Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/d5sc09883a

All publication charges for this article have been paid for by the Royal Society of Chemistry

TeLLAgent: a dual-agent framework for reliable scientific discovery with tool-enhanced LLMs

Jinyu Sun, [a] Jibin Zhou,[b] Huabei Wang,[a] Wei Liu,[a] Jun Yuan, [a] Yue Wang,[a] Ting Xie,[a] Lin Tan,[a] Hailiang Zhang,[a] Yingping Zou, [a] Zhimin Zhang *[a] and Hongmei Lu *[a]

Large language model agents hold immense promise for automating scientific discovery, yet their real-world application is hindered by an inability to reliably orchestrate tools and execute complex, multi-step plans without encountering hallucinations or logical inconsistencies. Here, we present TeLLAgent, a novel supervisor-executor dual-agent framework that explicitly separates strategic reasoning from precise tool operation to overcome these limitations. The global planning agent, powered by DeepSeek-R1, performs iterative chain-of-thought reasoning to decompose problems and formulate dynamic plans. The local execution agent, leveraging DeepSeek-V3.1, then accurately invokes a curated suite of 30 specialized tools. A critical self-correction loop, mediated by the Model Context Protocol, allows the system to "rethink" and "recover" from failures, significantly enhancing robustness. When rigorously benchmarked on a suite of complex tool-calling tasks, TeLLAgent significantly outperformed GPT-5 and existing agent frameworks, achieving higher success rates in multi-step planning and demonstrating superior scaling with task complexity. Furthermore, TeLLAgent drastically reduced factual hallucinations in knowledge retrieval, as validated by both human experts and LLM judges, underscoring its enhanced reliability. We ultimately demonstrate the power of this approach by deploying TeLLAgent for autonomous discovery in the demanding domain of organic solar cell materials. From a single natural language query, it executed an end-to-end workflow, from molecular design and property prediction to the identification of a high-performance quasi-macromolecular acceptor. This AI-designed molecule was subsequently synthesized and validated, achieving a power conversion efficiency of 16.44%. TeLLAgent establishes a new paradigm for building reliable, autonomous AI systems, proving its potential to accelerate scientific discovery in materials science, drug discovery, and beyond.

Received 17th December 2025
Accepted 19th May 2026

DOI: 10.1039/d5sc09883a

rsc.li/chemical-science

Introduction

Large language models (LLMs) have transcended their origins in natural language processing (NLP) to emerge as powerful engines for scientific discovery. Building upon their unprecedented capabilities in few-shot learning and cross-task generalization,[1,2] LLMs are now being explored as flexible, explainable interfaces for a range of scientific challenges, from protein folding prediction to molecular property estimation.[3,4] This potential is particularly transformative in chemical science, where the vast and complex design space makes LLM-driven solutions for accelerated molecule discovery both imperative and timely.

These emergent capabilities suggest that LLMs could serve as the central brains for autonomous scientific systems. However, the direct application of general-purpose LLMs in high-performance molecule discovery faces several fundamental challenges. First, the autoregressive, token-by-token generation paradigm struggles to capture the complex structural logic of molecules, often resulting in invalid structures or flawed reasoning when handling chemical representations like SMILES.[5,6] Second, the severe scarcity of large, high-quality labeled chemical datasets makes it impractical to rely solely on fine-tuning to inject domain knowledge, leaving LLMs with significant knowledge gaps.[7,8] Third, and most critically, without access to external tools and real-time knowledge, even fine-tuned LLMs lack the mechanism to perform precise domain-specific computations or verify facts, making their outputs prone to hallucinations, a fatal flaw in precision-demanding fields like material science.[9–11]

A promising pathway forward is the use of LLM-based autonomous agents, which extend the capabilities of standalone models by enabling interaction with external tools and environments through structured reasoning and action loops. Pioneering systems, such as ChatMOF,[12] ChemCrow,[13]

[a]College of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, P. R. China

[b]Dalian Institute of Chemical Physics, Chinese Academy of Science, Dalian, 116023, P. R. China. E-mail: zmzhang@csu.edu.cn; hongmeilu@csu.edu.cn

SciToolAgent,[14] AtomAgents,[15] and LLM-RDF,[16] have demonstrated the potential of these augmented models for tasks like knowledge retrieval and drug discovery. Nonetheless, these approaches remain constrained by limited multimodal integration, static knowledge bases, and insufficient support for multi-step tool invocation and collaboration.[17–19] Single-agent systems, which aim to integrate all capabilities within one LLM, often struggle with the competing demands of high-level strategic reasoning and low-level precise tool execution on complex real-world tasks.[20–23] While multi-agent frameworks have recently emerged to address some of these shortcomings through specialized role allocation and collaborative problem-solving.[24,25] A generalized, scalable, and highly adaptable agent architecture that explicitly decouples strategic planning from precise tool operation for molecular science has remained elusive.[26]

To address these gaps, we propose TeLLAgent, a Tool-enhanced LLM Agent framework built upon a novel supervisor-executor collaborative dual-agent architecture. This design is the cornerstone of our approach, directly tackling the core limitation of existing systems by explicitly separating complex, iterative reasoning from precise, localized tool execution. In this architecture, a global planning agent, powered by DeepSeek-R1, acts as the "supervisor", performing macro-strategy formulation and dynamic plan decomposition through chain-of-thought (CoT) reasoning. A local execution agent, leveraging DeepSeek-V3.1, then functions as the precise "executor", accurately invoking a curated suite of 30 specialized tools. This division of labor ensures robust performance on long-horizon, multi-step tasks. The framework further incorporates the Model Context Protocol (MCP) for robust tool integration and employs retrieval-augmented generation (RAG) and CoT reasoning to mitigate hallucinations and enhance response reliability.

We demonstrate the capabilities of TeLLAgent in the demanding domain of organic solar cells (OSCs), where it autonomously executes end-to-end workflows from molecular design and screening to device-performance prediction. Through rigorous human and LLM-based evaluation, we show that TeLLAgent outperforms GPT-5 and existing agent frameworks in knowledge retrieval, multimodal understanding, and factual accuracy. Finally, we demonstrate the extensibility of the framework by seamlessly adapting it to drug discovery, showcasing its potential as a universal AI-driven scientific innovation platform.

In this paper, we first present the TeLLAgent framework and its core components, detailing the supervisor-executor architecture and the integrated toolkit. We then rigorously benchmark its performance against state-of-the-art models, demonstrating superior accuracy and reduced hallucination. Subsequently, we showcase its application in autonomously discovering high-performance OSC materials, a process culminating in the synthesis and experimental validation of an AI-designed molecule. Finally, we highlight the cross-domain adaptability by applying it to a drug discovery task, screening for small-molecule inhibitors of a miRNA-protein complex.

Results

TeLLAgent framework

TeLLAgent is built upon a novel supervisor-executor dual-agent architecture, a design that explicitly decouples high-level strategic reasoning from low-level tool operation to overcome the limitations of single-agent systems in complex scientific domains. This division of labor ensures the system maintains a strategic overview while executing each step with precision, a critical factor for the success of long-horizon tasks.

The framework integrates three core components: Agents, LLMs, and a curated suite of Tools, orchestrated through a standardized MCP for seamless communication and context management. The Global Planning Agent, powered by DeepSeek-R1, serves as the strategic "supervisor", interpreting user queries, performing iterative chain-of-thought (CoT) reasoning to decompose complex problems into actionable sub-tasks, and supervising the precise execution of tasks. The Local Execution Agent, leveraging DeepSeek-V3.1, functions as the precise "executor", specializing in accurately invoking the tools of the framework. The functionality of TeLLAgent is extended through an extensible toolkit of 30 specialized tools (Fig. 1b), organized into four categories: molecular informatics, domain-specific applications, multimodal processing, and knowledge enhancement.

The operational workflow of TeLLAgent forms a self-correcting feedback loop that is fundamental to its autonomy (Fig. 1a). This "rethink-and-recover" mechanism operates as follows. The global agent first analyses a query and formulates a plan; the local agent then executes this plan by invoking tools. Critically, after the executor completes a tool call, the supervisor evaluates the result against the current sub-goal. If the result is unsatisfactory, the supervisor logs the failure and enters a "recovery mode". In recovery mode, the supervisor analyzes the failure reason and formulates a revised plan. This loop iterates until a satisfactory result is achieved, enabling dynamic recovery from failures without human intervention and ensuring robust autonomy.

To optimize performance, we implement advanced context engineering based on the MCP (Table S1). As a universal adapter between LLM agents and diverse tools, MCP offers a consistent interface that supports automated schema validation, error handling, and retry logic, distinguishing it from direct function calling. It also enables dynamic tool discovery, allowing the supervisor to query available tools at runtime, which is essential for long-horizon tasks where the exact tool sequence cannot be predetermined. This approach provides agents with a structured context comprising a system prompt that defines their role and capabilities, short-term and long-term memory, precise definitions of all available tools, and structured output constraints. Such curated context is critical for robust tool invocation and minimizing hallucination.

Evaluation and comparison

As illustrated in Fig. S1, TeLLAgent enhances knowledge retrieval by synthesizing results from multiple specialized

