Explainable Molecular Concept Learning with Large Language Models

Qianli Wu[1*], Shichang Zhang[2*], Botao Xia[2], Zimin Zhang[1], Fang Sun[2], Ziniu Hu[3], Yizhou Sun[2]

[1]Department of Mathematics, UCLA
[2]Department of Computer Science, UCLA
[3]Department of Computer Science, Caltech
Los Angeles, CA, USA
qianliwu@ucla.edu

Abstract

In the field of molecular science, artificial intelligence (AI), especially deep learning models like Graph Neural Networks (GNNs), has shown remarkable effectiveness in predicting molecular properties. However, their ability to facilitate novel scientific breakthroughs is often limited by the lack of explainability. This paper introduces an innovative approach for explainable concept learning via Large Language Models (LLMs) to overcome this limitation. Our method leverages LLMs, such as GPT and Claude, for automated molecular concept generation and value assignment. Our framework streamlines the concept learning process by eliminating the need for predefined concepts and concept labels in regular concept-based methods. Our iterative refinement step also greatly enhances explainability by providing concepts with improved qualities. Through experiments on MoleculeNet datasets for molecule property prediction, we demonstrate that concepts learned with our framework can achieve accuracy comparable to advanced GNNs even with only simple models. We also propose future work directions to learn better concepts via function generation with LLMs and domain knowledge incorporation. [1]

Introduction

Artificial intelligence (AI) has been a driving force behind several groundbreaking scientific discoveries, particularly in the domain of molecular science. A prime example is the utilization of deep learning by the MIT Jameel Clinic, leading to the identification of halicin – the first antibiotic discovered in three decades that is effective against a broad spectrum of 35 bacteria (Stokes et al. 2020). In molecular science, AI models like Graph Neural Networks (GNNs) have also shown promising capabilities for learning complex atomic structures and making accurate predictions of molecular properties (Wu et al. 2018). However, a major challenge with these advanced AI models, particularly deep learning-based models like GNNs, is their lack of explainability. While these models are powerful in terms of their predictive capabilities, they are often applied as "black boxes", offering only limited insight into how their predictions are derived. This lack of explainability can be a significant hurdle in fields where understanding the prediction process is crucial, like in molecular science, a prediction process could have profound implications in new scientific discoveries. Consequently, there is a growing need for the development of explainable AI (XAI) methods in molecular science.

Concept-based models emerge as a promising XAI solution in this context (Koh et al. 2020). Instead of directly making predictions like other black-box deep-learning models, concept-based models first produce human-interpretable concepts from the data, and then make predictions from these concepts. This approach is especially beneficial for AI models applied to science problems, because it translates non-explainable vector representations in general deep-learning models into meaningful concepts domain experts can work with, increasing the chance of new scientific discoveries. However, current concept-based models have not provided a perfect solution on molecule problems. Popular works like Concept Bottleneck Models (CBMs) (Koh et al. 2020), though effective for certain tasks, requires predefined concepts and training dataset with concept labels, which limits their flexibility. Follow up work like the label-free CBM (Oikarinen et al. 2023) tries to bypass the need for predefined concepts and labels by automatic concept generation, but their primarily focus on vision tasks and the concepts are often only qualitative, e.g., color of a fruit. In contrast, the desired concepts for molecules can be more quantitative, e.g., number of aromatic rings. The concept-based XAI has also been combined with GNNs on molecule data. They employ neuron-level grouping algorithms in activation layers to discern subgraphs as concepts (Magister et al. 2021, 2022; Xuanyuan et al. 2023). This approach, while innovative, cannot capture concepts beyond subgraphs, and some interpretation of subgraph concepts can be ad hoc.

In response to these challenges, our approach innovates by leveraging the capabilities of Large Language Models (LLMs) like GPT (Brown et al. 2020) and Claude (Bai et al. 2022) for automated, extensive concept generation and value assignment. We show that with proper prompts and iterative refinements, concepts and their grounded values generated by LLMs can actually be leveraged to achieve surprisingly good performance for predicting molecule properties using simple explainable models. The underlying intuition is founded on the idea that LLMs can be treated as exten-

[*]These authors contributed equally.

