A Multi-Modal Large Language Model for Free-Form, Open-Ended, and Interactive Prediction of Properties and Mechanisms of Candidate Drug Molecules

Youwei Liang [1] **Ruiyi Zhang** [1] **Zinnia Ma** [2] **Digvijay Singh** [3] **Yongce Li** [1] **Mingjia Huo** [1] **Chengzhan Gao** [1] **Hamidreza Rahmani** [4] **Satvik Bandi** [2] **Li Zhang** [1] **Robert Weinreb** [5] **Atul Malhotra** [6] **Danielle A. Grotjahn** [4] **Linda Awdishu** [7] **Trey Ideker** [6] **Michael Gilson** [7] **Pengtao Xie** [1 6]

Abstract

Accurately predicting the mechanisms and properties of candidate drug molecules is critical for advancing drug discovery. However, existing models are often limited to structured outputs, fixed task sets, and static, one-shot predictions. We present DrugChat, a multi-modal large language model that addresses these limitations through three key capabilities: (i) free-form text generation for predicting complex drug attributes such as indications, pharmacodynamics, and mechanisms of action; (ii) generalization to an open-ended set of tasks via prompt-based multi-task learning; and (iii) interactive, multi-turn dialogue for dynamic exploration for molecules. DrugChat integrates a molecular graph encoder, a molecular image encoder, and an instruction-tuned large language model. Pretrained on 248 million molecule-bioactivity records, DrugChat outperforms existing baselines across both unstructured and structured tasks, demonstrating strong zero-shot generalization.

1. Introduction

Accurate prediction of the mechanisms and properties of small molecules is crucial for advancing pharmaceutical research and facilitating drug discovery (Kirchmair et al., 2015). A significant body of work has focused on developing quantitative structure-activity relationship (QSAR) models (Tropsha, 2010). QSAR models enable virtual screening of large chemical libraries (Kitchen et al., 2004), reducing the need for costly synthesis and animal testing in pharmacology and toxicology (Luechtefeld et al., 2018; Madden et al., 2020). Deep learning has been used in QSAR modeling (Chen et al., 2018). Although substantial progress has been made in deep learning approaches for molecular property prediction, existing methods still face critical limitations that restrict their effectiveness and flexibility. Many key attributes of candidate drug molecules—such as indications, pharmacodynamics, and mechanisms of action—are inherently complex and context-dependent, making them better suited for description in free-form texts rather than rigid categorical or numerical formats. Despite their importance, most current methods are limited to producing structured outputs, such as discrete classes, scalar values, and dose-response curves (Yang et al., 2019a; Lu et al., 2021; Chen et al., 2020; Zeng et al., 2019).

In addition to their reliance on structured outputs, many existing multi-task deep learning methods for molecular property prediction (Dahl et al., 2014; Simões et al., 2018; Allenspach et al., 2024; Qian et al., 2023) face another fundamental limitation: they restrict predictions to a fixed set of tasks defined during training. These models rely on task-specific output heads, which restrict them to making predictions only for tasks they were explicitly trained on. As a result, they lack the flexibility to generalize to new, previously unseen tasks that arise at inference time—such as predicting activities for novel assays, emerging biological targets, or new therapeutic contexts.

To address these challenges, we propose DrugChat, a multi-modal large language model (LLM) framework designed to advance molecular property prediction beyond the limitations of traditional methods. First, DrugChat is capable of generating free-form textual predictions, enabling rich and nuanced descriptions of complex drug attributes such as indications, pharmacodynamics, and mech-

[1]Department of Electrical and Computer Engineering, UC San Diego [2]Department of Bioengineering, UC San Diego [3]School of Biological Sciences, UC San Diego [4]Department of Integrative Structural and Computational Biology, The Scripps Research Institute [5]Viterbi Family Department of Ophthalmology, UC San Diego [6]Department of Medicine, UC San Diego [7]School of Pharmacy and Pharmaceutical Science, UC San Diego. Correspondence to: Pengtao Xie <pxie1@ucsd.edu>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025.

anisms of action. Second, DrugChat enables open-ended, prompt-based multi-task learning. By combining a large, instruction-tuned language model with specialized molecular encoders, DrugChat can generalize to a large set of prediction tasks, including tasks not encountered during training. Finally, DrugChat enables dynamic, interactive exploration of molecules through multi-turn dialogue. Users can ask follow-up questions, refine their inquiries, and iteratively investigate different aspects of a molecule's properties and mechanisms within a continuous conversational flow. Together, these capabilities allow DrugChat to provide a more flexible, expressive, and powerful framework for molecular property prediction and analysis.

We conducted extensive experiments to evaluate DrugChat's capabilities across a range of tasks. For free-form prediction tasks involving drug indications, pharmacodynamics, and mechanisms of action, DrugChat significantly outperformed Galactica and GPT-4 in both human expert evaluations and automatic metrics. For structured QSAR tasks, including cytotoxicity prediction, administration route classification, and molecular property prediction on the MoleculeNet (Wu et al., 2018) benchmark, DrugChat consistently outperformed GNN-based models, multi-modal models, and scientific and general-purpose LLMs. Notably, on the challenging FS-Mol benchmark (Stanley et al., 2021)—where tasks and compounds are both unseen during training—DrugChat achieved strong zero-shot generalization. These results demonstrate DrugChat's ability to flexibly handle both unstructured and structured tasks, generalize to unseen molecular prediction tasks, and support interactive exploration of drug properties.

2. DrugChat Overview

DrugChat integrates a molecular graph encoder, a molecular image encoder, an instruction-tuned large language model (LLM), and two adapters that map molecular features into the LLM's latent space (Fig.1b). The graph encoder is a Graph Isomorphism Network (GIN)(Xu et al., 2018), the image encoder is a ResNet (He et al., 2016), and the LLM is Vicuna (Chiang et al., 2023). A molecule is input as a SMILES string, which is transformed into a molecular graph and image. The GIN encodes the graph using node/edge features and iterative message passing (Goodfellow et al., 2016), followed by attention-based pooling (Wu et al., 2020). It is pretrained on two million ZINC15 molecules (Sterling & Irwin, 2015) via self-supervised learning (Hu et al., 2020). The ResNet, trained on ten million PubChem molecules (Kim et al., 2023), processes the image using convolutional layers and average pooling (Krizhevsky et al., 2012) to generate a representation vector. Linear adapters convert graph and image vectors into a unified molecule token, compatible with the LLM. This token is inserted into the LLM's language token stream, enabling Vicuna-13B (Chiang et al., 2023) to generate outputs via autoregressive decoding (Vaswani et al., 2017). Each output token includes a probability score, supporting confidence estimation and performance metrics like AUROC and ΔAUPRC.

DrugChat is trained in two stages: large-scale pretraining on 4M compounds and 248M activity records from PubChem (Kim et al., 2023), followed by finetuning on 500K compound-activity records across >5,000 tasks. Finetuning datasets include QSAR benchmarks (Stanley et al., 2021; Wu et al., 2018), DrugBank (Wishart et al., 2006), and ChEMBL (Mendez et al., 2019), covering >230,000 compounds with both structured and free-text activity labels. Training data consist of molecule-prompt-answer triplets, with answers ranging from yes/no to descriptive text. DrugChat is optimized by minimizing the negative log-likelihood between its predictions and the annotated answers (Sutskever et al., 2014). The details of our method can be found in Sec. B.

3. Experiments

3.1. DrugChat generates free-form predictions for drug indications, pharmacodynamics, and mechanisms of action

We evaluated DrugChat's ability to predict drug attributes such as indications, pharmacodynamics, mechanisms of action, and drug overviews using free-form prompts (e.g., "what is its indication?"). These attributes were curated from DrugBank (Wishart et al., 2006), which includes 5,846 approved and experimental compounds. Each drug also has an expert-written overview, which DrugChat was tasked with generating. We employed nested 5-fold cross-validation with scaffold-based splits (Bemis & Murcko, 1996) to avoid compound-series bias (Baumann & Baumann, 2014), ensuring structurally similar molecules were not in both training and test sets. DrugChat was compared against Galactica (Taylor et al., 2022), a 6.7B-parameter scientific LLM, and GPT-4 (Achiam et al., 2023), a general-purpose model likely exposed to chemical data via CommonCrawl. All models received SMILES strings as input (Methods).

Human experts scored model outputs (0 = incorrect, 1 = partially correct, 2 = correct) based on DrugBank annotations. DrugChat outperformed GPT-4 and Galactica across all tasks, with average scores of 1.025 (indication), 1.001 (pharmacodynamics), 0.846 (mechanism), and 0.925 (overview). GPT-4 and Galactica scored significantly lower ($p < 1.6 \times 10^{-5}$ and $p < 3.2 \times 10^{-4}$, respectively). For example, DrugChat had 40% correct predictions for indications, versus 12.5% for GPT-4 and 6.2% for Galactica. Fig. 3 shows qualitative differences. DrugChat correctly identified a diabetes drug's indication and pharmacodynam-

ics, while GPT-4 and Galactica failed. Galactica often gave irrelevant outputs due to lack of instruction tuning, and GPT-4 sometimes contradicted itself—likely due to its limited grounding in molecular structure. In contrast, DrugChat's outputs remained consistent and chemically grounded. Automated evaluations using semantic similarity, BLEU (Papineni et al., 2002), and METEOR (Lavie & Denkowski, 2009) confirmed DrugChat's superiority. It scored 0.460 in semantic similarity (vs. 0.317 for GPT-4, 0.238 for Galactica), 0.311 in BLEU (vs. 0.175 and 0.120), and 0.219 in METEOR (vs. 0.141 and 0.092). DrugChat's advantage lies in its molecule-aware architecture. While GPT-4 and Galactica treat SMILES as plain text, DrugChat uses a GIN and ResNet, pretrained on large-scale chemical data, to extract structural and visual features. This enables better understanding of molecular context, critical for accurate drug attribute prediction.

