nature machine intelligence

Article

https://doi.org/10.1038/s42256-025-00994-z

Large language models for scientific discovery in molecular property prediction

Received: 24 October 2023

Accepted: 15 January 2025

Published online: 25 February 2025

Check for updates

Yizhen Zheng[1,4] ✉, Huan Yee Koh[1,2,4], Jiaxin Ju[3,4], Anh T. N. Nguyen[2], Lauren T. May[2], Geoffrey I. Webb[1,5] ✉ & Shirui Pan[3,5] ✉

Large language models (LLMs) are a form of artificial intelligence system encapsulating vast knowledge in the form of natural language. These systems are adept at numerous complex tasks including creative writing, storytelling, translation, question-answering, summarization and computer code generation. Although LLMs have seen initial applications in natural sciences, their potential for driving scientific discovery remains largely unexplored. In this work, we introduce LLM4SD, a framework designed to harness LLMs for driving scientific discovery in molecular property prediction by synthesizing knowledge from literature and inferring knowledge from scientific data. LLMs synthesize knowledge by extracting established information from scientific literature, such as molecular weight being key to predicting solubility. For inference, LLMs identify patterns in molecular data, particularly in Simplified Molecular Input Line Entry System-encoded structures, such as halogen-containing molecules being more likely to cross the blood–brain barrier. This information is presented as interpretable knowledge, enabling the transformation of molecules into feature vectors. By using these features with interpretable models such as random forest, LLM4SD can outperform the current state of the art across a range of benchmark tasks for predicting molecular properties. We foresee it providing interpretable and potentially new insights, aiding scientific discovery in molecular property prediction.

Scientific productivity is facing a notable decline, with progress in many fields roughly halving every 13 years[1]. As scientific discovery becomes increasingly complex and challenging, traditional methodologies struggle to keep pace, necessitating innovative approaches. Scientific discovery relies on building on existing knowledge to analyse experimental data, recognize data patterns and formulate well-reasoned hypotheses[2]. This process requires two essential abilities: prior knowledge understanding and reasoning abilities. Large language models (LLMs) have demonstrated capabilities in both areas[3–7]. Pretrained on extensive scientific literature, LLMs possess a deep understanding of scientific concepts such as chemistry[8] and demonstrate effective reasoning skills as part of their emergent abilities. Consequently, LLMs have great potential to accelerate scientific discovery.

Molecular property prediction is crucial for advancing drug design and materials discovery. Understanding molecular properties helps identify key factors that drive chemical behaviour, providing deeper insights into chemistry. In this context, LLMs possess extensive prior knowledge of molecular property prediction tasks. For instance, when predicting the solubility of molecules, they recognize that molecular weight and the number of aromatic rings are important

[1]Department of Data Science and Artificial Intelligence, Monash University, Melbourne, Victoria, Australia. [2]Drug Discovery Biology, Monash Institute of Pharmaceutical Sciences, Monash University, Clayton, Victoria, Australia. [3]School of Information and Communication Technology, Griffith University, Southport, Queensland, Australia. [4]These authors contributed equally: Yizhen Zheng, Huan Yee Koh, Jiaxin Ju. [5]These authors jointly supervised this work: Shirui Pan, Geoffrey I. Webb. ✉e-mail: yizhen.zheng1@monash.edu; geoff.webb@monash.edu; s.pan@griffith.edu.au

factors—knowledge derived from the literature they are trained on. Additionally, LLMs can understand formal scientific languages, such as the Simplified Molecular Input Line Entry System (SMILES), which describes molecular structures and is used for storing molecular property data. For example, to represent a molecule's solubility, its SMILES string and logP score can be stored together. Given these two key abilities—understanding molecular property prediction tasks and interpreting SMILES—we are motivated to explore two questions: can LLMs leverage their prior knowledge and reasoning abilities to facilitate scientific discovery? Can LLMs be effectively used to help predicting the properties of molecules?

In this work, we propose LLM4SD (LLMs for Scientific Discovery). LLM4SD functions by performing two main tasks: synthesizing knowledge from existing literature and inferring knowledge by observing experimental data. First, LLM4SD retrieves known rules to predict molecular properties based on its pretrained literature, such as molecules with molecular weight under 500 Da being more likely to pass the blood–brain barrier (BBB). Second, using its understanding of SMILES notation and chemistry knowledge, LLM4SD identifies patterns from experimental data, such as molecules containing halogens being more likely to pass the BBB. These rules are then used to create interpretable feature vectors for each molecule. By training an interpretable machine learning model using these vectors, we show that this pipeline achieves the current state of the art on molecular property prediction across 58 benchmark tasks from the MoleculeNet dataset curated by the Stanford PANDE group[9]. These tasks, encompassing both classification and regression, span four domains: physiology, biophysics, physical chemistry and quantum mechanics.

Although these molecular property predictions address complex challenges, such as predicting a molecule's BBB permeability, they represent only a small corner of the breadth of scientific endeavour. The findings of LLM4SD highlight the broader potential of using LLMs for scientific discovery.

Results

LLM4SD pipeline

Our scientific discovery pipeline, LLM4SD, shown in Fig. 1, consists of four main components: (1) knowledge synthesis from the scientific literature, (2) knowledge inference from data, (3) model training and (4) interpretable insights.

In the knowledge synthesis from literature phase (Fig. 1a), LLMs use their pretrained understanding from extensive scientific literature[5,10,11] to synthesize rules for predicting molecular properties. For example, in predicting BBB permeability (BBBP), LLMs can apply established knowledge, such as the Lipinski rule of five[12], which assesses drug-likeness based on rules like molecular weight being under 500 Da or there being fewer than five hydrogen bond donors. In the knowledge inference from data phase (Fig. 1b), LLMs can utilize their inferential and analytical abilities to identify patterns in scientific data: for example, SMILES strings and their corresponding labels. For instance, LLMs can detect that molecules containing halogens are more likely to pass the BBB, as this trait is frequently observed in the provided data for molecules that successfully cross the barrier.

In both the knowledge synthesis and inference stages, we require that the identified rules have either a numerical or categorical measure associated with them. This ensures that the rules can be readily transformed into corresponding code functions, which in turn can convert each molecule into a vector of values. Prompts used for both knowledge synthesis and inference in LLM4SD are shown in Extended Data Tables 1–3.

To convert these rules into corresponding executable code functions, we utilize GPT-4 to generate Python code using cheminformatics software such as RDKit[13]. For example, the rule 'if the molecular weight is smaller than 500 Da, the molecule is likely more permeable to the blood–brain barrier' can be converted into code as 'rdMolDescriptors.CalcExactMolWt (mol) < 500', which will return a binary result. Here, rdMolDescriptors is a module in the RDKit[13] library. If the result is 0, it means the molecule has a weight greater than 500 Da and vice versa. In practice, human experts can review these generated rules to drop duplicate or non-functional parts. However, the experimental results presented were obtained without any human intervention or modification.

With these rules and functions, molecules can be transformed into vectorized representations: that is, features. These rule-based features can then be used to train an interpretable model, such as a random forest or linear classifier (Fig. 1c). Remarkably, we noted that when enhanced with LLM4SD, these traditional interpretable models can surpass state-of-the-art baselines.

Once the interpretable models are trained (Fig. 1d), we can gain valuable insights. In the BBBP prediction task, we can see the prediction results, the rules content, how each molecule aligns with those rules by looking into their vector representation and how important each rule is for the final prediction. For example, a random-forest model builds decision trees, and if these trees rely heavily on certain features, those features are assessed to have higher importance for the task. This helps us understand which rules are most critical in the model's predictions. As shown in Fig. 1d, the input molecule has a molecular weight under 500 Da, and molecular weight is a key factor in predicting BBBP. To improve usability for researchers, we have created a web-based application that offers knowledge synthesis, inference and prediction with interpretation (Supplementary Information Section 3).

Experimental results

In this section, we offer a synopsis of LLM4SD's performance spanning the four domains of physiology, biophysics, quantum mechanics and physical chemistry. Subsequently, we investigate the key components of LLM4SD, examining its performance across various LLM backbones of differing scales and pretraining datasets.

Overall performance on four domains. To evaluate LLM4SD's versatility, we conducted a comprehensive analysis of its performance across 58 molecular prediction tasks across four domains (Fig. 2). We compared LLM4SD's performance with nine specialized, state-of-the-art supervised machine learning models. These are advanced graph or geometric neural networks (GNNs): AttrMask[14], GraphCL[15], MolCLR[16], 3DInfomax[17], GraphMVP[18], MoleBERT[19], Grover[20] and UniMol[21]. Each model was pretrained on large datasets with diverse molecular knowledge and then fine-tuned for specific tasks (Methods). As a standard baseline, we implemented random forest with ECFP4 (ref. 22) as input set features.

Benchmarking LLM4SD against the baseline, LLM4SD demonstrated its superior efficacy and performance (Fig. 2). This performance spanned 58 diverse tasks, from physiology (Extended Data Figs. 1–3) and biophysics (Extended Data Fig. 4) to physical chemistry (Extended Data Fig. 5) and quantum mechanics (Extended Data Fig. 6). The detailed descriptions of these tasks is illustrated in the Methods section (Methods, 'Datasets').

In both physiology and biophysics, our model outperformed all existing baselines (Fig. 2a). Notably, we attained state-of-the-art results in physiology, raising the area under the receiver operating characteristic curve (AUC-ROC) from a previous best of 74.53% to 76.60%, a gain of 2.07%. In biophysics, our model also achieved the best performance. These advancements in physiology and biophysics emphasize the robustness and precision of LLM4SD in tasks that demand intricate biological understanding and modelling.

