Chemically Interpretable Molecular Representation for Property Prediction

M S B Roshan[+†*], Nirav Bhatt[+†*]
[+]BioSystems Engineering and Control Group, Department of Biotechnology, IIT Madras
[†]Robert Bosch Centre for Data Science and Artificial Intelligence (RBCDSAI), IIT Madras
[*]Centre for Integrative Biology and Systems medicinE (IBSE), IIT Madras

Abstract

Molecular property prediction using a molecule's structure is a crucial step in drug and novel material discovery, as computational screening approaches rely on predicted properties to refine the existing design of molecules. Although the problem has existed for decades, it has recently gained attention due to the advent of big data and deep learning. On average, one FDA drug is approved for 250 compounds entering the preclinical research stage, requiring screening of chemical libraries containing more than 20000 compounds. In-silico property prediction approaches using learnable representations increase the pace of development and reduce the cost of discovery. We propose developing molecule representations using functional groups in chemistry to address the problem of deciphering the relationship between a molecule's structure and property. Functional groups are substructures in a molecule with distinctive chemical properties that influence its chemical characteristics. These substructures are found by (i) curating functional groups annotated by chemists and (ii) mining a large corpus of molecules to extract frequent substructures using a pattern-mining algorithm. We show that the Functional Group Representation (FGR) framework beats state-of-the-art models on several benchmark datasets while ensuring explainability between the predicted property and molecular structure to experimentalists.

1 Introduction

Molecular property prediction is a task that finds applications in drug discovery, quantum mechanical attribute prediction of molecules, hydrophobicity prediction, material design and drug toxicity prediction. In the field of drug discovery and novel material discovery, computational approaches for predicting molecular properties can boost the processes of finding better drug candidates and materials [1, 2]. Characterising and predicting molecular properties is one of the most crucial problems in drug discovery. Numerous strategies are being used globally to enhance efficiency and improve the success of the drug discovery and development process. These strategies use a wide range of data such as genomics and proteomics, drug molecule structures and properties, and methods such as pharmaceutical modelling and artificial intelligence [3]. On average, one drug is approved by US FDA for five compounds entering clinical trials that, in turn, are the result of thorough preclinical testing of 250 compounds themselves selected by screening 5000–10000 compounds [4]. Experimentally testing many such compounds is both time and resource-consuming. In recent years, computational methods have significantly increased in the drug discovery domain [3]. The traditional computational approaches for in-silico molecular property prediction have relied on extracting fingerprints or hand-engineered features. Since these features are typically designed based on the property prediction task, it captures features only relevant to the particular task.

In contrast to traditional computational approaches, deep learning-based (DL) approaches can automatically learn features from molecules directly for the task at hand, and hence, it can reduce the time and cost for property prediction [5, 6]. Instantaneous molecular property prediction using deep learning algorithms can help generate novel molecules with desired profiles and engineer artificial synthesis pathways faster and cheaper. Graph neural networks (GNN) and their variants have been widely used for molecular property prediction tasks due to their ability to generate better molecular representations [7, 6, 8, 9, 10, 11, 12]. These approaches use the information on atoms, bonds, topology, interactions and molecular geometry (3D spatial structure) of molecules for learning molecular representation. However, GNN-based approaches require a large amount of labelled data for a particular task, and it is impossible to generate such a large number of labelled data for several applications. Several graph-based self-supervised learning approaches have been proposed to learn molecular representation from unlabelled molecular data to handle the problem of limited labelled data [9, 13, 14].

Although GNNs and self-supervised learning models have provided promising results on several property prediction tasks, the relationships between properties and molecule structures are challenging to interpret due to the complex molecular representations generated by these methods for chemists. For novel molecule discovery and drug repurposing applications, chemically interpretable molecular representation is essential for testing the generated molecules via wet-lab experiments by chemists. Hence, a chemistry-inspired representation of molecules can be vital in achieving interpretability and improved predictive performance of these models.

In this work, we propose a molecular representation learning framework that uses the concept of functional groups in chemistry. The functional groups are substructures in a molecule that are attributed to the chemical properties of the molecule, including its reactivity. This work proposes a functional group representation (FGR) framework that allows embedding molecules based on their substructures. Firstly, we introduce two approaches for the generation of the functional group vocabulary, namely, functional groups (FG) curated from the OCHEM database [15] and mined functional groups (MFG) from the PubChem database [16]. Then, we develop four different latent feature encodings using the FG- and MFG-based vocabulary generated in the first step for property prediction tasks. Further, we investigate the effect of pretraining using unlabelled molecules in the PubChem database on the property prediction tasks. We perform experiments on several benchmark datasets in the available literature and compare the results of the proposed FGR framework in this work with other state-of-the-art methods. We demonstrate that the FGR framework outperforms several property prediction tasks or provides comparable results on several other tasks compared to the state-of-the-art methods while providing interpretability to chemists and practitioners.

2 Objectives

O1 Generate a functional group vocabulary characterised by chemists and extract frequent sub-structures from a large chemical corpus.