3.2. DrugChat achieves competitive results in multi-task QSAR modeling

Beyond generating free-form responses, DrugChat performs traditional QSAR modeling by predicting binary outcomes (yes/no) with associated confidence scores. These token-level probabilities enable computation of standard metrics such as AUROC and ΔAUPRC (Methods).

We evaluated DrugChat on three benchmark datasets: (1) a cytotoxicity dataset (Wong et al., 2024) with 39,043 compounds tested against HepG2, HSkMC, and IMR-90 cell lines (117,129 measurements), (2) the ChEMBL dataset (Gaulton et al., 2012) with 3,462 compounds across four classification tasks (e.g., oral/parenteral/topical administration and prodrug potential), and (3) subsets from MoleculeNet (Wu et al., 2018), including BACE, BBBP, ClinTox, and SIDER. Cytotoxicity and ChEMBL evaluations used scaffold-based 5-fold cross-validation to reduce compound-series bias. MoleculeNet tasks followed standard splits, with scaffold-based splits for BACE/BBBP and random splits for ClinTox/SIDER. Results were averaged over five runs, reporting mean and standard deviation. Evaluation used AUROC and ΔAUPRC, which adjusts for dataset imbalance by subtracting the base rate of actives from AUPRC (Wu et al., 2018). We compared DrugChat against Galactica (Taylor et al., 2022), general-domain LLMs (ChatGLM (GLM et al., 2024), FastChat-T5 (Zheng et al., 2024), LLaMA v2 (Touvron et al., 2023b)), GNN baselines (MPNN (Gilmer et al., 2017), Chemprop (Yang et al., 2019b)), and multi-modal models (Text2Mol (Edwards et al., 2021), KV-PLM (Zeng et al., 2022b), CLAMP (Seidl et al., 2023)). All models received SMILES strings as input prompts and returned binary predictions.

DrugChat outperformed all baselines across the three cytotoxicity tasks in both AUROC and ΔAUPRC (Fig. 4a), with statistically significant improvements ($p < 0.03$). It also surpassed all baselines on ChEMBL tasks, including prodrug status and administration routes, with 70% of improvements being statistically significant ($p < 0.05$). On MoleculeNet datasets, DrugChat exceeded Galactica, Text2Mol, and KV-PLM, and matched or slightly outperformed Chemprop, MPNN, and CLAMP. DrugChat's edge stems from its dual molecular encoders—a GIN for graph structure and a ResNet for molecular images—enabling richer molecular representations than GNN-only (e.g., MPNN, Chemprop) or single-encoder models (e.g., KV-PLM). Unlike LLMs that treat SMILES as plain text, DrugChat's encoders are pretrained on large chemical datasets, allowing it to extract chemically meaningful features. Its instruction-tuned LLM further enhances understanding of biomedical prompts. Even compared to CLAMP, DrugChat's design enables better generalization and accuracy across diverse QSAR tasks.

3.3. DrugChat demonstrates strong zero-shot generalization to unseen compounds and tasks

We evaluated DrugChat's zero-shot generalization using the FS-Mol benchmark (Stanley et al., 2021), which includes 5,120 protein-target assays and 233,786 compounds. Each assay is a separate prediction task. Following the official split, we assessed performance on 157 test tasks and 27,520 compounds, including 58 entirely unseen tasks and 14,064 unseen compounds. Two evaluation settings were used: full zero-shot (only unseen tasks and compounds) and partial zero-shot (all test tasks, including seen and unseen compounds). DrugChat was compared against multi-modal models (Text2Mol, KV-PLM), Galactica, and the zero-shot model CLAMP. GNN-based models (e.g., Chemprop) were excluded as they can't generalize to new tasks without retraining. In the full zero-shot setting, DrugChat achieved the highest AUROC (0.598) and ΔAUPRC (0.106), outperforming CLAMP (AUROC 0.549, ΔAUPRC 0.078; $p < 10^{-5}$), as well as Galactica, Text2Mol, and KV-PLM (all $p < 2.6 \times 10^{-4}$). In partial zero-shot, DrugChat continued to outperform all but performed on par with CLAMP (Fig. 6b). DrugChat's superior generalization is driven by its dual encoders (GIN and ResNet), which extract rich chemical features, and its large language model, which understands complex biomedical contexts. Combined with pretraining on hundreds of millions of activity records, DrugChat shows strong zero-shot performance across novel molecular tasks.

3.4. Pretraining on large-scale bioassay data significantly improves DrugChat's generalization

To assess the value of pretraining, we trained DrugChat From Scratch (DC-FS) using only the cytotoxicity dataset, without initializing from the PubChem checkpoint. DC-FS achieved mean ΔAUPRCs of 0.210, 0.207, and 0.190

on the HepG2, HSkMC, and IMR-90 tasks, respectively. In comparison, pretrained DrugChat achieved significantly higher scores: 0.306, 0.318, and 0.235 ($p = 0.002$, 0.022, and 0.027; Fig. 7).

These results underscore the importance of large-scale pretraining. It equips DrugChat's GIN and ResNet encoders with stronger feature extractors and helps the adapters align molecular and textual features more effectively (Li et al., 2022). Pretraining also mitigates overfitting on small datasets (Hendrycks et al., 2019), providing generalizable molecular representations and reducing reliance on spurious correlations during fine-tuning.

3.5. DrugChat's integration of molecular graph and image modalities surpasses single-modality variants

To evaluate the role of different molecular encoders, we created two DrugChat variants: DrugChat-Graph, using only graph-based representations via a GNN, and DrugChat-Image, using only image-based features via a CNN (ResNet). On the cytotoxicity dataset, the original DrugChat—which integrates both modalities—consistently outperformed both variants (Fig. 7), confirming the advantage of combining graph and image features. Graphs capture atomic connectivity, while images highlight spatial and functional patterns. The fusion of these complementary views enables a richer, more comprehensive molecular representation, improving prediction accuracy. In contrast, single-modality models miss important information, limiting performance. Notably, DrugChat-Image (DC-I) performed comparably to GNN models like MPNN (Gilmer et al., 2017) and Chemprop (Yang et al., 2019b), demonstrating the strength of image-based encoders in extracting meaningful molecular features.

3.6. DrugChat enables dynamic, iterative exploration of drug mechanisms and properties

DrugChat supports multi-turn interactions, allowing users to ask follow-up questions about the same molecule. Starting with an initial query, users can engage in iterative dialogue to explore molecular properties in greater depth. Fig. 8 shows an example where DrugChat accurately answered questions about a molecule's indication (hypertension), mechanism of action (ACE inhibition), site of action (lungs), and the RAAS pathway. This dialogue demonstrates DrugChat's ability to interpret user intent and deliver precise, coherent biomedical responses, enabling rich and informative exploration.

4. Discussion and Future Work

DrugChat is a multi-modal large language model (LLM) that introduces three key capabilities: (1) Free-form Text Predictions: Unlike prior models that output structured labels or values (Yang et al., 2019a; Chen et al., 2020), DrugChat generates natural language descriptions for complex drug attributes such as indications, pharmacodynamics, and mechanisms of action. While Galactica (Taylor et al., 2022) supports text generation, it lacks molecular encoders and instruction tuning, both crucial to DrugChat's accuracy. (2) Prompt-based Multi-task Learning: DrugChat generalizes to an open-ended task space without retraining. By simply modifying prompts, it can handle both standard QSAR tasks and complex biomedical queries, including unseen tasks. In contrast, traditional GNNs and multi-task models (Yang et al., 2019b; Dahl et al., 2014) require fixed output heads and retraining. Existing multi-modal models like KV-PLM (Zeng et al., 2022b), CLAMP (Seidl et al., 2023), and Text2Mol (Edwards et al., 2021) support structured outputs for unseen tasks but not free-form responses. (3) Interactive Dialogue: DrugChat enables multi-turn conversations, allowing users to refine questions and explore molecular properties iteratively—a feature not available in other models.

DrugChat embeds structured (e.g., cytotoxicity) and unstructured (e.g., pharmacodynamics) molecular features into a shared latent space, allowing it to learn richer representations and generalize better than models trained on isolated tasks. This unified approach helps uncover cross-domain patterns missed by more narrowly focused models like Chemprop. While databases like DrugBank are authoritative, they are static. DrugChat can suggest new properties for known compounds, supporting drug repurposing efforts by identifying alternative indications or mechanisms of action. Crucially, DrugChat doesn't rely on simple chemical similarity. Using scaffold-based data splits for key datasets, it demonstrated strong generalization to structurally distinct molecules (Figs.2,4), suggesting it learns deeper structure–function relationships. DrugChat provides token-level confidence scores for its outputs, aiding uncertainty estimation in classification tasks. However, it is not intended for clinical use or public deployment. Instead, it serves as a research tool for experts who can interpret its probabilistic outputs responsibly. Future versions will include disclaimers and improved calibration and factuality checks to promote safe usage.

A known limitation is interpretability—DrugChat's reasoning is not transparent, typical of LLMs. This opacity can hinder trust in high-stakes contexts like drug risk assessment. Future work will focus on enhancing explainability. Future directions include expanding training data, integrating with simulation tools (e.g., docking), applying it in real-world pipelines (e.g., drug-drug interaction prediction), and extending support to complex molecules like biologics. These enhancements could significantly broaden DrugChat's impact in pharmaceutical research.

References

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

Allenspach, S., Hiss, J. A., and Schneider, G. Neural multi-task learning in drug design. *Nature Machine Intelligence*, 6(2):124–137, 2024.

Banerjee, P., Eckert, A. O., Schrey, A. K., and Preissner, R. Protox-ii: a webserver for the prediction of toxicity of chemicals. *Nucleic acids research*, 46(W1):W257–W263, 2018.