On tasks in quantum mechanics and physical chemistry, LLM4SD demonstrated substantial advancements (Fig. 2b). In the domain of quantum mechanics, it showed a profound improvement of 48.2% over the best performing baseline, registering an average mean absolute error (MAE) of 5.8233 across 12 tasks as opposed to 11.2450 achieved by

https://doi.org/10.1038/s42256-025-00994-z

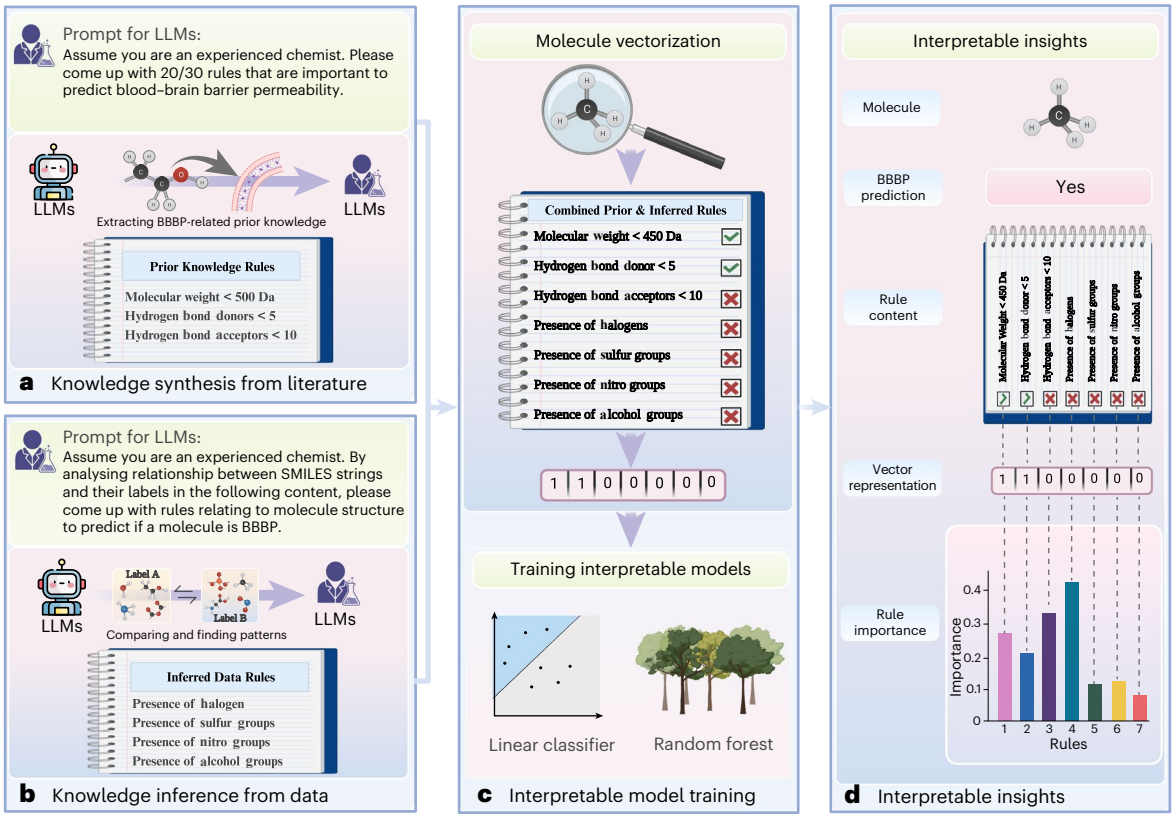


Fig. 1 | LLMs for scientific discovery in molecular prediction pipeline. **a**, Knowledge synthesis from the literature. In this phase, LLMs synthesized knowledge based on their pretrained literature for tasks like predicting BBBP. For example, molecules with a molecular weight under 500 Da are more likely to pass through the BBB. **b**, Knowledge inference from data. Here, LLMs analyse data, such as SMILES strings with labels (1 for BBB permeable, 0 for non-BBB permeable), to identify patterns. For instance, they may observe that molecules containing halogens have a higher chance of crossing the BBB. **c**, Model training. With synthesized and inferred rules, a molecule can be converted to vector representations based on its corresponding rule value. The vectorized representations can then be used to train interpretable models. **d**, Interpretable insights. Once the model is trained, it provides insights that explain how it makes its predictions. For example, in the context of BBBP prediction, the model can reveal the significance of each rule, showing which are important for the final prediction. Figure created with BioRender.com.

the second-best baseline GraphMVP. Similarly, in physical chemistry, LLM4SD observed a noteworthy enhancement, with the model reaching an MAE of 1.28, marking an 12.9% advancement over the baseline root mean square error (RMSE) of 1.47. These substantial improvements in regression tasks affirm the refined capability of our approach in continuous prediction.

Compared to these advanced GNN baselines, LLM4SD has two primary advantages. First, it leverages a wealth of prior knowledge accumulated from decades of scientific literature. Although GNNs can learn general patterns from extensive molecular datasets through pretraining, incorporating specific scientific knowledge typically requires explicit effort, such as the careful curation and integration of domain-specific features into the model's input, unless that knowledge is hand-coded for their use. Developers must manually decide both what knowledge to include and how to integrate it effectively into the model[23,24]. In contrast, LLMs, pretrained on vast amounts of literature across fields like chemistry, inherently embed substantial amounts of scientific knowledge that can be leveraged directly without additional intervention beyond natural language interaction. Second, GNNs, although effective in encoding molecules into embeddings, often lack interpretability. This limits their utility in generating clear scientific hypotheses because their interpretive mechanisms, such as attention mechanisms, tend to be opaque.

Study of key components. To delve deeper into the intricacies of the LLM4SD pipeline, we studied the influence of scale and pretraining datasets on its performance. In addition, we assessed the relative contributions of knowledge synthesis and inference. Our evaluation spanned across a spectrum of foundational LLM backbones, notably the GPT-4 (ref. 5), Falcon-7b (ref. 11), Falcon-40b (ref. 11), Galactica-6.7b (ref. 10), Galactica-30b (ref. 10), ChemLLM-7b (ref. 25) and ChemDFM-13b (ref. 26). These backbones can be categorized into two classes, general LLMs and domain-specific LLMs. In particular, GPT-4 and the Falcon models are trained for a broad range of applications, imbibing a more general context during their pretraining phase, whereas the Galactica[10] models, ChemLLM[25] and ChemDFM[26] are pretrained or fine-tuned on mainly scientific literature.

Effect of scale. The comparison of seven LLM4SD backbones revealed substantial differences among the different LLMs (Fig. 3a,b). Within the Falcon series, performance disparities were conspicuous. Falcon-7b, a smaller model, fell short compared to Falcon-40b in its range of domain expertise. Notably, it failed to conduct tasks in two key areas—physiology and quantum mechanics—indicating a weaker understanding of scientific challenges and data interpretation. Specifically, it produced gibberish responses when asked to conduct inference.

Conversely, the Galactica series painted a more nuanced picture. Unlike the Falcon series, a larger model did not necessarily translate to superior performance. In disciplines such as physiology, biophysics and physical chemistry, Galactica-6.7b rivalled the performance of Galactica-30b, despite the latter having more than four times the number of parameters. However, in the domain of quantum mechanics, the larger Galactica-30b surged ahead, outperforming Galactica-6.7b by a margin of 14%. This variance could be attributed to the intricate and abstract

https://doi.org/10.1038/s42256-025-00994-z

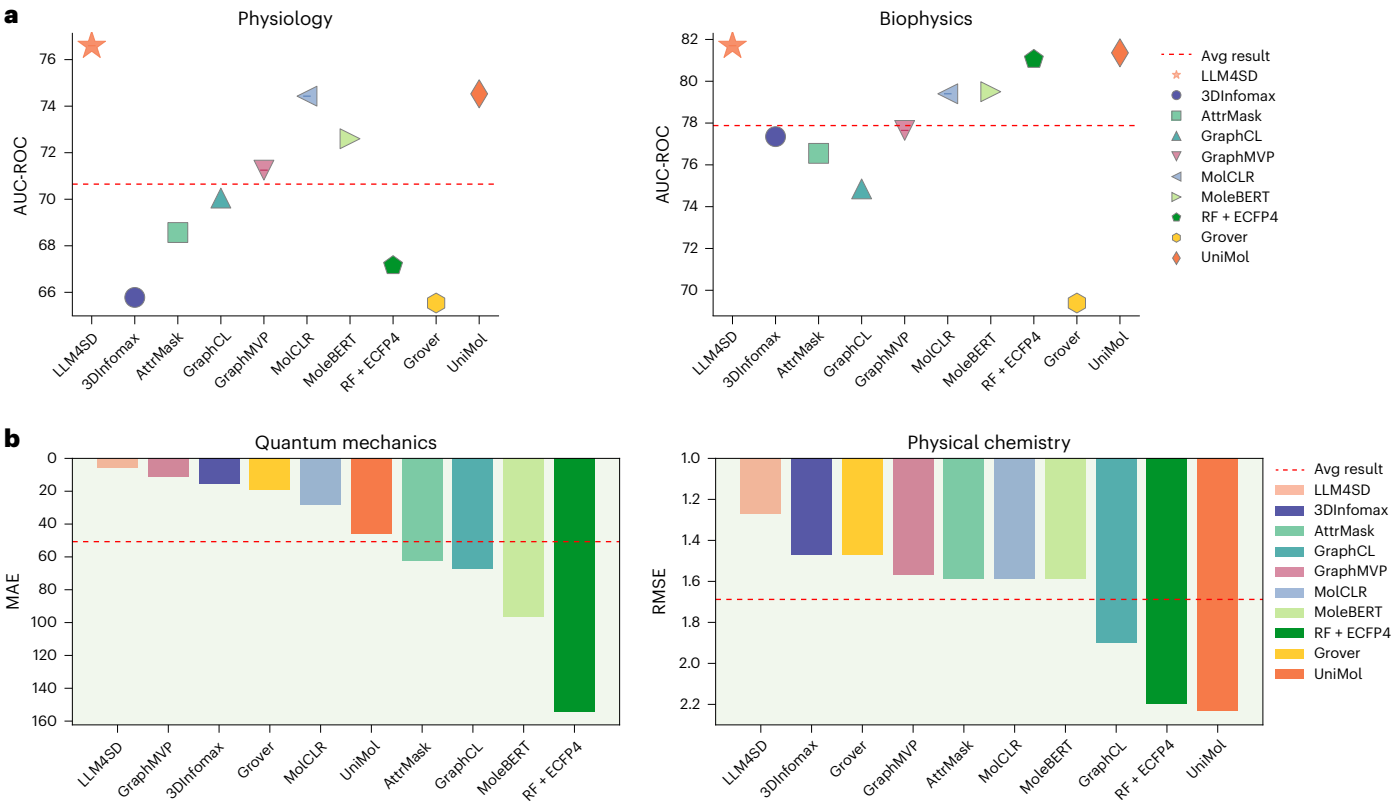


Fig. 2 | Comparison between LLM4SD and baselines across four domains. The red dashed line represents the average performance of all baselines. **a**, Comparative analysis of model performance versus baselines in physiology and biophysics. **b**, Comparative analysis of regression performance: LLM4SD versus baselines in quantum mechanics and physical chemistry.

nature of quantum mechanics, where the depth and breadth of knowledge encapsulated in the larger model might offer a discernible advantage.

