“‘"

Supplementary Figures

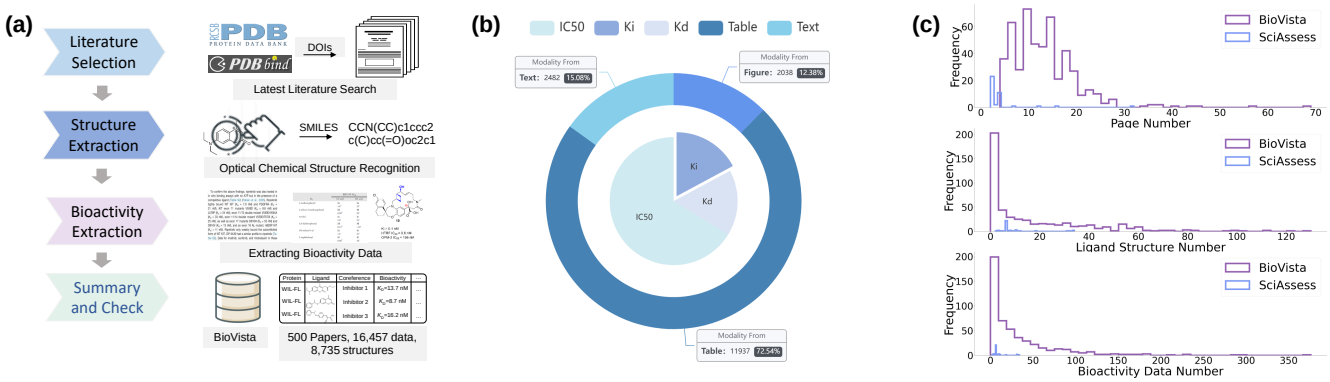


Figure S1. BIOVISTA construction process and analysis of data statistics. (a) The construction of BIOVISTA consists of four steps. (b) Distribution of bioactivity data. Only Ki, Kd, and IC50 data are collected in BIOVISTA. Most of data are come from tabular data. (c) Comparison between BIOVISTA and SciAssess, a previous dataset contains limited bioactivity data.