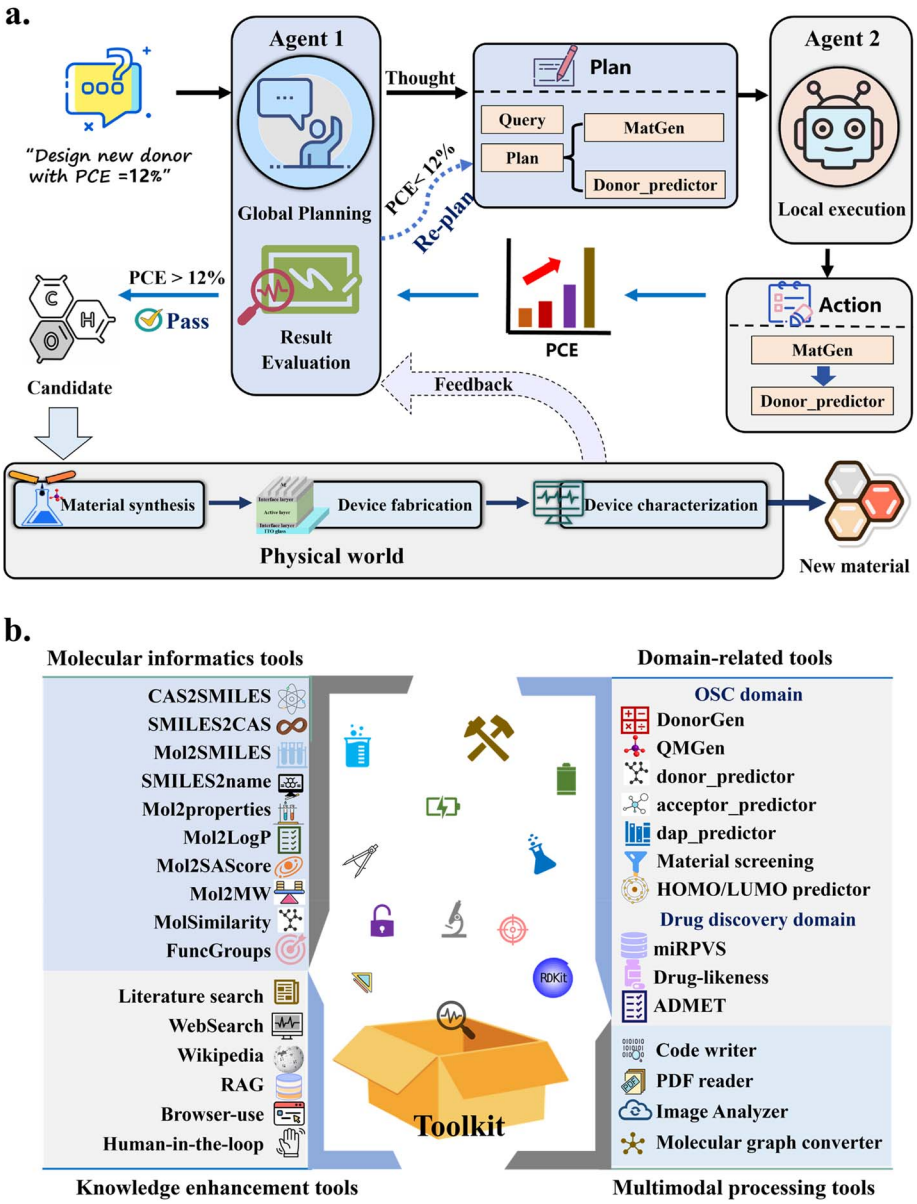


Fig. 1 The TeLLAgent framework and its integrated toolkit. (a) The supervisor-executor dual-agent architecture and its autonomous operation workflow. A global planning agent (supervisor, powered by DeepSeek-R1) interprets the user query and uses iterative CoT reasoning to formulate a strategic plan. This plan directs a local execution agent (executor, powered by DeepSeek-V3.1) to precisely invoke specialized tools. A critical self-correction loop enables the supervisor to evaluate results and dynamically re-plan or select alternative tools if the output is unsatisfactory, ensuring robust task completion. The framework supports human–AI collaboration for experimental guidance. (b) The curated suite of 30 specialized tools, organized into four categories, empowers TeLLAgent for scientific discovery. These tools provide capabilities in molecular informatics, domain-specific applications, multimodal processing, and knowledge enhancement, facilitating end-to-end tasks from knowledge acquisition and molecular design to property prediction and experimental validation.

knowledge enhancement tools and employing a chain-of-thought reasoning strategy, enabling cross-validation across diverse sources and substantially improving both the reliability and accuracy of information acquisition. To quantitatively benchmark TeLLAgent, we conducted a comprehensive evaluation against GPT-5-2025-08-07 and the recently published SciToolAgent across three critical dimensions, hallucination reduction, knowledge retrieval accuracy, and multimodal chemical information processing. A dual assessment framework comprising human experts (four material scientists) and LLM evaluators (Gemini 2.5 Pro and Claude 4 Sonnet) was used to conduct a double-blind review, ensuring objectivity and fairness. The evaluation prompts of the LLM evaluators are shown in SI note 1. As detailed in Table S2, twelve challenging questions spanning the OSC domain were designed by material experts.

We first investigated the propensity for hallucination, a critical failure mode for scientific models. The results revealed a stark contrast (Fig. 2a). The assessment rubrics of human experts are shown in Table S3. TeLLAgent achieved the highest

factual accuracy score (8.86), significantly reducing factuality hallucinations compared to GPT-5 (8.08) and SciToolAgent (8.57). It also maintained superior faithfulness (8.91), indicating that its responses are more strictly derived from the provided sources and query intent, whereas GPT-5 and SciToolAgent showed a more pronounced gap between their faithfulness (8.71, 8.73) and factuality scores. To ensure evaluation consistency, all human experts independently scored each response. The global Intraclass Correlation Coefficient (ICC) was used to evaluate inter-rater reliability among human experts. The overall ICC score is 0.812 (95% CI: [0.750, 0.870], $p = 1.28 \times 10^{-60} < 0.001$), indicating substantial agreement. The RAG-enhanced knowledge base and evidence-based reasoning of TeLLAgent proved decisive in mitigating these risks, ensuring its outputs are both factually correct and contextually faithful.

A critical finding emerged from this analysis that extends beyond the performance of TeLLAgent itself. The LLM evaluators (Gemini 2.5 Pro and Claude 4 Sonnet) failed to detect the severe factual hallucinations in the responses of GPT-5 on tasks 1, 5, and 6 (Fig. S2). This failure underscores the inherent risks of relying solely on LLM-based evaluators for validating specialized scientific content, as they may lack the precise domain knowledge to identify critical inaccuracies. This observation powerfully validates our decision to employ a dual human-LLM assessment framework and, more fundamentally, underscores the necessity of a tool-grounded, evidence-based approach of TeLLAgent to ensure scientific rigor.

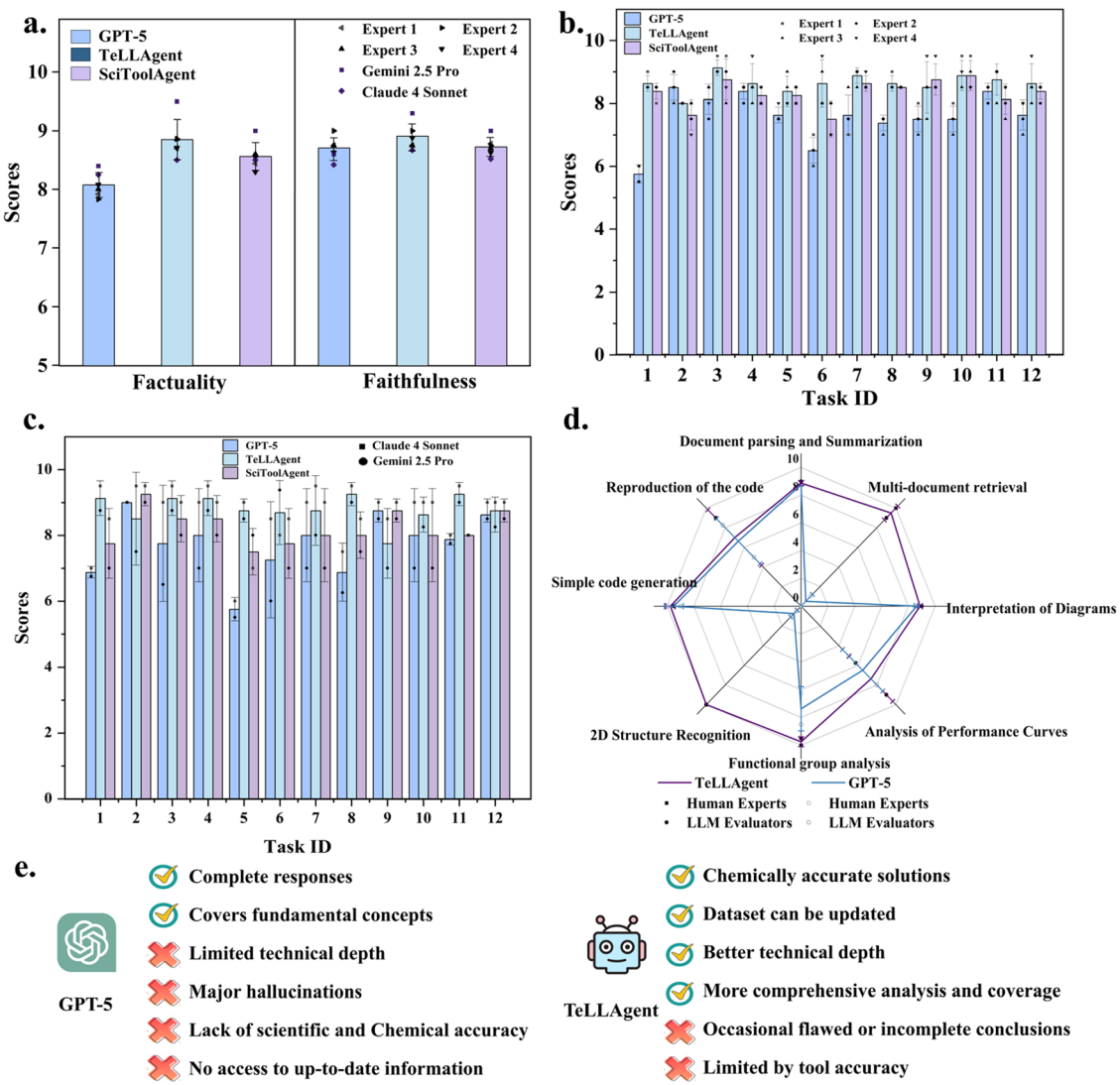


Fig. 2 Comprehensive benchmarking of TeLLAgent against GPT-5 and SciToolAgent. (a) Evaluation on twelve domain-specific knowledge retrieval tasks of factuality and faithfulness hallucinations (Table S3) by human expert material scientists and LLM evaluators. TeLLAgent demonstrates superior performance in mitigating both types of hallucinations. (b and c) Evaluation of the completeness, scientific rigor, and temporal relevance (Table S4) by human expert material scientists (b) and LLM evaluators (Fig. S3) (c). (d) Performance on eight multimodal chemical information processing tasks (Tables S5 and S6), including document parsing, diagram comprehension, and molecular structure recognition. (e) A synthesized summary of model capabilities and limitations, derived from the quantitative assessments in (a–d), highlighting the domain-specific technical depth and reduced hallucination rates of TeLLAgent.

We next specifically assessed the ability to retrieve and synthesize accurate, in-depth scientific knowledge. The dual assessment framework scored responses based on completeness, scientific rigor, and temporal relevance (Table S4 and Fig. S3). As evidenced in Fig. 2b and c, TeLLAgent consistently outperformed its counterparts. It achieved an average score of 8.63 from human experts, surpassing GPT-5 (7.57) and SciToolAgent (8.33). Similarly, in the LLM evaluator assessment, TeLLAgent scored 8.80, significantly higher than GPT-5 (7.74) and SciToolAgent (8.23).