Surprisingly, ChemLLM-7b and ChemDFM-13b substantially underperform compared to Galactica-6.7b, despite all of them being domain-specific models. Unlike Galactica-6.7b, which was built from scratch using 106 billion tokens of scientific literature, ChemLLM-7b and ChemDFM-13b are adapted from general LLMs—InternLM2-7b (ref. 27) and LLaMa-13b (ref. 28), respectively—through fine-tuning. Moreover, Galactica benefits from a rich training dataset, whereas ChemLLM-7b utilizes only 7 million tokens of chemical data and ChemDFM-13b is trained on 34 billion tokens. We conjecture that this discrepancy in data scale and training approach leads to the performance difference.

GPT-4 consistently performs well on all tasks, which is reasonable considering its enormous scale. It is said to have been trained on approximately 1.76 trillion tokens[29], which is more than 100 times larger than all baselines.

Effect of pretraining datasets of LLMs. It becomes evident that an LLM steeped in scientific literature, even if smaller in scale, exhibits a commendable prowess in scientific tasks (Fig. 3a,b). Conversely, the Falcon series, designed for general utility, necessitates a more substantial scale to effectively navigate scientific challenges. We postulate that this phenomenon is underpinned by the emergent capabilities[30] inherent in large-scale LLMs. These capabilities empower the more expansive Falcon-40b to bridge the knowledge gap and adapt to scientific tasks. In a broader perspective, despite their relatively modest scale, the Galactica models consistently outperformed the Falcon series, underscoring the pivotal role of proper domain-specific pretraining. This is further supported by the fact that, despite its smaller scale, the Galactica series achieves performance comparable to that of GPT-4.

Contributions of knowledge synthesis and inference. It is important to assess the relative contributions of the features synthesized from literature and those inferred from data. To this end, we trained each of the classifiers using just one or the other or both forms of feature. Overall values for these three types of models were obtained by averaging over results for all tasks in a domain (Fig. 3c).

Across the 58 tasks, the combination of synthesis and inference features consistently outperformed individual methods. Specifically, in the field of physiology, an average AUC-ROC of 73.62 was achieved using both methods, compared to 70.46 with synthesis alone and 69.25 with inference. Similarly, in biophysics, combining both methods yielded an average AUC-ROC of 79.10, surpassing the scores of 76.30 and 75.76 obtained from synthesis and inference features, respectively. In physical chemistry, the combined approach resulted in an average RMSE of 1.54, which is notably better than the 1.99 from synthesis features and 1.77 from inference features. Finally, in quantum mechanics, the use of both synthesis and inference features produced an MAE of 10.42, improving on the values of 37.67 and 68.17 recorded with synthesis and inference alone.

These observations highlight the value of combining knowledge synthesis from scientific literature with inference from data. Literature imparts foundational theoretical insights, whereas empirical data identifies further regularities. The fusion of these knowledge facets equips the models with a comprehensive understanding, empowering them to excel across varied tasks and domains.

Validation of established rules

With LLM4SD outperforming specialized, state-of-the-art methods, we further validated the rules generated by Galactica-6.7b due to its superior performance and ease of reproducibility. Individual rules were validated in two ways: statistical tests to confirm their association with

https://doi.org/10.1038/s42256-025-00994-z

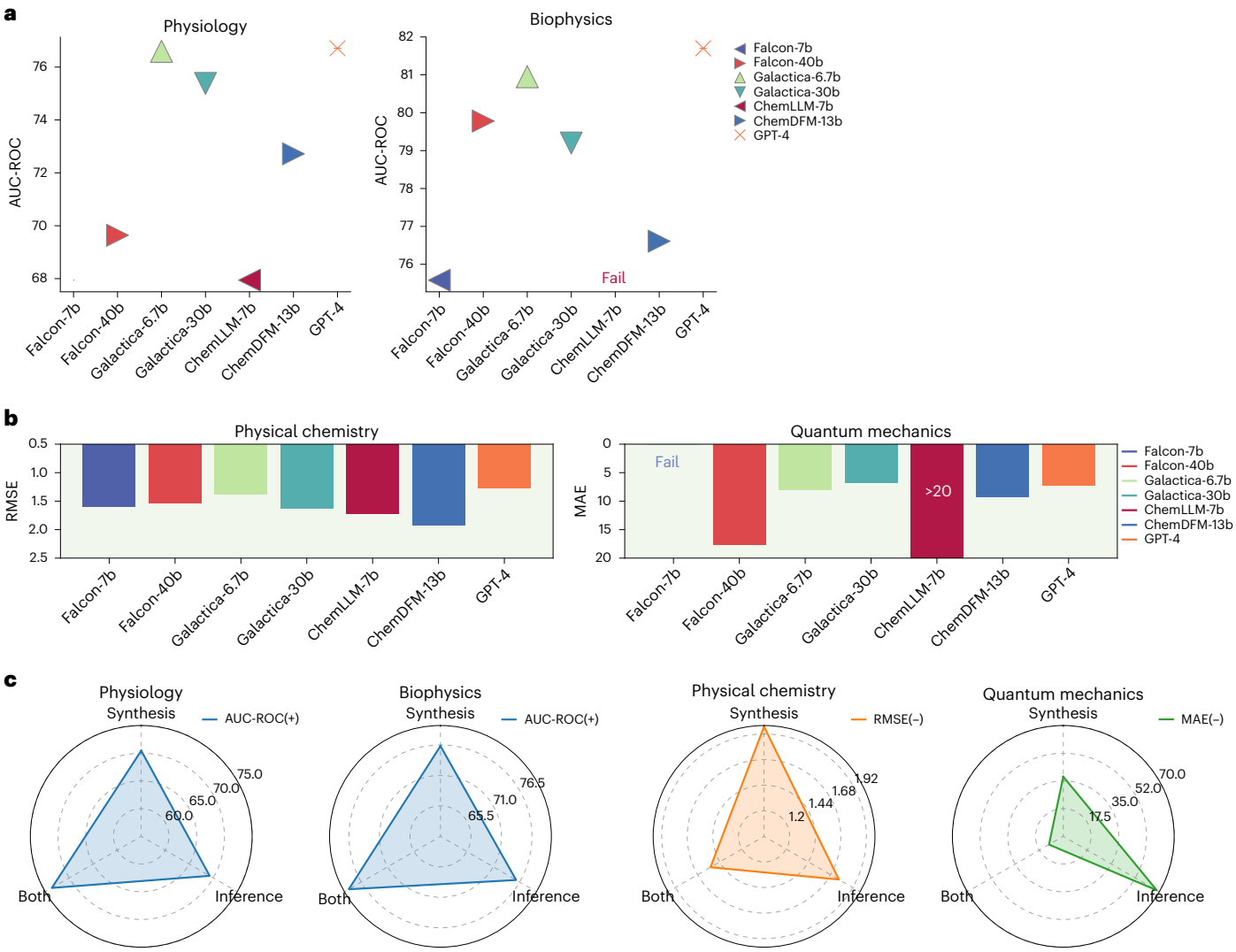


Fig. 3 | Study of LLM4SD's components. a, Performance comparison of LLM4SD using seven LLM backbones for physiology and biophysics. **b**, Performance comparison of LLM4SD using seven LLM backbones for physical chemistry and quantum mechanics. **c**, Examining the influence of both synthesized and inferred knowledge on the average (across all backbones) model performance across all four domains. The triangle's colour signifies the metric employed for domain-specific tasks. A (+) next to the metric name indicates that higher values yield better results, whereas a (−) suggests the contrary.

the target molecular attribute and a literature review to assess whether they are discussed in existing scientific literature.

We employed the Mann–Whitney *U*-test[31] of association for classification tasks and the linear correlation *t*-test for regression tasks. The Mann–Whitney *U*-test[31] compared the distributions of a chosen rule across the two classes of the target variable, thereby evaluating the statistical relevance of the rule's ability to distinguish classes. Conversely, the linear regression *t*-test treated the chosen rule as the independent variable and examined whether its coefficient significantly deviated from 0, reflecting whether the rule contributes to regression prediction.

We further carried out a comprehensive review with in-domain experts of statistically significant rules to evaluate whether they are already identified in existing literature (Supplementary Information Section 4). Specifically, the evaluation process follows a two-step approach. First, two pharmacologists independently perform a literature review using Google Scholar for any studies related to the identified statistically significant rules for the downstream task. After that, if both experts found no related articles, we categorize the rule as 'statistically significant and not found in literature'. If either identified the rule in the literature, it would be classified as 'statistically significant and found in literature'. Following this process, we categorized each rule into one of three classes: statistically significant and literature supported; statistically significant and not found in literature; or statistically insignificant (Fig. 4).

Knowledge synthesis from scientific literature. We discovered that most of the synthesized rules we examined are readily available in existing scholarly works. Notably, across all selected tasks, an overwhelming majority (85%) of these rules had statistically significant association with the target labels, affirming our pipeline's ability to summarize rules from scientific literature that were the most important for different tasks and domains.

Importantly, except for BACE[32] and Tox21-NR-Ahr9, we found no instances where statistically significant synthesized rules were absent from existing literature (Fig. 4). This aligns with the design of our pipeline: without analysing the data, LLMs tend to aggregate and summarize existing knowledge. To illustrate, in the context of BBBP, the rules generated by our pipeline were consistent with well-established determinants such as molecular weight, lipophilicity, distribution coefficient, topological polar surface area and hydrogen bonds[33–35]. These findings validate the robustness and reliability of our pipeline in leveraging LLMs to summarize existing scientific literature.