Baumann, D. and Baumann, K. Reliable estimation of prediction errors for qsar models under model uncertainty using double cross-validation. *Journal of cheminformatics*, 6:1–19, 2014.

Bemis, G. W. and Murcko, M. A. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928. PMID: 8709122.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250, 2018.

Chen, H., Cheng, F., and Li, J. idrug: Integration of drug repositioning and drug-target prediction via cross-network embedding. *PLoS computational biology*, 16(7): e1008040, 2020.

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J., and Tang, Y. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS computational biology*, 8(5):e1002503, 2012.

Cheng, F., Desai, R. J., Handy, D. E., Wang, R., Schneeweiss, S., Barabási, A.-L., and Loscalzo, J. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nature communications*, 9(1):2691, 2018.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Dahl, G. E., Jaitly, N., and Salakhutdinov, R. Multi-task neural networks for qsar predictions. *arXiv preprint arXiv:1406.1231*, 2014.

Daina, A., Michielin, O., and Zoete, V. Swissadme: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific reports*, 7(1):42717, 2017.

Edwards, C., Zhai, C., and Ji, H. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 595–607, 2021.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Rojas, D., Feng, G., Zhao, H., Lai, H., Yu, H., Wang, H., Sun, J., Zhang, J., Cheng, J., Gui, J., Tang, J., Zhang, J., Li, J., Zhao, L., Wu, L., Zhong, L., Liu, M., Huang, M., Zhang, P., Zheng, Q., Lu, R., Duan, S., Zhang, S., Cao, S., Yang, S., Tam, W. L., Zhao, W., Liu, X., Xia, X., Zhang, X., Gu, X., Lv, X., Liu, X., Liu, X., Yang, X., Song, X., Zhang, X., An, Y., Xu, Y., Niu, Y., Yang, Y., Li, Y., Bai, Y., Dong, Y., Qi, Z., Wang, Z., Yang, Z., Du, Z., Hou, Z., and Wang, Z. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, 7(1):496, 2011.

Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., and Kumar, P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular diversity*, 25:1315–1360, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *International conference on machine learning*, pp. 2712–2721. PMLR, 2019.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJlWWJSFDH.

Jastrzębski, S., Leśniak, D., and Czarnecki, W. M. Learning to smile (s). *arXiv preprint arXiv:1602.06289*, 2016.

Jiménez, J., Skalic, M., Martinez-Rosell, G., and De Fabritiis, G. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling*, 58(2):287–296, 2018.

Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., Jensen, N. H., Kuijer, M. B., Matos, R. C., Tran, T. B., et al. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, 2009.

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

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Kirchmair, J., Göller, A. H., Lang, D., Kunze, J., Testa, B., Wilson, I. D., Glen, R. C., and Schneider, G. Predicting drug metabolism: experiment and/or computation? *Nature reviews Drug discovery*, 14(6):387–404, 2015.

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, 3(11):935–949, 2004.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Lavie, A. and Denkowski, M. J. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23:105–115, 2009.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Liu, Z., Zhang, A., Fei, H., Zhang, E., Wang, X., Kawaguchi, K., and Chua, T.-S. ProtT3: Protein-to-text generation for text-based protein understanding. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5949–5966, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.324. URL https://aclanthology.org/2024.acl-long.324/.

Lo, Y.-C., Rensi, S. E., Torng, W., and Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug discovery today*, 23(8):1538–1546, 2018.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Lu, J., Bender, B., Jin, J. Y., and Guan, Y. Deep learning prediction of patient response time course from early data via neural-pharmacokinetic/pharmacodynamic modelling. *Nature machine intelligence*, 3(8):696–704, 2021.

Luechtefeld, T., Marsh, D., Rowlands, C., and Hartung, T. Machine learning of toxicological big data enables read-across structure activity relationships (rasar) outperforming animal test reproducibility. *Toxicological Sciences*, 165(1):198–212, 2018.

Madden, J. C., Enoch, S. J., Paini, A., and Cronin, M. T. A review of in silico tools as alternatives to animal testing: principles, resources and applications. *Alternatives to Laboratory Animals*, 48(4):146–172, 2020.

Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. Deeptox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3:80, 2016.

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D.-A., and Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical science*, 9(24):5441–5451, 2018.

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., et al. Chembl: towards direct

deposition of bioassay data. *Nucleic acids research*, 47 (D1):D930–D940, 2019.

Öztürk, H., Özgür, A., and Ozkirimli, E. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Pires, D. E., Blundell, T. L., and Ascher, D. B. pkcsm: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *Journal of medicinal chemistry*, 58(9):4066–4072, 2015.

Qian, Y., Li, Z., Tu, Z., Coley, C., and Barzilay, R. Predictive chemistry augmented with text retrieval. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12731–12745, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.784. URL https://aclanthology.org/2023.emnlp-main.784/.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Seidl, P., Vall, A., Hochreiter, S., and Klambauer, G. Enhancing activity prediction models in drug discovery with the ability to understand human language. In *International Conference on Machine Learning*, pp. 30458–30490. PMLR, 2023.

Simões, R. S., Maltarollo, V. G., Oliveira, P. R., and Honorio, K. M. Transfer and multi-task learning in qsar modeling: advances and challenges. *Frontiers in pharmacology*, 9: 74, 2018.

Stanley, M., Bronskill, J. F., Maziarz, K., Misztela, H., Lanini, J., Segler, M., Schneider, N., and Brockschmidt, M. Fs-mol: A few-shot learning dataset of molecules. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., and Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, 2018.

Sterling, T. and Irwin, J. J. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D. M., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A. S., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I. M., Korenev, A. V., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023b. URL https://api.semanticscholar.org/CorpusID:259950998.

Tropsha, A. Best practices for qsar model development, validation, and exploitation. *Molecular informatics*, 29 (6-7):476–488, 2010.

Tropsha, A., Isayev, O., Varnek, A., Schneider, G., and Cherkasov, A. Integrating qsar modelling and deep learning in drug discovery: the emergence of deep qsar. *Nature Reviews Drug Discovery*, 23(2):141–155, 2024.

Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Wegner, J. K., Ceulemans, H., and Hochreiter, S. Deep learning as an opportunity in virtual screening. In *Proceedings of the deep learning workshop at NIPS*, volume 27, pp. 1–9. MIT Press Cambridge, MA, United States, 2014.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NeurIPS*, 2017.

Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., and Langer, T. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1–12, 2020.

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl_1): D668–D672, 2006.

Wong, F., Zheng, E. J., Valeri, J. A., Donghia, N. M., Anahtar, M. N., Omori, S., Li, A., Cubillos-Ruiz, A., Krishnan, A., Jin, W., et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, 626(7997):177–185, 2024.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

Xiong, G., Wu, Z., Yi, J., Fu, L., Yang, Z., Hsieh, C., Yin, M., Zeng, X., Wu, C., Lu, A., et al. Admetlab 2.0: an integrated online platform for accurate and comprehensive predictions of admet properties. *Nucleic acids research*, 49(W1):W5–W14, 2021.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Xu, M., Yuan, X., Miret, S., and Tang, J. Protst: Multimodality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, pp. 38749–38767. PMLR, 2023.

Yang, J. H., Wright, S. N., Hamblin, M., McCloskey, D., Alcantar, M. A., Schrübbers, L., Lopatkin, A. J., Satish, S., Nili, A., Palsson, B. O., et al. A white-box machine learning approach for revealing antibiotic mechanisms of action. *Cell*, 177(6):1649–1661, 2019a.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019b.

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. deepdr: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, 35 (24):5191–5198, 2019.

Zeng, X., Xiang, H., Yu, L., Wang, J., Li, K., Nussinov, R., and Cheng, F. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016, 2022a.

Zeng, Z., Yao, Y., Liu, Z., and Sun, M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862, 2022b.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