Finally, we evaluated the models on eight multimodal chemical information processing tasks, including document parsing, diagram comprehension, molecular structure recognition, and code generation (Table S5). A direct comparison with SciToolAgent was not feasible for this benchmark, as it is architected primarily for text-based tool invocation and lacks native multimodal capabilities. As shown in Fig. 2d and Table S6, TeLLAgent demonstrates a clear advantage over the powerful multimodal baseline, GPT-5, with an average score of 8.65 compared to 6.08. It demonstrated particular strength in converting molecular imagery to SMILES strings and parsing complex, multi-document sources, capabilities that are critical for automated cheminformatics pipelines. This performance gap underscores the success of TeLLAgent in integrating and reasoning across diverse data modalities, a key advancement beyond text-centric agent frameworks. To evaluate the statistical significance of improvements, as shown in Table S7, a Friedman test was used and revealed highly significant overall differences among the models ($p < 0.001$), followed by pairwise Wilcoxon signed-rank tests with Benjamini–Hochberg correction for multiple comparisons. The results demonstrate that TeLLAgent significantly outperforms both GPT-5 and SciToolAgent (adjusted $p < 0.001$) in all evaluation metrics except faithfulness scores, confirming its effectiveness.

We also evaluated a version of the dual-agent framework powered entirely by GPT-5 and ChemCrow. As shown in Table S8, the dual-agent framework based on GPT-5 outperformed standalone GPT-5 and other common agents, and achieved nearly identical performance to TeLLAgent using DeepSeek models. This indicates that the dual-agent architecture, equipped with efficient tools, is the key factor in improving reliability. However, as shown in Fig. S5, TeLLAgent (DeepSeek) achieves a comparable score at a total cost of $0.11 for the 12-task benchmark, lower than standalone GPT-5 ($0.31) and TeLLAgent (GPT-5) ($0.63). As DeepSeek models are open-source and can be deployed locally at zero cost, they represent the most economical and reliable choice for academic and industrial research settings. Synthesizing these findings, Fig. 2e provides a holistic comparison of model capabilities. GPT-5 excels at addressing broad conceptual questions, achieving high scores in completeness for fundamental topics. However, its performance markedly declined when handling emerging theoretical frameworks and questions requiring professional depth. It exhibits high hallucination frequency, limited capacity for temporal knowledge updates, and offers superficial treatment of mechanistic details. In stark contrast, TeLLAgent exhibits a consistent and defining strength in domain-specific technical depth. It provides nuanced insights into the latest theoretical issues and research findings, underpinned by its ability to dynamically access and reason with specialized tools and current literature. This tool-enhanced approach is the direct cause of its significantly reduced hallucination rates and superior performance on complex, multimodal tasks.

This comparative study validates that TeLLAgent transcends the capabilities of a general-purpose LLM, evolving into a domain-specialized assistant. Its superiority is not merely incremental but qualitative, enabling a new level of precision and reliability in rapidly evolving research fields. The performance of TeLLAgent validates its core design. It achieves superior knowledge retrieval, multimodal understanding, and hallucination reduction through two key innovations: a supervisor-executor architecture that enables robust planning, and a tool-enhanced paradigm that ensures evidence-based, precise execution.

Ablation studies validate architectural choices

A pivotal aspect of our supervisor-executor architecture is the strategic pairing of LLMs with distinct capabilities to fulfill the specialized demands of the role of each agent. To create a comprehensive benchmark for evaluating these tool-calling abilities, we designed a suite of tasks of controlled complexity (Table S9). These tasks systematically varied the minimum number of sequential tool calls (1, 3, 5, 7, 9) and the size of the available toolset (5, 10, 15, 20, 25), creating a matrix of increasing difficulty for an extensive ablation study (Fig. 3).

We first validated the necessity of our dual-agent design. As shown in Fig. 3a, the collaborative framework significantly outperformed a single-agent counterpart across all task complexities. As shown in Table S10, the main root causes of the failures are planning and execution-related failures. The results reflect the difficulty in seamlessly switching between deciding what to do and actually doing it within a single model instance. To further demonstrate the generalizability of this architectural advantage beyond our final model choices, we tested the framework using GPT-4-0613 as a common backbone for both single and dual-agent setups. As shown in Fig. S4, the dual-agent configuration consistently enhanced the capabilities of GPT-4-0613, resulting in higher success rates in multi-step planning. This result confirms that the decoupling of planning from execution is a generally applicable and critical principle for successfully orchestrating complex tasks, independent of the underlying LLM.

We then systematically evaluated various LLM combinations for the two agent roles. The selection results for the global planning and local execution agents are shown in Fig. 3b and c, respectively. The combination of DeepSeek-R1 and DeepSeek-V3.1 has demonstrated the most robust multi-step task-solving capabilities. This outcome was not only an empirical victory but also a validation of our capability-driven selection strategy. The DeepSeek-R1 model, known for its strong reasoning and planning proficiencies, is inherently suited to the responsibility of global agent for iterative CoT reasoning and macro-strategy formulation. Conversely, the DeepSeek-V3.1

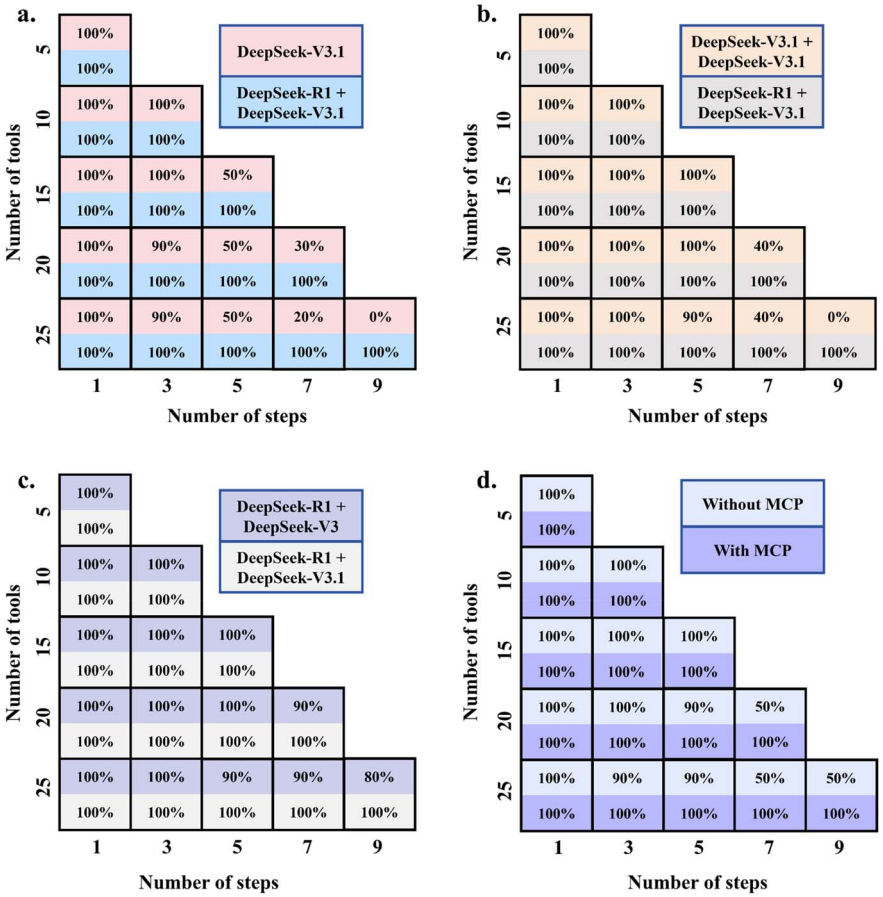


Fig. 3 Ablation studies validate key architectural design choices. (a) The dual-agent architecture (this work) robustly outperforms a single-agent baseline, confirming that separating planning from execution is critical for complex tasks. (b and c) LLM selection results for the global planner (b) and local executor (c) identify DeepSeek-R1 and DeepSeek-V3.1 as the optimal pairing, aligning model strengths with role demands. (d) MCP is essential for robust performance, as its removal severely degrades success rates. All evaluations systematically scale task complexity through the number of tool-calling steps and the toolset size.

model excels at precise instruction-following and structured output generation, making it an ideal executor for reliable tool invocation. This strategic pairing ensures that each specialized role is powered by the most appropriate underlying architecture.

Finally, an ablation study on the MCP underscored its importance as a core component (Fig. 3d). The full framework with MCP maintained high success rates. In contrast, the framework without MCP suffered significant performance degradation, highlighting the essential role of MCP in maintaining context and ensuring robust tool interoperability.

TeLLAgent for autonomous donor material discovery

To demonstrate the capacity of TeLLAgent for end-to-end autonomous discovery, we tasked the system with designing novel high-performance donor materials for organic solar cells starting from a single natural language query. As shown in Fig. 4a, a conditional donor generation model (DonorGen) was created. DonorGen employs an improved transformer decoder that conditions generation on a target PCE value. The PCE is embedded and concatenated with the token embeddings of the SELFIES string at each decoding step, allowing the model to steer the molecular structure toward the specified efficiency range. This approach is inspired by conditional text generation and has been adapted for molecular generation *via* SELFIES, which guarantees 100% validity. The swiGLU and RMSNorm are used to improve the novelty of generation results. The architecture and training details are shown in SI note 2 and Fig. S6.

DonorGen is used as a tool of TeLLAgent for autonomous donor material discovery. This process rigorously evaluates the framework's ability to independently formulate a strategy, orchestrate a sequence of specialized tools, and iteratively refine outputs through a self-correcting loop.