https://doi.org/10.1038/s42256-025-00994-z

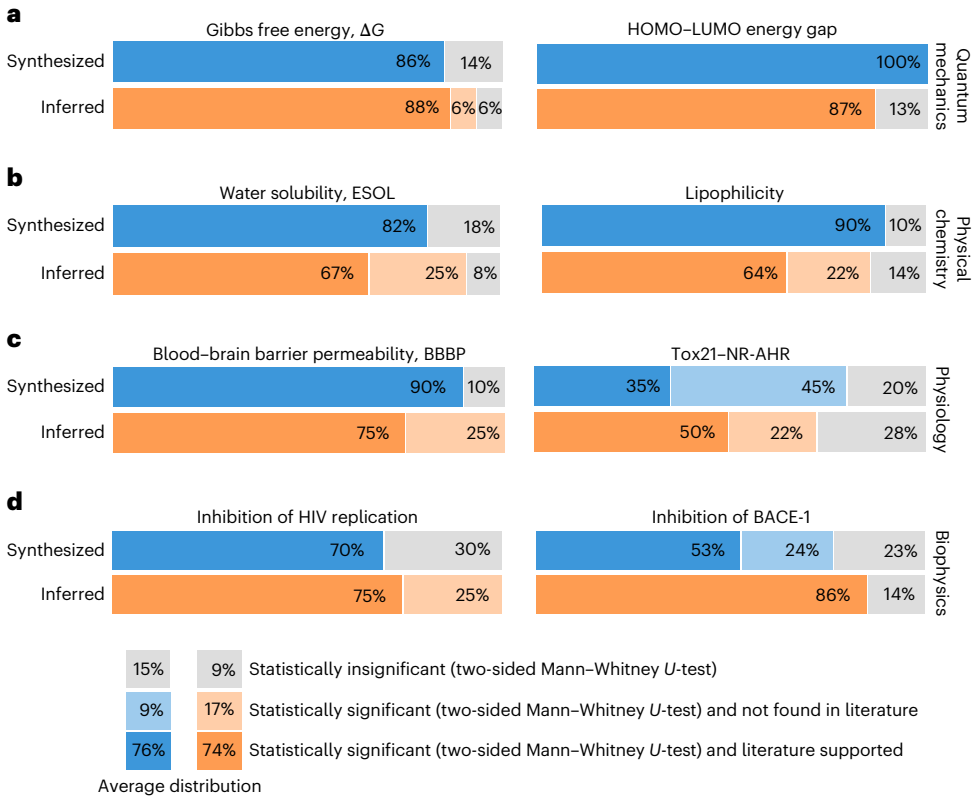


Fig. 4 | Literature review and statistical analysis of LLM rules. a–**d**, We conducted statistical analysis and a comprehensive literature review on rules generated by Galactica-6.7b across all four scientific domains, with two tasks evaluated for each domain: quantum mechanics (**a**), physical chemistry (**b**), physiology (**c**) and biophysics (**d**). In the statistical analysis, the significance of a rule is determined based on the task type: for classification, the two-sided Mann–Whitney *U*-test[31] compares the difference in distributions of chosen rule across the two classes of the target variable; and for regression, the two-sided linear regression *t*-test[44] treats the chosen rule as the independent variable and examined whether its coefficient significantly deviated from 0, reflecting whether the rule contributes to prediction. In both cases, we used a 0.05 *P* value threshold to determine rule significance. In the literature review, we assessed the prevalence of a rule in existing literature. With statistical analysis and literature review, each rule is categorized into one of three classes: statistically significant and literature supported; statistically significant and not found in literature; or statistically insignificant. Across all tasks, literature-synthesized knowledge rules were generally both prevalent in existing literature and statistically significant. In contrast, empirically inferred data rules yielded mixed results, with some easily found in existing literature and others not identified by the researchers.

Knowledge inference from data. We found that an average of 91.3% of the inferred rules were statistically significant, higher than synthesized rules (Fig. 4). Of these, an average of 74% rules were already documented in existing scientific literature, and we were unable to find prior mention of 17.3%. These latter rules were primarily associated with data patterns that are not widely discussed in the literature but can be inferred from our task dataset, such as obscure molecular substructures that influence target labels like the Gibbs free energy (Δ*G*) of a molecule.

In contrast to the knowledge synthesized from literature, we found that six out of eight tasks have statistically significant rules that we could not identify in the existing literature. This suggests that the inferred rules produced by LLM4SD are not merely a result of the LLM's textual memorization during pretraining. Instead, the inferred rules reflect a genuine capability to derive meaningful rules from data based on the specific task.

Our case studies further substantiated the utility of these unidentified but statistically significant rules. For instance, in BBBP, where 38% of rules are statistically significant but unidentified, Galactica-6.7b pinpointed the carbonyl functional group and fragment rings as key determinants of a molecule's BBBP. We hypothesize that these features are crucial for calculating a molecule's cross-sectional area, which in turn influences its orientation in lipid–water interfaces—factors vital for membrane partitioning and permeation[35]. Intriguingly, this suggests that our pipeline enables LLMs to infer 'second-order features'. These are features that may not be immediately obvious or widely recognized but are consistent with established scientific principles in literature. In doing so, LLMs not only corroborate existing knowledge but also apply existing knowledge in interpreting data, thereby enriching the current scientific discourse. An additional case study was conducted to assess the effectiveness of several lesser-explored rules (Supplementary Information Section 6).

By leveraging LLMs, our pipeline not only validates well-established scientific principles but also uncovers less documented and even potentially new rules. This facilitates a more effective and transparent interaction between scientists and the artificial intelligence (AI) system, enhancing both the quality and trustworthiness of the research output. Moreover, the statistically robust but underrepresented rules we identified could serve as promising avenues for future scientific exploration, thereby advancing the frontier of collaborative, AI-assisted research.

Discussion

Our exploration unveiled unexpected capabilities of LLM4SD through our specially designed pipeline, enabling LLMs to excel in scientific discovery in molecular prediction. Through seamless integration with our proposed architecture, LLMs exhibit state-of-the-art performance across 58 downstream tasks drawn from four domains in the MoleculeNet benchmark. The inherent versatility of LLM4SD stands as a testament to its potential for broader molecular applications across varied domains.

Scientific discovery, vast in scope, is constantly evolving with our expanding understanding of the universe. Our study only focused on molecular property prediction. However, LLM4SD's promising results in molecular property prediction hold promise for the direct application of LLMs in more advanced applications, such as protein sequence or gene sequence analysis. Protein and gene sequences are much more complicated than the SMILES strings in our study, which typically comprise up to dozens of characters. For example, protein sequences normally have 300–500 amino acids, whereas gene sequences typically contain many thousands of nucleotides. The complexity of these data poses great challenges for LLMs to understand long context domain-specific information, which requires extensive prior knowledge. In addition, LLMs need a longer and effective context window to process this data. Even for models that support large input, they often cannot use the long context input effectively in practice[36,37]. To enhance LLMs' ability to handle intricate biological data, pretraining them on vast, diverse datasets of protein or gene sequences may help the models understand the complex patterns more effectively. Additionally, incorporating retrieval-augmented generation with specialized biological knowledge bases, such as UniProt[38] and GenBank[39], can provide additional knowledge to improve understanding and contextual accuracy. Furthermore, developing efficient tokenization methods specifically tailored for biological sequences may enhance the model's ability to process and analyse this type of data, leading to more accurate and insightful results. We envision further expansion, integrating more tasks and domains, pushing LLMs to their full potential and reshaping the boundaries of scientific inquiry.

As we gaze towards the horizon, the potential trajectory for LLM4SD is compelling. We anticipate a future where the nexus between LLMs and advanced scientific toolkits deepens. As computational capabilities grow and scientific knowledge expands, our pipeline stands poised for evolutionary enhancements. Our steadfast goal is to harmoniously incorporate LLMs with myriad scientific arenas, unlocking insights and pioneering avenues previously unimagined.

Methods

Datasets

We conducted a thorough evaluation of LLM4SD, covering 58 subtasks across four unique domains for a robust assessment. The physiology domain included 41 tasks like BBBP, ClinTox and the 12-task Tox21, ranging from NR-AR to SR-p53, along with the 27-task SIDER suite covering various medical conditions. Biophysics offered two classification tasks: BACE and HIV. In physical chemistry, we addressed three regression tasks: ESOL, FreeSolv and Lipophilicity; and the quantum mechanics domain presented 12 regression tasks within the QM9 dataset, exploring properties from mu to G, providing a comprehensive insight into LLM4SD's capabilities.

Physiology

BBBP. The BBBP dataset contains 2,039 instances, each representing unique compounds labelled based on their permeability properties. Predicting which molecules can cross this barrier is paramount for drug development, especially for neurological conditions.

ClinTox. The ClinTox dataset, with 1,478 instances, provides comprehensive information on the toxicological properties of various compounds.

Tox21. With 7,831 instances, the Tox21 dataset is a collaborative effort to identify environmental toxicants. Its 12 classification tasks focus on specific biological targets or pathways. The nuclear receptor (NR) tasks, namely NR-AhR, NR-AR, NR-AR-LBD, NR-Aromatase, NR-ER, NR-ER-LBD and NR-PPAR-gamma, examine interactions with intracellular proteins influencing gene expression and potential toxic effects. The stress response (SR) tasks, including SR-ARE, SR-ATAD5, SR-HSE, SR-MMP and SR-p53, explore the impact of chemicals on stress-related cellular pathways.

SIDER. The SIDER dataset, with 1,427 instances, offers detailed data on medication side effects. Each task in this dataset relates to a specific adverse drug reaction, aiding researchers in understanding and predicting potential drug side effects. The 27 classification tasks are (1) hepatobiliary disorders; (2) metabolism and nutrition disorders; (3) product issues; (4) eye disorders; (5) investigations; (6) musculoskeletal and connective tissue disorders; (7) gastrointestinal disorders; (8) social circumstances; (9) immune system disorders; (10) reproductive system and breast disorders; (11) neoplasms benign, malignant and unspecified (including cysts and polyps); (12) general disorders and administration site conditions; (13) endocrine disorders; (14) surgical and medical procedures; (15) vascular disorders; (16) blood and lymphatic system disorders; (17) skin and subcutaneous tissue disorders; (18) congenital, familial and genetic disorders; (19) infections and infestations; (20) respiratory, thoracic and mediastinal disorders; (21) psychiatric disorders; (22) renal and urinary disorders; (23) pregnancy, puerperium and perinatal conditions; (24) ear and labyrinth disorders; (25) cardiac disorders; (26) nervous system disorders; and (27) injury, poisoning and procedural complications.

Biophysics

HIV. With a collection of 17,930 instances, the HIV dataset offers a comprehensive repository of molecules represented in the SMILES format. This dataset is instrumental in the classification of compounds based on their potential inhibitory effects against HIV.

BACE. The BACE dataset, comprising 11,908 instances, is a curated collection of molecules, each represented in the SMILES format. This dataset is tailored for classification tasks, aiming to discern molecules that can inhibit the BACE-1 enzyme. By analysing the molecules within this dataset, researchers can glean insights into the structural features that confer inhibitory properties against BACE-1.

Physical chemistry

ESOL. The ESOL dataset, comprising 1,128 instances, is a curated collection that delves into the solubility of molecules in water. By analysing the ESOL dataset, researchers can gain a deeper understanding of the molecular features that dictate solubility, thereby aiding in the design of compounds with optimal solubility profiles. Each entry in this dataset is represented using the SMILES notation, a universal language for describing the structure of chemical species.

FreeSolv. With 642 instances, the FreeSolv dataset provides comprehensive data on the hydration free energy of molecules. This dataset is pivotal for researchers aiming to predict how molecules interact with water, which has implications for drug solubility and stability. Each molecule in the FreeSolv dataset is also represented using the SMILES notation.

Lipophilicity. Lipophilicity is a fundamental property that influences the absorption, distribution, metabolism and excretion of drugs. The Lipophilicity dataset, with 4,200 compounds, offers a rich resource for understanding this property. Analysing this dataset allows researchers to discern the molecular attributes that contribute to a compound's lipophilicity, guiding the synthesis of molecules with desired pharmacokinetic properties. Like the other datasets in this domain, each entry is denoted using the SMILES notation.