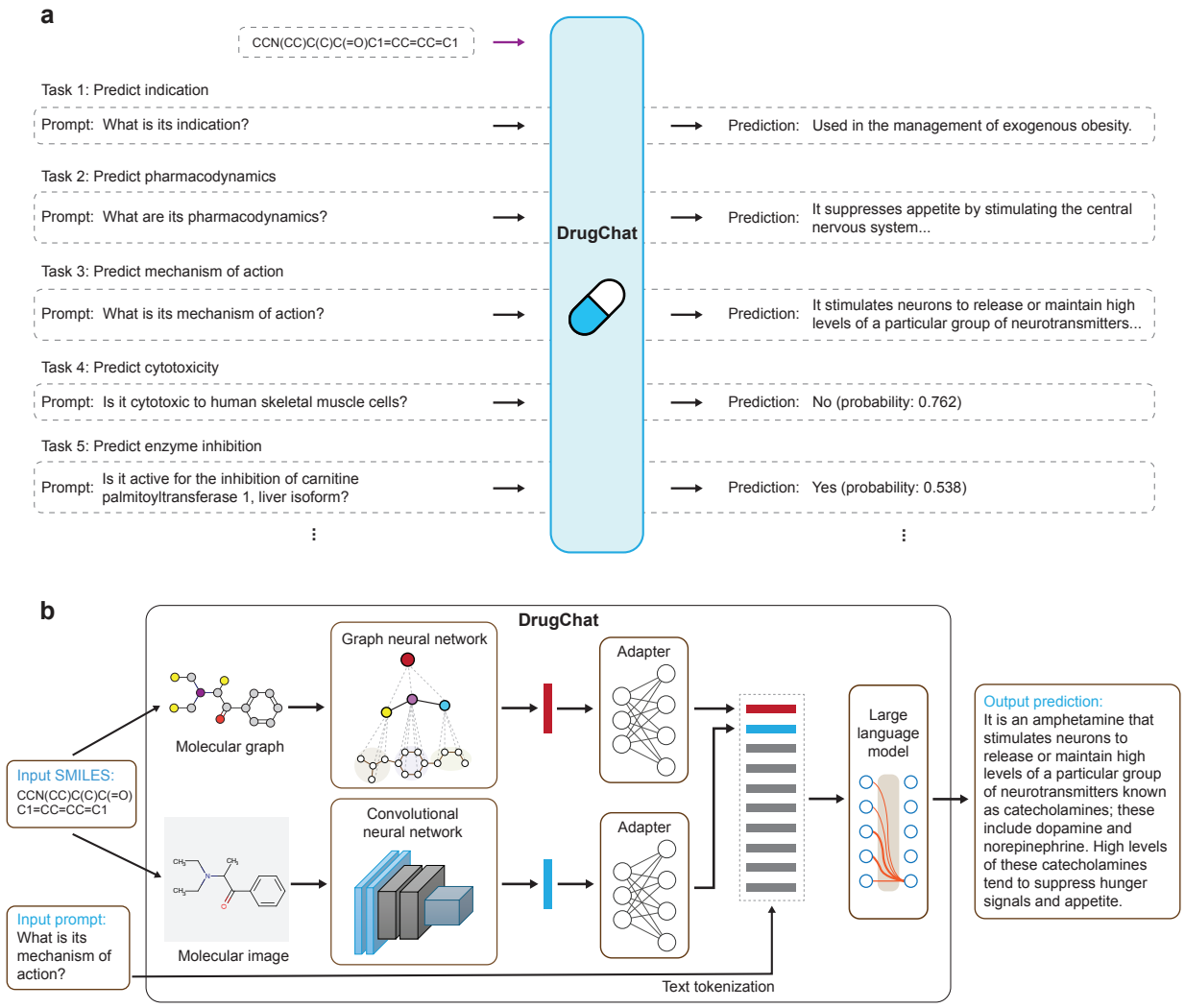

Figure 1. **DrugChat is a multi-modal LLM capable of predicting drug attributes in either free-form texts or as discrete categories.** **a**, DrugChat facilitates versatile predictions of drug attributes, enabling users to submit queries through flexible natural language (known as prompts). By using task-specific prompts, DrugChat can perform a wide range of prediction tasks within a unified framework, without requiring changes to model parameters. For classification tasks, DrugChat simultaneously outputs a probability associated with each prediction. **b**, Model architecture of DrugChat. It takes the SMILES representation of a molecule along with a prompt as inputs and generates a prediction in natural language. It comprises two molecular encoders—a graph neural network and a convolutional neural network—that learn representation vectors for the molecular graph and image derived from the SMILES input, two adapters that transform these representations into a format compatible with LLMs, and an LLM that generates a prediction based on the molecular representations and the prompt.

A. Related Work

In recent years, deep learning (DL) has emerged as a powerful tool for QSAR modeling and drug discovery (Chen et al., 2018; Vamathevan et al., 2019; Wieder et al., 2020; Tropsha et al., 2024), thanks to its capacity to analyze large-scale datasets and uncover complex patterns. DL methods have improved efficiency and accuracy across multiple stages of drug discovery (Lo et al., 2018; Gupta et al., 2021), from predicting protein-ligand binding affinity (Öztürk et al., 2018; Jiménez et al., 2018; Stepniewska-Dziubinska et al., 2018) and toxicity (Mayr et al., 2016; Banerjee et al., 2018), to enabling drug repositioning (Cheng et al., 2012; Keiser et al., 2009; Gottlieb et al., 2011; Cheng et al., 2018; Chen et al., 2020). For example, graph neural networks (GNNs) such as MPNN (Gilmer et al., 2017) and Chemprop (Yang et al., 2019b) have been

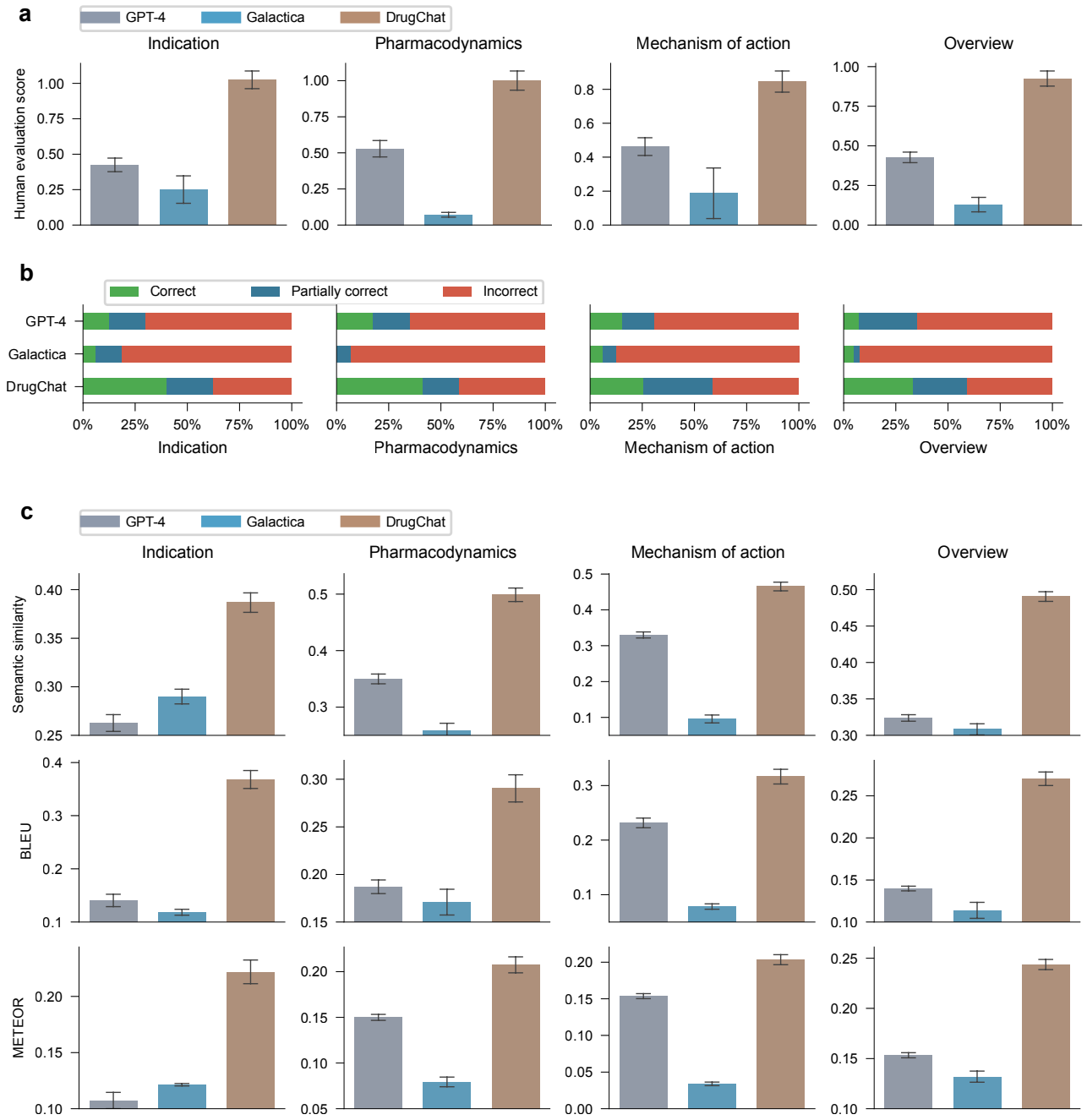

Figure 2. **DrugChat significantly outperformed GPT-4 and Galactica in predicting drug indications, pharmacodynamics, mechanisms of action, and overviews expressed in free-form texts.** **a**, Human expert evaluations demonstrated DrugChat's superior performance, with results reported as average scores and standard deviations from five-fold cross-validation. Predictions were scored on a 0–2 scale: 2 for fully correct, 1 for partially correct, and 0 for incorrect. **b**, DrugChat produced a higher proportion of correct predictions compared to GPT-4 and Galactica. **c**, In automated evaluations—using semantic similarity, BLEU, and METEOR metrics—DrugChat also substantially outperformed both baselines, with results reported as average scores and standard deviations from five-fold cross-validation.

widely adopted for molecular property prediction, leveraging molecular graph structures as input. Multi-task deep learning approaches (Dahl et al., 2014; Unterthiner et al., 2014; Ramsundar et al., 2015; Allenspach et al., 2024; Yang et al., 2019b; Mayr et al., 2018; Jastrzębski et al., 2016; Qian et al., 2023) have demonstrated strong efficacy in simultaneously predicting