XAI4Sci: Explainable machine learning for sciences, AAAI-24 (xai4sci.github.io)

[1]Access the source code here: https://github.com/Qianli-Wu/GPTConceptGen

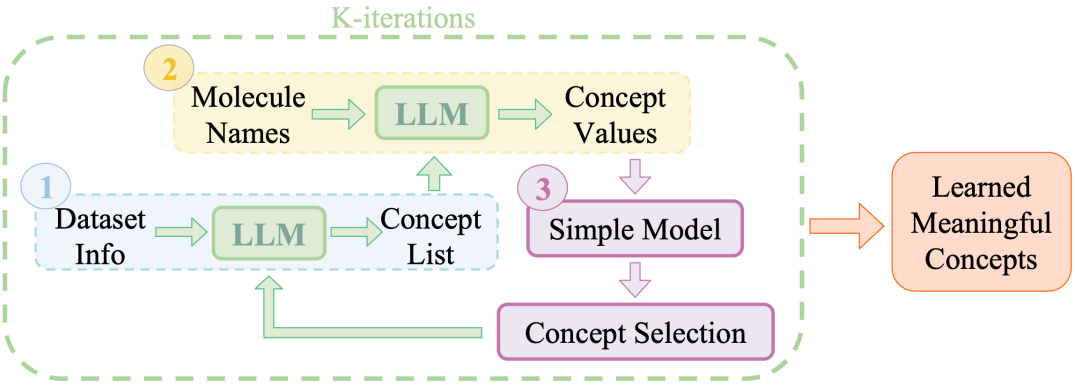

Figure 1: The concept learning framework with direct value assignment.

sive and integrated knowledge bases. Generating concepts from LLMs obviates the necessity for gathering information from various, often fragmentary, knowledge sources. Instead, our concept-learning framework is streamlined to a single, comprehensive interaction with LLMs, minimizing human error and bias for concept labeling and enhancing the efficiency and coherence of the concept learning process. Similarly to how Snorkel AI (Ratner et al. 2017) utilizes expert-derived rules to automate and enhance data labeling, our approach leverages LLMs to automate molecular concepts generation and value assignment. These concepts and their corresponding values form the foundation of our model's predictive capabilities and provide explanations that match domain knowledge. As verified by our experiments on MoleculeNet (Wu et al. 2018) datasets like FreeSolv and ESOL, our framework can match the performance of established GNN baselines for molecule property prediction while significantly enhancing explainability. This marks a notable advancement in molecular concept learning and points out a new direction of LLM-driven XAI for science. As a summarization, our contribution includes:

1. **Automatic Concept Generation and Value Assignment:** We propose an automated framework that leverages LLMs for concept generation and value assignment for molecules, which streamlines flexible concept learning and avoids concept labels in standard concept-based models.
2. **Accuracy and Explainability:** Our method produces lists of meaningful and explainable molecular concepts. Applying only simple models on these concepts achieves accuracy on par with powerful black-box GNNs.
3. **Exploration of LLM-driven XAI:** Our work highlights the potential of LLMs in addressing complex problems in molecular science and beyond. We introduce a novel perspective in concept learning, paving the way for future research that could further harness the capabilities of LLMs in scientific domains.

Related work

Predefined Concept Models: A paradigmatic example of models relying on predefined concepts is the Concept Bottleneck Model (CBM) (Koh et al. 2020). In CBMs, predictions are made through an intermediate layer of human-specified concepts, such as 'bone spurs' in medical imaging or 'wing color' in bird identification. This method enables interventions on the model's concept predictions, thereby improving accuracy and allowing for high-level concept-based interpretations. While CBMs offer structured, transparent decision-making, they are often constrained by the predefined nature of concepts, potentially limiting adaptability and application breadth. Moreover, CBMs have been shown to achieve competitive task accuracies comparable to standard end-to-end models, underscoring their utility in fields requiring interpretable models. Besides, there are several CBMs targeting particular tasks (De Fauw et al. 2018; Yi et al. 2018; Bucher, Herbin, and Jurie 2019; Losch, Fritz, and Schiele 2019; Chen, Bei, and Rudin 2020)

Automated Concept Generation Models: Advancing beyond predefined concepts, this approach employs Large Language Models for dynamic concept generation. A seminal work in this domain is the Label-free Concept Bottleneck Model (Oikarinen et al. 2023). Unlike traditional CBMs, this model automatically identifies and labels concepts directly from data, using GPT-3 for concept set creation and CLIP-Dissect for interpretability in image recognition tasks. This framework is scalable, efficient, and requires minimal human effort, marking a significant step towards flexible, AI-driven concept discovery. However, its primary application has been in the visual domain, indicating room for exploration in other data types like molecular structures.