Quantum mechanics

QM9. The quantum mechanics domain, central to understanding the fundamental properties of matter, is exemplified in our evaluation through the QM9 dataset. Comprising 133,885 instances, the QM9

dataset provides a comprehensive exploration of molecules' quantum mechanical attributes, essential for diverse applications from material science to pharmaceuticals. It includes 12 tasks: μ (dipole moment), α (polarizability), R^2 (squared radius), ZPVE (zero-point vibrational energy), C_v (heat capacity at constant volume), $\Delta\epsilon$ (energy gap), ϵ_{HOMO} (highest occupied molecular orbital energy), ϵ_{LUMO} (lowest unoccupied molecular orbital energy), U_0 (internal energy at 0 Kelvin), U (internal energy at standard state), H (enthalpy) and G (Gibbs free energy).

Baselines

We rigorously assessed our pipeline in comparison to specialized, state-of-the-art supervised machine learning methods. For conventional approaches, we employed random forest[40], using ECFP4 (ref. 22) as the input feature set. We also considered state-of-the-art GNNs, including attribute masking (AttrMask)[14], GraphCL[15], MolCLR[16], 3DInfomax[17], GraphMVP[18], MoleBERT[19], Grover[20] and UniMol[21]. Each of these models was initialized with pretrained weights and subsequently fine-tuned for specific tasks.

In summary, AttrMask pretraining involves teaching the GNN to predict randomly masked atom and bond attributes within molecular graphs. GROVER pretraining focuses on contextual predictions of an atom's surroundings and predicts graph-level motifs. GraphCL and MolCLR use graph augmentations for a contrastive learning objective, aimed at maximizing the similarity between augmentations originating from the same molecule while minimizing similarity between augmentations from different molecules. GraphMVP and 3DInfomax leverage existing three-dimensional (3D) molecular datasets to pretrain models capable of deducing 3D molecular geometry from two-dimensional graphs, by optimizing mutual information between 3D summary vectors and GNN graph representations. MoleBERT, the recent state-of-the-art method, employs a vector quantized variational autoencoder-based tokenizer to diversify atom vocabulary, thereby balancing dominant and rare atoms. It uses masked atoms modelling and triplet masked contrastive learning for node and graph-level pretraining, respectively. Finally, UniMol pioneers a universal 3D molecular representation learning method, which pretrained with 3D position recovery and masked atom prediction.

LLM4SD in the molecular prediction pipeline

In this section, we detail the proposed pipeline and the techniques used to align with the requirements of molecular property prediction tasks. Instead of merely prompting LLMs to generate scientific hypotheses[41] or training them for direct predictions[42], LLM4SD emulates how human experts conduct scientific research. This includes synthesizing knowledge from literature, inferring hypotheses from datasets, validating findings through experiments and elucidating the rationale behind predictions.

Knowledge synthesis from the scientific literature. LLMs are usually pretrained on large corpora of text data that include books, articles, websites and other written content. This extensive pretraining helps LLMs to learn the structure of the language, recognize patterns, understand context and acquire a wide-ranging knowledge of facts and concepts. Thus, the goal of the knowledge synthesis process is to extract relevant features from the vast pool of the knowledge that an LLM possesses from the pretraining stage.

To achieve this, we first instruct the LLM to adopt the persona of an experienced chemist and then engage it to identify pertinent features based on its existing knowledge. This form of role-playing prompt facilitates the knowledge mining process to mimic how human experts solve real-world challenges. For example, when a chemist needs to predict the bioactivity or BBBP of a molecule, they often apply feature-related rules such as number of hydrogen bond donors/acceptors, molecular weight and logP. We require that the features identified by LLMs can be measured with a numerical or categorical measure to enable their transcription into corresponding functions.

Knowledge inference from data. The objective of knowledge inference form data is to harness the powerful reasoning skills of LLMs to identify relevant features by analysing the given data. Given their impressive ability to solve mathematical problems and identify patterns, we conjecture that LLMs have the capacity to discern common patterns within groups of molecules based on their scientific understanding. To validate this hypothesis, we provide LLMs with an instruction and several batches of sampled instances with their corresponding class labels or instance property values. In the instruction, the LLM is tasked with analysing patterns from provided data to identify features that effectively discriminate between two classes of instances or predict their property values. As a result, LLMs will come up with rules distilled from the analysis for each batch. Because the generated rules in different batches may contain duplicates, we ultimately employ the LLMs' summarization capability to condense the rules and eliminate duplicates, resulting in the final list of features.

Model training. In this stage, all the features identified in the first two stages are transcribed into corresponding functions. All these functions take a scientific instance as input—for example, a SMILES string for molecules—and return a feature value. Consequently, the final representation of an instance resides in an r-dimensional space, where r is the number of features that have been identified.

These vector representations function as the feature vectors for the model training. Employing interpretable models like a linear layer or random forest enables quantification of each rule's importance in prediction, thus elucidating their contribution to the model's final decision. This transparency fosters an intuitive comprehension of the decision-making process, enhancing trust and usability among domain experts.

Interpretable insights. Following interpretable model training, our pipeline delivers clear insights, including prediction results, vector representations of molecules, rules and their corresponding importance scores. This information helps users identify key factors influencing the prediction, thereby improving interpretability.

Metrics

We assessed LLM4SD across 58 molecular property prediction tasks spanning four domains, utilizing distinct evaluation metrics tailored to each task's nature. For the domains of physiology and biophysics, the AUC-ROC metric was employed. AUC-ROC measures the ability of the model to distinguish between classes, with a range from 0 to 1, where a higher value indicates better performance. In the domain of physical chemistry, the RMSE was used. RMSE quantifies the difference between predicted and observed values, with a lower value indicating a closer fit to the true data. For quantum mechanics, we utilized the MAE metric. MAE measures the average magnitude of errors between predicted and true values, with smaller values denoting better accuracy.

Related work

Molecular property prediction. Research in molecular property prediction has explored a range of computational methods. Preliminary attempts use extended connectivity fingerprint[22] to encode molecular characteristics for traditional machine learning methods such as random forest[40]. Transitioning into neural network approaches, GROVER[20], GraphCL[15] and AttrMask[14] employ pretraining on large-scale unlabelled molecular graphs and then fine-tune for property prediction. Specifically, GROVER[20] focuses on contextual predictions of an atom's surroundings and predicts graph-level motifs, whereas GraphCL[15] uses a contrastive learning approach to maximize the similarity between

different augmented views of a molecule. AttrMask masks and predicts the attributes of nodes, such as atom type.

Recent advancements have introduced more sophisticated frameworks. MolCLR[16], for instance, extends contrastive learning with three molecular graph augmentations, atom masking, bond deletion and subgraph removal. Meanwhile, GraphMVP[18] and 3Dinfomax[17] enhance contrastive learning by incorporating 3D information about the molecules, contrasting two-dimensional and 3D views of molecules. UniMol[21] pioneers a universal 3D molecular representation learning method, which pretrains with 3D position recovery and masked atom prediction. Additionally, MolBERT[19] employs a vector quantized variational autoencoders framework to encode atoms into chemically meaningful discrete codes and trains through masked atom prediction.

Although previous methods like GNN that encode information into vectors are effective, they often lack interpretability. This makes them less useful for generating clear hypotheses, as their interpretative mechanisms, like attention, offer only a vague understanding. In contrast, our LLM4SD method produces comprehensible rules akin to human analytical processes, offering clearer and more actionable insights. Additionally, previous models cannot effectively harness prior knowledge of chemistry, which remains a challenging task. LLMs pretrained on vast amounts of knowledge, including chemistry, possess the potential to act as human domain-specific experts. Despite their promise, the application of LLMs in scientific discovery is still underexplored. Our proposed LLM4SD method leverages LLMs to drive scientific discovery in molecular property prediction, demonstrating improvements by synthesizing prior knowledge and inferring principles from data.

Experimental settings

The experimental setting consists of splitting datasets and evaluation.

Splitting datasets. We followed the MolCLR setup by using the code from MolCLR's GitHub repository to partition the dataset into an 80/10/10 split for training, validation and test sets. Specifically, for the physiology, biophysics and physical chemistry tasks, we employed a scaffold split for molecular compounds, whereas for quantum mechanics tasks, we used a random split. To ensure a fair comparison across all baselines, we reran all baseline methods using the exact same data split (that is, all models share the same SMILES-label pairs in the train, validation and test sets). This guarantees consistency in the training, validation and test sets across all baselines.

Evaluation. After constructing the dataset, molecules were converted into numerical features using the LLM4SD framework, leveraging four distinct LLM backbones: Falcon-7b, Falcon-40b, Galactica-6.7b and Galactica-30b. For each backbone, two separate sets of inferred rules were generated, corresponding to 30 and 50 iterations of data sampling, respectively. These numerical features served as input for training a random-forest model. To optimize the model's performance, a grid search was performed to identify the best hyperparameters. The optimized random-forest model was then employed to predict molecular properties on the test set. Each prediction was repeated ten times, and the final test results reflect the backbone and sampling times combination that achieved the best performance on the validation set.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The split datasets utilized in this study are entirely open source and have been made publicly available to ensure straightforward replication of our findings. We have provided presplit datasets for a variety of tasks, including BBBP, ClinTox, Tox21 (12 subtasks), SIDER (27 subtasks), HIV, BACE, ESOL, Lipophilicity, FreeSolv and QM9 (12 subtasks). These datasets are divided into training, validation and test sets based on our experimental settings. You can access them at the following GitHub repository: https://github.com/zyzisastudyreallyhardguy/LLM4SD/tree/main/scaffold_datasets. For the original raw datasets provided by MoleculeNet[9], the corresponding links are as follows: BBBP, https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/BBBP.csv; ClinTox, https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/clintox.csv.gz; Tox21, https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/tox21.csv.gz; SIDER, https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/sider.csv.gz; HIV, https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/HIV.csv; BACE, https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/bace.csv; ESOL, https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/delaney-processed.csv; FreeSolv, https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/SAMPL.csv; Lipophilicity, https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/Lipophilicity.csv; QM9, https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/qm9.csv. Source data are provided with this paper.

Code availability

In our commitment to transparency and reproducibility, we have released our code showing our implementation. This encompasses methodologies for literature knowledge mining, knowledge inference rule mining and interpretable model training. Throughout this work, we have employed several open-source libraries, including Hugging Face, numpy, rdkit, pytorch, scipy, bitsandbytes and accelerate. The GitHub link of the model is https://github.com/zyzisastudyreallyhardguy/LLM4SD (https://doi.org/10.5281/zenodo.13986921)[43]. Furthermore, we are in the process of deploying a website to facilitate scientists in utilizing LLM4SD. The site features three core functionalities for scientific users: knowledge synthesis, knowledge inference and prediction with explanations. Examples of user interactions with the website can be found in the Supplementary Information.