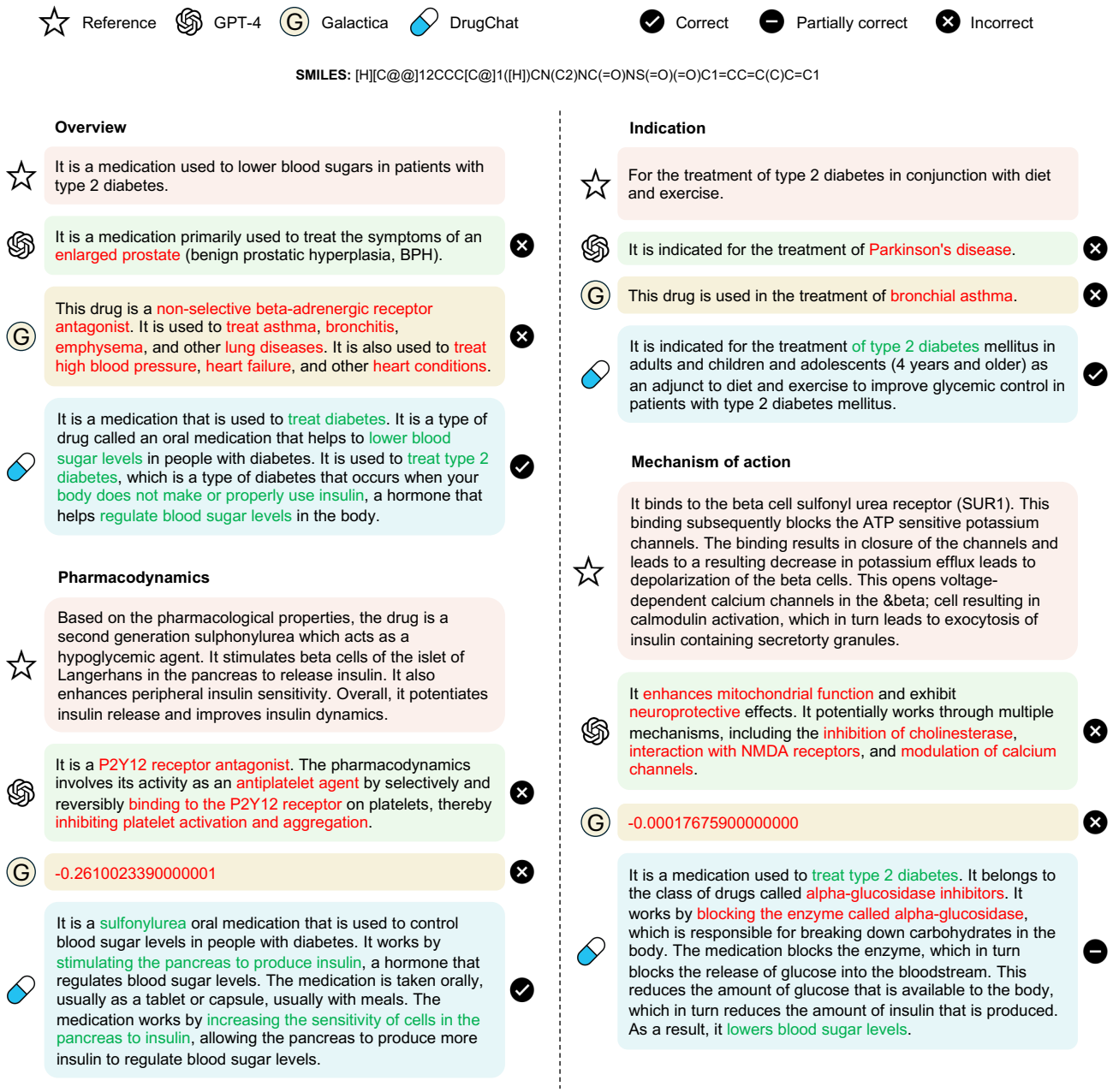

Figure 3. **Examples of free-form predictions generated by DrugChat, Galactica, and GPT-4, along with their corresponding human evaluation scores.** Text highlighted in green indicates correct predictions, while text in red denotes incorrect predictions.

ADME (Absorption, Distribution, Metabolism, Excretion) properties (Pires et al., 2015; Daina et al., 2017; Xiong et al., 2021). More recently, multi-modal models (Zeng et al., 2022b; Seidl et al., 2023; Edwards et al., 2021; Liu et al., 2024; Xu et al., 2023) such as KV-PLM (Zeng et al., 2022b), CLAMP (Seidl et al., 2023), and Text2Mol (Edwards et al., 2021) have been proposed to jointly model molecular structures and textual descriptions of prediction tasks, combining molecular encoders with language encoders. In parallel, molecule-focused scientific large language models (LLMs) (Thirunavukarasu et al., 2023) such as Galactica (Taylor et al., 2022) have been developed to leverage large-scale molecular and textual data for scientific question answering and property prediction.

Due to the structured outputs in most existing deep learning models for molecule property prediction, they are unable to generate rich, natural language descriptions that mirror the way these properties are discussed in biomedical literature and expert annotations. While KV-PLM (Zeng et al., 2022b), CLAMP (Seidl et al., 2023), and Text2Mol (Edwards et al., 2021) combine molecular features with textual inputs, they are primarily designed for classification or retrieval tasks, producing scalar similarity or likelihood scores for molecule-text pairs. These models are not capable of generating free-form textual responses. While Galactica (Taylor et al., 2022) supports free-form text generation, it relies on a single language model to process both SMILES strings and textual queries without incorporating dedicated molecular structure encoders. This design limits its ability to accurately capture chemical structures and spatial relationships, thereby impairing its capacity to predict pharmacological properties and mechanisms of action with high fidelity.

A further limitation of current molecular prediction models is that they generate one-shot predictions without supporting interactive, multi-turn exploration. Once a model provides an output for a given input, there is no mechanism for users to iteratively refine their questions, ask follow-up inquiries, or progressively explore different aspects of a molecule within a conversational workflow. This static approach to prediction limits the depth and flexibility of molecular analysis, hindering the discovery of nuanced or context-specific insights that could be revealed through dynamic, iterative interaction.

B. Methods

B.1. Data collection and processing

We curated training data for DrugChat from publicly available compound databases, including PubChem, ChEMBL, and DrugBank. The PubChem database[1] contains information on 66,469,244 chemical compounds. We downloaded the bioassay dataset, last updated on January 9, 2025, which includes 4,019,927 unique compounds and 248,667,695 activity records across 636,397 bioassays. The ChEMBL database[2] provides information on 2,354,965 chemical compounds. We downloaded the SQLite version of the dataset, last updated on February 28, 2023. From this, we selected 3,462 compounds that contain information about administration routes and prodrug status. The DrugBank database[3] (version 5.1.10, released on January 4, 2023) contains 16,428 drug entries. We selected 11,583 entries with available SMILES strings, focusing exclusively on small molecules and excluding biotech-classified compounds. After further filtering to retain only entries with annotations for drug indications, pharmacodynamics, or mechanisms of action, we curated 5,846 drug molecules for use in DrugChat. The distribution of drug categories is shown in Fig. 9. For each selected compound, we collected its SMILES string along with a variety of attributes, such as free-form descriptions (e.g., indications, pharmacodynamics, mechanisms of action) and structured bioassay activities. Additionally, the cytotoxicity, MoleculeNet, and FS-Mol datasets used in our work were curated by (Wong et al., 2024), (Wu et al., 2018), and (Stanley et al., 2021), respectively.

Using these drugs and their annotated attributes, we curated the training data for DrugChat. For each attribute a of a compound molecule m, we created a triplet consisting of the molecule's SMILES representation, a textual prompt querying the value of a, and the corresponding ground truth for a. Each attribute type had its own tailored prompt. For instance, for the attribute 'drug indication', the corresponding prompt is 'What is its indication?'. The answer is a textual description of the drug compound's indication, provided by human experts from the DrugBank database. As another example, for the attribute 'prodrug status', the corresponding prompt is 'Is the molecule a prodrug?' The ground truth answer is 'Yes' if the compound is a prodrug, and 'No' otherwise. The distribution of ground truth answers in the cytotoxicity and ChEMBL datasets is shown in Fig. 10. We created similar molecule-prompt-answer triplets from the bioassay activities in PubChem and FS-Mol. For each bioassay, we incorporated the textual description of the bioassay into the prompt using the following template: 'is the compound active in this bioassay: `<bioassay description>`'. The ground truth answer is 'Yes' if the compound is labeled as active in the bioassay, and 'No' otherwise. An example prompt for an assay in the FS-Mol dataset is: 'is the compound active in this bioassay: qHTS assay for identifying a potential treatment of Ataxia-Telangiectasia'. For a complete description of all prompts used across different datasets, please refer to the Prompts subsection below.

B.2. Model architecture

DrugChat is a multi-modal model that integrates information from three distinct modalities: graphs, images, and text. It consists of a Graph Isomorphism Network (GIN) (Xu et al., 2018), a ResNet (He et al., 2016), and a large language model

[1] https://pubchem.ncbi.nlm.nih.gov/
[2] https://www.ebi.ac.uk/chembl/
[3] https://go.drugbank.com/releases/latest

(LLM) (Chiang et al., 2023). For a given molecule, its SMILES string is converted into both a molecular graph and a molecular image using the RDKit software[4].

In the molecular graph, nodes correspond to the molecule's atoms, while edges represent the chemical bonds between them. Each atom is defined by its atom type and chirality, with 120 atom types in total, including a special 'Unknown' category for unidentified atoms. Atom chirality is categorized into four types: tetrahedral clockwise, tetrahedral counter-clockwise, unrecognized, and other. These atom type and chirality attributes serve as the initial features for each node. Chemical bonds are characterized by their type and direction, with bond types classified as single, double, triple, or aromatic, and bond directions as none, end upright, or end downright. These bond attributes are used as the initial features for each edge. All node and edge features are categorical, with each category encoded as a vector with learnable parameters. The molecular graph is input into the GIN to learn a representation vector for the entire graph. The GIN leverages the graph's connectivity, along with the initial node and edge features, to learn multi-layer representations for each node. Using a neighborhood aggregation approach, the GIN iteratively updates each node's vector by aggregating information from its neighbors and connecting edges (Kipf & Welling, 2017). After K layers of representation learning, information is propagated through K-hop paths across the graph. An average pooling operation is then applied to compute the mean of the final node representations, yielding a single vector that summarizes the entire graph. The GIN in DrugChat consists of five layers and approximately 1.9 million parameters. Both node and edge embeddings have a dimensionality of 300. The GIN was pretrained using a self-supervised context prediction approach (Hu et al., 2020), in which the model learns to predict a molecule's subgraphs based on its surrounding subgraphs. Pretraining was performed on 2 million unlabeled molecules from the ZINC15 database (Sterling & Irwin, 2015).