Concept Learning in Graphs/Molecules: This area explores concept learning within the complex structure of graph data, particularly in molecular studies. The evolution of methods in this field has seen significant advancements. Starting with GCExplainer (Magister et al. 2021), which introduced human-in-the-loop approaches for concept-based explanations in graph neural networks, subsequent works have refined this idea (Magister et al. 2022) (Xuanyuan et al. 2023). They progressed from using k-means clustering to more sophisticated similarity scoring algorithms in neuron-level grouping within activation layers. These methods exemplify the attempt to extract and interpret salient features in graph data, yet they often face challenges in fully captur-

Features	FreeSolv None	FreeSolv Reg	FreeSolv Add	ESOL None	ESOL Reg	ESOL Add
GIN	-	2.307	2.151	-	**1.026**	**0.998**
GCN	-	2.413	2.186	-	1.143	1.015
RDKit	3.124	2.887	2.634	1.408	1.182	1.192
GPT-3.5 turbo	2.520	2.546	2.285	**1.253**	1.154	1.153
Claude 2	**2.205**	**2.084**	**2.027**	1.290	1.140	1.120

Table 1: Regression RMSE on test set (smaller is better).

ing the nuanced complexity of molecular structures.

Method

In this section, we introduce our method for utilizing LLMs to learn meaningful molecular concepts. Our framework capitalizes on the capabilities of LLMs for the automated generation and value assignment of concepts in molecular science. The steps are depicted in Figure 1 and outlined in the following steps.

Step 1: Concept Generation Given a particular task on molecules, we first prompt LLMs to propose a diverse list of concepts that are potentially relevant to the task. This step functions like an extensive brainstorming session. Concepts range from general attributes like "Molecular Weight" to specific ones such as "Polar Surface Area" (PSA). The LLMs' capacity to comprehend and generate complex concepts is pivotal in this phase, yielding a wide spectrum of potentially relevant concepts for our analysis.

Step 2: Automated Value Assignment Following the concept generation, we proceed to value assignment. In this step, the LLMs directly assign numerical values to the generated concepts. This automation bypasses traditional, labor-intensive methods, reducing susceptibility to human error and bias, and is crucial for efficiently processing the large volumes of data encountered in molecular studies.

Step 3: Model Fitting and Concept Selection With concepts and their values at hand, we move on to integrate these concepts into simple, explainable models like logistic and linear regression. This step includes fitting the concept values as features in the model and applying the Akaike Information Criterion (AIC) (Akaike 1973, 1974) to identify the most effective subset of concepts.

Iterative Concepts Refinement We do an iterative refinement of the learned concepts by prompting LLMs again with the empirical performance of our simple model and the feature selection results from step 3. Including such information in an updated prompt allows LLMs to generate new concepts to replace the less useful ones from the previous iteration, ensuring that our model remains adaptable and up-to-date with the most relevant molecular features.

A potential improvement of the framework is to update step 2 to value assignment through function generation, which we discuss in the future work section.

Experiments

Datasets We use two regression molecule datasets from MoleculeNet, a benchmark suite for molecular machine learning (Wu et al. 2018). FreeSolv provides hydration free energy data for 642 molecules, while ESOL contains water solubility data for 1128 organic small molecules. We employed the same data splits as used in the Open Graph Benchmark (OGB) (Hu et al. 2020). Results are measured with Root Mean Square Error (RMSE). For dataset has three versions. 1) No node/edge features 2) Regular node feature (2 dim) and edge features (2 dim) 3) Additional node feature (9 dim) and edge features (3 dim).

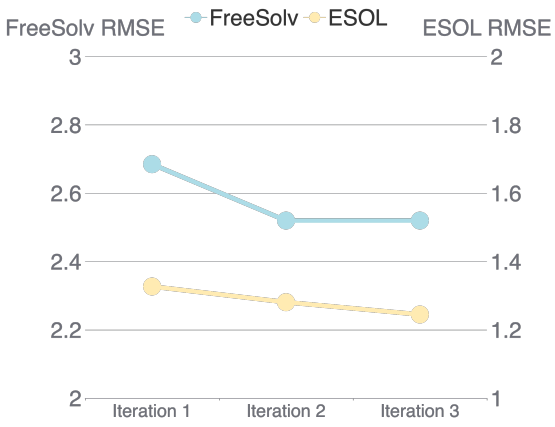

Figure 2: Prediction RMSEs get better over three iterations. The blue line is for FreeSolv using GPT-3.5 Turbo, and the yellow line is for ESOL using Claude 2.