References

1. Bloom, N., Jones, C. I., Reenen, J. & Webb, M. Are ideas getting harder to find? *Am. Econ. Rev.* **110**, 1104–1144 (2020).
2. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
3. Frank, M. Baby steps in evaluating the capacities of large language models. *Nat. Rev. Psychol.* **2**, 451–452 (2023).
4. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process Syst.* **33**, 1877–901 (2020).
5. Achiam, J. et al. Gpt-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
6. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
7. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
8. Mirza, A. et al. Are large language models superhuman chemists? Preprint at https://arxiv.org/abs/2404.01475 (2024).
9. Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
10. Taylor, R. et al. Galactica: a large language model for science. Preprint at https://arxiv.org/abs/2211.09085 (2022).
11. Almazrouei, E. et al. The Falcon series of open language models. Preprint at https://arxiv.org/abs/2311.16867 (2023).
12. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **64**, 4–17 (2012).
13. Landrum, G. et al. rdkit/rdkit: 2024_09_5 (Q3 2024) Release (Release_2024_09_5). *Zenodo* https://doi.org/10.5281/zenodo.14779836 (2025).

14. Hu, W. et al. Strategies for pre-training graph neural networks. In *Proc. International Conference on Learning Representations* (2020).
15. You, Y. et al. Graph contrastive learning with augmentations. *Adv. Neural Inf. Process. Sys.* **33**, 5812–5823 (2020).
16. Wang, Y., Wang, J., Cao, Z. & Farimani, A. B. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **4**, 279–287 (2022).
17. Stärk, H. et al. 3D Infomax improves GNNs for molecular property prediction. In *Proc. International Conference on Machine Learning* (eds Chaudhuri, K. et al.) 20479–20502 (PMLR, 2022).
18. Liu, S. et al. Pre-training molecular graph representation with 3D geometry. In *Proc. 10th International Conference on Learning Representations* (2022).
19. Xia, J. et al. Mole-bert: rethinking pre-training graph neural networks for molecules. In *Proc. 11th International Conference on Learning Representations* (2023).
20. Rong, Y. et al. Self-supervised graph transformer on large-scale molecular data. *Adv. Neural Inf. Process. Sys.* **33**, 12559–12571 (2020).
21. Zhou, G. et al. Uni-mol: a universal 3D molecular representation learning framework. In *Proc. 11th International Conference on Learning Representations* (2023).
22. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
23. Stokes, J. M. et al. A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702.e13 (2020).
24. Wong, F. et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature* **626**, 177–185 (2024).
25. Zhang, D. et al. Chemllm: a chemical large language model. Preprint at https://arxiv.org/abs/2402.06852 (2024).
26. Zhao, Z. et al. ChemDFM: a large language foundation model for chemistry. In *38th Conference on Neural Information Processing Systems, Foundation Models for Science: Progress, Opportunities, and Challenges* (NeurIPS, 2024).
27. Cai, Z. et al. Internlm2 technical report. Preprint at https://arxiv.org/abs/2403.17297 (2024).
28. Touvron, H. et al. Llama: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
29. Haque, M. & Li, S. Exploring ChatGPT and its impact on society. *AI Ethics* https://doi.org/10.1007/s43681-024-00435-4 (2024).
30. Wei, J. et al. Emergent abilities of large language models. *Transact. Mach. Learn. Res.* https://openreview.net/pdf?id=yzkSU5zdwD (2022).
31. McKnight, P. E. & Najab, J. in *The Corsini Encyclopedia of Psychology* (eds Weiner, I. B. & Craighead, W. E.) (Wiley, 2010).
32. Subramanian, G., Ramsundar, B., Pande, V. & Denny, R. A. Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches. *J. Chem. Inf. Model.* **56**, 1936–1949 (2016).
33. Wager, T. Defining desirable central nervous system drug space through the alignment of molecular properties, in vitro adme, and safety attributes. *ACS Chem. Neurosci.* **1**, 420–434 (2010).
34. Wager, T., Hou, X., Verhoest, P. & Villalobos, A. Moving beyond rules: the development of a central nervous system multiparameter optimization (cns mpo) approach to enable alignment of druglike properties. *ACS Chem. Neurosci.* **1**, 435–449 (2010).
35. Geldenhuys, W., Mohammad, A., Adkins, C. & Lockman, P. Molecular determinants of blood–brain barrier permeation. *Ther. Deliv.* **6**, 961–971 (2015).
36. Liu, N. F. et al. Lost in the middle: how language models use long contexts. *Trans. Assoc. Comput. Linguist.* **12**, 157–173 (2024).
37. Qin, G., Feng, Y. & Van Durme, B. The NLP task effectiveness of long-range transformers. In *Proc. 17th Conference of the European Chapter of the Association for Computational Linguistics* (eds Vlachos, A. & Augenstein, I.) 3774–3790 (ACL, 2023).
38. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
39. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
40. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
41. Park, Y. J. et al. Can chatgpt be used to generate scientific hypotheses? *J. Materiomics* **10**, 578–584 (2024).
42. Honda, S., Shi, S. & Ueda, H. R. Smiles transformer: pre-trained molecular fingerprint for low data drug discovery. Preprint at https://arxiv.org/abs/1911.04738 (2019).
43. zyzisastudyreallyhardguy & Ju, J. Code repository LLM4SD: release v.1.0. *Zenodo* https://doi.org/10.5281/zenodo.13986921 (2024).
44. Student. The probable error of a mean. *Biometrika* **6**, 1–25 (1908).

Acknowledgements

H.Y.K.'s scholarship is supported by the Australian Government Research Training Programme (RTP) Scholarship and Monash University as a cocontribution to Australian Research Council grant no. ARC DP210100072. L.T.M.'s, G.I.W.'s and A.T.N.N.'s research into AI applications for drug discovery is supported by a National Health and Medical Research Council (NHMRC) of Australia Ideas grant (grant no. APP2013629). L.T.M.'s research is also supported by the National Heart Foundation of Australia (grant no. 101857). L.T.M.'s and A.T.N.N.'s research is also funded by the NHMRC of Australia and the Department of Health and Aged Care through the Medical Research Future Fund (MRFF) Stem Cell Therapies Mission (grant no. MRF2015957). Computational resources were generously provided by the Nectar Research Cloud, a collaborative Australian research platform supported by the NCRIS-funded Australian Research Data Commons (ARDC) and the MASSIVE HPC facility. We also gratefully acknowledge the support of the Griffith University eResearch Service & Specialized Platforms Team and the use of the High-Performance Computing Cluster 'Gowonda'. S.P. is supported by ARC Future Fellowship (grant no. FT210100097) and ARC grant no. DP240101547.

Author contributions

S.P. and G.I.W. supervised the project. Y.Z., H.Y.K. and J.J. contributed to the conception and design of the work. Y.Z., H.Y.K. and J.J. contributed to the technical implementation. Y.Z., H.Y.K. and J.J. prepared the figures. Y.Z. contributed to the design of the web-based application. A.T.N.N. and L.T.M. provided domain expertise for the literature review and validation of rules. Y.Z., H.Y.K., A.T.N.N. and L.T.M. contributed to the design of the rule validation test. All authors edited and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

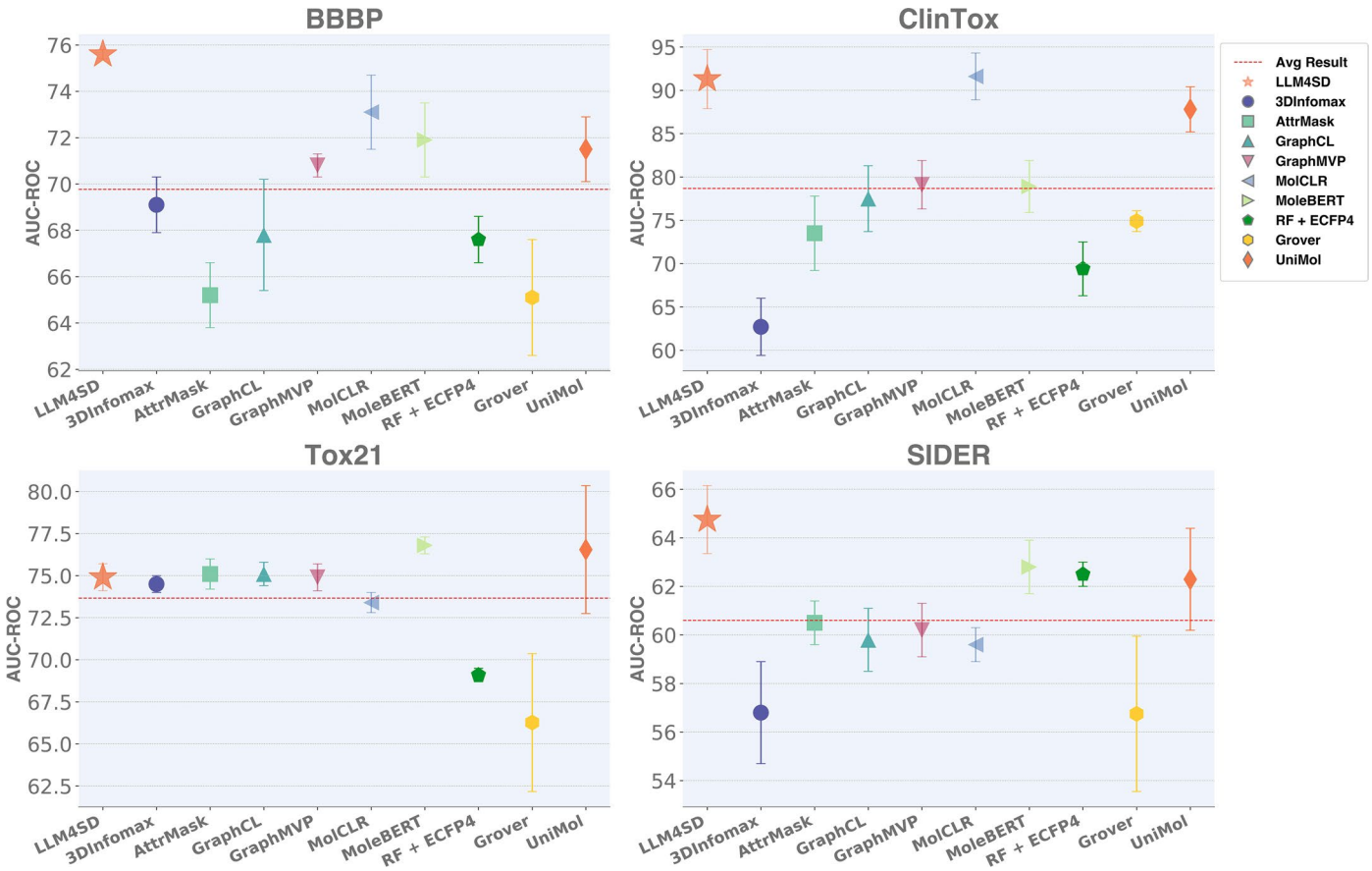
Extended data is available for this paper at https://doi.org/10.1038/s42256-025-00994-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-025-00994-z.