For the molecular image, we employed a ResNet to extract a representation vector. The ResNet processes the input image through multiple layers of 2D convolution, where each layer applies convolutional filters to detect specific patterns in the image or in the feature maps from the previous layer. We used a ResNet-18 (He et al., 2016) model, which consists of 18 convolutional layers and approximately 11 million parameters. A global average pooling layer converts the output of the final convolutional layer into a molecular image representation vector with a dimensionality of 512. Following the approach in ImageMol (Zeng et al., 2022a), the ResNet was pretrained on 10 million unlabeled images of drug-like, bioactive molecules from the PubChem database using self-supervised learning (SSL) techniques, such as molecular image reconstruction and contrastive learning. These SSL methods enable the ResNet to map structurally similar molecules to nearby points in the embedding space, allowing it to learn molecular features at scale without requiring human-annotated labels.

After extracting representation vectors from the molecule using the GIN and ResNet, we apply two separate linear layers—each consisting of a matrix multiplication followed by a bias addition—referred to as adapters, to transform these molecular representations into a format compatible with the LLM. LLMs typically use Transformer decoders (Vaswani et al., 2017) to model natural language as sequences of tokens, with each token represented as a vector (Brown et al., 2020). In DrugChat, the transformed molecular representations are treated as tokens and appended to the sequence of language tokens derived from the input prompt. This combined sequence is then passed into the LLM, which uses multi-head self-attention (Vaswani et al., 2017) to generate new tokens. These generated tokens constitute the final prediction. DrugChat employs Vicuna-13B (Chiang et al., 2023) as its LLM, which has 13 billion parameters. Vicuna-13B was fine-tuned from LLaMA-13B (Touvron et al., 2023a) using a dataset of 70K user-shared conversations from ShareGPT.com (containing interactions between humans and ChatGPT). It retains the architecture of LLaMA-13B, including 40 Transformer layers, 40 attention heads, and an embedding dimension of 5120. The base model was pretrained on a multi-terabyte text corpus—including Wikipedia and various web sources—to predict the next token given preceding context. In our experiments, the LLM weights were frozen and not updated during training. The adapter that converts molecular graph representations into LLM-compatible tokens is a linear layer with input dimension 300 and output dimension 5120, totaling 1.5 million parameters. The adapter for molecular image representations is similarly a linear layer, with an input dimension of 512 and output dimension of 5120, totaling 2.6 million parameters.

For a target answer T that has L text tokens, DrugChat computes the probability of generating T as follows:

$$p(T \mid M, P) = p_W\left(T_1 \mid M, P\right) \prod_{i=2}^{L} p_W\left(T_i \mid M, P, T_{<i}\right), \tag{1}$$

where M represents the input molecule and P is the input prompt. We denote the i-th token as T_i and all preceding tokens as $T_{<i}$. Model parameters are denoted by W. The generated token sequence is compared to the ground truth tokens to compute

[4] https://www.rdkit.org

the negative log-likelihood (NLL). The parameters W are optimized by minimizing the sum of NLL over all training data.

B.3. Model training

We first pretrained DrugChat using 248 million molecule-prompt-answer triplets derived from the bioassay activities in PubChem. To avoid data leakage, we removed all compounds present in the DrugBank, ChEMBL, cytotoxicity, MoleculeNet, and FS-Mol datasets from the pretraining data. Pretraining was conducted for one epoch with a batch size of 48. We used the AdamW optimizer (Loshchilov & Hutter, 2019) with a constant learning rate of 10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.001. Due to the strong class imbalance in the PubChem bioassay labels (with only approximately 3% of samples labeled as active), we performed class balancing by applying a 1:30 weighting scheme for the inactive and active samples when computing the loss function. After pretraining, we finetuned DrugChat using the training splits of the MoleculeNet and FS-Mol datasets. Finetuning was conducted for 5 epochs with a batch size of 64, using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.01. A linear warmup was applied during the first epoch, gradually increasing the learning rate from 10^{-5} to 5×10^{-4}, followed by cosine decay to 2×10^{-5} over the remaining epochs. For evaluation on the DrugBank, cytotoxicity, and ChEMBL datasets, we performed nested 5-fold cross-validation using scaffold-based splits. For each held-out fold, hyperparameters (learning rate and number of epochs) were tuned on the remaining four folds via 2-fold cross-validation. The model was then retrained on all four folds and evaluated on the held-out fold. All experiments were conducted using NVIDIA A100 GPUs with 80 GB of memory.

B.4. Baselines

For the DrugBank dataset, we experimented with different prompt templates for GPT-4 and Galactica and selected those that produced the most coherent and relevant responses. GPT-4 was prompted using the following template:

```
Given the SMILES of a molecule:
<SMILES>, <question>
```

where `<SMILES>` is replaced with the SMILES string of the molecule and `<question>` is replaced with the specific inquiry, such as 'what is its indication?', 'what are its pharmacodynamics?', or 'what is its mechanism of action?'. Following (Taylor et al., 2022), Galactica was prompted with the following template:

```
Question: <question> [START_SMILES]
<SMILES>[END_SMILES] Answer:
```

On the cytotoxicity dataset, we used the following prompt template for ChatGLM, FastChat-T5, and Llama:

```
Given the SMILES of a molecule:
<SMILES>, is the molecule  cytotoxic
to <cell>?
```

where `<SMILES>` is replaced with the SMILES representation of the molecule and `<cell>` is replaced with the target cell line (HepG2, HSkMC, or IMR-90). Galactica's prompt template was:

```
Question: Is the molecule
[START_SMILES]<SMILES>[END_SMILES]
cytotoxic to <cell> Answer:
```

Similar to DrugChat, we extracted the probabilities of generating the 'Yes' and 'No' tokens from the outputs of ChatGLM, FastChat-T5, Llama, and Galactica, and used the likelihood ratio $P(\text{Yes})/P(\text{No})$ as the decision value for calculating AUROC and ΔAUPRC. For the other baselines, including MPNN, Chemprop, Text2Mol, KV-PLM, and CLAMP, we utilized their official codebases for training and evaluation, where the decision values needed to compute AUROC and ΔAUPRC were directly provided by the models' outputs. On the ChEMBL dataset, we reused the prompt templates described above but replaced the questions with: 'Is the drug taken orally?', 'Is the drug administered parenterally?', 'Is the drug applied topically?', or 'Is the molecule a prodrug?'. For the other baselines, including MPNN, Chemprop, Text2Mol, KV-PLM, and CLAMP, we utilized their official codebases to perform training and evaluation. On the MoleculeNet and FS-Mol datasets, we adopted the following prompt template for Galactica:

```
Question: Is the molecule
[START_SMILES]<SMILES>[END_SMILES]
active in <task>? Answer:
```

where `<task>` is replaced with the description of the target bioassay in FS-Mol or the task in MoleculeNet. For KV-PLM and CLAMP, we cited their evaluation results on MoleculeNet and FS-Mol directly from the CLAMP paper (Seidl et al., 2023). For Text2Mol, we evaluated its performance using its official codebase.

B.5. Prompts

The prompts used by DrugChat to predict cytotoxicity to HepG2, HSkMC, and IMR-90 were 'Is the molecule cytotoxic to human liver carcinoma cells (HepG2)?', 'Is the molecule cytotoxic to primary skeletal muscle cells (HSkMC)?', and 'Is the molecule cytotoxic to human lung fibroblast cells (IMR-90)?'. The prompts used by DrugChat on the ChEMBL dataset were: 'Is the drug taken orally?', 'Is the drug administered parenterally?', 'Is the drug applied topically?', and 'Is the molecule a prodrug?'. The prompts used by DrugChat on the DrugBank dataset were 'What is its indication?', 'What are its pharmacodynamics?', 'What is its mechanism of action?', and 'What is its overview?'. For the MoleculeNet dataset, DrugChat used the prompt template: 'Is the compound active in `<task description>`?' For the FS-Mol dataset, DrugChat used the prompt template: 'Is the compound active in this bioassay: `<bioassay description>`?'

B.6. Model evaluation

We evaluated the free-form predictions of drug indications, pharmacodynamics, and mechanisms of action using both human assessment and automated metrics.

Human evaluation. In the human evaluation, experts specializing in drug molecules assessed the model's predictions using a 3-point Likert scale with an additional 'Unknown' option. The scales were defined as follows: 1) [Correct]—The prediction is mostly consistent with the ground truth or a subset of the ground truth, possibly extending it with additional plausible details; 2) [Partially Correct]—The prediction includes some correct descriptions but also introduces conflicting elements when compared to the ground truth or domain knowledge; and 3) [Incorrect]—The prediction is incorrect, irrelevant, or incomplete. Evaluators were asked to choose one of the three options and they did not know which model generated the predictions.

Automatic evaluation metrics. We conducted automatic evaluations using three metrics: semantic similarity, BLEU (Papineni et al., 2002), and METEOR (Lavie & Denkowski, 2009) scores. Semantic similarity was calculated as the cosine similarity between the sentence embeddings of the ground-truth and model-predicted texts, with embeddings generated using a pretrained sentence Transformer model All-MiniLM-L6-v2 (Reimers & Gurevych, 2019). Let SE represent the sentence embedding model. The embedding of the ground-truth text t_g is denoted as $e_g = \mathrm{SE}(t_g)$, while the embedding of the predicted text t_p is $e_p = \mathrm{SE}(t_p)$. The cosine similarity between these embeddings is defined as

$$\cos(e_g, e_p) = \frac{\langle e_g, e_p \rangle}{\|e_g\|_2 \|e_p\|_2},$$

where $\langle e_g, e_p \rangle$ represents the dot product between the two vectors, and $\|e_g\|_2$ denotes the L2 norm of the vector e_g. With this sentence Transformer, the embeddings of semantically similar sentences are positioned closer in the embedding space, resulting in higher cosine similarity values. For BLEU scores (Papineni et al., 2002), we used the BLEU-1 score without applying the brevity penalty (Papineni et al., 2002). The BLEU-1 score is a special case of the general BLEU-n metric, which measures the modified precision of n-grams (sequences of n consecutive words). Let $\hat{y}$ represent the predicted sentence, and y the ground-truth sentence. Define $G_n(\hat{y})$ as the set of all n-grams in the predicted sentence. Let $C(s, y)$ be an indicator function, which equals 1 if the n-gram s appears in y; otherwise, $C(s, y) = 0$. The BLEU-n score is then calculated as follows:

$$\mathrm{BLEU}_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}.$$

The METEOR score (Lavie & Denkowski, 2009) is viewed as an enhancement over the BLEU score. It is calculated using the harmonic mean of unigram precision and recall, with greater emphasis on recall. Additionally, it accounts for stemming and synonym matching, in addition to exact word matching.