O2 Learn functions $f_{\mathbf{x}_G} : \mathbf{x}_G \rightarrow \mathbf{z}_G$ using autoencoders [17] where $\mathbf{x}_G$ is a multi-hot vector of appropriate dimension (say p) depending on the input representation and $\mathbf{z}_G \in \mathbb{R}^l$ is the learnt latent vector.

O3 Decode the predicted property and molecular structure relationship using gradient-based model agnostic interpretability methods.

3 Methodology

In this work, a set of SMILES strings for n molecules, $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ which might be associated with a property y is considered. Furthermore, we also incorporate 2D global molecular descriptors to augment the learnt representation (FGR-Desc) and increase the performance of downstream property prediction tasks. The methods are summarised in Figure 1.

- **Generation of Functional Group vocabulary:** In this study, we use the OCHEM [15] database, which has a collection of 2786 functional groups (FG) characterised by chemists and frequent sub-structures are recognised using a sequential pattern mining algorithm applied on $\mathcal{S}$ from the PubChem database ($n > 114$ million). Based on the frequency threshold η, 3000 mined functional groups are identified (MFG). Then, any molecule $S_i \in \mathcal{S}$ can be represented by a multi-one-hot encoded vector, $[x_1, x_2, \ldots, x_b]^T$ where $x_i = 1$ if $FGR_i \in S_i$ and $x_i = 0$, if $FGR_i \notin S_i$.

- **Pretraining and Property Prediction:** Pretraining is decoupled from the downstream property prediction to develop a global representation capable of interpreting the chemical space that can be applied to any task. For the pretraining step, the autoencoder is trained separately from the downstream property prediction task. The reconstruction loss of the training phase in is minimized for all the molecules in the database for the pretraining purpose. One of the preliminary challenges of the encoder-decoder pretraining is the determination of the dimension of the latent feature vector. Hyper-parameter optimization is performed to obtain the dimension of the latent feature vectors for all four types of encodings. A fully connected neural network is used to compute a probability score $p(\mathbf{x}_G) \in [0, 1]$ based on $\mathbf{z}_G$ (latent feature vector) for property prediction.

- **Interpretability:** We evaluate each input feature's contribution to the model's output using primary attribution methods like feature permutation, integrated gradients and gradient SHAP [18, 19]. The goodness of

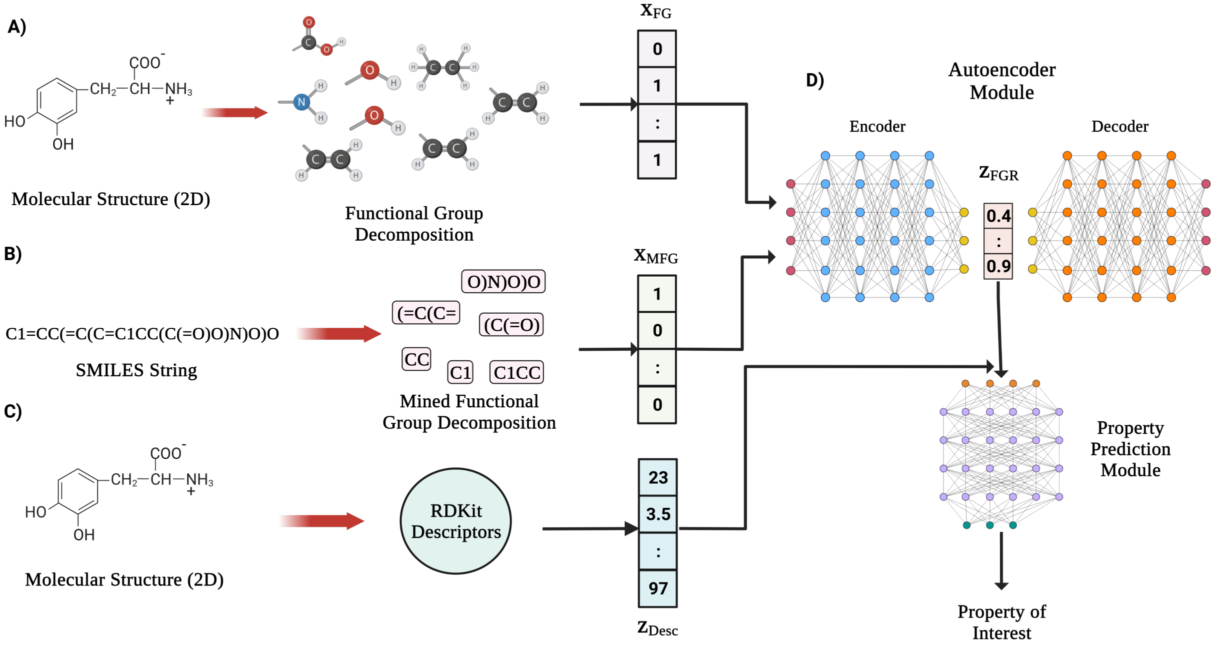

Figure 1: Overview of the Proposed Methodology: A) FG Representation, **B)** MFG Representation, **C)** Descriptor Representation, **D)** Latent Representation for FGR and Property Prediction Module