The process began with a user-specified target: "Design a donor material with a PCE greater than 15%". The global planning agent, serving as a strategic supervisor, analyzed the query and decomposed it into an executable plan. Its first action was to invoke DonorGen to produce a library of novel molecular structures generated *de novo*. Following this generation phase, TeLLAgent autonomously screened the library using the Donor_predictor tool to identify donors likely to achieve the target PCE. A multi-faceted evaluation cycle (Fig. 4b) was then initiated, in which the planning agent sequentially deployed a suite of physical and performance predictors, such as

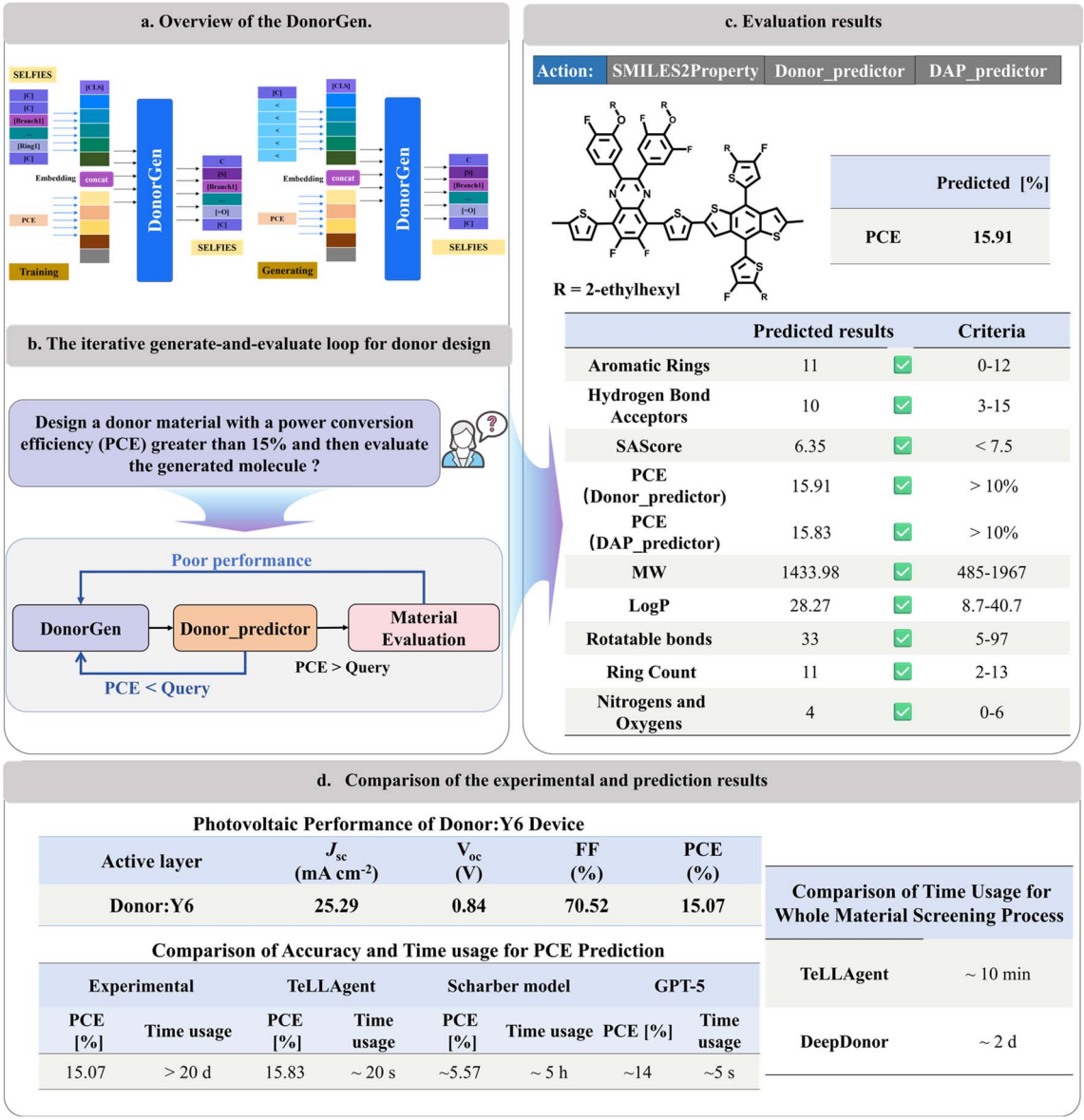


Fig. 4 TeLLAgent-enabled autonomous workflow for donor material design. (a) Overview of DeepDonor. The molecular SELFIES and PCEs are used as input during the training process. These tokens are embedded and concatenated as the input embedding of the model. The SELFIES of molecules are used as the target. The PCEs are used as constraints to generate molecules with desired properties during the generation process. The SELFIES of scaffolds together with properties are used for molecular optimization. (b) The autonomous "generate-and-evaluate" loop for *de novo* design of polymer donors. TeLLAgent generates candidate molecules based on a predefined target PCE (>15%). It then coordinates a suite of tools to rapidly evaluate the candidates' physical properties and performance metrics (*e.g.*, SAScore, log *P*, molecular weight, and PCE). (c) Multi-faceted evaluation results for the candidate donor that successfully passed the screening loop. All its properties meet the predefined criteria, and the predicted PCE of the donor: Y6-based device is 15.83%. (d) Comparison of experimental and prediction results. The discovered donor material, when fabricated into a device with Y6, achieved an experimental PCE of 15.07%. This result shows excellent agreement with the predicted value from TeLLAgent (15.83%), with an absolute error of only 0.76%. The whole autonomous donor discovery workflow of TeLLAgent demonstrates a significant efficiency advantage (~10 min) over the screening method (*e.g.*, DeepDonor, ~2 days).

synthetic accessibility (SAScore), key physicochemical properties (log *P*, molecular weight), and critical device performance metrics (DAP_predictor). This iterative generate-and-screen loop continued autonomously until a candidate satisfying all predefined criteria was identified. As shown in Fig. 4c, the PCE of the donor: Y6-based device is predicted to be 15.83%.

The robustness of the autonomous workflow was ultimately confirmed through experimental synthesis and device fabrication of the top-ranked donor material (Fig. 4d and Table S11). TeLLAgent exhibited remarkable predictive accuracy: for a device based on the well-known acceptor Y6, the model predicted a PCE of 15.83%, closely matching the experimental value of 15.07%, and yielding an absolute error (AE) of only 0.76%. This significantly outperformed recent machine learning models, GPT-5 (prediction: 14.0%, AE: 1.07%) and the widely used Scharber model based on predicted HOMO/LUMO

levels (prediction: 5.57%, AE: 9.5%). The results show that TeLLAgent can predict PCE with high accuracy in 20 seconds. This demonstration of human–AI collaborative-based autonomous discovery marks a key advance: TeLLAgent seamlessly integrates generative AI with precise evaluation tools to compress design cycles from days to minutes.

While TeLLAgent demonstrates strong capabilities in autonomous materials discovery, training the generative model (DonorGen) on PCE values predicted by DeepDonor inevitably introduces a degree of algorithmic error propagation. Furthermore, real-world device performance is highly sensitive to dynamic fabrication conditions that *in silico* models cannot perfectly capture. However, the generative framework can effectively map the chemical space and generate structurally novel materials bounded near the high-performance target distribution. These inherent computational and environmental margins of error underscore the fundamental necessity of subsequent screening and the human–AI collaborative mode. Integrating expert supervision into the automated loop remains critical to verify computational candidates, manually filter out anomalies, mitigate propagation errors, and ultimately maximize the experimental success rate. By autonomously translating a performance target into a synthetically viable, high-performance molecule with high predictive accuracy, it establishes a new paradigm for accelerated discovery of functional materials.

Human–AI collaboration for quasi-macromolecule acceptor discovery

Although full automation is powerful, many real-world scientific problems require the nuanced judgment of domain experts. To demonstrate the flexibility of our framework, we next engaged TeLLAgent in a human–AI collaborative mode to tackle a more complex challenge, the design of high-performance quasi-macromolecular acceptors (QMs), a promising but structurally intricate class of materials for OSCs. This experiment highlights how the framework seamlessly integrates human expertise into its autonomous reasoning loop, creating a synergistic partnership for exploration.

The workflow, as illustrated in Fig. 5a–c and S7, commenced when a researcher prompted TeLLAgent to design a high-efficiency quasi-macromolecular acceptor using specified π-bridges and acceptor units. The global planning agent first determined whether additional information was required by proactively invoking the "Human" tool. Once all necessary input is obtained, the agent immediately formulates a plan and activates the QMGen tool to construct candidate quasi-macromolecular structures using 6 distinct common-used π units, generating an initial set of 6 quasi-macromolecules.

Notably, the standard self-correction loop was enhanced by introducing a human-in-the-loop validation cycle. After about 2–3 cycles of generation and property prediction of candidate structures, the results were systematically presented to a materials expert for comprehensive evaluation. The expert assessed the candidates not only on predicted performance metrics but also on critical aspects beyond the model's immediate scope, including chemical feasibility, synthetic complexity, and structural novelty. Based on this assessment, the expert could either approve a candidate for further development or direct TeLLAgent to initiate a new generation cycle with refined design criteria. This iterative process of AI-driven generation and expert-guided curation continued until a promising acceptor candidate was successfully identified. To identify an optimal donor partner for the newly discovered acceptor, TeLLAgent autonomously screened an existing database containing 1285 experimental donors using its DAP_screen tool, presenting a ranked list of top-3 recommendations. The expert then made the final selection, considering practical experimental factors such as synthetic accessibility and material compatibility, thereby grounding the AI suggestions in real-world constraints.

The ultimate measure of a scientific AI framework lies in the accuracy of its predictions and their agreement with empirical reality. The effectiveness of this collaborative approach was confirmed by experimental validation. The synthesized acceptor material exhibited favorable optoelectronic properties. The absorption spectrum in the film state showed a distinct red shift compared to that in solution (Fig. 5d), indicating strong intermolecular π–π stacking, a crucial characteristic for high-performance organic photovoltaics. To rigorously benchmark the predictive fidelity of TeLLAgent, we systematically compared its predictions for AI-designed molecules, generated *via* both autonomous and collaborative workflows, against experimental measurements, as well as predictions from established computational methods and general-purpose LLMs (Fig. 5e and f).

We first evaluated the accuracy of electronic property predictions, a critical determinant of OSC performance. As summarized in Fig. 5e, for the quasi-macromolecular acceptor discovered through human–AI collaboration, TeLLAgent predicted HOMO and LUMO energy levels of −5.54 eV and −3.72 eV, respectively. These results were in close agreement with experimental values (−5.70 eV and −3.85 eV) and demonstrated comparable accuracy to resource-intensive density functional theory (DFT) calculations (−6.02 eV and −3.93 eV). In stark contrast, the predictions of GPT-5 (−5.65 eV and −4.15 eV) exhibited a larger deviation, particularly for the LUMO level. Crucially, TeLLAgent achieved this high accuracy in seconds, offering a speed-up of several orders of magnitude over DFT and experimental characterization.

The most critical benchmark is predicting device PCE. As shown in Fig. 5f, the performance of TeLLAgent was exceptional. For the donor material designed autonomously, it achieved an absolute prediction error of only 0.76%. For the collaboratively discovered acceptor, its prediction (16.07%) was remarkably close to the experimental outcome (16.44%). This consistent accuracy across different material classes and discovery modes underscores the robustness of its tool-enhanced prediction paradigm.