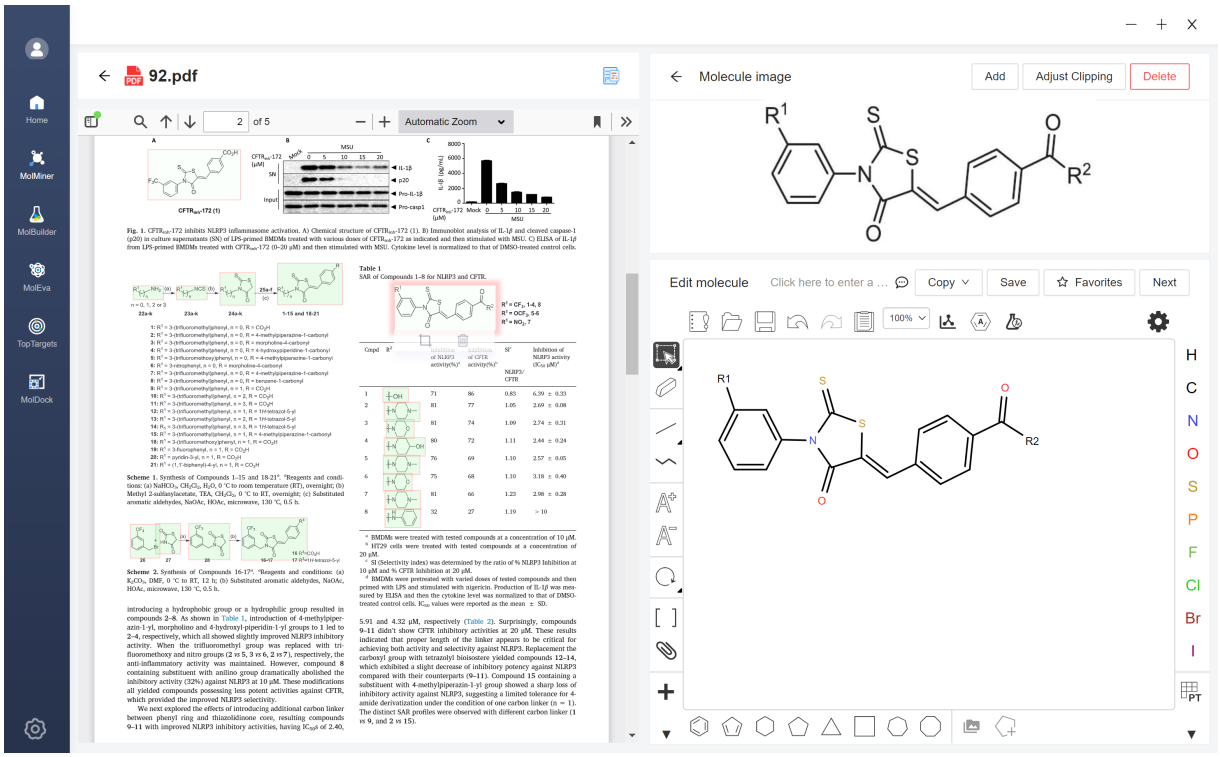


Figure S2. The interface of MolMiner platform. The molecule detection and OCSR results are checked in the MolMiner in the human-in-the-loop bioactivity data extraction workflow.

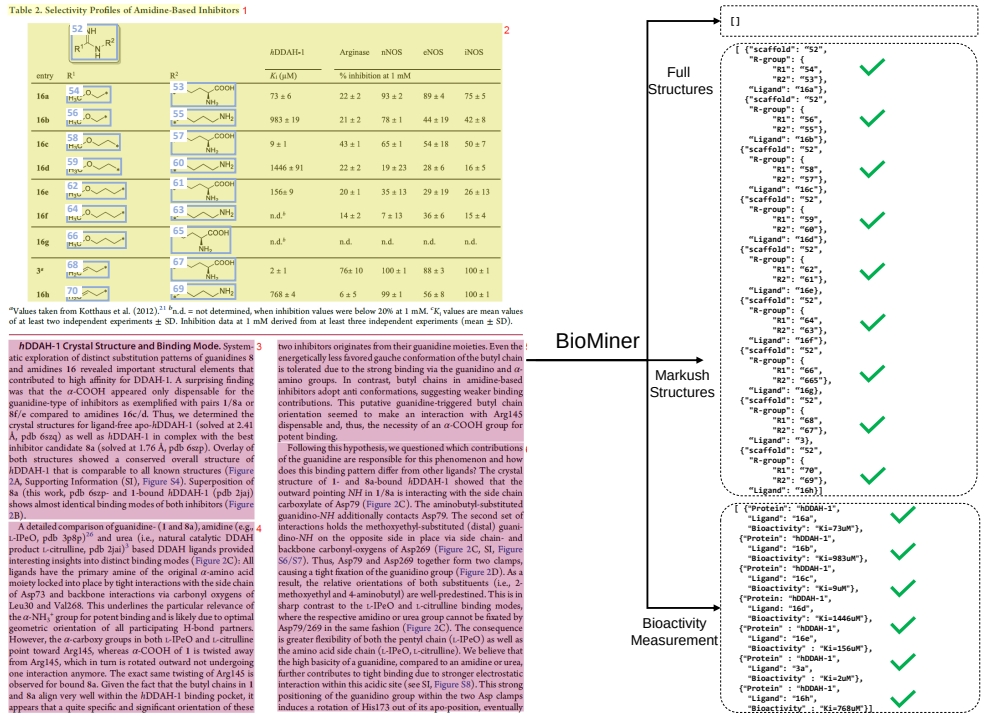


Figure S3. Success case study of BioMiner. (a) The input page for BioMiner with layout analysis result. Particularly, the indexes and color paint for layouts are only for visualization and are not included in the input page. (b) The extraction result of BioMiner. In this case, BioMiner extracts the bioactivity data and enumerates Markush structures successfully.