On the ChEMBL, cytotoxicity, MoleculeNet, and FS-Mol datasets, we evaluated performance using AUROC and ΔAUPRC. These metrics were computed based on the probabilities of generating the 'Yes' and 'No' tokens. During token generation, DrugChat's LLM produces a probability distribution over the entire vocabulary at each step and selects the token with the highest probability. We extracted the probabilities associated with the 'Yes' and 'No' tokens directly from this distribution without requiring additional computation. The ratio $P(\mathrm{Yes})/P(\mathrm{No})$ was then used as the decision score to compute AUROC and ΔAUPRC.

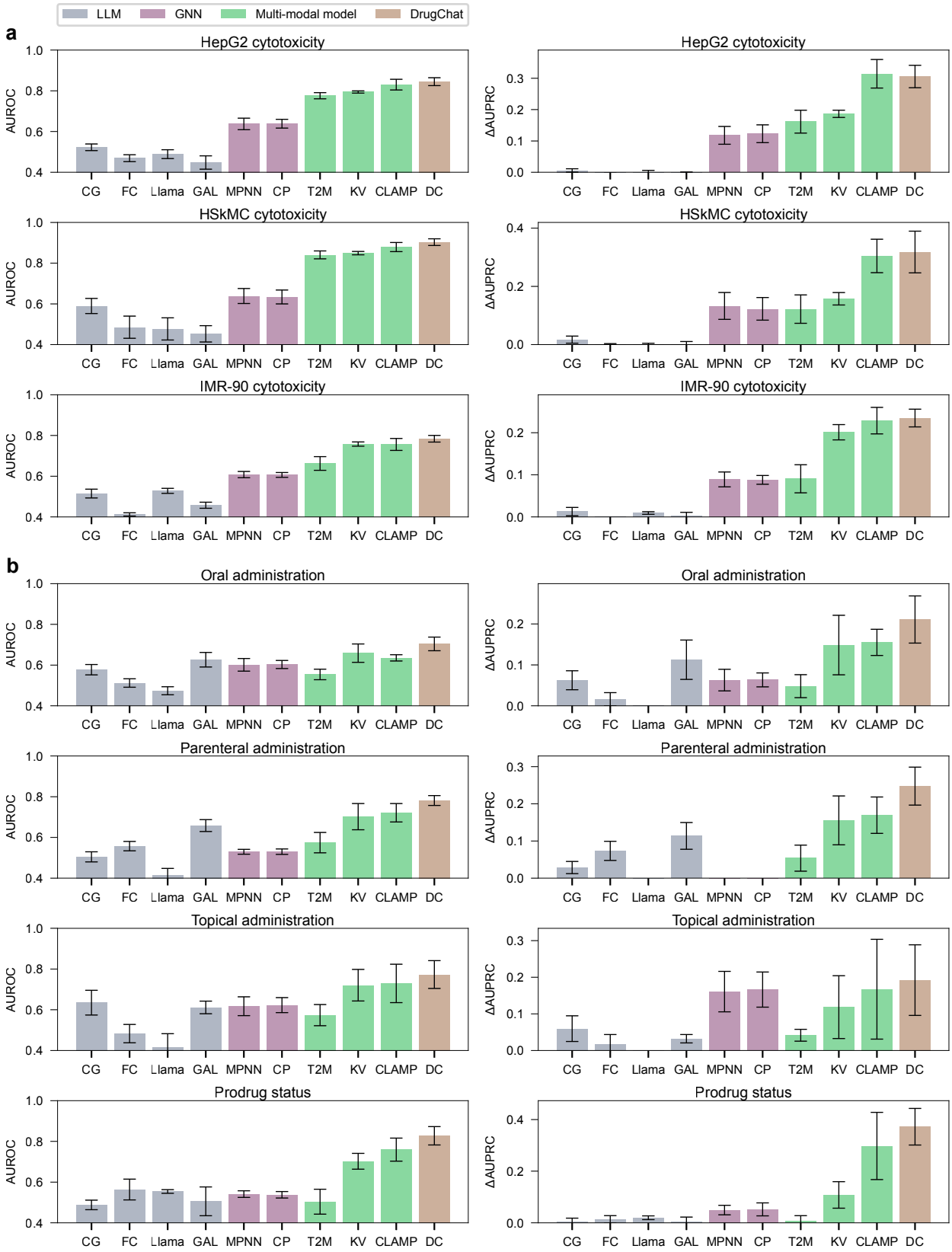

Figure 4. **DrugChat demonstrated strong performance across a variety of structured QSAR tasks.** **a**, For cytotoxicity prediction (Wong et al., 2024), DrugChat (DC) significantly outperformed (1) graph neural network (GNN)-based models including MPNN and Chemprop (CP), (2) multi-modal models including Text2Mol (T2M) and KV-PLM (KV), (3) the molecule-focused scientific LLM, Galactica, and (4) general-purpose LLMs including ChatGLM (CG), FastChat-T5 (FC), and LLaMA. DrugChat also slightly outperformed CLAMP in most cases. **b**, For administration route and prodrug status prediction on the ChEMBL dataset, DrugChat outperformed all baselines across all four tasks.

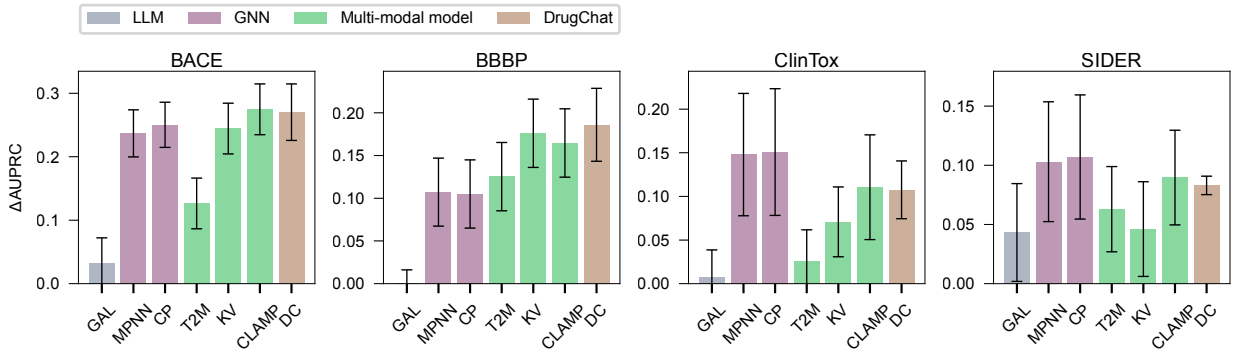

Figure 5. **On the MoleculeNet datasets, DrugChat (DC) consistently outperforms Galactica (GAL), Text2Mol (T2M), and KV-PLM (KV), while achieving performance comparable to MPNN, Chemprop (CP), and CLAMP.** Some baseline results were obtained from (Seidl et al., 2023), where only ΔAUPRC was reported; therefore, we report ΔAUPRC for consistency in these comparisons.

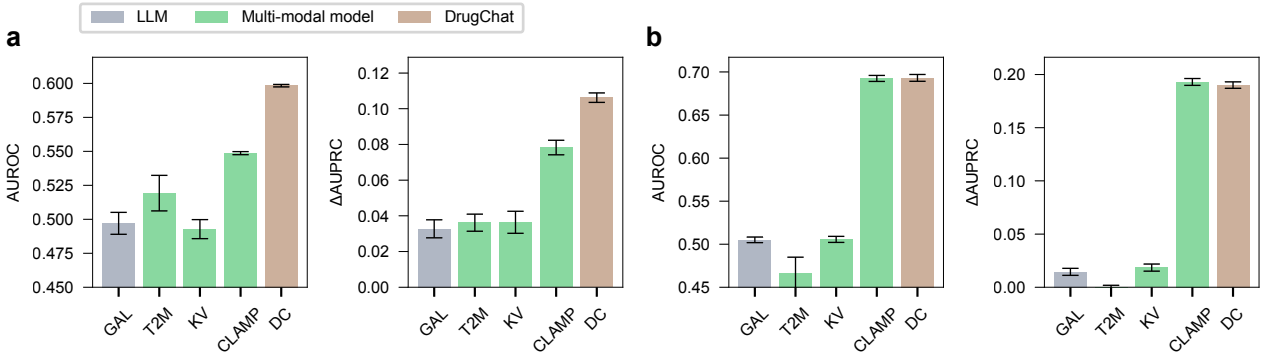

Figure 6. **DrugChat demonstrated strong zero-shot generalization on the FS-Mol dataset. a**, In the full zero-shot setting—where both the tasks and compounds in the test set were unseen during training—DrugChat (DC) significantly outperformed all baselines, including CLAMP, Text2Mol (T2M), KV-PLM (KV), and Galactica (GAL). **b**, In the partial zero-shot setting—where all test tasks were unseen and a subset of test compounds were unseen during training—DrugChat significantly outperformed Text2Mol, KV-PLM, and Galactica, and performed comparably to CLAMP.