This stands in sharp contrast to the alternatives. As shown in Table S11 and Fig. 5f, the ML-based models and Scharber model proved inadequate for these complex materials, yielding a gross underestimation. GPT-5, despite its broad knowledge, produced an over-optimistic and inaccurate forecast (18.00%), highlighting

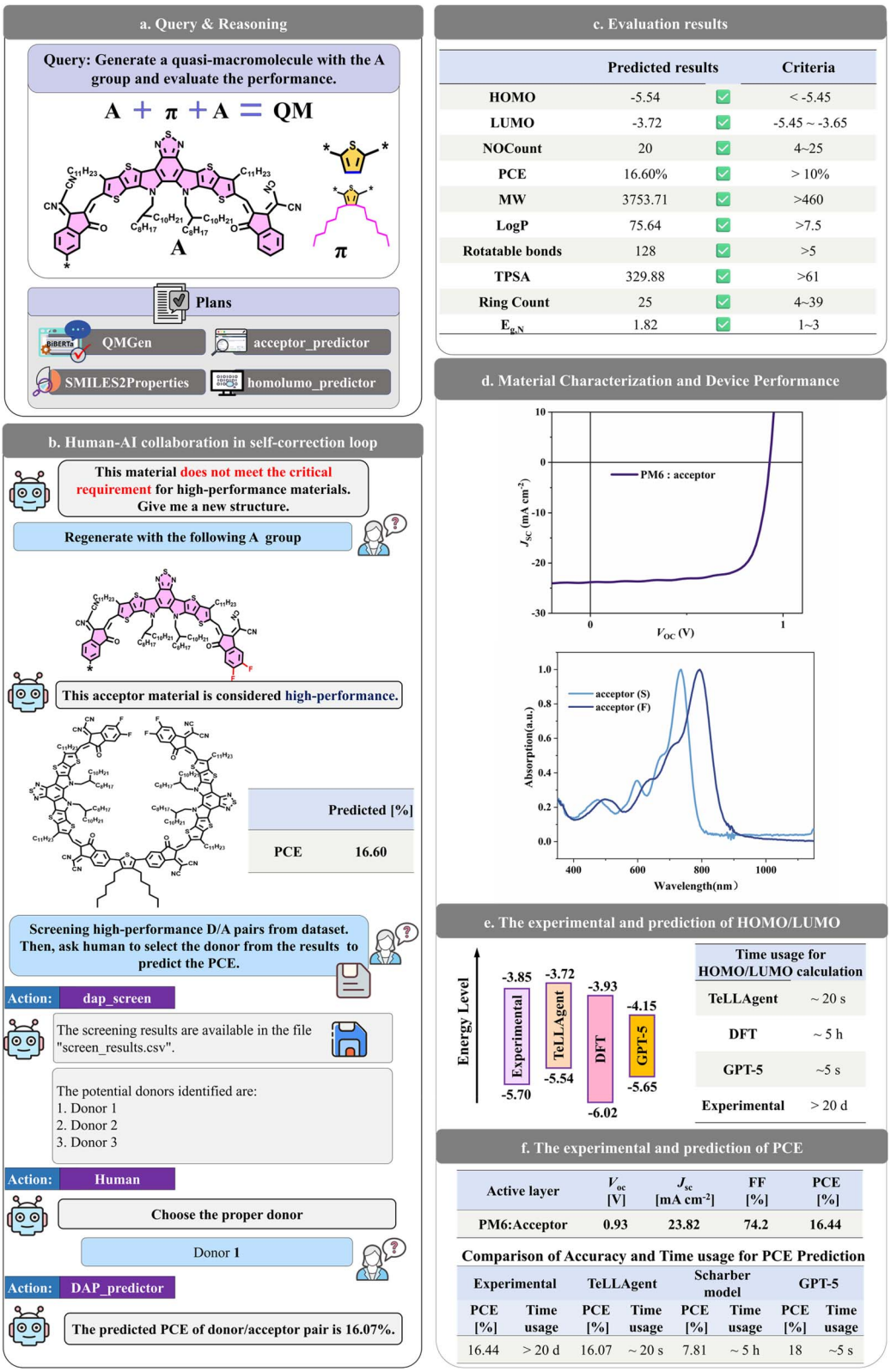


Fig. 5 Human–AI collaborative discovery of a quasi-macromolecular acceptor. (a) Initial query input and the action plan formulated by the global planning agent. (b) Schematic of the human–AI collaboration workflow for acceptor optimization and matching donor discovery, featuring iterative cycles of AI generation and expert curation. (c) Evaluation results of the collaboratively designed acceptor material. (d) Absorption spectra of the acceptor material in chloroform (CF) solution and film state, and the current density–voltage (J–V) characteristic curve of the corresponding photovoltaic device. The red shift in the film state indicates strong intermolecular π–π stacking. (e) Comparison of predicted HOMO and LUMO energy levels from TeLLAgent, GPT-5, and DFT calculations against experimental measurements. The predictions of TeLLAgent are highly concordant with experiments. (f) Comparison of predicted PCE from TeLLAgent, GPT-5, and the Scharber model against the experimental outcome. The predictions of TeLLAgent (16.07%) were strikingly accurate compared to the experimental result (16.44%).

the perils of relying on parametric knowledge without domain-specific tool grounding for precise quantitative tasks.

In summary, this consolidated validation demonstrates that TeLLAgent delivers not only novel molecular designs but also highly reliable quantitative predictions. The current limitations in achieving even higher PCEs are attributable to two factors: (1) the training data for the property prediction tools predominantly covers molecules with PCEs below 17%, limiting extrapolation; and (2) additional device engineering (*e.g.*, morphology optimization, interfacial layers) would be required to realize the full potential of the designed molecule.

TeLLAgent web server for accessible and reproducible materials discovery

To ensure the accessibility and reproducibility of our research, we have deployed TeLLAgent as a publicly accessible web server. This platform provides an intuitive interface that allows researchers, regardless of their computational expertise, to

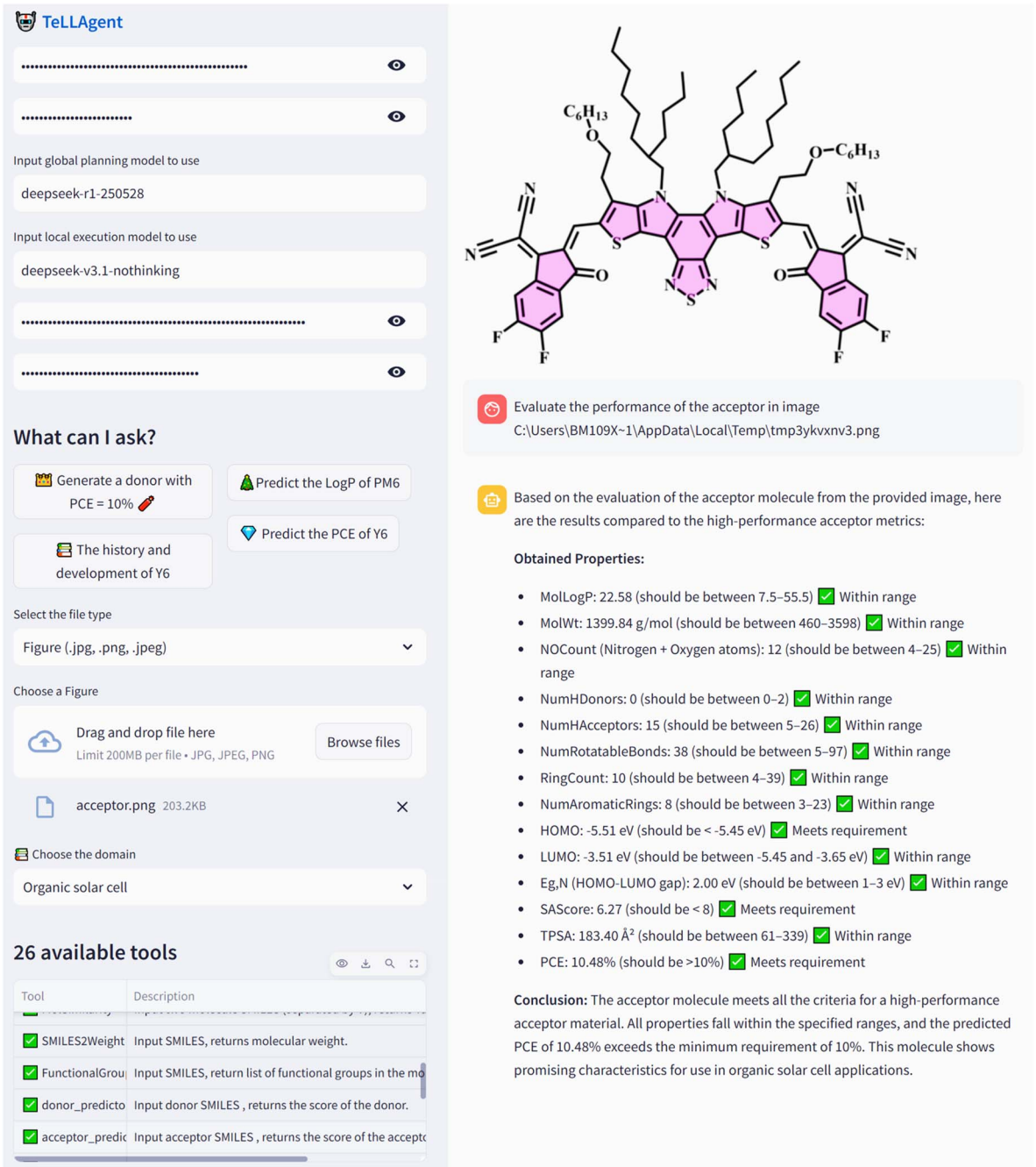


Fig. 6 Overview of the TeLLAgent web server for autonomous materials discovery. The public web platform allows users to input API keys and flexibly select LLMs for both the global and local agents, including popular open-source models. A scrollable list of available tools with usage information is provided. Example queries showcase the interactive query-response mechanism. The server supports multimodal input, including PDFs, CSV files, and images, for chemical information processing.

leverage the full power of the TeLLAgent framework for autonomous materials discovery (Fig. 6).

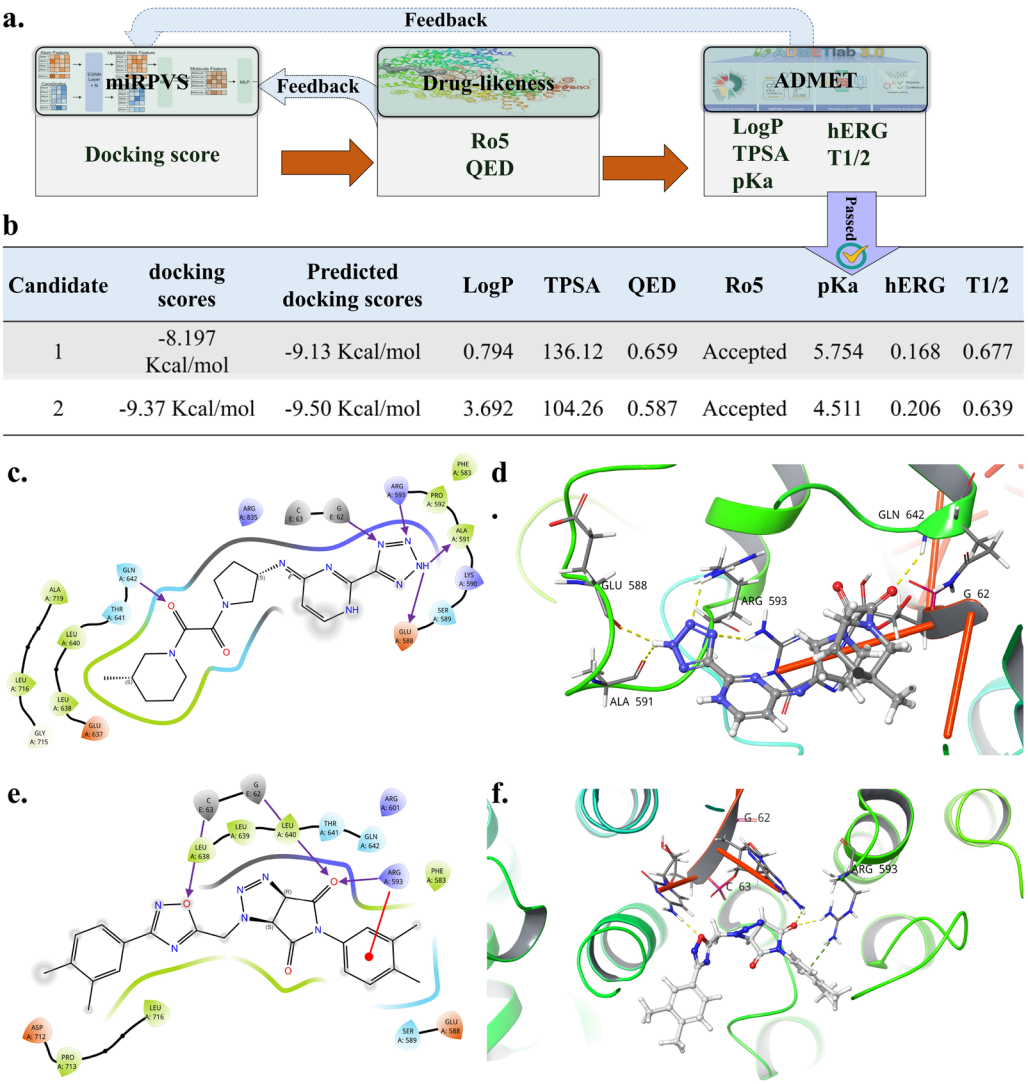
The server is designed with flexibility and openness at its core. Users can configure the system by providing their own API keys and selecting the most suitable open-source or proprietary LLMs for the global planning and local execution agents, enabling performance optimization and cost control for specific tasks. A central feature of the interface is a dynamic panel that lists the currently available tools along with concise descriptions of their functionality according to the chosen domain, guiding users in formulating their queries.

Interaction with TeLLAgent is facilitated through a natural language input field, lowering the barrier for complex scientific task specification. The multimodal capability of the platform is a key strength, allowing users to upload relevant files, such as PDF publications for literature mining, CSV datasets for high-throughput screening, or molecular images for structure recognition, alongside their textual queries. This seamless integration of diverse data types enables the execution of sophisticated, end-to-end workflows directly from the browser.

By providing this centralized, user-friendly portal, we not only demonstrate the capabilities of TeLLAgent in solving OSC-related challenges but also establish a verifiable and accessible benchmark for the future development of LLM-powered scientific assistants. The TeLLAgent web server stands as a resource for the scientific community, aiming to accelerate adoption and collaborative innovation in AI-driven materials science.

Cross-domain validation through drug discovery

To rigorously assess the generalizability of TeLLAgent beyond materials science, we reconfigured the framework for



b

Candidate	docking scores	Predicted docking scores	LogP	TPSA	QED	Ro5	pKa	hERG	T1/2
1	-8.197 Kcal/mol	-9.13 Kcal/mol	0.794	136.12	0.659	Accepted	5.754	0.168	0.677
2	-9.37 Kcal/mol	-9.50 Kcal/mol	3.692	104.26	0.587	Accepted	4.511	0.206	0.639

Fig. 7 TeLLAgent-assisted drug discovery for a miRNA-protein complex. (a) Workflow and tools used for the discovery of small-molecule drugs targeting complex 3a6p, including miRPVS for docking score prediction, a drug-likeness tool (Lipinski's Ro5 and QED), and ADMETlab3.0 for property prediction. (b) Prediction results (by TeLLAgent) and calculation results (by AutoDock Vina) for the two top candidate molecules, showing excellent affinity and properties. (c and e) Two-dimensional depictions of candidate 1 (c) and candidate 2 (e) within the binding pocket of complex 3a6p, highlighting key molecular interactions (hydrogen bonds in purple, pi–cation interactions in red). (d and f) Docking poses of candidate 1 (d) and candidate 2 (f) in the binding pocket, with corresponding 2D interaction diagrams.

a fundamentally different task, the virtual screening of small-molecule drugs. This migration underscored the modularity and ease of adaptation of the platform. By substituting the OSC-specific toolkit with a suite of drug discovery tools, including predictors for drug-likeness (Lipinski's rule of five and QED), molecular docking (miRPVS[27]), and ADMET properties (ADMETlab 3.0),[28] and updating the RAG knowledge base, we successfully repurposed the agent for the new domain. Critically, this entire integration process was completed within a single workday, demonstrating the low overhead and practical efficiency of applying TeLLAgent to novel scientific problems.

We deployed the adapted TeLLAgent to screen the ZINC20 database for inhibitors of the pre-miR-30a/Exp-5 complex (PDB: 3a6p), a miRNA-protein complex implicated in various diseases.[29] As illustrated in Fig. 7a, TeLLAgent autonomously orchestrated a multi-stage screening workflow. It first leveraged the miRPVS tool to predict docking scores across the library, then applied the drug-likeness and ADMET tools to evaluate the top-ranking candidates. The self-correction loop of the system was engaged when promising candidates failed subsequent filters, prompting it to recommend and evaluate new molecules from the database iteratively until it identified compounds satisfying all criteria.

This process yielded two candidate molecules with excellent predicted affinity and drug-like properties (Fig. 7b). The TeLLAgent predictions were subsequently validated using AutoDock Vina, which confirmed the high binding affinity of both candidates with minimal error. Further analysis of the binding poses revealed that these candidates form stable interactions within the binding pocket of the complex, including hydrogen bonds and pi–cation interactions with both the protein and RNA components (Fig. 7c–f). The application of molecular docking, ADMET profiling, and drug-likeness filtering as cross-validation methods without experimental confirmation does not guarantee that the identified molecules are effective inhibitors. Nevertheless, the successful application to an independent domain confirms TeLLAgent as a versatile and robust platform for generative molecular science, capable of accelerating discovery pipelines far beyond its original design scope.

Conclusion

In this work, we present TeLLAgent, a supervisor-executor dual-agent framework that leverages tool-enhanced LLMs to create an autonomous system for intelligent molecule discovery. Built upon the collaboration between two specialized agents, a global planner for strategic reasoning and a local executor for precise tool invocation, TeLLAgent overcomes the inherent limitations of single-agent systems in handling complex, multi-step scientific tasks. Powered by open-source LLMs (DeepSeek-R1 and DeepSeek-V3.1) and orchestrated through the MCP, our framework establishes a robust, multimodal platform capable of executing a wide array of tasks, from literature analysis and code generation to material inverse design and property prediction.

Rigorous evaluation in the domain of OSCs demonstrates the significant advantages of TeLLAgent over general-purpose LLMs like GPT-5. It achieved higher scores in knowledge retrieval and multimodal chemical information processing, as assessed by both human experts and LLM evaluators. Crucially, by dynamically retrieving information from up-to-date sources and employing an iterative CoT reasoning strategy, TeLLAgent effectively mitigates factual hallucinations, delivering more reliable and scientifically robust outputs. The experimental validation of an AI-designed quasi-macromolecular acceptor, which achieved a power conversion efficiency of 16.44%, stands as a testament to the predictive accuracy of the framework and its potential to bridge the gap between *in silico* design and real-world performance. The accompanying public web server enhances the accessibility of this powerful tool for the research community.

Although TeLLAgent represents a significant advancement, its current limitations indicate directions for future work. First, the framework's efficacy depends on the quality of its underlying LLMs. Additionally, while the computational overhead of the dual-agent architecture is justified by its performance gains, optimizing agent interaction protocols remains a future objective. Second, expanding the framework to new domains requires expert-driven integration of specialized toolkits. Crucially, the system is constrained by the physical boundaries of its tools and the training distributions of its models. Consequently, it struggles with out-of-distribution targets and analysis (*e.g.*, generating donors with PCEs > 22%).

Despite these challenges, the modular and extensible design of TeLLAgent makes it a highly versatile platform. Its successful application to drug discovery, achieving performance comparable to traditional methods with remarkable speed, underscores its powerful cross-domain capabilities and its potential to serve as a foundational technology for a new era of AI-driven scientific discovery.

Methods

The architecture of TeLLAgent

The TeLLAgent framework is architected around a supervisor-executor paradigm, which explicitly decouples high-level strategic reasoning from precise, low-level tool operation. This design is implemented through three cores, interacting components: the Agents, the LLMs that power them, and a curated, extensible Toolkit.

Agent roles and LLM selection

The system employs two distinct agent types with specialized roles. The global planning agent acts as the supervisor, responsible for comprehending the user's query, performing CoT reasoning to decompose the problem, and formulating a dynamic, high-level plan comprising sequential sub-tasks. The local execution agent functions as the executor, tasked with precisely carrying out the individual sub-tasks specified by the global agent. Its primary function is the accurate invocation of specialized tools with the correct parameters.

The selection of underlying LLMs for these roles was a deliberate process informed by their documented capabilities.

The global planning agent is powered by DeepSeek-R1, a model specifically recognized for its strong performance in complex reasoning and planning tasks, making it ideally suited for macro-strategy formulation.[30] The local execution agent leverages DeepSeek-V3.1, which excels in precise instruction-following and structured output generation, critical for reliable tool invocation and parameter passing.[31] This strategic pairing was empirically validated as the optimal configuration in our ablation studies (Fig. 3b and c).

Tool integration *via* model context protocol

The framework integrates a diverse suite of 30 specialized tools (detailed in the Toolkit section below) through a standardized MCP server. Each tool, whether developed in-house (*e.g.*, DonorGen, DAP_predictor, HomoLomo_predictor) or interfacing with a third-party API or library (*e.g.*, RDKit, Wikipedia), is wrapped as a unified MCP resource with a well-defined JSON schema for its inputs and outputs (SI note 3). This abstraction creates a consistent, discoverable interface for the local execution agent, effectively decoupling the complexity of the underlying software from the reasoning process of LLM. The MCP server manages all tool executions, ensuring robust, secure, and interoperable communication between the agents and the computational resources. The critical role of MCP in maintaining this interoperability and overall system performance is demonstrated by the significant performance drop observed when it is ablated (Fig. 3d).

Workflow

The operational workflow of TeLLAgent is an automated, iterative process that implements a reasoning-and-acting loop. This loop dynamically interleaves logical reasoning with tool execution to solve complex tasks. Upon receiving a user query, the process is initiated by the global planning agent. This agent first analyses the overall problem context and then engages in iterative CoT reasoning (Fig. S8) to decompose the query into a sequence of actionable sub-tasks, formulating a strategic plan. This plan is then executed step by step in a cycle managed by the local execution agent. For each sub-task, the workflow proceeds as follows. The local execution agent analyses the current sub-task and the overall context to determine the most appropriate tool from the available toolkit. It then formally invokes the selected tool, providing the necessary input parameters as specified by the schema of the tool. The output from the tool execution is returned to the agent.

The results from this tool invocation are then passed back to the global planning agent for evaluation. The global agent assesses the relevance and sufficiency of output in addressing the current sub-task and the overall query. If the result is deemed inadequate, the framework triggers a re-planning and recovery mechanism. This involves dynamically revising the strategy or selecting an alternative tool, and then re-initiating the execution loop for the sub-task.

This iterative cycle of planning, execution, and validation continues until the global planning agent determines that a satisfactory result has been achieved for all sub-tasks. Finally, the collective results are synthesized and formatted into a coherent, human-interpretable answer by the LLM. Some examples of the running process are shown in SI note 4.

Context engineering

To ensure robust task execution and minimize hallucination, the behavior of TeLLAgent is guided by a structured context, engineered in accordance with the MCP. This curated context provides the agents with a comprehensive framework for reasoning and action, and is composed of the following key elements. System prompt, an initial instruction set, defines the agent's identity, role, core capabilities, and core behavioral constraints (*e.g.*, "You are an AI system called TeLLAgent, and your task is to respond to the question or solve the problem to the best of your ability using the provided tools"). Memory, the operational state of the agent, comprises both short-term memory (the immediate conversation history, including previous user inputs and model responses) and long-term memory (persistent, predefined guidelines and facts that the agent can reference across conversations). Tool definitions, a complete specification of all available tools, include their names, detailed descriptions of their utility, and the expected format for their inputs and outputs. This schema, as shown in SI note 3, is automatically provided by the MCP server, making tools discoverable and usable by the agent. Structured output constraints, explicit instructions, mandate the agent to format its responses according to a predefined template, ensuring consistency and machine-readability of its final answers.

This structured input framework critically shapes the generation process of the agent, transforming it from an open-ended dialogue into a controlled, goal-oriented procedure. By supplying all necessary role, memory, capability, and format information within the context window, this engineering practice directly contributes to the reliability, predictability, and overall performance of the TeLLAgent system.

Toolkit

The toolkit comprises ten molecular informatics tools, four multimodal processing tools, six knowledge enhancement tools, seven organic solar cell (OSC)-specific tools, and three drug-domain tools.

Molecular informatics tools

Mol2SMILES

Converts molecular names (common or IUPAC names, such as Y6,[32] PM6) into standardized SMILES strings by querying PubChem or a custom database. In this work, the custom dataset was constructed based on prior research, encompassing small-molecule acceptors, donors, and polymer donors for organic photovoltaic materials.

Query2SMILES

Retrieves SMILES strings from Chemical Abstracts Service (CAS) numbers *via* PubChem.

Query2CAS

Obtains CAS registry numbers from IUPAC names or SMILES strings using PubChem.

SMILES2Name

Converts SMILES representations into IUPAC names using PubChem.

FuncGroups

Identifies functional groups and substructures in a molecule by matching against a predefined set of 102 SMiles ARbitrary Target Specification (SMARTS) patterns. This output aids in understanding molecular reactivity, properties, and potential applications.

SMILES2SAScore

Computes synthetic accessibility scores (0–1 range) from SMILES strings using SAScore,[33] enabling preliminary synthesizability assessment in material discovery.

SMILES2LogP

Predicts log *P* values from SMILES strings using RDKit (v2023.9.5), a key metric in organic material efficiency evaluation.

SMILES2Weight

Calculates molecular weight from SMILES representations *via* RDKit.

SMILES2Properties

Computes a suite of physicochemical properties, including SAScore, molecular weight, counts of nitrogen and oxygen atoms, hydrogen bond acceptors/donors, log *P*, rotatable bonds, ring counts, aromatic rings, and topological polar surface area (TPSA) by using RDKit. These properties are linked to material performance.[34,35]

MolSimilarity

Quantifies structural similarity between two molecules using the Tanimoto coefficient derived from ECFP2 fingerprints. The output score informs on molecular analogies, supporting property prediction in materials research.

Multimodal processing tools

Codewriter. Automates Python script generation for chemical workflows (*e.g.*, RDKit operations) using DeepSeek-V3.1, facilitating computational protocol development for non-programmers.

Graphconverter. Converts molecular graphs into SMILES strings using DECIMER,[36] a deep learning-based optical chemical structure recognition tool that processes both hand-drawn and printed structures.

Imageanalysis. Leverages the multimodal capabilities of Qwen3-VL to interpret image and text inputs, enhancing cross-modal comprehension and enabling contextual image understanding.

PDFreader. Processes PDF files using retrieval-augmented generation (RAG). Documents are segmented, embedded *via* OpenAI embeddings,[37] and indexed in a Faiss vector database (v1.9.0).[38] User queries are processed with DeepSeek-V3.1 to retrieve relevant information.

Knowledge enhancement tools

RAG

Integrates a FAISS vector database with a large language model (LLM) to retrieve information and mitigate hallucinations. The database, exemplified with 1000 OSC-related publications in this study, is updateable with recent research, enabling personalized and current knowledge access. The details are shown in Fig. S9 and SI note 5.

LiteratureSearch

Extracts scientific insights from literature using PaperQA[39] and paperscraper.[40] The tool embeds documents *via* OpenAI Embeddings and FAISS, with DeepSeek-V3.1 generating answers from retrieved contexts.

WebSearch

Accesses current web information *via* SerpAPI, collecting snippets from Google search results to enrich scientific knowledge and verify information accuracy.

Wikipedia

Retrieves and parses text, sections, and categories from Wikipedia using its API.

Browseruse

Automates web queries on chemical databases (*e.g.*, PubChem, ChemSpider) *via* Browser-use (https://github.com/browser-use/browser-use), enabling real-time information retrieval.

Human

Facilitates human–computer interaction, allowing the agent to request user input when encountering uncertainties. It queries the user for input, allowing the supervisor to obtain high-level guidance or resolve underspecified constraints. The agent can leverage external human expertise as a strategic resource without compromising its capacity for independent planning and execution.

OSC-domain tools

DonorGen

Performs inverse design of polymer donors for OSCs using an improved Transformer model. Trained on repeat units and

PCEs predicted by DeepDonor, it generates novel donors based on target PCE inputs (Fig. S6).

DAP_predictor

Predicts PCE of donor–acceptor pairs using BiBERTa[41] from SMILES inputs, aiding in high-performance pair discovery.

Donor_predictor

Evaluates PCE of small-molecule and polymer donors *via* DeepDonor[34] using SMILES inputs.

Acceptor_predictor

Assesses performance of small-molecule acceptors using DeepAcceptor[35] from SMILES inputs.

HomoLumo_predictor

Predicts HOMO and LUMO levels of organic photovoltaic materials using a random forest model trained on the Clean Energy Project Database (2.3 million molecules with DFT-calculated properties).

DAP_screen

Screens donor–acceptor datasets for PCE using BiBERTa. Accepts CSV inputs and outputs predictions.

QMGen

Generates quinoidal molecules (QMs) for OSCs by linking A–π–A units *via* BiBERTa, using SMILES of π-bridge and acceptor units as input.

Drug-domain tools

miRPVS

Conducts virtual screening of small-molecule drugs targeting miRNA-protein complexes using miRPVS,[27] predicts binding affinity from SMILES after pocket identification.

Drug-likeness

Evaluates drug-likeness *via* Lipinski's Rule of Five (Ro5) and quantitative estimate of drug-likeness (QED).[42] Inputs SMILES and outputs compliance with Ro5 and QED scores.

ADMET

Predicts absorption, distribution, metabolism, excretion, and toxicity properties using ADMETlab3.0.[28] Inputs SMILES and returns descriptors such as log *P*, TPSA, pK_a, hERG, and T1/2.

Author contributions

Z. Z. and H. L. conceived the idea and initiated this project. J. S. collected the data and designed the computational workflow. W. L., J. Y., Y. Z. synthesized the organic photovoltaic materials and fabricated OSC devices. H. W., Y. W., T. X., L. T., H. Z performed the validation. J. S. wrote the manuscript. All authors read and approved the final manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

The data supporting the findings of this study are available in the GitHub repository at https://github.com/JinYSun/TeLLAgent or from the corresponding authors on a reasonable request. An open-source version of the TeLLAgent platform has been released at https://github.com/JinYSun/TeLLAgent, which includes the main agent setup and a subset of tools used in the original implementation. The publicly accessible platform is available at https://huggingface.co/spaces/jinysun/TeLLAgent.

Supplementary information (SI): context engineering and prompt details; evaluation tasks; evaluation rubrics of human experts; performance comparisons and statistical tests; detailed descriptions of tools like DonorGen and RAG; failure analysis of single-agent baselines; multimodal evaluation results; cost and token analysis; workflow illustrations and example outputs. See DOI: https://doi.org/10.1039/d5sc09883a.