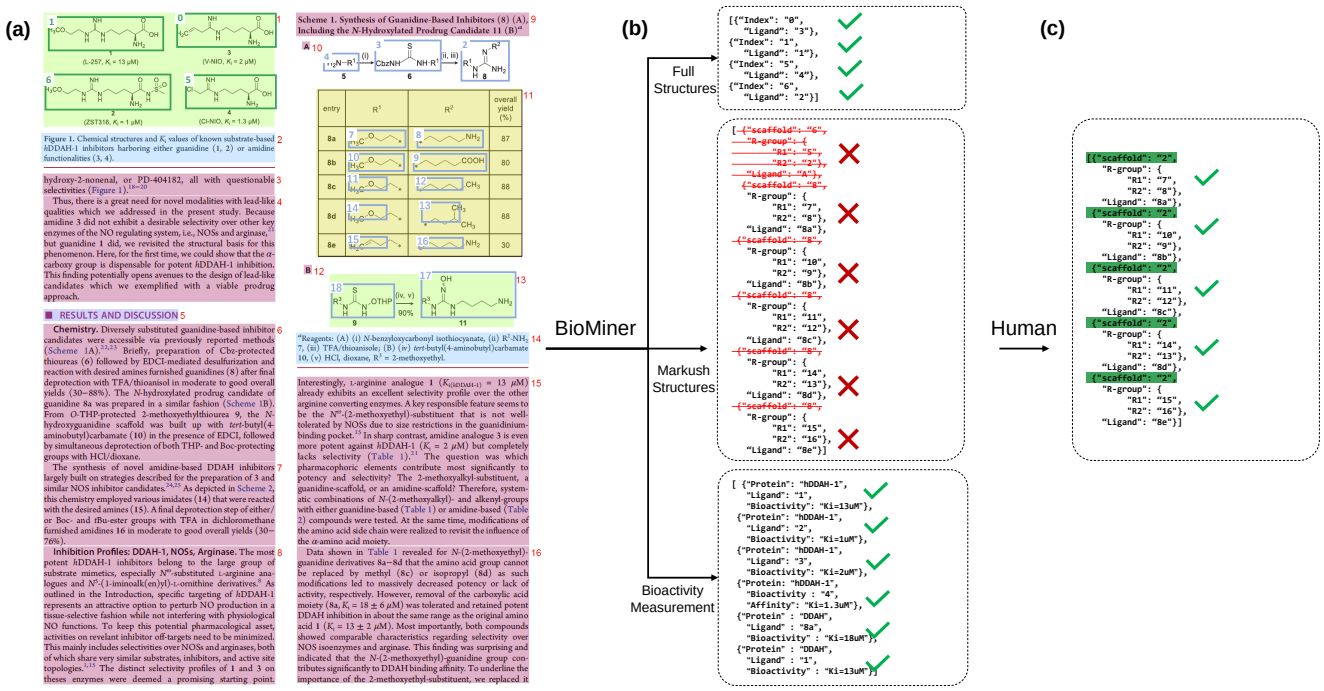


Figure S4. Failed case study of BioMiner. (a) The input page for BioMiner with layout analysis result. Particularly, the indexes and color paint for layouts are only for visualization and are not included in the input page. (b) The extraction result of BioMiner. In this case, BioMiner fails to determine the artificial indices of the Markush scaffold. (c) The corrected result. If in the human-in-loop workflow, human experts will check these intermideate results and correct error cases.