Baselines We consider three baseline models, including two GNNs used in OGB benchmark: Graph Isomorphism Network (GIN) and Graph Convolutional Network (GCN), which provides a well-established standard and represents the state-of-the-art in the field. Note that GNNs can only work when the node/edge features are provided Additionally, we include a baseline linear model utilizing five molecule properties that can be precisely generated with RDKit (Landrum 2010) as features, which represents a more traditional feature-engineering approach.

Experiment Setting We use GPT-3.5 turbo and Claude 2 as our backbone LLMs for concept generation and value assignment. After collecting the concept values, we use them to predict the labels using linear regression models, implemented with the scikit-learn package (Pedregosa et al. 2011). Our method can run without relying on the provided node/edge features (the first column for each dataset in Table 1). For a fair comparison with GNNs, we also consider summing the node/edge features into graph-level features and adding them as input to the linear regression model. Since this will greatly increase the feature dimension, we apply a LASSO regression with a regularizing coefficient chosen from [0.01, 0.1, 1]. We then report the best model, where the best alpha is 0.1 for FreeSolv and 0.01 for ESOL.

Molecule Property Prediction The results in Table 1 compare our approach against the baselines. For the FreeSolv dataset, our method with either GPT-3.5 turbo and Claude 2 achieves competitive RMSE scores to GNN mod-

Dataset	LLM	Molecular Weight	Number of Rings	Polar Surface Area	Lipophilicity	Number of Rotatable Bonds
FreeSolv	GPT-3.5 turbo	0.914	0.645	0.633	0.161	0.549
FreeSolv	Claude 2	0.924	0.875	0.540	0.766	0.377
ESOL	GPT-3.5 turbo	0.832	0.462	0.519	0.692	0.491
ESOL	Claude 2	0.841	0.518	0.525	0.679	0.381

Table 2: RDKit benchmarking results measured in R^2 (ranging from 0 to 1, higher the better).

ESOL	FreeSolv
* Polar surface area	* refractivity
* Number of aromatic rings	* polar surface area
* Presence of charged groups	* heavy atom count
* Octanol-water partition coefficient	* ring count
* Melting point	* aromatic ring count
* Polarity	
* Hydration energy	
* Lipophilicity	

Table 3: Learned concepts by our framework with Claude 2.

els and outperforms the RDKit baseline, respectively. In particular, our results with Claude 2 achieve the best among all methods across three settings. For the ESOL dataset with features, our results are not as good as GNNs, but the difference is small We hypothesize the results are affected by the direct value assignment being not perfectly accurate, as we checked with RDKik benchmarking in the next part. Also, summing node/edge features into graph-level features is a rough way of incorporating them. By employing the function value assignment idea discussed in future work, the results have room for improvement. We also show in Figure 2 that the prediction gets better as we run more iterations.

Benchmarking Concept Values with RDKit The concept value assignment from step 2 is crucial for our method. To evaluate its accuracy and reliability, we use RDKit to benchmark the assigned values by LLMs. This is particularly pertinent for concepts where RDKit can calculate ground-truth values for some molecular properties, such as Molecular Weight and Polar Surface Area. We compare these ground-truth values with those derived from our models using R-square (R^2). This helps us to better understand how accurate the generated concepts are. The RDKit benchmarking, as shown in Table 2, reveals the correlation between our model's concept values and RDKit's ground-truth values. The R-square values across FreeSolv and ESOL datasets indicate an acceptable level of accuracy, especially for concepts like Molecular Weight and Polar Surface Area. These findings affirm the viability of our approach in generating concept values, and at the same time point to potential directions for improvement on concepts with lower correlation scores. We discuss such improvements in future work.

Concept Interpretation The learned concepts of our method are shown in Table 3. The selection of refractivity, polar surface area, heavy atom count, ring count, and aromatic ring count as essential factors for the FreeSolv dataset is consistent with their influence on hydration free energy. Refractivity, indicating how light interacts with molecules in water (Foerst et al. 1967), and polar surface area, correlating directly with solubility (Pajouhesh and Lenz 2005), are pivotal in understanding molecular behavior in aqueous environments. Heavy atom count, a proxy for molecular size and complexity, influences how molecules interact with water (Sheridan et al. 2014). Additionally, ring and aromatic ring counts, highlighting structural rigidity and electronic properties, are significant for predicting a molecule's behavior in water (Ritchie and Macdonald 2009).