Correspondence and requests for materials should be addressed to Yizhen Zheng, Geoffrey I. Webb or Shirui Pan.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-025-00994-z



Extended Data Fig. 1 | Detailed performance comparison between 'LLM4SD' and nine baselines in the physiology domain. The red dashed line shows the average result across all methods. Each marker's error bar denotes the method's standard deviation, which is obtained via 10 runs. LLM4SD outperformed other models in 3 out of 4 datasets using the AUC-ROC metric and consistently surpassing the average across all datasets. The results for Tox21 and SIDER are average scores from 12 and 27 tasks respectively (see Extended Data Figs. 2 and 3 for detailed breakdown).

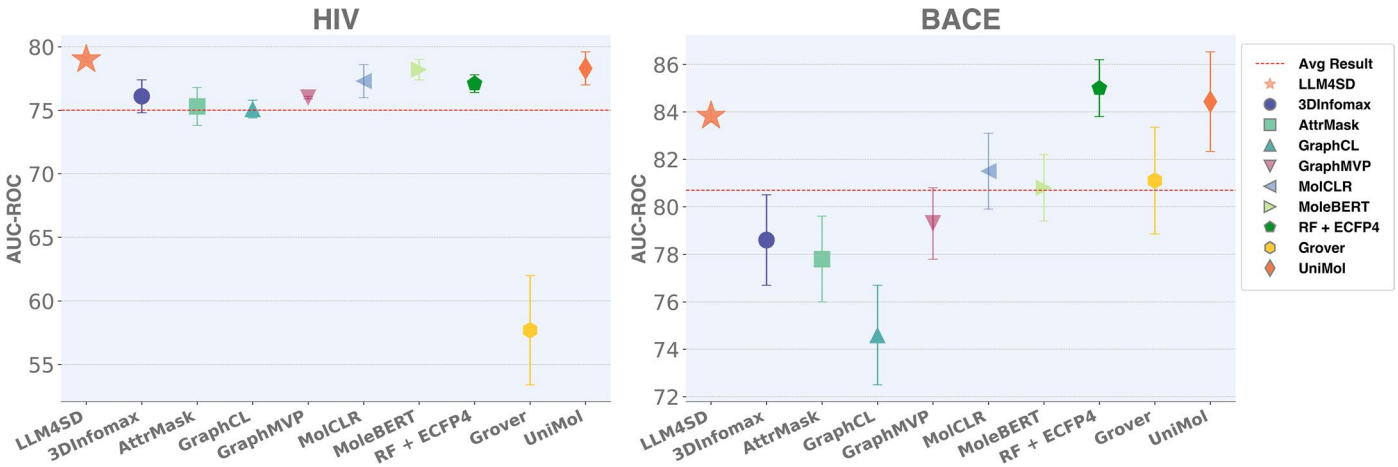
https://doi.org/10.1038/s42256-025-00994-z

Extended Data Fig. 2 | Detailed performance comparison between 'LLM4SD' and nine baselines on Tox21 Dataset. The red dashed line shows the average result across all methods. Each marker's error bar denotes the method's standard deviation, which is obtained via 10 runs. LLM4SD ranks among the top three methods in 8 out of 12 tasks and consistently outperformed the average in all tasks.

https://doi.org/10.1038/s42256-025-00994-z

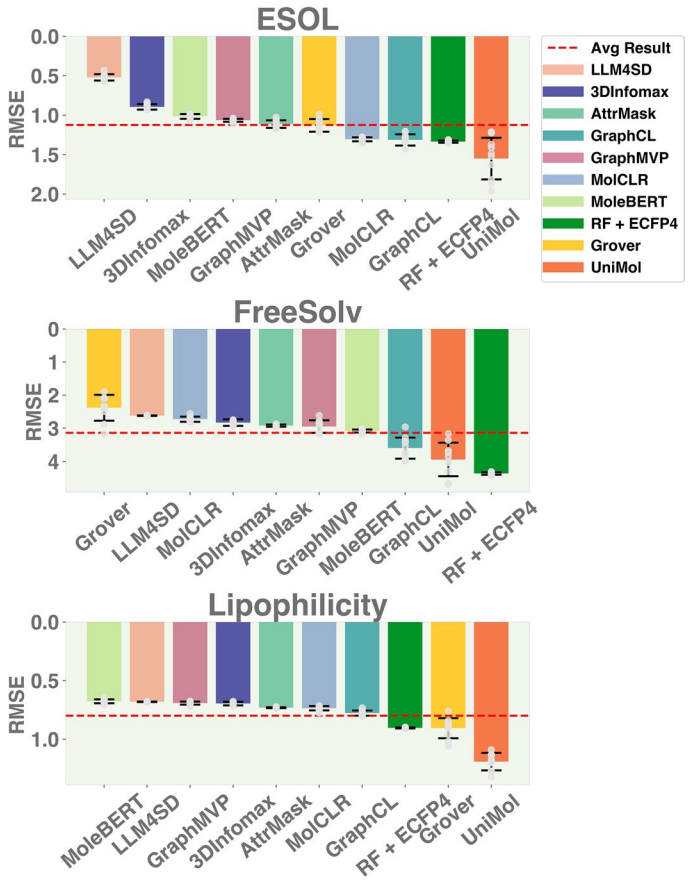
Extended Data Fig. 3 | Detailed performance comparison between 'LLM4SD' and nine baselines on Sider Dataset. The red dashed line shows the average result across all methods. Each marker's error bar denotes the method's standard deviation, which is obtained via 10 runs. LLM4SD ranks among the top three methods in 22 out of 27 tasks, and consistently outperforms the average in all tasks with the exception of the 'Psychiatric disorders' task.

https://doi.org/10.1038/s42256-025-00994-z



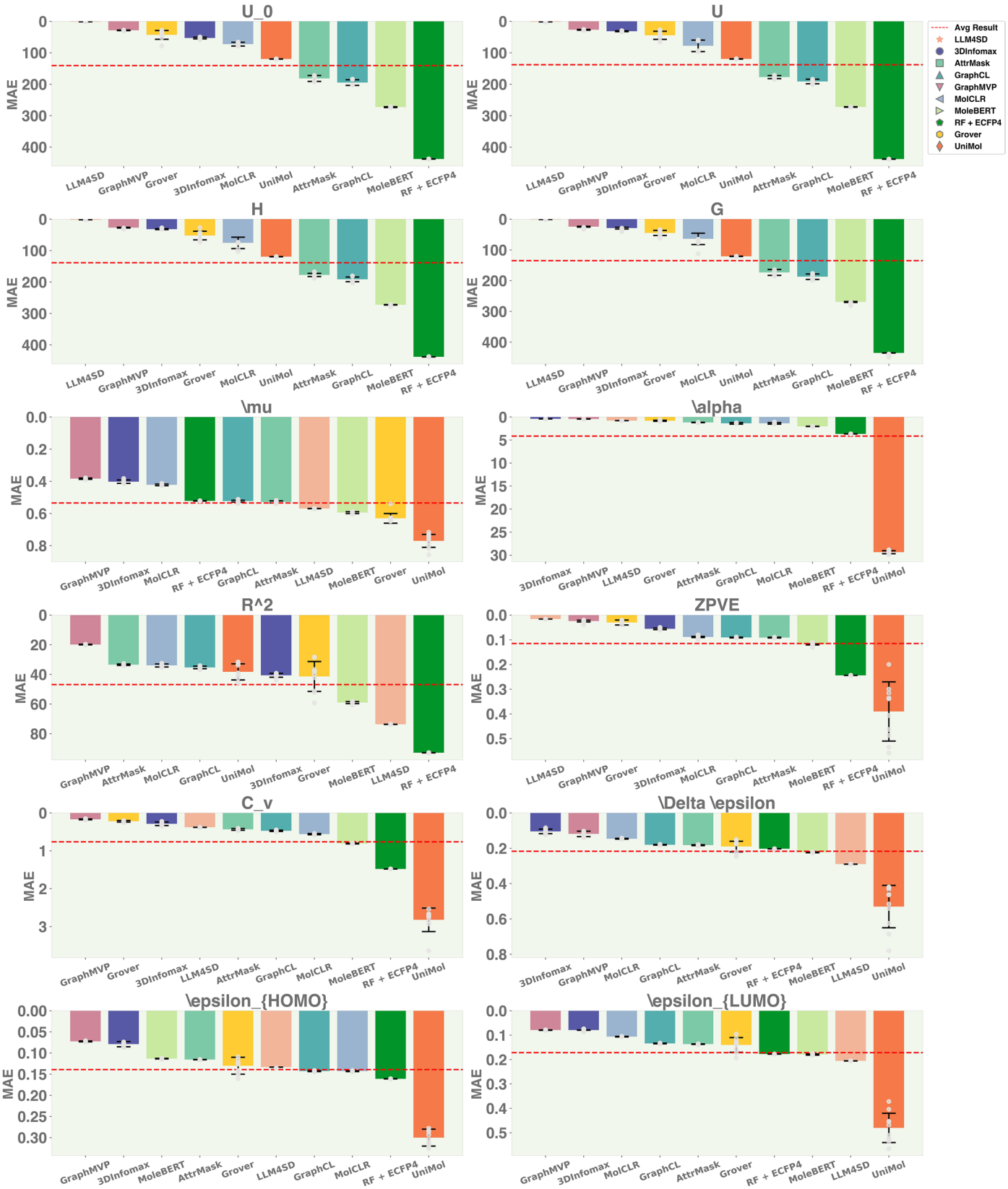
Extended Data Fig. 4 | Detailed performance comparison between 'LLM4SD' and nine baselines in the biophysics domain. The red dashed line shows the average result across all methods, in terms of AUC-ROC. Each marker's error bar denotes the method's standard deviation, which is obtained via 10 runs. LLM4SD outperformed the top-performing baseline by roughly 1% on the HIV dataset and closely matched the best performing method, UniMol. In both cases, LLM4SD delivered a visibly superior outcome compared to the average performance.