Acknowledgements

This work is funded by National Natural Science Foundation of China (22473116, 22373117, 21873116, and 22273120) and Natural Science Foundation of Hunan Province in China (2024JJ2068). We gratefully acknowledge the High Performance Computing Center of Central South University for its computational resources.

References

1 J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2019, pp. 4171–4186.

2 T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, in *Proceedings of the 34th International Conference on Neural Information Processing Systems Article 159*, Curran Associates Inc., 2020.

3 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer,

S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.

4 R. Irwin, S. Dimitriadis, J. He and E. J. Bjerrum, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 015022.

5 D. Christofidellis. G. Giannone, J. Born, O. Winther, T. Laino and M. Manica, Unifying molecular and textual representations via multi-task language modelling, Proceedings of the 40th International Conference on Machine Learning, 2023, vol. 202, pp. 6140–6157.

6 C. Edwards, T. Lai, K. Ros, G. Honke, K. Cho and H. Ji, Abu Dhabi, United Arab Emirates, in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2022, pp. 375–413.

7 K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi, S. Cox, W. A. de Jong, M. L. Evans, N. Gastellu, J. Genzling, M. V. Gil, A. K. Gupta, Z. Hong, A. Imran, S. Kruschwitz, A. Labarre, J. Lála, T. Liu, S. Ma, S. Majumdar, G. W. Merz, N. Moitessier, E. Moubarak, B. Mouriño, B. Pelkie, M. Pieler, M. C. Ramos, B. Ranković, S. G. Rodriques, J. N. Sanders, P. Schwaller, M. Schwarting, J. Shi, B. Smit, B. E. Smith, J. Van Herck, C. Völker, L. Ward, S. Warren, B. Weiser, S. Zhang, X. Zhang, G. A. Zia, A. Scourtas, K. J. Schmidt, I. Foster, A. D. White and B. Blaiszik, *Digital Discovery*, 2023, **2**, 1233–1250.

8 J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson and A. Jain, *Nat. Commun.*, 2024, **15**, 1418.

9 M. C. Ramos, C. J. Collison and A. D. White, *Chem. Sci.*, 2025, **16**, 2514–2572.

10 H. T. Mai, C. X. Chu and H. Paulheim, in *23rd International Semantic Web Conference*, Springer-Verlag, 2024, pp. 126–143.

11 C. M. Castro Nascimento and A. S. Pimentel, *J. Chem. Inf. Model.*, 2023, **63**, 1649–1655.

12 Y. Kang and J. Kim, *Nat. Commun.*, 2024, **15**, 4705.

13 A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, *Nat. Mach. Intell.*, 2024, **6**, 525–535.

14 K. Ding, J. Yu, J. Huang, Y. Yang, Q. Zhang and H. Chen, *Nat. Comput. Sci.*, 2025, **5**, 962–972.

15 A. Ghafarollahi and M. J. Buehler, *Proc. Natl. Acad. Sci. U. S. A.*, 2025, **122**, e2414074122.

16 Y. Ruan, C. Lu, N. Xu, Y. He, Y. Chen, J. Zhang, J. Xuan, J. Pan, Q. Fang, H. Gao, X. Shen, N. Ye, Q. Zhang and Y. Mo, *Nat. Commun.*, 2024, **15**, 10160.

17 Y. Inoue, T. Song, X. Wang, A. Luna and T. Fu, in *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, ICLR, 2025.

18 A. Ghafarollahi and M. J. Buehler, *Adv. Mater.*, 2025, **37**, 2413523.

19 T. Song, M. Luo, X. Zhang, L. Chen, Y. Huang, J. Cao, Q. Zhu, D. Liu, B. Zhang, G. Zou, G. Zhang, F. Zhang, W. Shang, Y. Fu, J. Jiang and Y. Luo, *J. Am. Chem. Soc.*, 2025, **147**, 12534–12545.

20 Y. Du, S. Li, A. Torralba, J. B. Tenenbaum and I. Mordatch, in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, 2024, 235, pp. 11733–11763.

21 W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C.-M. Chan, H. Yu, Y. Lu, Y.-H. Hung, C. Qian, Y. Qin, X. Cong, R. Xie, Z. Liu, M. Sun and J. Zhou, in *The Twelfth International Conference on Learning Representations*, ICLR, 2024.

22 Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger and C. Wang, in *First Conference on Language Modeling*, COLM, 2024.

23 C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu and Z. Liu, in *The Twelfth International Conference on Learning Representations*, ICLR, 2024.

24 X. Hou, M. Yang, W. Jiao, X. Wang, Z. Tu and W. X. Zhao, *arXiv*, 2024, preprint arXiv:2406.13381, DOI: 10.48550/arXiv.2025.2406.13381.

25 Z. Shi, S. Gao, X. Chen, Y. Feng, L. Yan, H. Shi, D. Yin, P. Ren, S. Verberne and Z. Ren, in *Findings of the Association for Computational Linguistics: EMNLP 2024 10642-10657*, Association for Computational Linguistics, 2024.

26 T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest and X. Zhang, in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence Article 890*, International Joint Conferences on Artificial Intelligence Organization, 2024.

27 H. Wang, Z. Zhang, G. Zhang, M. Wen and H. Lu, *J. Pharm. Anal.*, 2026, **16**, 101339.

28 L. Fu, S. Shi, J. Yi, N. Wang, Y. He, Z. Wu, J. Peng, Y. Deng, W. Wang, C. Wu, A. Lyu, X. Zeng, W. Zhao, T. Hou and D. Cao, *Nucleic Acids Res.*, 2024, **52**, W422–W431.

29 X. Zhang, S. Dong, Q. Jia, A. Zhang, Y. Li, Y. Zhu, S. Lv and J. Zhang, *Biosci. Rep.*, 2019, **39**, BSR20190788.

30 D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Xu, H. Ding, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Chen, J. Yuan, J. Tu, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. You, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Zhou, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao,

Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang and Z. Zhang, *Nature*, 2025, **645**, 633–638.

31 DeepSeek-AI, A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Zhang, H. Ding, H. Xin, H. Gao, H. Li, H. Qu, J. L. Cai, J. Liang, J. Guo, J. Ni, J. Li, J. Wang, J. Chen, J. Chen, J. Yuan, J. Qiu, J. Li, J. Song, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Xu, L. Xia, L. Zhao, L. Wang, L. Zhang, M. Li, M. Wang, M. Zhang, M. Zhang, M. Tang, M. Li, N. Tian, P. Huang, P. Wang, P. Zhang, Q. Wang, Q. Zhu, Q. Chen, Q. Du, R. J. Chen, R. L. Jin, R. Ge, R. Zhang, R. Pan, R. Wang, R. Xu, R. Zhang, R. Chen, S. S. Li, S. Lu, S. Zhou, S. Chen, S. Wu, S. Ye, S. Ye, S. Ma, S. Wang, S. Zhou, S. Yu, S. Zhou, S. Pan, T. Wang, T. Yun, T. Pei, T. Sun, W. L. Xiao, W. Zeng, W. Zhao, W. An, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, X. Q. Li, X. Jin, X. Wang, X. Bi, X. Liu, X. Wang, X. Shen, X. Chen, X. Zhang, X. Chen, X. Nie, X. Sun, X. Wang, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yu, X. Song, X. Shan, X. Zhou, X. Yang, X. Li, X. Su, X. Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Y. Zhang, Y. Xu, Y. Xu, Y. Huang, Y. Li, Y. Zhao, Y. Sun, Y. Li, Y. Wang, Y. Yu, Y. Zheng, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Tang, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Wu, Y. Ou, Y. Zhu, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Zha, Y. Xiong, Y. Ma, Y. Yan, Y. Luo, Y. You, Y. Liu, Y. Zhou, Z. F. Wu, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Huang, Z. Zhang, Z. Xie, Z. Zhang, Z. Hao, Z. Gou, Z. Ma, Z. Yan, Z. Shao, Z. Xu, Z. Wu, Z. Zhang, Z. Li, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Gao and Z. Pan, *arXiv*, 2025, preprint arXiv:abs/2412.19437, DOI: 10.48550/arXiv.2412.19437.

32 J. Yuan, Y. Zhang, L. Zhou, G. Zhang, H.-L. Yip, T.-K. Lau, X. Lu, C. Zhu, H. Peng, P. A. Johnson, M. Leclerc, Y. Cao, J. Ulanski, Y. Li and Y. Zou, *Joule*, 2019, **3**, 1140–1151.

33 P. Ertl and A. Schuffenhauer, *J. Cheminf.*, 2009, **1**, 8.

34 J. Sun, D. Li, Y. Wang, T. Xie, Y. Zou, H. Lu and Z. Zhang, *J. Mater. Chem. A*, 2024, **12**, 21813–21823.

35 J. Sun, D. Li, J. Zou, S. Zhu, C. Xu, Y. Zou, Z. Zhang and H. Lu, *npj Comput. Mater.*, 2024, **10**, 181.

36 K. Rajan, H. O. Brinkhaus, M. I. Agea, A. Zielesny and C. Steinbeck, *Nat. Commun.*, 2023, **14**, 5045.

37 A. Neelakantan, T. Xu, R. Puri, A. Radford, J. M. Han, J. Tworek, Q. Yuan, N. Tezak, J. W. Kim, C. Hallacy, J. Heidecke, P. Shyam, B. Power, T. E. Nekoul, G. Sastry, G. Krueger, D. Schnurr, F. P. Such, K. Hsu, M. Thompson, T. Khan, T. Sherbakov, J. Jang, P. Welinder and L. Weng, CoRR, *arXiv*, 2022, preprint arXiv:abs/2201.10005, DOI: 10.48550/arXiv.2201.10005.

38 J. Johnson, M. Douze and H. Jégou, *IEEE Trans. Big Data*, 2021, **7**, 535–547.

39 J. Lála, O. O'Donoghue, A. Shtedritski, S. Cox, S. G. Rodriques and A. D. White, *arXiv*, 2023, preprint arXiv:abs/2312.07559, DOI: 10.48550/arXiv.2312.07559.

40 J. Born and M. Manica, *Curr. Med. Chem.*, 2021, **28**, 7862–7886.

41 J. Sun, D. Li, J. Zou, X. Tan, Y. Wang, H. Zhang, Y. Zou, Z. Zhang and H. Lu, *J. Mater. Chem. A*, 2025, **13**, 23570–23580.

42 G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan and A. L. Hopkins, *Nat. Chem.*, 2012, **4**, 90–98.