For the ESOL dataset, which focuses on water solubility data, the selected features are highly pertinent. Polar surface area and the number of aromatic rings are critical in influencing a molecule's solubility and interaction with water (Pajouhesh and Lenz 2005). The presence of charged groups is vital in solubility determination in polar solvents like water due to their strong interactions with water molecules (Burke 1984). Moreover, the octanol-water partition coefficient directly reflects a molecule's hydrophobicity and thus its solubility in water (Sangster 1997).

Future work

Value Assignment Through Function Generation As direct value assignment for concepts can be challenging for LLMs in some cases, we also introduce a variant of our method with an updated step 2 for value assignment through function generation. The idea is shown in Figure 3 in Appendix, where we prompt LLMs to generate Python functions f_i for each concept generated in step 1. Then we compute the concept values by plug in molecule graphs G_j into the each function. There are two advantages for function generation compared to direct value assignment. One is that functions allow a better usage of the available molecule information, like the molecule structure in terms of a graph adjacency matrix and atom and bond information in terms of node and edge features. In contrast to text information like molecule names, such matrices and numeric values are hard to be processed by LLMs directly. The second advantage is that for some concepts, especially quantitative ones like molecular weight, functions can compute their values more precisely, avoid errors in direct value assignment.

Concept Refinement with Domain knowledge In our experiments, Claude 2 suggested "pKa" as a relevant concept for the ESOL dataset. However, pKa is a property specific to acidic compounds, and not all compounds in ESOL possess this attribute. This led to a limitation where LLMs struggled to assign pKa values universally across the dataset. This highlights a current constraint in our model's ability to discern applicability of certain properties to specific compounds. On the other hand, this shows the potential to fur-

ther improve the performance of our method by employing domain knowledge.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, 267–281. Akademiai Kiado.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6): 716–723.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bucher, M.; Herbin, S.; and Jurie, F. 2019. Semantic bottleneck for computer vision tasks. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14*, 695–712. Springer.

Burke, J. 1984. Solubility parameters: theory and application.

Chen, Z.; Bei, Y.; and Rudin, C. 2020. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12): 772–782.

De Fauw, J.; Ledsam, J. R.; Romera-Paredes, B.; Nikolov, S.; Tomasev, N.; Blackwell, S.; Askham, H.; Glorot, X.; O'Donoghue, B.; Visentin, D.; et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9): 1342–1350.

Foerst, W.; et al. 1967. Chemie für Labor und Betrieb. *Chemie für Labor und Betrieb*, 3: 32–34.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33: 22118–22133.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International conference on machine learning*, 5338–5348. PMLR.

Landrum, G. 2010. RDKit: Open-source cheminformatics. https://www.rdkit.org. Accessed: Nov 22, 2023.

Losch, M.; Fritz, M.; and Schiele, B. 2019. Interpretability beyond classification output: Semantic bottleneck networks. *arXiv preprint arXiv:1907.10882*.

Magister, L. C.; Barbiero, P.; Kazhdan, D.; Siciliano, F.; Ciravegna, G.; Silvestri, F.; Jamnik, M.; and Lio, P. 2022. Encoding concepts in graph neural networks. *arXiv preprint arXiv:2207.13586*.

Magister, L. C.; Kazhdan, D.; Singh, V.; and Liò, P. 2021. Gcexplainer: Human-in-the-loop concept-based explanations for graph neural networks. *arXiv preprint arXiv:2107.11889*.

Oikarinen, T.; Das, S.; Nguyen, L.; and Weng, L. 2023. Label-free Concept Bottleneck Models. In *International Conference on Learning Representations*.

Pajouhesh, H.; and Lenz, G. 2005. Medicinal Chemical Properties of Successful Central Nervous System Drugs. *NeuroRx*, 2(4): 541–553.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Ratner, A.; Bach, S. H.; Ehrenberg, H.; Fries, J.; Wu, S.; and Ré, C. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, 269. NIH Public Access.

Ritchie, T. J.; and Macdonald, S. J. 2009. The impact of aromatic ring count on compound developability–are too many aromatic rings a liability in drug design? *Drug discovery today*, 14(21-22): 1011–1020.

Sangster, J. M. 1997. *Octanol-water partition coefficients: fundamentals and physical chemistry*, volume 1. John Wiley & Sons.

Sheridan, R. P.; et al. 2014. Modeling a crowdsourced definition of molecular complexity. *Journal of Chemical Information and Modeling*, 54: 1604–1616.

Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; et al. 2020. A deep learning approach to antibiotic discovery. *Cell*, 180(4): 688–702.

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

Xuanyuan, H.; Barbiero, P.; Georgiev, D.; Magister, L. C.; and Liò, P. 2023. Global concept-based interpretability for graph neural networks via neuron analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10675–10683.

Yi, K.; Wu, J.; Gan, C.; Torralba, A.; Kohli, P.; and Tenenbaum, J. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31.

Appendix

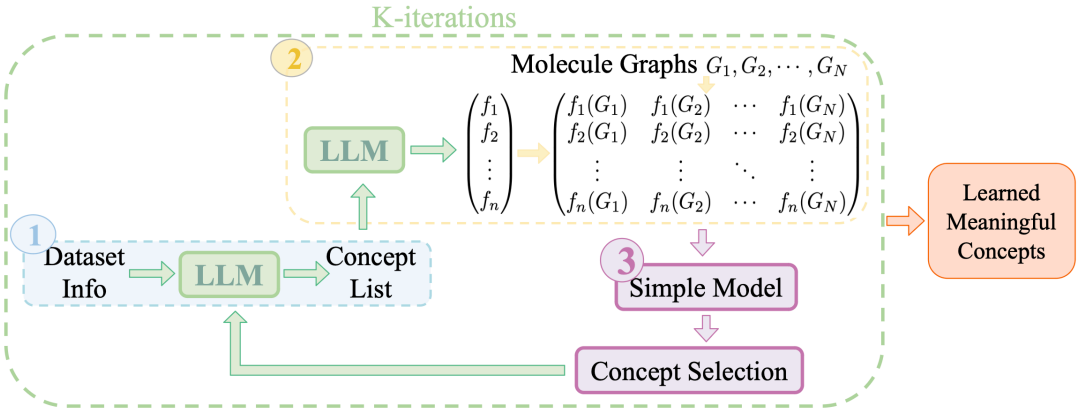

Figure 3: The concepts learning framework with value assignment through functions generation, which allows us to include more information of molecule graphs.

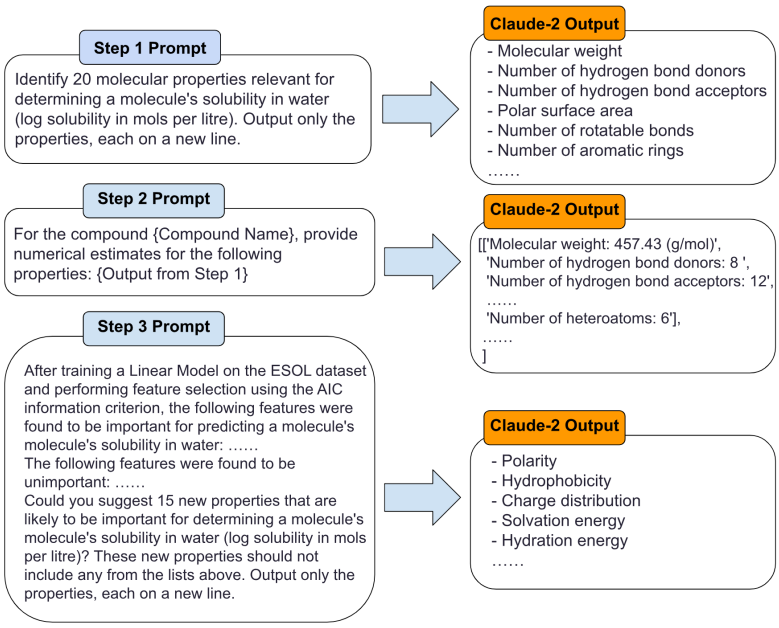

Figure 4: Prompts for concept generation and value assignment with Claude 2.

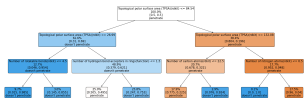

Figure 5: Enter Caption