Beyond Chemical Language: A Multimodal Approach to Enhance Molecular Property Prediction

Eduardo Soares
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
eduardo.soares@ibm.com

Emilio Vital Brazil
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
evital@br.ibm.com

Karen Fiorella Aquino Gutierrez
IBM Research Brazil
São Paulo, SP, Brazil
fiorella.aquino@ibm.com

Renato Cerqueira
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
rcerq@br.ibm.com

Dan Sanders
IBM Research Almaden
San Jose, CA, USA
dsand@us.ibm.com

Kristin Schmidt
IBM Research Almaden
San Jose, CA, USA
schmidkr@us.ibm.com

Dmitry Zubarev
IBM Research Almaden
San Jose, CA, USA
dmitry.zubarev@ibm.com

Abstract

We present a novel multimodal language model approach for predicting molecular properties by combining chemical language representation with physicochemical features. Our approach, MULTIMODAL-MOLFORMER, utilizes a causal multi-stage feature selection method that identifies physicochemical features based on their direct causal effect on a specific target property. These causal features are then integrated with the vector space generated by molecular embeddings from MOLFORMER. In particular, we employ Mordred descriptors as physicochemical features and identify the Markov blanket of the target property, which theoretically contains the most relevant features for accurate prediction. Our results demonstrate a superior performance of our proposed approach compared to existing state-of-the-art algorithms, including the chemical language-based MOLFORMER and graph neural networks, in predicting complex tasks such as biodegradability and PFAS toxicity estimation. Moreover, we demonstrate the effectiveness of our feature selection method in reducing the dimensionality of the Mordred feature space while maintaining or improving the model's performance. Our approach opens up promising avenues for future research in molecular property prediction by harnessing the synergistic potential of both chemical language and physicochemical features, leading to enhanced performance and advancements in the field.

1 Introduction

Machine learning has become a popular and efficient tool for predicting molecular properties in drug discovery and material engineering [1]. While traditional models were trained on predefined descriptors like molecular fingerprints or geometric features [2], recent models now focus on chemical language representations as they can facilitate the creation of molecules that exhibit specific desired characteristics [3]. One example of chemical molecule representations are the SMILES, which is a character string representation of a molecule generated by flattening the molecular graph through a depth-first pre-order spanning tree traversal [4]. It represents important structural features of

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

molecules, including branches, cyclic structures, and chirality information, which are not explicitly represented in the graph representation. However, string-based representations lack topological awareness [5], which may hinder the ability of deep chemical language models to capture the implicit topological structure of molecular graphs [6].

Moreover, in the domain of molecular property prediction such as the estimation of the biodegradability of organic molecules and the toxicity of per- and polyfluoroalkyl substances (PFAS), the scarcity of labeled data [7] is a major challenge that limits the effectiveness of supervised training for large language models [8]. The cost associated with labeling molecules and the vastness of the space of plausible chemicals that require labeling further exacerbates this issue [9]. Therefore, there is a growing need for molecular representation learning that can be generalized to various property prediction tasks in an unsupervised or self-supervised setting [10].

Overall, chemical language-based machine learning has become a widely adopted approach for predicting molecular properties due to its efficiency and ability to accurately represent important structural features of molecules [11]. Recent advances in large transformers-based foundation models have demonstrated promising results in learning task-agnostic based on chemical language representations through pre-training on large unlabeled corpora and subsequent fine-tuning on downstream tasks of interest [12, 13]. While pre-trained Language Models (LMs) have emerged as viable options for predicting molecular properties [13, 14, 15], they are still in their early stages of development, and there remains a need for further research to improve their performance and address limitations such as generalization and sample efficiency [8, 16].

Here we present a novel multimodal language model (MULTIMODAL-MOLFORMER) approach for predicting molecular properties, which combines chemical language representation and physicochemical features. Our approach employs a causal multi-stage feature selection method that selects physicochemical features based on their direct causal-effect on a specific target property to predict. Specifically, we use Mordred descriptors as physicochemical features and Markov blanket causal graphs as the inference algorithm to identify the most relevant features. Our results demonstrate that our proposed approach outperforms existing state-of-the-art algorithms, including the chemical language-based MOLFORMER and graph neural networks, in predicting complex tasks such as the biodegradability of general compounds and PFAS toxicity estimation. Furthermore, we show the effectiveness of our feature selection method in reducing the dimensionality of the feature space while maintaining or improving performance. Our approach provides a promising direction for future research in molecular property prediction, as it leverages both chemical language and physicochemical features for improved performance.

2 Related Work

Traditional chemical fingerprints, such as ECFP [17], provide an efficient method to represent chemical molecules in a vector space. The algorithm explores the substructure connectivity around each atom of a molecule within a given radius, generating numeric representations for each substructure identified [11]. These representations are then combined to create a fixed-length bit string that captures the topological chemical space of the molecule. However, while ECFP fingerprints provide useful topological information about molecules, they lack essential details about key molecular properties such as electronegativity, polarizability, and surface area [5]. To overcome this limitation, molecular descriptors like Mordred have been developed, enabling the computation of physicochemical information about molecules [18], which can significantly enhance the discovery process through machine learning algorithms [19].

Recently, the success of language representation models in downstream natural language processing tasks has sparked interest in extending this paradigm to other domains [20]. Advanced neural networks, such as transformers, can be leveraged to pre-train on large unlabeled corpora and contextual language models, resulting in domain-specific "language" embeddings [21]. This embedding becomes the exclusive input for several downstream tasks, achieving superior performance and accuracy [20].

For example, advanced language models trained on protein sequences have been successful in understanding the language of life [22], where features extracted by the models directly from single protein sequences achieve state-of-the-art performance in downstream prediction tasks, even when used without evolutionary information [23]. Similar large-scale language models pre-trained on chemical strings sequences have been explored for molecular property prediction [13]. However,

chemical language-based models may lack awareness of topological and physicochemical properties, which may hinder their ability to capture implicit properties of such molecules [19].

To address this limitation, we propose a multimodal language model, MULTIMODAL-MOLFORMER, that uses string-based chemical representation and physicochemical features based on Mordred descriptors and selected based on their causal effect on the properties to predict. By combining these two modalities, our model can capture both the topological and physicochemical properties of molecules, improving performance in downstream prediction tasks.

3 Overview of the proposed approach

The architecture of the proposed multimodal language model for molecular prediction is shown in Fig. 1. The details of the base model is described in section 3.1. The tokenization employed is described in section 3.2. The multi-stage causal-based feature selection of molecular properties is presented in section 4. Section 4.5, describes the concatenation process of string-based features and physicochemical features derived from the causal selection.

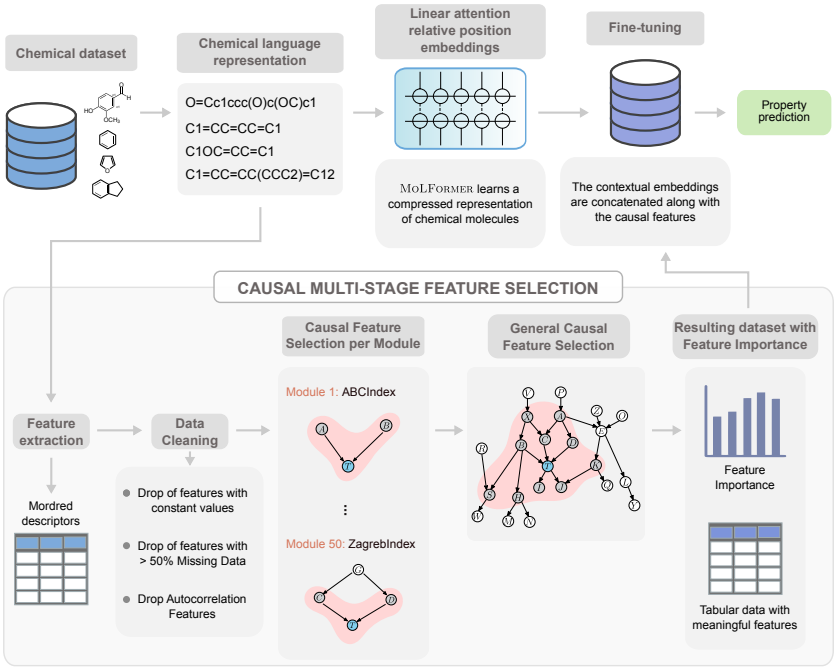

Figure 1: The diagram above depicts the proposed architecture of MULTIMODAL-MOLFORMER, a multimodal language model designed for predicting molecular properties. This model is equipped with a causal multi-stage feature selection method, which enhances its ability to accurately predict these properties.

3.1 Details of the base model

Our approach is built on MOLFORMER [13], a state-of-the-art transformer-based model for chemical language representations. MOLFORMER is a large-scale masked language model that processes inputs through a series of blocks that alternate between self-attention and feed-forward connections.

MOLFORMER's self-attention mechanism allows the network to construct complex representations that incorporate context from across the sequence. By transforming the sequence features into queries (q), keys (k), and value (v) representations, attention mechanisms can weigh the importance of different elements within the sequence. This enables the model to learn highly informative representations of the input data, making it a powerful tool for predicting molecular properties.

Recent studies have demonstrated that incorporating relative position embeddings between tokens results in enhanced performance [14]. To optimize the relative encoding through position-dependent

rotations R_m of the query and keys at position m, the model uses a modified version of the RoFormer [24] attention mechanism. These rotations can be implemented efficiently as pointwise multiplications and do not significantly increase computational complexity as shown in Eq (1).

$$Attention_m(Q, K, V) = \frac{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle \, v_n}{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle} \tag{1}$$

where Q,K,V are the query, key, and value respectively, and φ is a random feature map.

3.2 Tokenization process and vocabulary construction

The approach uses a tokenization process, as detailed in [12], to create its vocabulary. This process utilizes 1.1 billion molecules from PubChem and ZINC datasets to generate 2362 unique vocabulary tokens. These tokens are then employed to fine-tune or retrain the models with a fixed embedding capacity and vocabulary size. To reduce computation time, the sequence length has been limited to a range of 1 to 202 tokens, including special tokens since over 99.4% percent of all 1.1 billion molecules contain less than 202 tokens.

4 Causal multi-stage feature selection

The multi-stage feature selection method we propose follows a causal-based approach and is divided into four distinct modules (see Fig. 1). The first module is the feature extraction block, which extracts relevant features from the data. The second module is the data cleaning block. Next, we have the causal feature selection per module block, which is the core of our method. In this module, we use Markov Blanket causal inference to identify and select features that have a causal relationship with the outcome of interest. Finally, we have the general feature selection block, which further refines the feature selection by considering other factors such as redundancy and predictive power.

4.1 Feature extraction

To manipulate the chemical language representation for each molecule the RDKit [25] package has been used. This allowed us to index and merge data from different sources. The RDKit was used to standardize the chemical language representations, mapping different strings belonging to the same molecule to a single molecule object.

We then extracted molecular features using the Mordred descriptor [5]. This open-source library boasts 1826 features organized into 50 different modules, including both two- and three-dimensional descriptors.

4.2 Data cleaning

Following the extraction of Mordred features from the chemicals, we initiated the data cleaning process as outlined below:

a) **Drop of features with constant values:** Features with constant values provide little or no relevant information to machine learning models and may introduce unnecessary noise. Thus, we removed such features during the data cleaning process.

b) **Drop of features with more than 50% of missing values:** Features with more than 50% of missing values may not provide meaningful insights as the missing values could mask important information.

c) **Drop of Autocorrelation features:** There are more than 600 Autocorrelation descriptors. As per the findings of [26], these do not correspond to structural, physical, and chemical properties of the molecule. Therefore, we opted to remove these features during the cleaning step.

After this process, the per-module causal feature selection is initiated.

4.3 Per-module causal feature selection

In contrast to correlation analysis, causal-effect analysis delves deeper into the cause-and-effect relationship between a feature and a target variable [27]. Keeping this in mind, we developed a per-module causal feature selection approach for Mordred descriptors. Since the Mordred descriptor generates over 1800 features across 50 different modules, adopting a causal approach per module proved to be an efficient strategy, optimizing computational resources.

The Markov Blanket algorithm has been used to built the causal graphs. The Markov Blanket (MB) algorithm theoretically selects a optimal set of features considering their performance both individually and in group to predict the variable target T (see [28]). Therefore, with MB(T) it is enough to know the distribution of T and all the values of the other variables that do not belong to MB(T) become superfluous. Next, in Fig 2, we present an example of a Markov Blanket graph that is used to represent the joint probability distribution of a set of features $\{A, B, C, D, E, T\}$ in the form of a directed acyclic graph (DAG). The nodes of the graph form the features and the directed edges form the conditional dependencies between these features. The MB(T) consists of its parents (direct causes), children (direct effects), and spouses (other parents of this variable's children) of T [27]. In the graph the variables inside the gray circle are in the MB of the target T and these are used for their prediction.

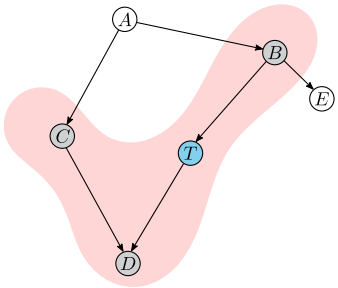

Figure 2: The figure illustrate a Markov Blanket Graph where the target T is direct affected by the feature B and causes feature D.

In this study, we considered the Predictive Permutation Feature Selection (PPFS) algorithm [29], which employs a Markov Blanket graph for feature selection. PPFS stands out as a versatile method since it enables subset selection for datasets encompassing both categorical and continuous features, making it applicable to both classification and regression tasks.

Markov blankets are the minimal set of features that can predict the target variable [30]. Therefore, in the end of this process we choose the MB with the highest causal values for each module contained in the set of features descriptors.

4.4 General causal feature selection

At this step, we run again the Markov Blanket based algorithm to the resulting set after joining all features descriptors selected in each module in previous step. In this sense, we filter the most important features descriptors once and minimize the redundant information between modules. At the end of this process, we have a parsimonious set of features with their feature importance which can be analyzed by experts. Moreover, we also have a cause-effect graph which can be used to gain deeper insights about the problem treated.

4.5 Concatenation of chemical language-based features and physicochemical features

Here, we present the concatenation layer of the MULTIMODAL-MOLFORMER. At this step, the chemical language representation embeddings are combined with the physico-chemical features derived from the causal multi-stage feature selection process.

Let (x, y) denote a feature-target pair where $x = (x_{\text{CL}}, x_{\text{causal}})$. The x_{CL} denotes all the features based on chemical language representations, and x_{causal} refers to physicochemical features selected during the multi-stage causal-based feature selection process. Where, each x_{CL} can have a sequence of tokens with length limited to a range of 1 to 202 tokens, and 768 embedding per token. Fig. 3 illustrates the concatenation architecture of the MULTIMODAL-MOLFORMER architecture. The resulting embeddings per chemical language string, after the data transformation (Transformers layers, see Fig. 3a), is concatenated along with the physicochemical features resulting in a learning vector of dimension $(d \times e + c)$. Where, d is the dimension of the dataset, e is the size of the resulting embeddings and c is the number of physicochemical features. The resulting vector is then

passed to a learning algorithm to calculate the loss function. In this case, the learning algorithm is a Feed-Forward network with 2 fully connected layers.

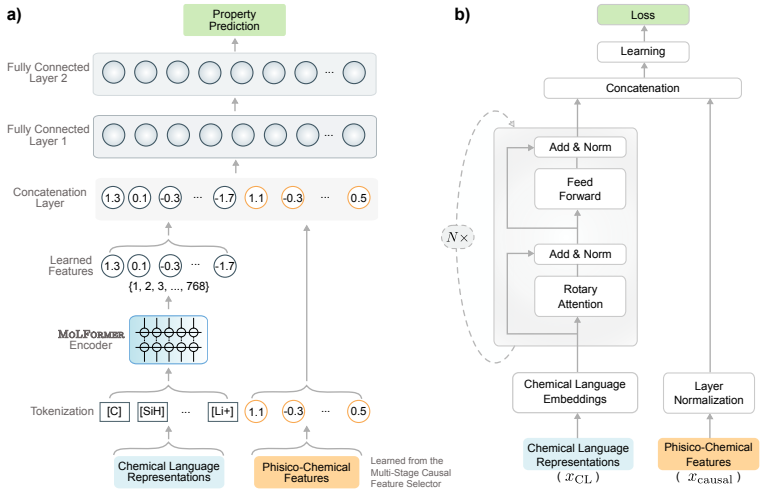

Figure 3: The Fig.3a illustrates the general architecture of the learning process of the MULTIMODAL-MOLFORMER with the concatenation layer. Fig.3b emphasizes the feature learning of chemical language representations.

5 Experiments

To evaluate this proposed methodology, we assessed its efficacy in tackling two demanding tasks: predicting the toxicity of PFAS substances [31], and classifying molecules by their biodegradability [32]. The study of PFAS toxicity presents a captivating case, as it requires a thorough understanding of their harmful effects to effectively address and remediate their environmental impact [33]. In the same token, the evaluation of the biodegradability of molecules offers an intriguing study case, since it is critical to establish effective assessment techniques for categorizing the possible biodegradability of organic materials [34]. The datasets utilized in this study are described below.

5.1 Dataset for toxicity estimation

To predict acute toxicity of chemical compounds we considered the LDToxDB dataset [35] which is part of the AI4PFAS project[1]. The LDToxDB dataset is composed of 13,329 unique compounds with oral rat LD_{50} aggregated from the measurements aggregated from the EPA Toxicity Estimation. Acute toxicity refers to a chemical's propensity to cause adverse health effects within a short period.

A more generalized approach for assessing the relative toxicity of a group of compounds, without delving into specific biological pathways, involves comparing the median lethal doses (LD_{50}) measured in milligrams per kilogram of body weight (mg/kg). The LD_{50} metric quantifies the minimum amount of a compound required to cause fatality in 50The United States Environmental Protection Agency (EPA) employs a toxicity ranking system based on the quantity of the substance relative to the mass of a laboratory rat's body weight. Under this classification, EPA designates Class 1 for high toxicity, where $LD_{50} < 50$; Class 2 for moderate toxicity, encompassing LD_{50} values between 50 and 500; Class 3 for low toxicity, covering LD_{50} values ranging from 500 to 5000; and Class 4 for compounds exhibiting minimal toxicity, with LD_{50} values exceeding 5000.

5.2 Dataset for biodegradability estimation

In this study, we utilize the "All-Public set" to predict the biodegradability of compounds [36]. The "All-Public set" is a composite dataset compiled from five distinct sources: I) the ECHA database; II) the NITE database; III) VEGA database; IV) the EPI Suite; V) and the OPERA suite. The dataset

[1]AI4PFAS project `https://github.com/AI4PFAS/AI4PFAS`.

provides information on the biodegradability of compounds, represented by two distinct categories, "ready-biodegradability" (RB) and "not ready-biodegradability" (NRB), each containing 1097 and 1733 data samples, respectively.

6 Results and Discussion

The Results and Discussion section of this paper is structured into two subsections. Firstly, we detail the outcomes of our acute toxicity prediction and compare our findings with state-of-the-art methodologies. Secondly, we present the findings of our biodegradability prediction task.

6.1 Acute toxicity prediction

Here, we present a comparative analysis of state-of-the-art results and our proposed MULTIMODAL-MOLFORMER approach for machine learning-based LD_{50} predictions using the LDToxDB dataset. Table 1 displays the benchmark results from the literature for state-of-the-art models, including details such as the employed machine learning methods, feature selection techniques (if applicable), input size, and performance metrics for the top-performing models. In addition to these benchmarks, we also report the results of our proposed method and the base MOLFORMER approach.

Table 1: Toxicity prediction considering the LDToxDB dataset. R^2, MAE, and $RMSE$ metrics are considered to evaluate the performance of the ML methods to predict the LD_{50} dosage, *acc* is used to evaluate the classification of EPA categories. For each model, the feature selection employed (if applicable) , as well as the number of input features is shown. Best results are highlighted.

Method	**Feature Selection**	**Input Size**	R^2	**MAE**	**RMSE**	***acc***
Multimodal-MoLFormer	**Causal Multi-stage Feature Selection**	797	0.641	**0.277**	0.526	**0.84**
MoLFormer (base model)	–	768	0.615	0.315	0.549	0.75
XGBoost	**Causal Multi-stage Feature Selection**	29	**0.692**	0.334	**0.484**	0.72
Random Forest	Pearson correlation coefficient [37]	106	0.660	0.337	0.496	0.70
XGBoost	Pearson correlation coefficient [37]	106	0.645	0.360	0.535	0.71
Random Forest	Boruta [38]	320	0.525	0.440	0.593	0.67
Random Forest	RFE [39]	10	0.233	0.590	0.791	0.68
DNN [35]	Pearson correlation coefficient	300	0.658	0.342	0.516	0.68
GP [35]	RF Gini	10 Mordred / 200 ECFP bits	0.627	0.376	0.538	0.65
Random Forest [35]	–	1826	0.647	0.372	0.523	0.66

The results presented in Table 1 demonstrate that the proposed multi-stage causal feature selection approach selected only 29 features that exhibit a causal relationship with the toxicity target, out of the 1826 features provided by the Mordred descriptors. The selected features are illustrated in Fig. 4, which demonstrates their causal importance in predicting the toxicity of the compounds. The XGBoost model equipped with the proposed feature selection method achieved better performance than the state-of-the-art algorithms proposed in [35]. This indicates the effectiveness of the proposed feature selection method in improving the performance of the model.

Incorporating the 29 physicochemical features selected during the causal multi-stage process in the MULTIMODAL-MOLFORMER model resulted in a significant improvement in the classification accuracy for EPA categories, from 0.75 to 0.84, when compared to the base MOLFORMER approach. While the MOLFORMER base model has the ability to learn from chemical language representations, including task-specific physicochemical features can enhance the model's understanding and performance for specific tasks, such as toxicity estimation.

It is worth noting that the $RMSE$ and R^2 metrics suggest that the MULTIMODAL-MOLFORMER model may miss outliers values, but the MAE metric indicates that the model performs well on average, leading to better classification accuracy. Although further investigation is required to address the potential limitation of the model in handling outliers, our results suggest that incorporating physicochemical features can improve the overall performance of the MULTIMODAL-MOLFORMER model for predicting molecular properties.

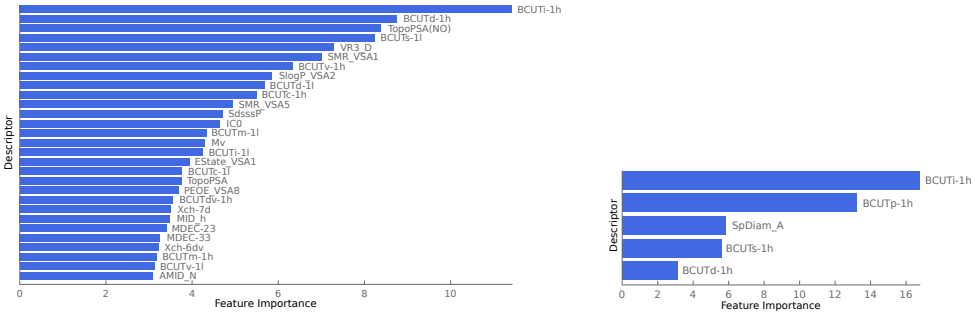

Figure 4: The figure illustrates the Markov Blanket feature importance histogram for the selected features in the case of compounds toxicity estimation (left) and biodegradability (right). These scores differ from classifier-based ratings as they derive from the causal connection between features and the target. They provide valuable pre-classification insights for experts, aiding in the creation of an optimal or near-optimal dataset tailored to specific tasks.

We conducted an experiment to evaluate the performance of our multi-modal approach across datasets of varying sizes. We kept the test dataset constant and generated training datasets with sample sizes of 128, 256, 512, 1024, 2048, 4096, and 8192. For each dataset size, we trained a MULTIMODAL-MOLFORMER model and calculated its accuracy. Figure 5 visually represents the accuracy evolution of the MULTIMODAL-MOLFORMER for toxicity estimation. Our analysis revealed that the proposed MULTIMODAL-MOLFORMER required 4096 samples to achieve an accuracy of approximately 75%, a performance comparable to the original MOLFORMER approach, which used a full training set of more than 10,000 samples. This observation underscores the positive impact of incorporating physicochemical features in enhancing the learning process of our approach, even when dealing with a smaller training dataset.

Overall, these results demonstrate the effectiveness of the proposed MULTIMODAL-MOLFORMER approach for predicting the toxicity of compounds, and the importance of incorporating physicochemical features in improving the model's performance.

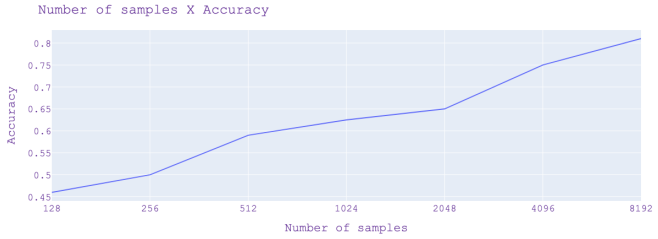

Figure 5: This figure illustrates the MULTIMODAL-MOLFORMER learning evolution with different training data samples sizes.

6.2 Biodegradability classification

The biodegradability classification task results are presented in Table 2. This task aims to classify compounds into one of two classes: "ready-biodegradability" (RB) or "not ready-biodegradability" (NRB). The RB class denotes compounds that are expected to be readily biodegradable in the environment, while the NRB class denotes compounds that are not. The table displays the performance metrics for the state-of-the-art machine learning models evaluated on the "All-Public set" dataset. Specifically, the Table 2 includes the name of each model, the feature selection method (if any), the input size, and the evaluation metrics.

The experimental results demonstrate the effectiveness of the proposed approach for biodegradability classification, surpassing the performance of previous benchmark models in terms of accuracy, precision, and sensitivity. This is a significant contribution to the development of sustainable and environmentally friendly materials, as the ability to accurately classify compounds based on

Table 2: Biodegradability classification considering the "All-Public set". *acc*, *Specificity*, and *Sensitivity* metrics are considered to evaluate the performance of the studied methods to evaluate the classification of biodegradability classes. Best results are highlighted.

Method	Feature Selection	Input Size	*acc*	*Specificity*	*Sensitivity*
Multimodal-MoLFormer	**Causal Multi-stage Feature Selection**	773	**0.94**	**0.94**	**0.94**
MoLFormer (base model)	–	768	0.84	0.81	0.87
XGBoost	**Causal Multi-stage Feature Selection**	**5**	0.86	0.85	0.86
*k*NN [36]	Genetic Algorithm	642	0.83	0.85	0.81
SVM [36]	Genetic Algorithm	50	0.84	0.82	0.86
Random Forest [36]	Genetic Algorithm	832	0.84	0.81	0.86
Gradient Boosting [36]	Genetic Algorithm	470	0.84	0.78	0.90

their biodegradability potential can help in the design and selection of materials with minimal environmental impact.

Interestingly, the proposed causal feature selection approach was able to achieve better results with just 5 features (Fig. 4-right), compared to the more than 50 features used by the state-of-the-art models to obtain a classification accuracy of 84%. This not only improves the interpretability of the models, but also provides experts with greater insights into the underlying causal mechanisms involved in biodegradability classification. Moreover, it enhances the performance of MULTIMODAL-MOLFORMER compared to the base MOLFORMER model, by aggregating features with relevant causal effects over the biodegradability classes.

These results highlight the importance of considering causal relationships between the features and the target variable in machine learning applications, particularly in domains where interpretability and explainability are critical. The proposed approach provides a promising direction for future research in the field of biodegradability classification and other related environmental applications.

In both experiments, we could observe that the proposed feature select approach helped to improve the metrics over these challenging tasks. We could note also, an improvement in terms of model's compactness, which is useful to experts to understand the process, and obtain deeper insights about such tasks. Moreover, the method proposed here provides feature importance about the causal effect of selected features over the target, which is helpful when building a dataset for a such specific task.

7 Conclusion

In this paper, we have introduced a novel multimodal approach for predicting molecular properties that leverages both chemical language representations and physicochemical features. Our proposed approach incorporates a unique causal multi-stage feature selection method that selects relevant physicochemical features based on their direct causal-effect on a specific target property to predict. Using Mordred descriptors as physicochemical features and identifying the Markov blanket of the target property for feature selection, our approach has demonstrated superior performance in predicting complex tasks such as general compounds biodegradability and PFAS toxicity estimation.

The effectiveness of our feature selection method in reducing the dimensionality of the Mordred feature space while maintaining or improving performance is noteworthy. Our proposed approach has outperformed existing state-of-the-art algorithms, including the chemical language-based only MOLFORMER and graph neural networks. Although further investigation is needed to address the potential limitation of the model in handling outliers, our results suggest that incorporating physicochemical features can improve the overall performance of the MULTIMODAL-MOLFORMER model for predicting molecular properties.

These findings offer a promising direction for enhancing the accuracy and effectiveness of molecular property prediction. By incorporating both chemical language and physicochemical features, our approach has the potential to make significant contributions to drug discovery, materials science, and other fields that depend on accurate prediction of molecular properties. Future research may build upon our approach to further improve molecular property prediction and advance the understanding and application of chemical language and physicochemical features.

References

[1] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," *Nature*, vol. 559, no. 7715, pp. 547–555, 2018.

[2] L. Pattanaik and C. W. Coley, "Molecular representation: going long on fingerprints," *Chem*, vol. 6, no. 6, pp. 1204–1207, 2020.

[3] J. Born and M. Manica, "Regression transformer enables concurrent sequence regression and generation for molecular language modelling," *Nature Machine Intelligence*, pp. 1–13, 2023.

[4] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, pp. 31–36, 1988.

[5] H. Moriwaki, Y.-S. Tian, N. Kawashita, and T. Takagi, "Mordred: a molecular descriptor calculator," *Journal of cheminformatics*, vol. 10, no. 1, pp. 1–14, 2018.

[6] W. X. Shen, X. Zeng, F. Zhu, Y. L. Wang, C. Qin, Y. Tan, Y. Y. Jiang, and Y. Z. Chen, "Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations," *Nature Machine Intelligence*, vol. 3, no. 4, pp. 334–343, 2021.

[7] P. Kirkpatrick and C. Ellis, "Chemical space," *Nature*, vol. 432, no. 7019, pp. 823–824, 2004.

[8] S. Wang, Y. Guo, Y. Wang, H. Sun, and J. Huang, "Smiles-bert: large scale unsupervised pre-training for molecular property prediction," in *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 2019, pp. 429–436.

[9] O. A. von Lilienfeld and K. Burke, "Retrospective on a decade of machine learning for chemical discovery," *Nature communications*, vol. 11, no. 1, p. 4895, 2020.

[10] X.-C. Zhang, C.-K. Wu, J.-C. Yi, X.-X. Zeng, C.-Q. Yang, A.-P. Lu, T.-J. Hou, and D.-S. Cao, "Pushing the boundaries of molecular property prediction for drug discovery with multitask learning bert enhanced by smiles enumeration," *Research*, vol. 2022, p. 0004, 2022.

[11] D. S. Wigh, J. M. Goodman, and A. A. Lapkin, "A review of molecular representation in the age of machine learning," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 12, no. 5, p. e1603, 2022.

[12] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction," *ACS central science*, vol. 5, no. 9, pp. 1572–1583, 2019.

[13] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das, "Large-scale chemical language representations capture molecular structure and properties," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1256–1264, 2022.

[14] Y. Liu, R. Zhang, T. Li, J. Jiang, J. Ma, and P. Wang, "Molrope-bert: An enhanced molecular representation with rotary position embedding for molecular property prediction," *Journal of Molecular Graphics and Modelling*, vol. 118, p. 108344, 2023.

[15] M. A. Skinnider, R. G. Stacey, D. S. Wishart, and L. J. Foster, "Chemical language models enable navigation in sparsely populated chemical space," *Nature Machine Intelligence*, vol. 3, no. 9, pp. 759–770, 2021.

[16] M. Moret, F. Grisoni, P. Katzberger, and G. Schneider, "Perplexity-based molecule ranking and bias estimation of chemical language models," *Journal of chemical information and modeling*, vol. 62, no. 5, pp. 1199–1206, 2022.

[17] A. C. Mater and M. L. Coote, "Deep learning in chemistry," *Journal of chemical information and modeling*, vol. 59, no. 6, pp. 2545–2559, 2019.

[18] J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, and A. Tkatchenko, "Combining machine learning and computational chemistry for predictive insights into chemical systems," *Chemical reviews*, vol. 121, no. 16, pp. 9816–9872, 2021.

[19] Q. Yang, Y. Liu, J. Cheng, Y. Li, S. Liu, Y. Duan, L. Zhang, and S. Luo, "An ensemble structure and physicochemical (spoc) descriptor for machine-learning prediction of chemical reaction and molecular properties," *ChemPhysChem*, vol. 23, no. 14, p. e202200255, 2022.

[20] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[21] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[22] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos Jr, C. Xiong, Z. Z. Sun, R. Socher *et al.*, "Large language models generate functional protein sequences across diverse families," *Nature Biotechnology*, pp. 1–8, 2023.

[23] N. Ferruz and B. Höcker, "Controllable protein design with language models," *Nature Machine Intelligence*, vol. 4, no. 6, pp. 521–532, 2022.

[24] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *arXiv preprint arXiv:2104.09864*, 2021.

[25] G. Landrum, "Rdkit documentation," *Release*, vol. 1, no. 1-79, p. 4, 2013.

[26] B. Hollas, "An analysis of the autocorrelation descriptor for molecules," *Journal of mathematical chemistry*, vol. 33, pp. 91–101, 2003.

[27] I. Guyon, C. Aliferis *et al.*, "Causal feature selection," in *Computational methods of feature selection*. Chapman and Hall/CRC, 2007, pp. 79–102.

[28] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab, Tech. Rep., 1996.

[29] A. Hassan, J. H. Paik, S. Khare, and S. A. Hassan, "Ppfs: Predictive permutation feature selection," *arXiv preprint arXiv:2110.10713*, 2021.

[30] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation." *Journal of Machine Learning Research*, vol. 11, no. 1, 2010.

[31] M. G. Evich, M. J. Davis, J. P. McCord, B. Acrey, J. A. Awkerman, D. R. Knappe, A. B. Lindstrom, T. F. Speth, C. Tebes-Stevens, M. J. Strynar *et al.*, "Per-and polyfluoroalkyl substances in the environment," *Science*, vol. 375, no. 6580, p. eabg9065, 2022.

[32] F. Cheng, Y. Ikenaga, Y. Zhou, Y. Yu, W. Li, J. Shen, Z. Du, L. Chen, C. Xu, G. Liu *et al.*, "In silico assessment of chemical biodegradability," *Journal of chemical information and modeling*, vol. 52, no. 3, pp. 655–669, 2012.

[33] S. E. Fenton, A. Ducatman, A. Boobis, J. C. DeWitt, C. Lau, C. Ng, J. S. Smith, and S. M. Roberts, "Per-and polyfluoroalkyl substance toxicity and human health review: Current state of knowledge and strategies for informing future research," *Environmental toxicology and chemistry*, vol. 40, no. 3, pp. 606–630, 2021.

[34] S. Su and P. M. Kang, "Systemic review of biodegradable nanomaterials in nanomedicine," *Nanomaterials*, vol. 10, no. 4, p. 656, 2020.

[35] J. Feinstein, G. Sivaraman, K. Picel, B. Peters, Á. Vázquez-Mayagoitia, A. Ramanathan, M. MacDonell, I. Foster, and E. Yan, "Uncertainty-informed deep transfer learning of perfluoroalkyl and polyfluoroalkyl substance toxicity," *Journal of chemical information and modeling*, vol. 61, no. 12, pp. 5793–5803, 2021.

[36] M. Lee and K. Min, "A comparative study of the performance for predicting biodegradability classification: the quantitative structure–activity relationship model vs the graph convolutional network," *ACS omega*, vol. 7, no. 4, pp. 3649–3655, 2022.

[37] A. E. Comesana, T. T. Huntington, C. D. Scown, K. E. Niemeyer, and V. H. Rapp, "A systematic method for selecting molecular descriptors as features when training models for predicting physiochemical properties," *Fuel*, vol. 321, p. 123836, 2022.

[38] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta–a system for feature selection," *Fundamenta Informaticae*, vol. 101, no. 4, pp. 271–285, 2010.

[39] X.-w. Chen and J. C. Jeong, "Enhanced recursive feature elimination," in *Sixth international conference on machine learning and applications (ICMLA 2007)*. IEEE, 2007, pp. 429–435.