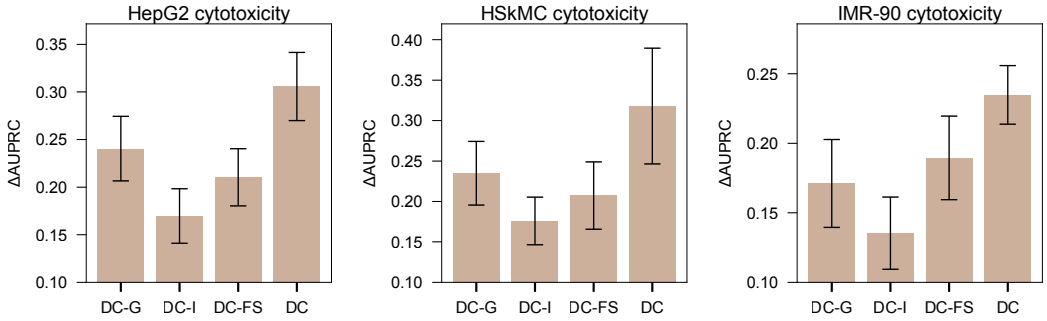

Figure 7. **Ablation studies of DrugChat.** Integrating both molecular graph and image modalities, DrugChat (DC) outperformed its single-modality variants—DC-G (graph only) and DC-I (image only)—on the cytotoxicity dataset. Additionally, the version of DrugChat pretrained on PubChem bioassay data outperformed the variant trained from scratch (DC-FS).

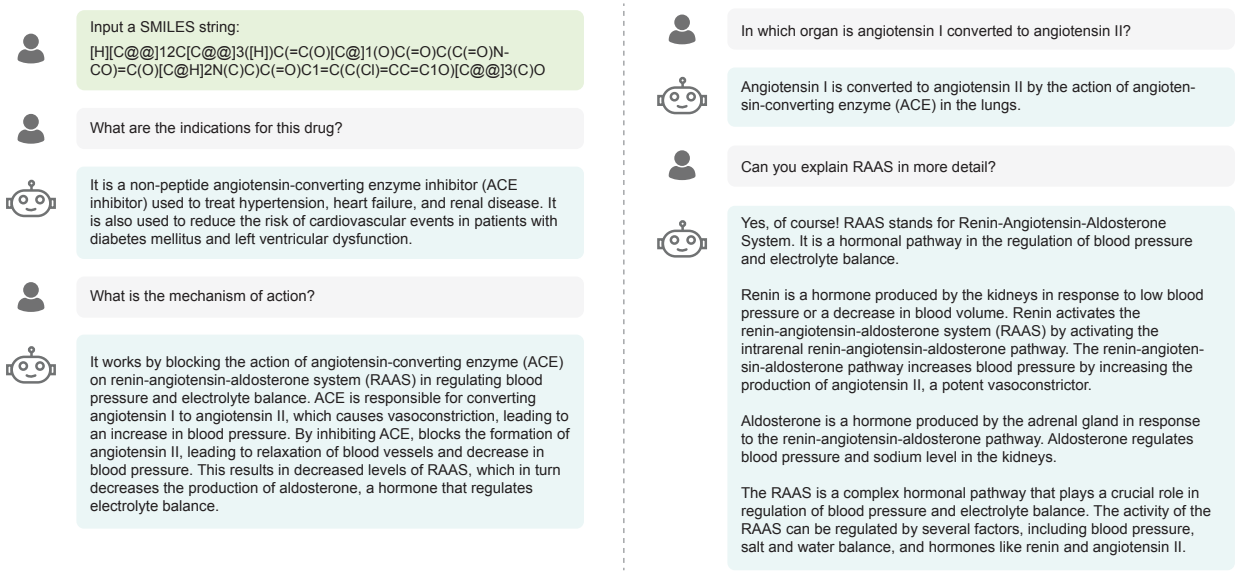

Figure 8. **An exemplar multi-turn dialogue between DrugChat and a user regarding the same molecule.**

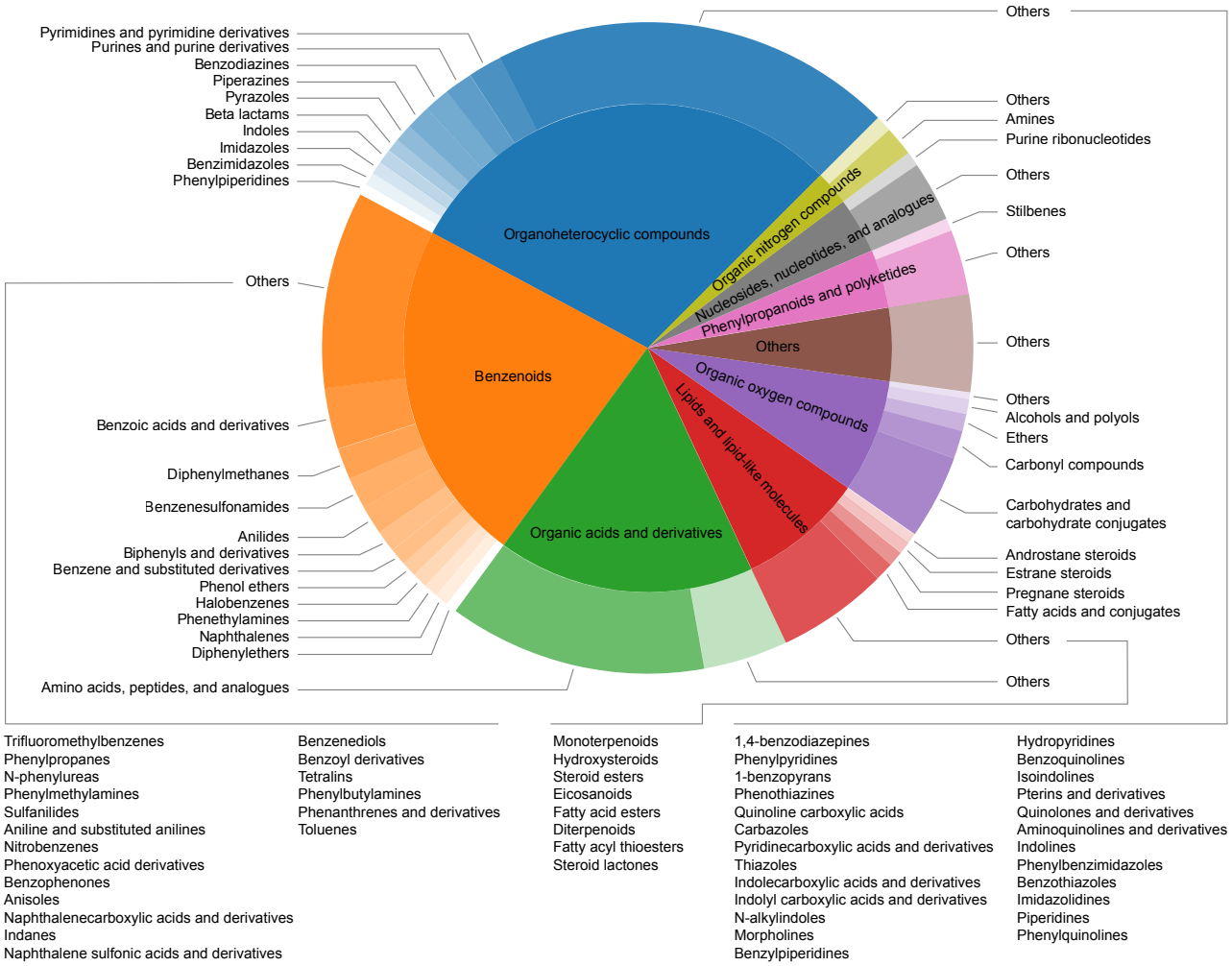

Figure 9. **Our curated dataset features a diverse distribution of compound categories.** The inner disk represents the compound superclasses, while the outer ring shows the corresponding subclasses.

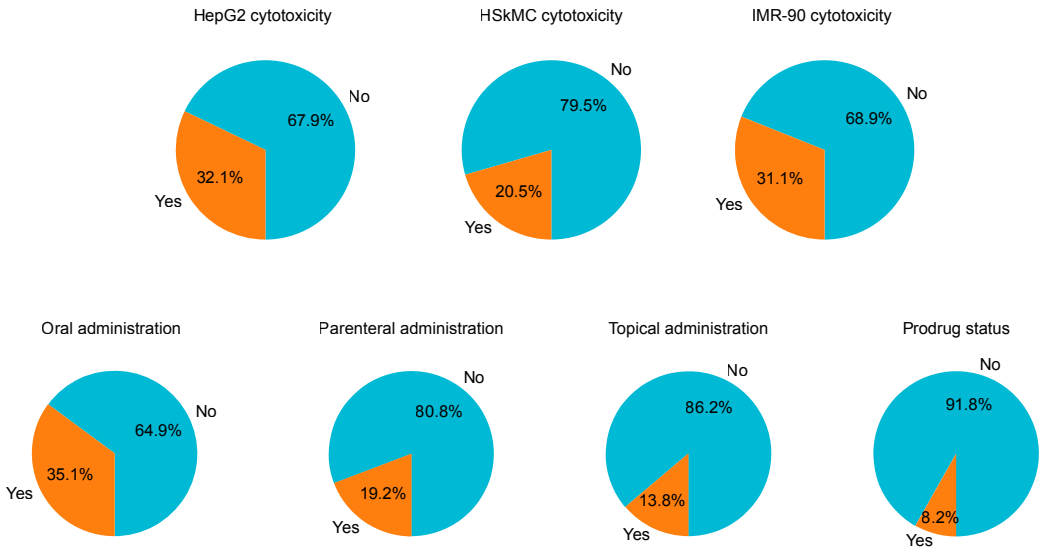

Figure 10. The distribution of ground-truth answers in the test set across the tasks of predicting cytotoxicity, administration routes, and prodrug status.