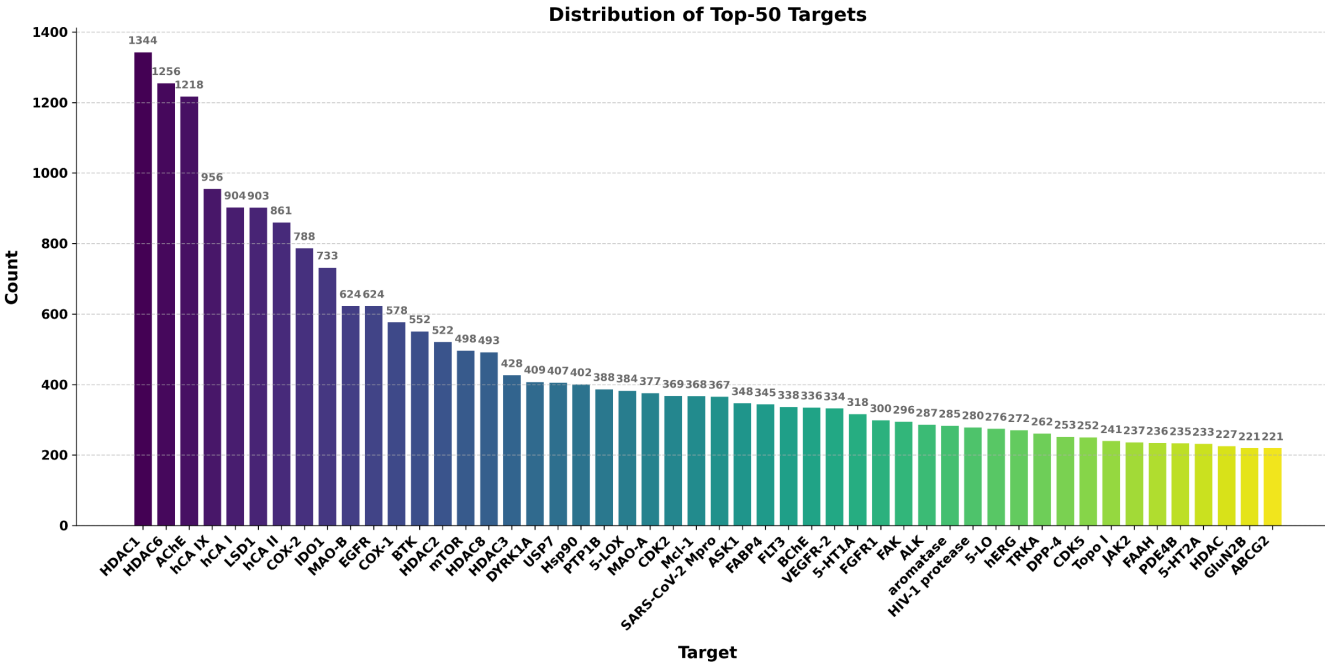


Figure S5. Distribution of Top-50 protein targets in 67,953 bioactivity data extracted from 11,683 papers.

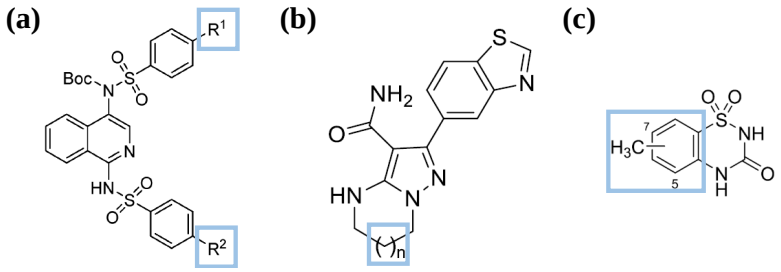


Figure S6. Different types of commonly used Markush structures. The Markush parts are marked in blue boxes. (a) Variable substitutions (R-groups). (b) Repeating substructures. (c) Variable attachment position. In this work, we focus on handle the first type Markush structure with variable substitutions.

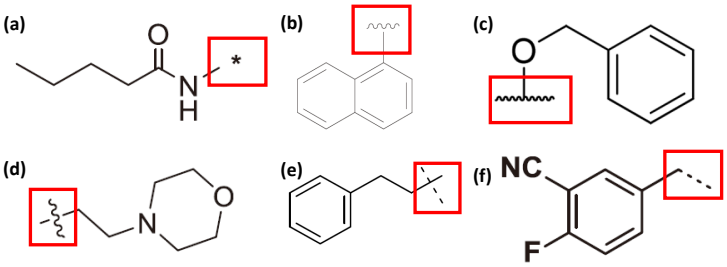


Figure S7. Different commonly used plotting styles of attachment points in Markush R-groups. Attchment points are marked in red boxes. These various plotting styles propose challenges for OCSR methods to accurately recognize attachment points, especially for type (b),(c),(d),(e), and (f).

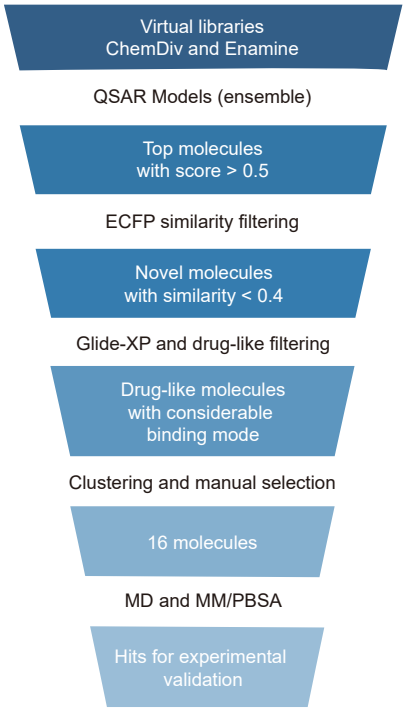


Figure S8. The workflow of NLRP3 inhibitor screening. The structures of 16 molecules are depicted in Table S7

Supplementary Tables

Table S1. Evaluation metrics, and data statistics of BIOVISTA's six evaluation tasks

Tasks	Metric	Input	Ground-truth Labels
Bioactivity Triplet Extraction	F1, Precision, Recall	500 Papers	16,457 bioactivity
Structure-Bioactivity Annotation	Recall@N	500 structure-paper pairs	500 structure-bioactivity pairs
Molecule Detection	Average Precision	500 Papers	11,212 boundary boxes
OCSR	Accuracy	8,861 2D molecule structure depictions	8,861 SMILES
Full Structure Coreference Recognition	F1, Precision, Recall	962 Augmented Images	5,105 full structure-coreference pairs
Markush Enumeration	F1, Precision, Recall	355 Augmented Images	3,513 Markush Scaffold-R Group-Coreference Pairs

Table S2. Performance of different molecule detections models on BIOVISTA. Different evaluation metrics are listed. The best results are **bolded**.

Metrics	YOLO	MolMiner	**BIOMINER**
mAP	0.648	**0.878**	0.857
AP_{50}	0.846	0.899	**0.929**
AP_{75}	0.752	0.876	**0.892**
AP-Small	0.000	0.000	**0.185**
AP-Large	0.720	**0.932**	0.922

Table S3. Performance of different OCSR models on BIOVISTA. Different chemical structure types are differentiated. Here the evaluation metric are accuracy. The best results is **bolded**.

Structure Types	MolMiner	MolScribe	MolNexTR	DECIMER	**MolParser**
Full	**0.774**	0.703	0.695	0.545	0.669
Chiral	**0.497**	0.481	0.419	0.326	0.352
Markush	0.185	0.156	0.045	0.000	**0.733**
All	0.507	0.455	0.401	0.298	**0.703**

Table S4. Ablation study of BIOMINER. Explanation of these variant are presented in previous section.

Method	Structure w/ Coreference			Bioactivity Measurement			Bioactivity Triplet			Bioactivity-Structure Matching		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Top-1	Top-3	Top-10
BIOMINER	0.440	0.454	0.447	0.594	0.486	0.535	0.244	0.195	0.217	0.346	0.544	0.686
w/o MinerU	0.428	0.440	0.434	0.609	0.462	0.525	0.240	0.179	0.205	0.314	0.508	0.708
w/o YOLO for Tables	0.442	0.461	0.451	0.594	0.486	0.535	0.242	0.193	0.215	0.350	0.548	0.698
w/o 1D R-Group Processing	0.412	0.461	0.435	0.594	0.486	0.535	0.225	0.178	0.199	0.334	0.524	0.668
w/o Divide-and-conquer	0.426	0.440	0.433	0.594	0.486	0.535	0.243	0.194	0.212	0.352	0.542	0.698
w/o Prompt Update	0.423	0.452	0.437	0.556	0.490	0.521	0.226	0.199	0.212	0.280	0.460	0.612

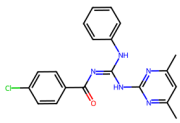
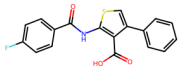
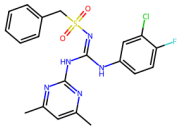
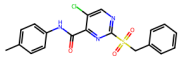
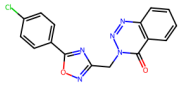
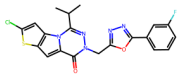
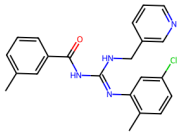
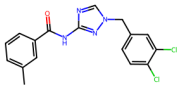
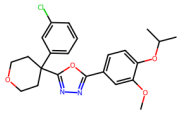
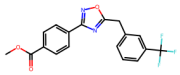
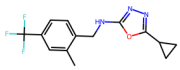
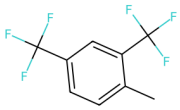
Table S5. Performance of classification QSAR models with different training data and features. The best results are **bolded**. Models trained on BIOMINER set achieves generally better performance compared to trained on ChEMBL dataset.

Models	Feature	Training Data	Acc	AUROC	BEDROC	EF1%	Precision	Recall	MCC
SVM	CATS	BIOMINER	0.9724	0.8763	0.6116	47.7778	0.1509	0.5532	0.2792
Random Forest	CATS	BIOMINER	**0.9967**	0.9493	**0.8528**	**76.4444**	**0.9350**	0.6117	**0.7548**
Logistic Regression	CATS	BIOMINER	0.9640	0.8884	0.6028	40.8765	0.1239	0.6011	0.2618
Decision Tree	CATS	BIOMINER	0.9728	0.8977	0.6684	55.2099	0.1722	0.6596	0.3280
XGBoost	CATS	BIOMINER	0.9949	**0.9530**	0.8329	71.1358	0.6684	0.6755	0.6694
Multi-layer Perceptron	CATS	BIOMINER	0.9915	0.9297	0.7980	69.0123	0.4662	0.6968	0.5659
K-Nearest Neighbors	CATS	BIOMINER	0.9293	0.9029	0.6799	49.3704	0.0784	**0.7553**	0.2289
SVM	CATS	ChEMBL	0.9779	0.8189	0.5284	39.2840	0.1651	0.4574	0.2658
Random Forest	CATS	ChEMBL	**0.9947**	0.8413	**0.6089**	**52.0247**	**0.8315**	0.3936	**0.5700**
Logistic Regression	CATS	ChEMBL	0.9659	0.8453	0.5208	37.1605	0.1166	**0.5160**	0.2339
Decision Tree	CATS	ChEMBL	0.9766	0.7986	0.4780	35.5679	0.1523	0.4415	0.2499
XGBoost	CATS	ChEMBL	0.9915	**0.8544**	0.5636	45.1235	0.4509	0.4149	0.4283
Multi-layer Perceptron	CATS	ChEMBL	0.9844	0.8435	0.5866	48.3086	0.2526	**0.5160**	0.3541
K-Nearest Neighbors	CATS	ChEMBL	0.9370	0.7962	0.4901	33.4444	0.0649	0.5319	0.1698
SVM	ECFP	BIOMINER	**0.9982**	0.9774	**0.9377**	**90.2469**	**0.9000**	0.8617	**0.8797**
Random Forest	ECFP	BIOMINER	0.9968	0.9805	0.9240	84.9383	0.7707	0.8404	0.8032
Logistic Regression	ECFP	BIOMINER	0.9951	**0.9808**	0.9215	83.3457	0.6434	0.8351	0.7307
Decision Tree	ECFP	BIOMINER	0.9868	0.9354	0.8372	79.0988	0.3515	0.8245	0.5334
XGBoost	ECFP	BIOMINER	0.9968	0.9805	0.9240	84.9383	0.7707	0.8404	0.8032
Multi-layer Perceptron	ECFP	BIOMINER	0.9965	0.9736	0.9227	84.9383	0.7429	0.8298	0.7834
K-Nearest Neighbors	ECFP	BIOMINER	0.9803	0.9692	0.9195	82.8148	0.2724	**0.9202**	0.4949
SVM	ECFP	ChEMBL	0.9956	**0.9656**	0.8301	71.6667	0.8203	0.5585	**0.6749**
Random Forest	ECFP	ChEMBL	**0.9957**	0.9537	**0.8489**	**78.0370**	**0.8673**	0.5213	0.6705
Logistic Regression	ECFP	ChEMBL	0.9928	0.9446	0.7326	61.0494	0.5317	0.5798	0.5516
Decision Tree	ECFP	ChEMBL	0.9876	0.8470	0.6231	45.1235	0.3353	0.6064	0.4453
XGBoost	ECFP	ChEMBL	0.9951	0.9288	0.7331	62.6420	0.7329	0.5691	0.6434
Multi-layer Perceptron	ECFP	ChEMBL	0.9916	0.9287	0.7053	56.2716	0.4649	0.5638	0.5078
K-Nearest Neighbors	ECFP	ChEMBL	0.9861	0.9421	0.7936	65.8272	0.3205	**0.7074**	0.4705

Table S6. Performance of regression QSAR models with different training data and features. The best results are **bolded**

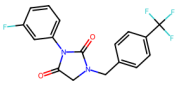
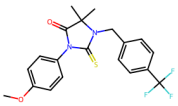
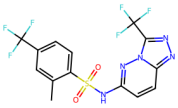
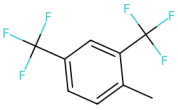
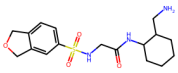
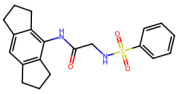
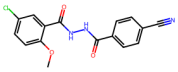
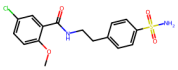
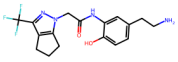
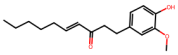
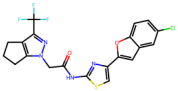
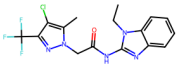
Models	Feature	Training Data	RMSE	MSE	Pearson	Spearman
Linear Regression	CATS	BioMiner	1.4463	2.0918	0.3296	0.2811
Ridge Regression	CATS	BioMiner	1.1529	1.3293	0.3766	0.3231
Lasso Regression	CATS	BioMiner	0.9706	0.9421	0.2862	0.3002
ElasticNet Regression	CATS	BioMiner	1.0126	1.0253	0.2981	0.3126
Support Vector	CATS	BioMiner	0.9288	0.8627	0.4449	0.4665
K-Nearest Neighbors	CATS	BioMiner	0.9435	0.8902	0.4456	0.4247
Decision Tree	CATS	BioMiner	0.9400	0.8836	0.4532	0.4813
Random Forest	CATS	BioMiner	0.8256	0.6816	0.5679	0.5636
Gradient Boosting	CATS	BioMiner	0.8437	0.7118	0.5599	0.5526
XGBoost	CATS	BioMiner	**0.7944**	**0.6310**	**0.5999**	**0.5947**
Multi-layer Perceptron	CATS	BioMiner	1.3748	1.8901	0.3359	0.3282
Linear Regression	CATS	ChEMBL	1.7879	3.1964	0.3035	0.1562
Ridge Regression	CATS	ChEMBL	1.3388	1.7924	0.3566	0.2753
Lasso Regression	CATS	ChEMBL	0.9589	0.9195	0.2645	0.3129
ElasticNet Regression	CATS	ChEMBL	1.0000	1.0000	0.2985	0.3470
Support Vector	CATS	ChEMBL	**0.9092**	**0.8266**	0.3849	0.4181
K-Nearest Neighbors	CATS	ChEMBL	1.1530	1.3294	0.2547	0.2685
Decision Tree	CATS	ChEMBL	1.0678	1.1403	0.4524	0.4308
Random Forest	CATS	ChEMBL	0.9716	0.9440	**0.4861**	**0.4458**
Gradient Boosting	CATS	ChEMBL	0.9856	0.9713	0.4466	0.3819
XGBoost	CATS	ChEMBL	0.9993	0.9986	0.4684	0.4406
Multi-layer Perceptron	CATS	ChEMBL	1.4640	2.1434	0.3144	0.2448
Linear Regression	ECFP	BioMiner	1.2769	1.6304	0.3123	0.3054
Ridge Regression	ECFP	BioMiner	0.8150	0.6643	0.5870	0.5447
Lasso Regression	ECFP	BioMiner	0.9821	0.9646	nan	nan
ElasticNet Regression	ECFP	BioMiner	0.9821	0.9646	nan	nan
Support Vector	ECFP	BioMiner	0.7256	0.5265	0.6855	0.6759
K-Nearest Neighbors	ECFP	BioMiner	0.7620	0.5807	0.6570	0.6303
Decision Tree	ECFP	BioMiner	0.7995	0.6391	0.5909	0.585
Random Forest	ECFP	BioMiner	0.7573	0.5734	0.6337	0.6181
Gradient Boosting	ECFP	BioMiner	0.7804	0.6090	0.6184	0.5916
XGBoost	ECFP	BioMiner	**0.7130**	**0.5084**	**0.6861**	**0.6514**
Multi-layer Perceptron	ECFP	BioMiner	1.0278	1.0564	0.5242	0.5294
Linear Regression	ECFP	ChEMBL	1.1000	1.2099	0.4037	0.3213
Ridge Regression	ECFP	ChEMBL	1.0406	1.0828	0.4430	0.3745
Lasso Regression	ECFP	ChEMBL	0.9637	0.9287	nan	nan
ElasticNet Regression	ECFP	ChEMBL	0.9637	0.9287	nan	nan
Support Vector	ECFP	ChEMBL	**0.8524**	**0.7265**	**0.5500**	**0.5232**
K-Nearest Neighbors	ECFP	ChEMBL	1.0046	1.0092	0.4758	0.4124
Decision Tree	ECFP	ChEMBL	1.0069	1.0138	0.4009	0.3567
Random Forest	ECFP	ChEMBL	0.9065	0.8217	0.4825	0.4284
Gradient Boosting	ECFP	ChEMBL	0.9363	0.8767	0.4839	0.4234
XGBoost	ECFP	ChEMBL	0.9548	0.9116	0.4759	0.4234
Multi-layer Perceptron	ECFP	ChEMBL	1.4290	2.0422	0.3152	0.3261

Table S7. MM/PBSA results of screened inhibitors. Experiments are independently performed three times to calculate ΔG(kcal/mol)(mean±SEM).

Type	Compounds	ΔG(kcal/mol)	Structure	QSAR Model	Max ECFP Similarity	Nearest Reported Structure
Hits	4477-1943	-33.46 ± 1.69		RF-CATS	0.186	
	6173-0372	-39.97 ± 0.19		RF-CATS	0.232	
	G851-0706	-32.08 ± 0.61		RF-CATS	0.221	
	L087-0119	-35.88 ± 2.95		RF-CATS	0.218	
	Z1166111962	-35.91 ± 1.48		RF-CATS	0.156	
	Z1798181064	-29.87 ± 0.94		SVM-ECFP	0.211	

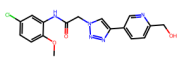
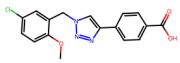
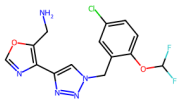
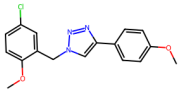
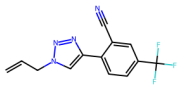
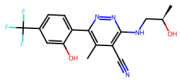
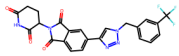
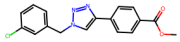
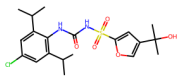
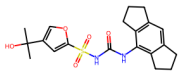
Continued on next page...

Table S7. MM/PBSA results (Continued)

Type	Compounds	ΔG(kcal/mol)	Structure	QSAR Model	Max ECFP Similarity	Nearest Reported Structure
	Z1868696094	-25.89 ± 0.97		RF-CATS	0.232	
	Z2038261672	-36.60 ± 1.27		SVM-ECFP	0.170	
	Z2440402593	-35.32 ± 1.66		SVM-ECFP	0.195	
	Z27873501	-33.34 ± 2.81		RF-CATS	0.303	
	Z4212990401	-28.28 ± 1.33		SVM-ECFP	0.189	
	Z5232931194	-43.33 ± 2.52		SVM-ECFP	0.188	

Continued on next page...

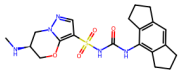
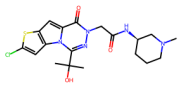
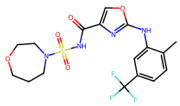
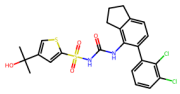
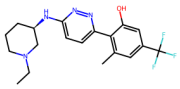
Table S7. MM/PBSA results (Continued)

Type	Compounds	ΔG(kcal/mol)	Structure	QSAR Model	Max ECFP Similarity	Nearest Reported Structure
	Z6739821465	-36.30 ± 2.73		SVM-ECFP	0.288	
	Z6739936901	-38.95 ± 0.02		SVM-ECFP	0.318	
	Z7765335303	-28.63 ± 0.70		SVM-ECFP	0.195	
	Z8148596279	-46.99 ± 2.48		SVM-ECFP	0.275	
	NP3-146 (7ALV)	-47.11 ± 3.38		-	-	-
	MCC950 (7PZC)	-37.76 ± 0.90		-	-	-

PDB

Continued on next page...

Table S7. MM/PBSA results (Continued)

Type	Compounds	ΔG(kcal/mol)	Structure	QSAR Model	Max ECFP Similarity	Nearest Reported Structure
	GDC-2394 (8ETR)	-52.44 ± 1.46		-	-	-
	NP3-562 (8RI2)	-36.75 ± 0.68		-	-	-
	Compound 32 (8WSM)	-44.28 ± 0.94		-	-	-
	SN3-1 (8ZEM)	-48.91 ± 3.90		-	-	-
	NP3-253 (9GU4)	-47.93 ± 0.71		-	-	-