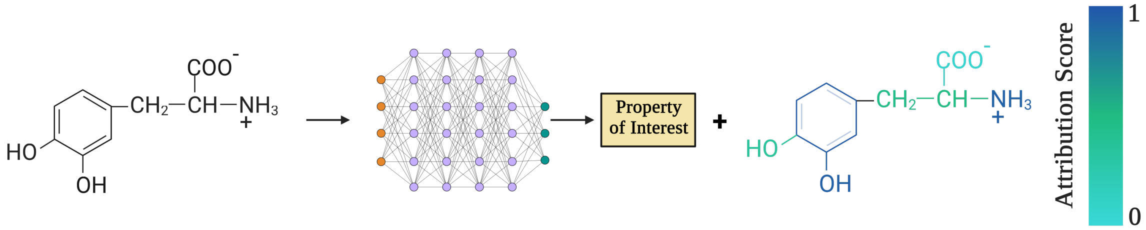

Figure 2: Overview of Interpretability Analysis: For any given property, attribution scores for input features are calculated and the substructures can be visualised overlapped with the scores

explanations is quantified using infidelity and sensitivity metrics. A visualisation tool is also developed to highlight essential substructures that contribute to predicting desired properties, as shown in Figure 2.

4 Results

Extensive evaluation of the model was done for robustness and generalizability on classification and regression tasks using five-fold random and scaffold splits. The results are summarized in Table 1 and Table 2.

5 Conclusion

This work presents a functional group representation (FGR) framework using functional groups in chemistry for molecular representation learning. The framework allows four types of molecular representations: FG, MFG, FG-MFG-based and FG-MFG-descriptors-based representation. The proposed FGR framework-based molecular embeddings have been evaluated on several benchmark datasets. The proposed framework performs at par and sometimes better than the state-of-the-art algorithms in classification tasks. The FGR framework also provides chemically interpretable encoding as it is inspired by rules of chemistry to maintain explainability with the encoding. In the proposed framework, autoencoders are used to learn latent representations. Also, we demonstrated that the pretraining in the FGR framework could be performed due to decoupling between the latent representation learning

Scaffold Split Classification (ROC-AUC) ↑			
Dataset	**FGR**	**DMPNN**	**GEM**
BACE	**0.89 ± 0.01**	0.86 ± 0.05	0.86 ± 0.01
BBBP	**0.96 ± 0.008**	0.92 ± 0.02	0.72 ± 0.00
Tox21	0.71 ± 0.01	0.69 ± 0.01	**0.78 ± 0.001**
ClinTox	**0.99 ± 0.002**	0.88 ± 0.03	0.90 ± 0.01
SIDER	**0.72 ± 0.07**	0.63 ± 0.03	0.67 ± 0.004

Table 1: Comparison of ROC-AUC scores for FGR, DMPNN [6], and GEM [8]

Scaffold Split Regression (RMSE) ↓			
Dataset	**FGR**	**DMPNN**	**GEM**
ESOL	**0.62 ± 0.06**	1.05 ± 0.008	0.79 ± 0.02
FreeSolv	**0.78 ± 0.19**	2.08 ± 0.082	1.87 ± 0.094
Lipo	**0.64 ± 0.035**	0.68 ± 0.016	0.66 ± 0.008

Table 2: Comparison of RMSE scores for FGR, DMPNN [6], and GEM [8]

task and the property prediction task. It is envisaged to extend the FGR framework for building pre-trained models with explainability using self-supervised learning on large-scale molecular data.

References

[1] W Patrick Walters and Regina Barzilay. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research*, 54(2):263–270, 2020.

[2] Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1–12, 2020.

[3] Geoffrey Kabue Kiriiri, Peter Mbugua Njogu, and Alex Njoroge Mwangi. Exploring different approaches to improve the success of drug discovery and development projects: a review. *Future Journal of Pharmaceutical Sciences*, 6(1):1–12, 2020.

[4] Jie Shen and Christos A Nicolaou. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies*, 32:29–36, 2019.

[5] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical science*, 9(24):5441–5451, 2018.

[6] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 8 2019.

[7] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[8] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

[9] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

[10] Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *International Conference on Learning Representations*, 2019.

[11] Chengqiang Lu, Qi Liu, Chao Wang, Zhenya Huang, Peize Lin, and Lixin He. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1052–1060, 2019.

[12] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

[13] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

[14] Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34:15870–15882, 2021.

[15] Iurii Sushko, Sergii Novotarskyi, Robert Körner, Anil Kumar Pandey, Matthias Rupp, Wolfram Teetz, Stefan Brandmaier, Ahmed Abdelaziz, Volodymyr V Prokopenko, Vsevolod Y Tanchuk, et al. Online chemical modeling environment (ochem): web platform for data storage, model development and publishing of chemical information. *Journal of computer-aided molecular design*, 25:533–554, 2011.

[16] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

[17] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[18] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[19] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.