https://doi.org/10.1038/s42256-025-00994-z



Extended Data Fig. 5 | Detailed performance comparison between 'LLM4SD' and nine baselines in the physical chemistry domain. The red dashed line shows the average result across all methods. The physical chemistry domain encompasses three datasets: the ESOL dataset with 1,128 instances, the FreeSolv dataset with 642 instances, and the Lipophilicity dataset comprising 4,200 compounds. Each marker's error bar represents the method's standard deviation, calculated based on 10 independent runs (n=10). These data points are overlaid on the plot in grey colour. LLM4SD substantially outperformed all baseline methods on ESOL, demonstrating a 57% improvement over the average outcome for that dataset, and achieved state-of-the-art results on the additional datasets, FreeSolv and Lipophilicity.

https://doi.org/10.1038/s42256-025-00994-z



Extended Data Fig. 6 | Detailed performance comparison between 'LLM4SD' and nine baselines in the quantum mechanics domain. The red dashed line shows the average result across all methods. The quantum mechanics domain includes the QM9 datasets with 12 subtasks, comprising 133,885 instances. Each marker's error bar denotes the method's standard deviation, which is obtained via 10 runs (n=10). These data points are overlaid on the plot in grey colour. LLM4SD excelled in predicting properties such as U0, U, H, and G, showing substantial enhancements. In other tasks, the results from LLM4SD were comparable to the average of all methods.

 https://doi.org/10.1038/s42256-025-00994-z

Extended Data Table 1 | General prompt for Knowledge Synthesis from the Scientific Literature and Knowledge Inference from Data

Prompts for Knowledge Synthesis from the Scientific Literature
Classification Tasks (General): Assume you are an experienced biologist/chemist. Please come up with **[20 rules/30 rules]** that you think are very important to predict **[Task Description]**. Each rule is either about the structure or property of a molecule.
Regression Tasks (General): Assume you are an experienced biologist/chemist. Please come up with **[20 rules/30 rules]** that you think are very important to predict **[Task Description]**. Each rule is either about the structure or property of a molecule **(without access to 3D information)**. **Note:** • "20 rules" is for smaller LLMs prompt while "30 rules" is for larger LLMs prompt. • "without access to 3D information" is added only for QM9 dataset.
Prompts for Knowledge Inference from Data
Classification Tasks (General): Assume you are a very experienced biologist/chemist. In the following data, with label 1, it means **[Task Description]**. With label 0, it means it is not. Please infer step-by-step to come up with 3 rules that directly relate the properties/structures of a molecule.
Regression Tasks (General): Assume you are a very experienced biologist/chemist. The following data includes molecules and their corresponding value **[Task Description]**. Please infer step-by-step to come up with 3 rules that directly relate the properties/structures of a molecule **(without access to 3D information)**. **Note:** • "without access to 3D information" is added only for QM9 dataset.

https://doi.org/10.1038/s42256-025-00994-z

Extended Data Table 2 | Classification task descriptions for the general prompt in Extended Data Table 1

Classification Task Name		Task Description for Table X General Prompt
Note: add 'if' for Knowledge Synthesis from the Scientific Literature prompt		
BBBP		**(if)** a molecule is blood brain barrier permeable (BBBP)
ClinTox		**(if)** a molecule will be approved by the FDA
BACE		**(if)** a molecule can inhibit human β-secretase 1(BACE-1)
HIV		**(if)** a molecule can inhibit HIV replication.
Note: add 'it is related to' for Knowledge Inference from Data prompt		
Tox21		**(it is related to)** the toxicity activity of a molecule against the **[subtask description]** in the **[nuclear receptor (NR)/ stress response (SR)]** signalling pathway.
Subtask Description	nr-ar	androgen receptor
	nr-ar-lbd	androgen receptor ligand-binding domain
	nr-ahr	aryl hydrocarbon receptor
	nr-aromatase	aromatase
	nr-er	estrogen receptor
	nr-er-lbd	estrogen receptor ligand-binding domain
	nr-ppar-gamma	peroxisome proliferator activated receptor
	sr-are	nuclear factor (erythroid- derived 2)-like 2 antioxidant responsive element
	sr-atad5	genotoxicity indicated by ATAD5
	sr-hse	heat shock factor response element
	sr-mmp	mitochondrial membrane potential
	sr-p53	DNA damage p53-pathway
Sider		**(it is related to)** the side-effect activity of a molecule in causing **[subtask name].**

Subtask names					
	respiratory, thoracic and mediastinal disorders	metabolism and nutrition disorders	product issues	eye disorders	investigations
	musculoskeletal and connective tissue disorders	blood and lymphatic system disorders	immune system disorders	social circumstances	hepatobiliary disorders
	general disorders and administration site conditions	surgical and medical procedures	cardiac disorders	vascular disorders	endocrine disorders
	skin and subcutaneous tissue disorders	congenital, familial and genetic disorders	infections and infestations	renal and urinary disorders	psychiatric disorders
	pregnancy, puerperium and perinatal conditions	reproductive system and breast disorders	ear and labyrinth disorders	gastrointestinal disorders	nervous system disorders
	injury, poisoning and procedural complications	neoplasms benign, malignant and unspecified (incl cysts and polyps)			

https://doi.org/10.1038/s42256-025-00994-z

Extended Data Table 3 | Regression task descriptions for the general prompt in Extended Data Table 1

Regression Task Name	Task Description for Table X General Prompt
ESOL	the water solubility data (log solubility in mols per litre)
FreeSolv	the octanol/water distribution coefficient (logD at pH 7.4)
Lipophilicity	the hydration free energy of a molecule in water
Quantum Mechanics	
μ	dipole moment (Mu) of a molecule
α	Isotropic polarizability of a molecule
R^2	electronic spatial extent of a molecule
$ZPVE$	Zero-Point Vibrational Energy (ZPVE) of a molecule
C_v	the heat capacity at constant volume of a molecule
Δ_ϵ	the HUMO-LUMO gap of a molecule
ϵ_{HOMO}	the highest occupied molecular orbital (HOMO) energy of a molecule
ϵ_{LUMO}	Lowest Unoccupied Molecular Orbital (LUMO) energy of a molecule
U_0	internal energy at absolute zero temperature (0 Kelvin), U0, of a molecule
U	internal energy of a molecule at a specific temperature, specifically at 298.15 Kelvin (approximately room temperature), (U) of a molecule
H	enthalpy of the molecule at a specific temperature, specifically at 298.15 Kelvin (approximately room temperature), (U) of a molecule
G	Gibbs free energy of the molecule at a specific temperature, specifically at 298.15 Kelvin (approximately room temperature), (U) of a molecule

nature portfolio

Corresponding author(s): Shirui Pan

Last updated by author(s): 22/12/2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: Code is available at https://github.com/zyzisastudyreallyhardguy/LLM4SD.

Data analysis: All software description used in the study is shown below:

accelerate==0.20.3
numpy==1.24.3
pandas==1.5.3
scipy==1.11.1
six==1.16.0
future==0.18.3
tqdm==4.65.0
torch==2.0.1
transformers==4.30.2
tensorboard==2.13.0
scikit-learn==1.2.2
dask==2023.7.1
nltk==3.8.1
textblob==0.17.1
sentencepiece==0.1.99
uvicorn==0.22.0
websocket-client==1.6.1

websockets==11.0.3
httpx==0.24.1
rdkit==2023.3.2
nltk==3.8.1
textblob==0.17.1
sentencepiece==0.1.99
openai==0.28
mordred
tiktoken
einops

All baseline methods' corresponding Github Repository are shown below:

3DInfomax: https://github.com/HannesStark/3DInfomax
AttrMask: https://github.com/snap-stanford/pretrain-gnns
GraphCL: https://github.com/Shen-Lab/GraphCL
GraphMVP: https://github.com/chao1224/GraphMVP
MolCLR: https://github.com/yuyangw/MolCLR
MoleBERT: https://github.com/junxia97/Mole-BERT
GROVER: https://github.com/tencent-ailab/grover
UniMol: https://github.com/deepmodeling/Uni-Mol/tree/main/unimol
RandomForest: sklearn.ensemble.RandomForestRegressor/sklearn.ensemble.RandomForestClassifier (scikit-learn==1.2.2)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Splitted Datasets:
The splitted datasets utilized in this study are entirely open-source and have been made publicly available to ensure straightforward replication of our findings. We have provided pre-split datasets for a variety of tasks, including BBBP, ClinTox, Tox21 (12 subtasks), SIDER (27 subtasks), HIV, BACE, ESOL, Lipophilicity, FreeSolv, and QM9 (12 subtasks). These datasets are divided into training, validation, and test sets based on our experimental settings. You can access them at the following GitHub repository:
https://github.com/zyzisastudyreallyhardguy/LLM4SD/tree/main/scaffold_datasets.

Raw Datasets:
For the original raw datasets provided by MoleculeNet (Wu et al., 2018), their corresponding links are shown below:

- BBBP: https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/BBBP.csv
- ClinTox: https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/clintox.csv.gz
- Tox21: https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/tox21.csv.gz
- SIDER: https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/sider.csv.gz
- HIV: https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/HIV.csv
- BACE: https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/bace.csv
- ESOL: https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/delaney-processed.csv
- FreeSolv: https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/SAMPL.csv
- Lipophilicity: https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/Lipophilicity.csv
- QM9: https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/qm9.csv est-1.amazonaws.com/datasets/qm9.csv)

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size of the datasets used in this study varied, ranging from approximately 1,000 to over 100,000 data points. The sample sizes were determined by the datasets provided, as we utilized them in their entirety without any modifications. No additional sample size calculations were performed.
Data exclusions	No data were excluded from the anlaysis.
Replication	We conduct experiments 10 times for each task and record to ensure the results are reproducible.
Randomization	We adopted the same dataset splitting method as MolCLR [1] framework, using the code from MolCLR's GitHub repository https://github.com/yuyangw/MolCLR to divide the dataset into an 80/10/10 split for training, validation, and test sets. Specifically, for all datasets except QM9, instead of random splitting, we used a scaffold split, which organizes data based on molecular substructures. This approach introduces a more rigorous and realistic challenge compared to random splitting, making the prediction task more challenging yet realistic as suggested in [1,2]. For QM9, we adopted random splitting, as it is a standard practice in related studies, ensuring consistent benchmarking[1]. [1] Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. Nature Machine Intelligence, 4(3), 279–287. [2] Hu, W. et al. Strategies for pre-training graph neural networks. In International Conference on Learning Representations (2020).
Blinding	The model was blinded to the group allocation. During model training, the training and test sets were strictly separated to ensure that the model did not access test set data, preventing information leakage. This approach effectively maintains the integrity of the evaluation process and avoids potential bias in the results.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging