**Advancing PFAS Biodegradation Prediction: Integrating Large Language Models with Environmental Cheminformatics**

Wufan Ding†, Aofan Liu†Zhengkai Chen†, Haodong Ji†,\*

†Eco-environment and Resource Efficiency Research Laboratory, School of Environment and Energy, Peking University Shenzhen Graduate School, Shenzhen 518055, PR China

、结论：

1. 低数据或零数据的情况下，大模型好，RMSE，MAE，R2，意味着什么，（大部分情况环境数据量都是很小，所以用这个很好，用机器学习会过拟合）
2. 根据PFAS分析一些结论，对比不同链长，有没有规律，或者哪个biodegradable最长或最短

3）我们训练的模型，非常适用于PFAS类污染物，（欧式距离）

**\*Corresponding authors**

H. Ji, Tel: +86-0755-26035343, E-mail: jihaodong@pku.edu.cn

**ABSTRACT**

The environmental challenges and biodegradation issues of per- and polyfluoroalkyl substances (PFAS) are critical. PFAS are of global environmental concern due to their persistence, bioaccumulation, and potential ecotoxicity. Their stability renders them resistant to natural degradation, impacting water bodies, soil, and organisms over the long term. Biodegradation testing is a vital component of PFAS regulation, encompassing classification, labeling, environmental risk assessment, and hazard evaluation. However, current laboratory methods for assessing biodegradability and persistence are costly, time-consuming, and yield inconsistent results. To date, approximately 145,299 chemicals have been pre-registered under the EU REACH regulations, but only 4,214 have been reliably screened for biodegradability. Testing the remainder is a cumbersome, expensive, and potentially unfeasible task. Additionally, the factors influencing biodegradability are poorly understood and under-researched. While traditional machine learning and deep learning models have advanced in predicting PFAS biodegradation half-lives, they are often constrained by insufficient data and feature extraction limitations. To address these challenges, we utilize the Llama 2-13B-chat Large Language Model (LLM) to predict the biodegradation half-lives of PFAS. By employing a novel fine-tuning technique—Low-Rank Adaptation Matrix Decomposition (LoRA)—we achieve predictive capabilities under limited graphics processing unit resources. Furthermore, integrating a Chain of Thought (CoT) process enhances the interpretability and reliability of our predictions, demonstrating higher predictive accuracy in zero-shot or low-shot learning scenarios. Additionally, we innovatively use an Euclidean distance-based algorithm to calculate the similarity between compounds within the domain of applicability, showing the potential of large models as tools for environmental monitoring and risk assessment.

**KEYWORDS:** Per- and Polyfluoroalkyl Substances (PFAS); biodegradation half-lives; Large Language Models (LLMs); Chain of Thought (CoT); processing Euclidean-distance-based algorithm

**Environment Implication**

This research utilizes the advanced Llama 2-13B-chat Large Language Model (LLM) to predict the biodegradation half-lives of per- and polyfluoroalkyl substances (PFAS), Addressed the issue of data scarcity in PFAS experimentation.. Traditional biodegradability tests are costly and unreliable, highlighting the need for efficient alternatives like our novel Euclidean-distance-based algorithm, which offers improved chemical similarity assessment over the Tanimoto index. Enhanced with Low-Rank Adaptation (LoRA) and Chain of Thought processing, our model provides precise, interpretable predictions, informing regulatory policies and risk assessments. This methodology could extend to other persistent pollutants, advancing environmental science and AI-driven environmental management.

**INTRODUCTION**

In recent years, the persistence, bioaccumulation, and potential ecotoxicity of per- and polyfluoroalkyl substances (PFAS) have become a global environmental concern.1, 2 The stability of PFAS makes them resistant to degradation in the natural environment, thereby posing long-term effects on water bodies, soil, and even organisms.2 Biodegradation testing is an indispensable step in PFAS regulation; it determines the classification and labeling of PFAS, environmental risk assessment, and hazard evaluation.3 However, current laboratory methods for assessing biodegradability/persistence are expensive, time-consuming, and have poor reproducibility.4 To date, an estimated 145,299 chemicals (unique substances pre-registered under the EU REACH regulation) exist, of which only 4,214 have been reliably screened for biodegradability (OECD, 2017). Testing the remaining existing and new chemicals in this manner is a tedious, expensive, and potentially unfeasible task. Furthermore, the factors affecting biodegradability are poorly understood and seldom studied.5 Traditional machine learning and deep learning models have made some progress in predicting the biodegradation half-lives of PFAS, but these models are often limited by insufficient data and constraints in feature extraction.6 In the field of machine learning, the development of Large Language Models (LLMs), also known as foundation models, is one of the most notable recent advancements.7 These models have garnered widespread attention primarily due to their concise yet powerful processing capabilities: for any text prompt, whether a phrase or a sentence, these models can generate corresponding text responses in natural language. It is noteworthy that the quality of the generated text is so high that it is often indistinguishable from human writing, a reality in many cases. Currently, we are just beginning to perceive the impact of this technology, with numerous startups focusing on developing specific applications. From a scientific standpoint, one of the significant applications of these foundational models is their ability to write coherent summaries for scientific articles and even code for specific programming tasks. Recent studies have also shown that these models can handle relatively simple tabular regression and classification tasks,8–10 despite not being explicitly trained for these tasks, a point that remains particularly striking. These models demonstrate the ability to solve simple tasks, even without targeted training, raising a question: can these models also address unanswered questions in the scientific domain? Considering that most questions in the environmental field can be expressed in text form, we should be able to train these models to answer questions posed by scientists. For instance, questions like 'Can this pollutant be degraded?' or 'What is the Chemical Oxygen Demand (COD) of these waste?' are typically not answerable by theory alone or require highly complex simulations or experiments.Inspired by this,we adopted a novel approach: by fine-tuning large language models, we achieved a significant breakthrough in predicting the biodegradation half-lives of PFAS. We also employed other machine learning methods to predict the biodegradation half-lives of PFAS as a benchmark test. We found that, in zero-shot or low-shot learning scenarios, this method could surpass traditional models, providing more accurate and reliable prediction results. In addition, we discussed the applicability and interpretability of the model. This study focuses on the biodegradation of PFAS, marking the first application of large language models in the field of environmental science. By appropriately fine-tuning the model, we were able to leverage its powerful semantic understanding and pattern recognition capabilities to capture the complex biochemical processes associated with PFAS degradation. Furthermore, the ability of large language models to handle sparse datasets reveals their significant potential in data-limited applications in environmental science. In summary, this study not only provides a new technological approach for the accurate prediction of PFAS biodegradation half-lives but also opens up new research directions in using large language models to address environmental science issues. In future work, we will further explore the potential of this approach in fields such as environmental monitoring, pollutant tracking, and ecological risk assessment.

**Material and method**

**Data set and data preprocessing.** We obtained the DTXSIDs (Distributed Structure-Searchable Toxicity Substance Identifier ) and biodegradation half-life data of 7,562 PFAS from the United States Environmental Protection Agency (EPA) as the ground-truth for this study.11 As we used a language model, performing regression in the real number domain is impractical (since, strictly speaking, real numbers have an infinite number of decimal places), but we can approximate regression by retaining nine decimal places. To further enhance the performance of large language models in predicting biodegradation half-lives, an effective method is fine-tuning with domain-specific data. In this case, the dataset primarily comprises three key components: DTXSIDs , SMILES (Simplified Molecular Input Line Entry System, i.e., the textual representation of their molecular structures),12 and the actual measured values of biodegradation half-lives. For example, the molecular structure of DTXSID70790800 is 'FC(F)(F)C(Br)(C(Br)=O)C(F)(F)F,' with a half-life of 4.570881896days. This process specifically includes the following four steps: The first step concerns *Data Cleaning*, this process will ensure all data points are complete and correct, deleting or amending any erroneous or incomplete records. For SMILES strings, this may include standardizing molecular structure representations. Following that, the second step concerns *Data Transformation*, this process will convert the raw data into a question-and-answer (QA) pair format, which aids the language model in understanding and generating predictions. For instance, transforming SMILES strings and half-life values into a question about molecular half-life and an answer containing the specific half-life time. The third step concerns *Linguistic Diversity*, this process will generate multiple different phrasings for the same query rather than just one when constructing QA pairs. This objective can be achieved by using synonyms, changing sentence structures, or employing different expressions, thereby increasing the diversity of language expressions encountered by the model. The forth step concerns *Dataset Partitioning*, it will divide the data into training and test sets, ensuring an even distribution of data in both sets to prevent overfitting and ensure the model performs well on unseen data." By employing the aforementioned methods, we can effectively expand and optimize the dataset and its preprocessing. This provides a solid foundation for fine-tuning large language models, thereby enhancing their performance and accuracy in predicting the biodegradation half-lives of substances. **Figure 1** shows the workflow illustrating the integration of machine learning and chemical informatics to predict the half-life of the molecule represented by SMILES notation FC(F)(F)C(Br)(C(Br)=O)C(F)(F)F. The process begins with the molecular structure and associated half-life data (4.57 days), which is input into an inference chain of thought (CoT).13 The CoT enhanced by dialogue generation capabilities, interfaces with the Llama13B-chat AI, which has been fine-tuned using the LoRA method to incorporate annotated domain-specific knowledge.14 This integrated system leads to an updated prediction of the half-life as 3.5 days, demonstrating the potential of AI-assisted predictive modeling in chemical research.

**Model design and fine-tuning.** Data transformation is implemented through either ChatGPT-constructed or manually written code. For efficiency and effectiveness, the Llama 2-13B-chat model is chosen for fine-tuning. For fine-tuning Llama 2-13B-chat, the LoRA (Low-Rank Adaptation of Large Language Models) approach can be applied.14 LoRA is a fine-tuning technique for large language models, particularly suited for tasks like fine-tuning large language models to predict biodegradation half-lives. Its core concept is based on low-rank matrix decomposition, enabling adjustments to the model's behavior through small-scale yet effective modifications to specific internal parameters. The key advantage of LoRA lies in its ability to effectively fine-tune large models with a minimal increase in parameters. In LoRA, low-rank decomposition is used in the fine-tuning process of the model. Instead of altering all parameters of the entire model, LoRA focuses on modifying key parts, such as the weights of attention layers. By applying low-rank decomposition to these crucial parts, LoRA can effectively modify the model's behavior with minimal additional parameters. Low-rank matrix decomposition is a mathematical method for approximating large matrices. A 'matrix' can represent the weight or parameter set of a large language model. The goal of low-rank decomposition is to find two or more smaller matrices that, when multiplied together, approximate the original large matrix. The 'rank' of these smaller matrices is much lower than that of the original matrix. Fine-tuning large language models to adapt to specific domains, such as through the LoRA method, is notably effective and efficient due to several key factors. When applying LoRA for fine-tuning, we select the second-to-last layer of the Transformer in Llama 2-13B-chat. LoRA then modifies these parameters by adding low-rank matrices, whose rank is much smaller relative to the rank of the original parameter matrices. Specifically for the task of predicting biodegradation half-lives, LoRA can be used to fine-tune the language model for better understanding and prediction of the degradation process of chemical substances. LoRA makes this process both precise and efficient, as it allows the model to make detailed adjustments for specific tasks while maintaining a vast knowledge base.

**LoRA** LoRA (Low-Rank Adaptation) represents a significant leap in the fine-tuning of Large Language Models (LLMs) for specialized tasks, standing out for its ability to optimize model performance with minimal computational resource increases.14 This technique specifically targets the weights of crucial layers within LLMs, such as the attention mechanisms of the Llama 2-13B-chat model, applying low-rank matrix decomposition to effectuate precise, yet efficient modifications. LoRA's strength lies in its strategic approach to model adjustment, concentrating on altering pivotal parameters rather than overhauling the entire model structure. This method offers a balanced solution to enhancing model specificity and performance while conserving computational resources, making it an ideal choice for tasks requiring nuanced understanding and prediction, like the environmental science applications discussed. By maintaining the model's extensive knowledge base and leveraging low-rank adjustments, LoRA paves the way for advanced predictive analytics in complex domains with reduced training times and lower energy consumption, exemplifying the fusion of cutting-edge AI techniques with sustainable computing practices.

**Evaluation metrics.** Three key quantitative metrics are utilized to evaluate the efficacy of our developed regression model: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R²).15 RMSE quantifies the standard deviation of the prediction errors or residuals, reflecting the discrepancy between the predicted and observed values (as delineated in Equation 1). MAE provides an average magnitude of the absolute discrepancies between predicted and observed values (refer to Equation 2), and R², which spans from 0 to 1, serves as a statistical measure of fit, with values closer to 1 denoting a superior fit to the data (as defined in Equation 3). A model exhibiting optimal performance will demonstrate low RMSE and MAE scores, coupled with an R² value nearing 1, indicating high predictive accuracy and model reliability.

(1)

(2)

(3)

**Chain of Thought and Interpretability.** Chain of Thought (CoT) is a technique to enhance the reasoning abilities and interpretability of large language models by simulating the step-by-step reasoning process of humans.13 This method excels in handling complex problems, particularly suitable for deep understanding tasks such as predicting the biodegradation half-life of chemical substances. Using CoT, the model can demonstrate a more in-depth and detailed reasoning process, not only improving the accuracy of the predictions but also enhancing the interpretability of the results and the user's trust. Under the CoT framework, the model is guided to follow several key principles, thereby enhancing its predictive and reasoning capabilities in complex tasks. Here is a detailed analysis of these principles: Step-by-step reasoning process: The CoT method emphasizes that the model should adopt a step-by-step reasoning approach, similar to how humans solve problems. This method helps in understanding and breaking down complex problems and clearly displaying the specific steps to solve them. For instance, in predicting the biodegradation half-life of the chemical substance DTXSID70790800, CoT first requires the model to analyze the compound's chemical structure in detail, identify key functional groups (such as 2 Br, 4 F), and then assess the specific impact of these groups on the compound's stability and reactivity. This process not only helps the model to more accurately predict the compound's half-life but also enhances the understanding of the reasoning process. Transparency and Interpretability: The CoT method emphasizes generating explanatory text during the reasoning process, making the model's decision-making process more transparent. This transparency greatly enhances the user's understanding and trust in the model's predictions. For example, in the prediction of the half-life of the chemical substance DTXSID70790800, the model not only provides the prediction result (5.32 days) but also explains the reasoning, including the analysis of the impact of multiple halogens (2 Br, 4 F) and the diketone group. Such explanations provide users with a deep understanding of the logic behind the prediction results, increasing trust in the predictions. Error Detection and Self-correction: Under the CoT framework, the model can self-check each step in the reasoning process, identifying and correcting potential errors or inconsistencies. This means that the model can continuously optimize its reasoning strategy during the prediction process, thereby increasing the accuracy of the answers. In the prediction of biodegradation half-lives, this self-correction mechanism is particularly important as the reactivity and stability of chemical substances can vary greatly due to minor structural differences. By continuously checking and adjusting its reasoning process, CoT enables the model to more accurately predict the outcomes of complex chemical reactions. The Chain of Thought method not only improves the accuracy of the model in handling complex scientific problems but also enhances the interpretability and transparency of the model's predictions, thereby providing a more reliable and in-depth tool in fields like scientific research and data analysis. Through this method, we can better understand and predict complex chemical and biological processes.

**Applicability domain.** In the current study, we put forth a novel algorithm poised to enhance the assessment of chemical similarity. While the Tanimoto index has been the cornerstone in previous research for evaluating such similarity, it is tailored for binary data and may not capture the nuances of continuous variables effectively. Recognizing this limitation, we pivot towards a Euclidean distance-based metric, which promises a more nuanced approach for continuous data in cheminformatics.

Our methodology leverages the Scikit-learn library in Python,16 transforming chemical structural expressions into feature vectors and then utilizing the `euclidean\_distances` function to compute the direct distances between entities in our datasets.17 We convert these distances into similarity scores using the formula(4), yielding a score where values approaching 1 imply greater similarity.

To fine-tune our model's accuracy, we adopt a detailed approach towards adjusting the similarity threshold. Starting at 0.5 and increasing in increments of 0.05, we not only monitor the RMSE of the test set at each threshold but also employ a nuanced validation method. For each threshold, we identify the most similar sample in the training set for each test sample—that is, the training sample that has the highest similarity score above the current threshold with the test sample. If a test sample's maximum similarity with any training sample falls below the current threshold, that test sample is excluded from further analysis. Consequently, we calculate the RMSE only for the remaining test samples, those which have a counterpart in the training set that meets the threshold criteria. This process ensures that our evaluation of RMSE is grounded on test samples that the model is most likely to predict accurately, thereby providing a more precise measure of our model's performance.

The proposed Euclidean-based approach, therefore, stands as a robust alternative to the traditional Tanimoto index,18 especially in applications requiring precise similarity measures, such as drug discovery, chemical categorization, and toxicology studies. This iterative thresholding and validation method fine-tunes our model to achieve an optimal balance between predictive accuracy and model reliability, marking a significant advancement in the field of cheminformatics.

**RESULTS AND DISCUSSION**

**Seven-Fold Cross-Validation Results for the Llama 2-13B-chat Model. Table 1** shows the results of a seven-fold cross-validation of the Llama 2-13B-chat model’s predictive performance on datasets from seven different groups.15 It enumerates the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R²) for each group during both training and testing phases. During the training phase, the models exhibit substantial predictive accuracy, reflected by consistent RMSE and MAE values across the folds. In the testing phase, there is a notable decrease in performance; however, the model corresponding to Group 7 (R-7) maintains lower RMSE and MAE, alongside a higher R², suggesting a robust generalizability. This comprehensive cross-validation approach underscores the reliability of the Llama 2-13B-chat model and informs potential enhancements for its application on diverse chemical datasets.

**Comparison of Fine-tuning the LLAMA Model and Traditional Machine Learning Models Based on Molecular Descriptors at Different Data Volumes.** In the field of environmental science, acquiring high-quality data is often constrained by factors such as sampling difficulties, cost, and time, making data-driven model training particularly challenging. In this context, our research aims to evaluate and compare the performance of different computational models under data-limited conditions, especially considering the scarcity and value of data in the environmental domain. We conducted a series of experiments using the LLAMA model and traditional machine learning algorithms based on RDKit, including XGBoost, LightGBM, GradientBoost and CatBoost.19–22 We used Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as evaluation metrics to quantify the predictive accuracy of the models at different training sample sizes. The specific results are shown in **Figure 3.** The experimental results reveal that the LLAMA model performs significantly better than the RDKit-based algorithms when trained with very few training samples (zero-shot or few-shot learning scenarios). Specifically, when the sample size is 50, MAE and RMSE LLAMA algorithm achieved are 0.6023 and 0.2072. However, the other four algorithms function with MAE (0.6257-0.6834) and RMSE (0.2738-0.4253). However, as the number of training samples increases the performance improvement rate of LLAMA gradually slows down, while traditional machine learning methods continue to improve. Particularly, when the sample size approaches 1000, MAE and RMSE of LLAMA are 0.5088 and 0.1535. Comparatively, XGBoost and CatBoost show performance improvements, with MAE (XGBoost 0.4701; CatBoost 0.4821) and RMSE (XGBoost 0.0631; CatBoost 0.0516). These observations emphasize the applicability of the LLAMA model in data-limited environmental science research. Its strong zero-shot and few-shot learning capabilities allow it to provide relatively accurate predictions in scenarios with scarce samples, which is particularly valuable for environmental applications. Although the performance gains of LLAMA may be limited in data-rich scenarios, considering the typically high cost and complexity of acquiring environmental data, an accurate few-shot learning model like LLAMA can provide an effective solution to support environmental decision-making and policy formulation. Therefore, we recommend prioritizing the use of the LLAMA model in environmental applications, especially in cases of high-quality data scarcity. Furthermore, we advocate for further optimization and customization of the LLAMA model to adapt to the specific characteristics of environmental data, further enhancing its application value in environmental science.

**Chain of Thought and Interpretability.** In this study, we employed the Llama 2-13B-chat model and conducted targeted fine-tuning using LoRA to enhance the prediction accuracy of the biodegradation half-life of specific chemical compounds. To further augment the model's reasoning capabilities, we introduced a Chain of Thought (CoT) process to generate more detailed prediction explanations.13 Comparative analysis of the half-life predictions for the molecule DTXSID70790800 is shown in **Figure 2.** The left panel presents an initially incorrect simulation suggesting a half-life of 1 month. The right panel shows the refined results after a guided thought process, indicating a half-life of 5.3 days, which is deemed almost correct.Through this approach, the model not only provides a prediction for the half-life but also offers a step-by-step reasoning path, demonstrating how it arrived at this prediction. To gain a deeper understanding of the model's decision-making process, we focused on analyzing feature importance and the model's decision pathways. Firstly, we assessed the impact of input features, especially structural elements in the SMILES representation, on the model's predictions. By identifying the features that have the greatest influence on prediction results, we can discern which chemical structural attributes are more closely associated with the biodegradability of compounds. Additionally, we utilized the intermediate steps generated by CoT to trace the model's reasoning path. This not only aids in comprehending how the model progressively constructs its predictions but also reveals the assumptions or rules upon which the model may rely in specific scenarios. **Figure 4** illustrates the Chain of Thought (CoT) process for estimating the half-life of two chemical compounds. The left panel shows the CoT for Decane (DTXSID8020606), highlighting the posed question, the chemical rationale based on its non-polarity and lack of functional groups, the estimated half-life, and the actual half-life for validation. The right panel shows a similar CoT for the compound with SMILES notation FC(F)(F)C(Br)(C(Br)=O)C(F)(F)F (DTXSID70790800), discussing the impact of halogens and a dicarbonyl group on reactivity, the estimated half-life based on these features, and the corresponding actual half-life. This figure demonstrates the application of CoT in deriving educated predictions that are then compared with empirical data.

根据PFAS分析一些结论，对比不同链长，有没有规律，或者哪个biodegradable最长或最短

**Applicability domain.** In establishing the applicability domain, **Table2.** indicates that as the threshold for maximum similarity increases, more compounds in the test dataset fall outside the domain. However, the RMSE\_test first decreases and then increases, with a threshold of 0.86 yielding the minimum RMSE\_test of 0.226, suggesting optimal model prediction performance. Although the differences in RMSE\_test are marginal when the threshold exceeds 0.74, the number of compounds outside the applicability domain also rises. Lower RMSE\_test and fewer compounds outside the applicability domain are preferred when defining an optimal domain. Our analysis suggests that a threshold value of 0.86 is considered optimal, ensuring robust predictions across a diverse range of organic compounds.

**ASSOCIATED CONTENT**

**Supporting Information**.

**CRediT authorship contribution statement**

**Wufan Ding:** Investigation, Data curation, Formal analysis, Writing-original draft; **Haodong Ji:** Supervision, Conceptualization, Methodology, Funding acquisition, Writing-review & editing.

**AUTHOR INFORMATION**

**Corresponding Authors**

**Haodong Ji-** *Eco-environment and Resource Efficiency Research Laboratory, School of Environment and Energy, Peking University Shenzhen Graduate School, Shenzhen, 518055, PR China; orcid.org/0000-0002-7804-798X;* Email: jihaodong@pku.edu.cn

**Authors**

**Wufan Ding-** *Eco-environment and Resource Efficiency Research Laboratory, School of Environment and Energy, Peking University Shenzhen Graduate School, Shenzhen, 518055, PR China;*

**Aofan Liu-** *Shenzhen Key Lab of Information theory & Future Internet architecture, School of Information Engineering,* P*eking University Shenzhen Graduate School, Shenzhen, 518055, PR China.*

**Zhengkai Chen-** *Eco-environment and Resource Efficiency Research Laboratory, School of Environment and Energy, Peking University Shenzhen Graduate School, Shenzhen, 518055, PR China.*

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGEMENTS**

This work has been supported by the National Natural Science Foundation of China (No. 52100069), the Shenzhen Science and Technology Program (JCYJ20220531093205013).

**REFERENCES**

(1) Glüge, J.; Scheringer, M.; Cousins, I. T.; DeWitt, J. C.; Goldenman, G.; Herzke, D.; Lohmann, R.; Ng, C. A.; Trier, X.; Wang, Z. An Overview of the Uses of Per- and Polyfluoroalkyl Substances (PFAS). *Environ. Sci. Process. Impacts* **2020**, *22* (12), 2345–2373. https://doi.org/10.1039/D0EM00291G.

(2) Ghisi, R.; Vamerali, T.; Manzetti, S. Accumulation of Perfluorinated Alkyl Substances (PFAS) in Agricultural Plants: A Review. *Environ. Res.* **2019**, *169*, 326–341. https://doi.org/10.1016/j.envres.2018.10.023.

(3) Rücker, C.; Kümmerer, K. Modeling and Predicting Aquatic Aerobic Biodegradation – a Review from a User’s Perspective. *Green Chem.* **2012**, *14* (4), 875. https://doi.org/10.1039/c2gc16267a.

(4) Kowalczyk, A.; Martin, T. J.; Price, O. R.; Snape, J. R.; Van Egmond, R. A.; Finnegan, C. J.; Schäfer, H.; Davenport, R. J.; Bending, G. D. Refinement of Biodegradation Tests Methodologies and the Proposed Utility of New Microbial Ecology Techniques. *Ecotoxicol. Environ. Saf.* **2015**, *111*, 9–22. https://doi.org/10.1016/j.ecoenv.2014.09.021.

(5) Wang, Y.; Lai, A.; Latino, D.; Fenner, K.; Helbling, D. E. Evaluating the Environmental Parameters That Determine Aerobic Biodegradation Half-Lives of Pesticides in Soil with a Multivariable Approach. *Chemosphere* **2018**, *209*, 430–438. https://doi.org/10.1016/j.chemosphere.2018.06.077.

(6) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J.-J.; Zhang, H. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.* **2021**, acs.est.1c01339. https://doi.org/10.1021/acs.est.1c01339.

(7) Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.; Chen, A.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y.; Ruiz, C.; Ryan, J.; Ré, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; Liang, P. On the Opportunities and Risks of Foundation Models. arXiv July 12, 2022. http://arxiv.org/abs/2108.07258 (accessed 2024-02-24).

(8) Dinh, T.; Zeng, Y.; Zhang, R.; Lin, Z.; Gira, M.; Rajput, S.; Sohn, J.; Papailiopoulos, D.; Lee, K. LIFT: Language-Interfaced Fine-Tuning for Non-Language Machine Learning Tasks.

(9) Hocky, G. M.; White, A. D. Natural Language Processing Models That Automate Programming Will Transform Chemistry Research and Teaching. *Digit. Discov.* **2022**, *1* (2), 79–83. https://doi.org/10.1039/D1DD00009H.

(10) Edwards, C.; Lai, T.; Ros, K.; Honke, G.; Cho, K.; Ji, H. Translation between Molecules and Natural Language. arXiv November 3, 2022. http://arxiv.org/abs/2204.11817 (accessed 2024-02-24).

(11) Rebecca S. Katz; Rebecca S. Katz. Environmental Protection Agency. *Grants Regist. 2022* **///**. https://doi.org/10.4135/9781412914260.n159.

(12) David Weininger; David Weininger. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**. https://doi.org/10.1021/ci00057a005.

(13) Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv January 10, 2023. http://arxiv.org/abs/2201.11903 (accessed 2024-02-24).

(14) Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. arXiv October 16, 2021. http://arxiv.org/abs/2106.09685 (accessed 2024-02-24).

(15) Raschka, S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. arXiv November 10, 2020. http://arxiv.org/abs/1811.12808 (accessed 2024-02-23).

(16) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. arXiv June 5, 2018. http://arxiv.org/abs/1201.0490 (accessed 2024-02-24).

(17) Liberti, L.; Lavor, C.; Maculan, N.; Mucherino, A. Euclidean Distance Geometry and Applications. *SIAM Rev.* **2014**, *56* (1), 3–69. https://doi.org/10.1137/120875909.

(18) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminformatics* **2015**, *7* (1), 20. https://doi.org/10.1186/s13321-015-0069-3.

(19) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD ’16; Association for Computing Machinery: New York, NY, USA, 2016; pp 785–794. https://doi.org/10.1145/2939672.2939785.

(20) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2017; Vol. 30.

(21) Natekin, A.; Knoll, A. Gradient Boosting Machines, a Tutorial. *Front. Neurorobotics* **2013**, *7*. https://doi.org/10.3389/fnbot.2013.00021.

(22) Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2018; Vol. 31.

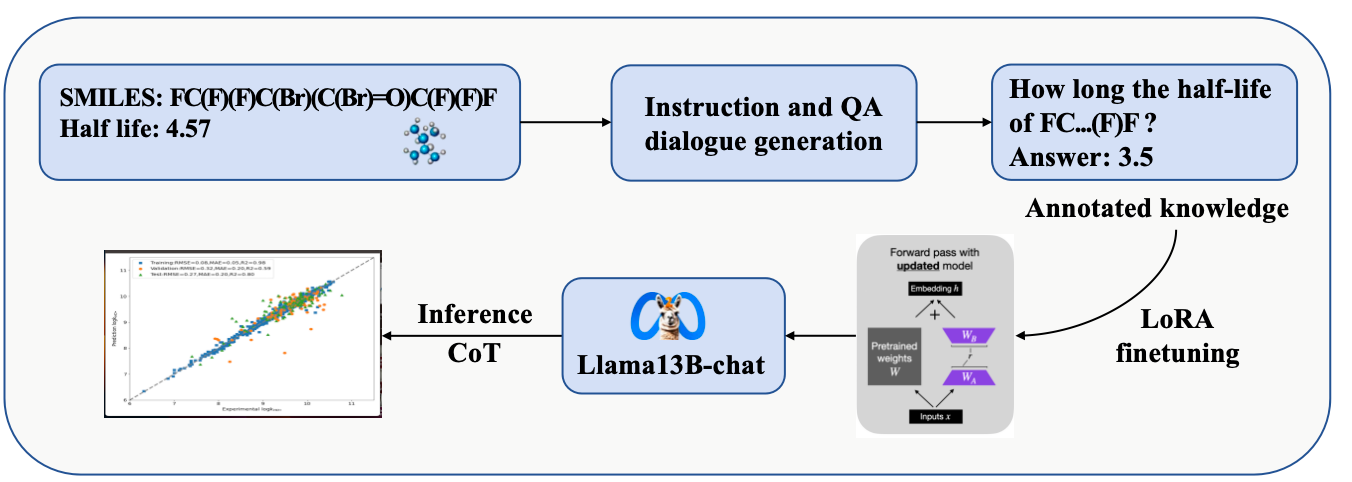
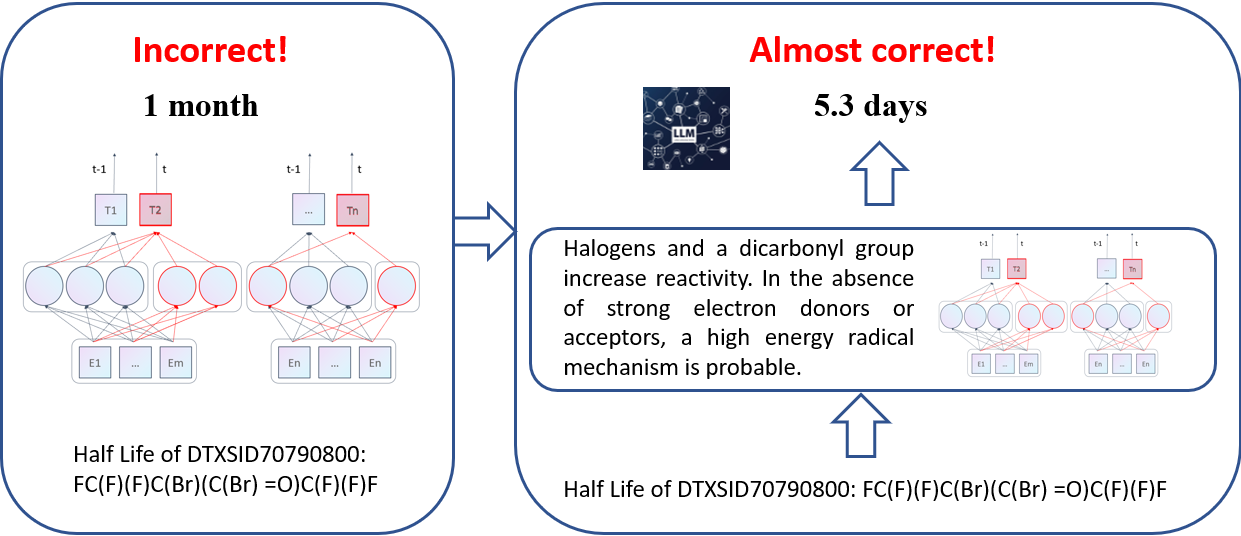
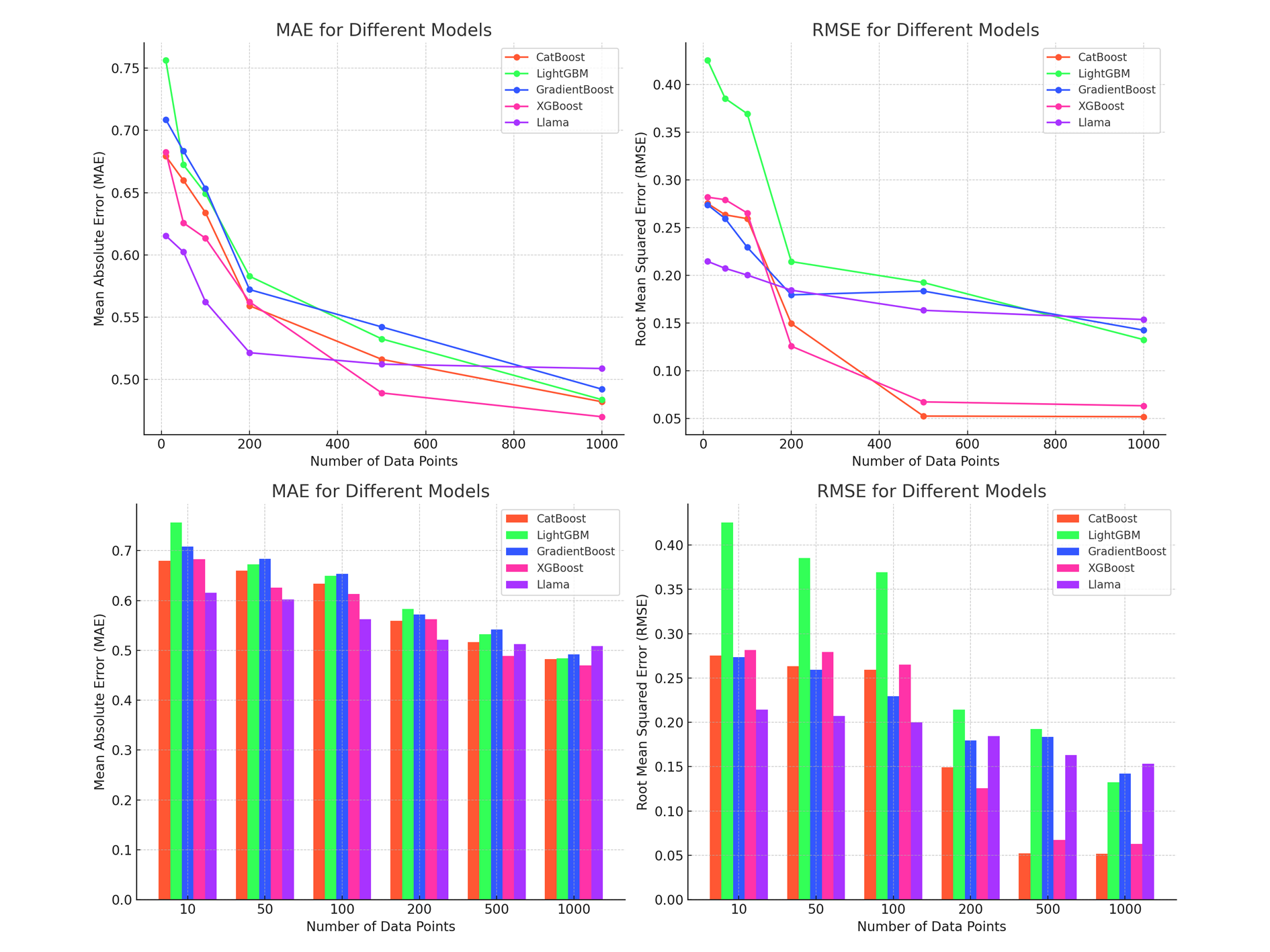


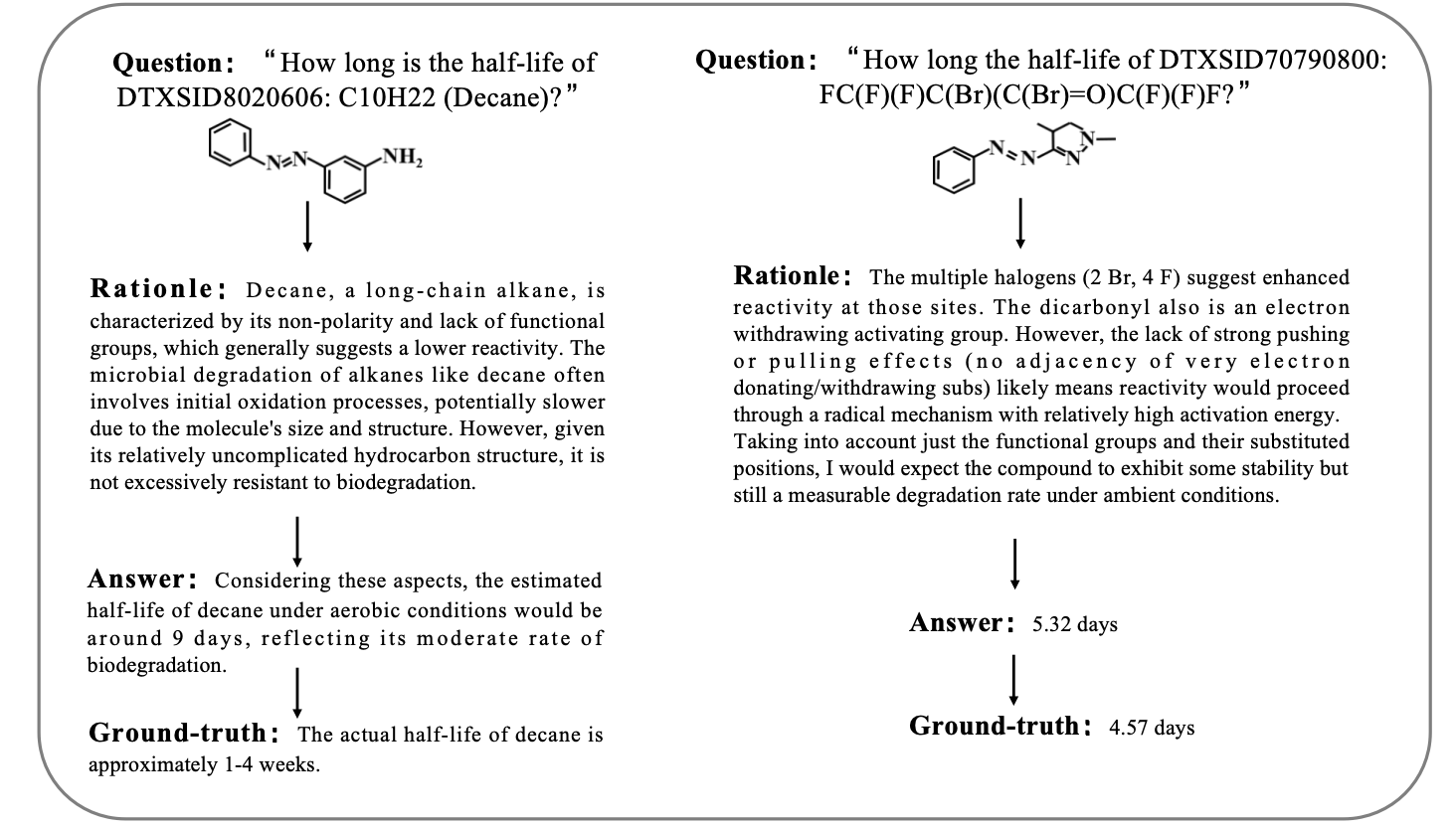
Figure 1. Flowchart of this study.



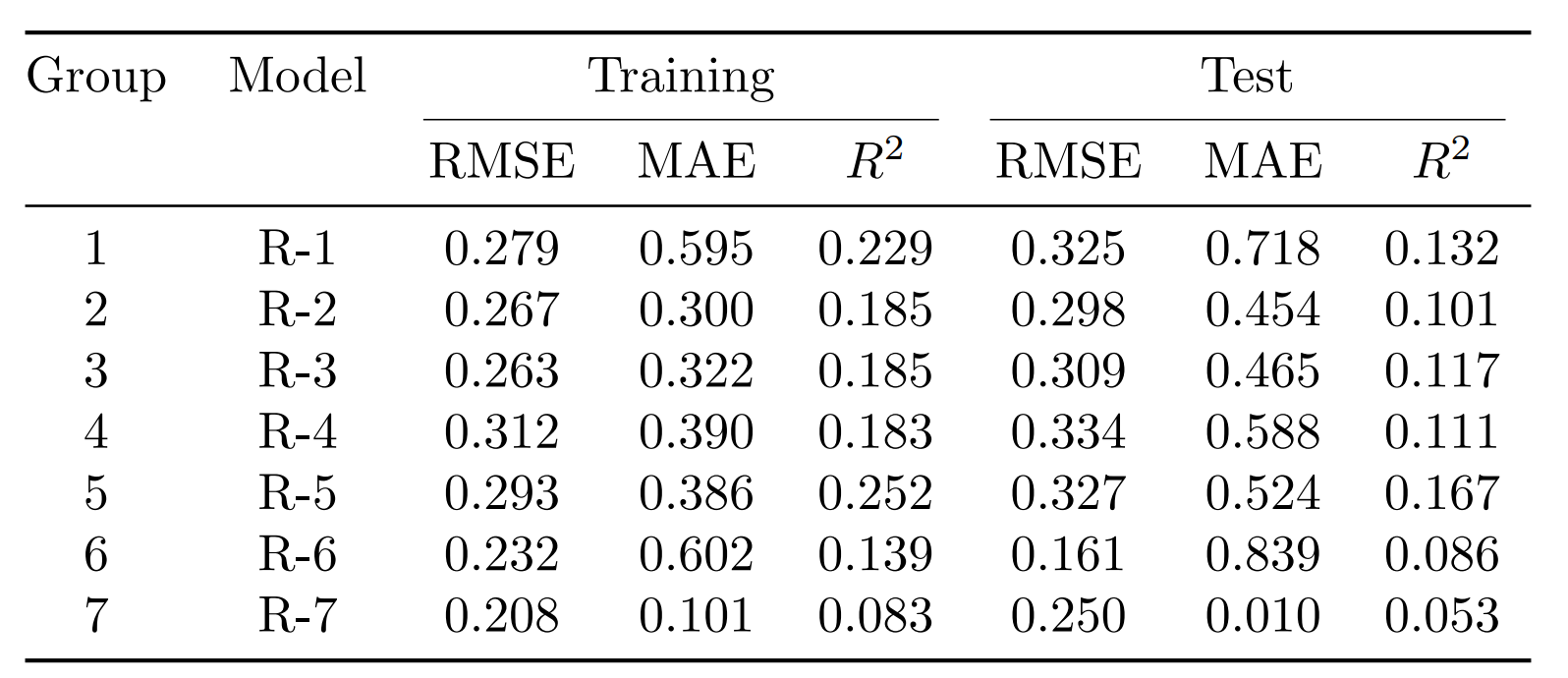
**Figure 2.** Comparative analysis of the half-life predictions for the molecule DTXSID70790800.



**Figure 3.** Performance Comparison of Machine Learning Models Over Varying Data Points. The top two plots illustrate the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) against the number of data points for five different models: CatBoost, LightGBM, GradientBoost, XGBoost, and Llama. The bottom two bar graphs provide a more detailed view of the MAE and RMSE for these models at specific data point intervals (10, 50, 100, 200, 500, 1000). Each model's performance is evaluated based on how well they minimize the MAE and RMSE, which are critical metrics for model accuracy.



**Figure 4.** Illustration of the Chain of Thought (CoT) process for estimating the half-life of two chemical compounds.

****

**Table1.Seven-Fold Cross-Validation Results for the Llama 2-13B-chat Model.**

|  |  |  |
| --- | --- | --- |
| Threshold value | # of Compounds outside the applicability domain | *RMSE*test |
| 0.74 | 0 | 0.252 |
| 0.79 | 2 | 0.243 |
| 0.84 | 3 | 0.231 |
| **0.86** | **6** | **0.226** |
| 0.88 | 7 | 0.233 |
| 0.91 | 10 | 0.242 |
| 0.96 | 12 | 0.249 |

**Table2.AD of the Model**