# Enabling Accurate and Interpretable Property Prediction with TDiMS in Large Molecules

**Lisa Hamada**[1,*], **Akihiro Kishimoto**[1], **Kohei Miyaguchi**[1], **Masataka Hirose**[2], **Junta Fuchiwaki**[2], **Indra Priyadarsini**[1], **Seiji Takeda**[1], **Sina Klampt**[1], **Takao Moriyama**[1]

[1]IBM Research – Tokyo, Japan [2]JSR Corporation, Japan, Lisa.Hamada@ibm.com

## Abstract

In materials discovery, descriptors that are both accurate and interpretable are essential for predicting molecular properties. However, existing descriptors, including neural network-based approaches, often struggle to capture long-range interactions between substructures. We analyze the previously proposed descriptor TDiMS, which models nonlocal structural relationships via average topological distances between substructure-pairs. While TDiMS has shown strong performance, its size dependence had not been systematically assessed. Our analysis reveals that TDiMS is particularly effective for larger molecules, where long-range interactions are critical and conventional descriptors underperform. SHAP-based analysis highlights that its predictive power derives from distant substructure-pair features. In addition to improved accuracy, TDiMS offers interpretable features that provide chemical insight, potentially accelerating molecular design and discovery.

## 1 Introduction

Machine learning (ML) has accelerated materials discovery by enabling efficient prediction of molecular properties [24]. A critical component of this process is the representation of molecules through informative descriptors that encode their structural and chemical characteristics. Conventional Quantitative Structure–Property Relationship (QSPR)-based descriptors [4, 3, 2, 18] enumerate substructures or physicochemical properties, while neural network-based approaches [6, 20, 22, 1] learn data-driven representations from molecular graphs or SMILES. Despite their success, both types of descriptors often struggle to capture nonlocal relationships among intra-molecular substructures. Moreover, the interpretability of these models remains limited, posing challenges in applications that demand chemical insight and design rationale.

We previously proposed the Topological Distance of intra-Molecular Substructures (TDiMS) descriptor [8], which captures long-range topological relationships between substructure-pairs while maintaining interpretability, even with a potentially vast number of features. It also supports flexible substructure definitions, including data-driven and domain-specific motifs, making it adaptable to a wide range of prediction tasks. Prior studies showed that TDiMS outperforms conventional and neural descriptors on datasets such as Chromophore [8] and MolNet [9], particularly where long-range interactions are critical.

In this study, we further investigate the effectiveness of TDiMS with a particular focus on molecular size, examining the conditions under which it offers strong advantages over existing descriptors. We find that TDiMS demonstrates particularly strong performance for larger molecules, especially those with heavy atom counts of 25 or more, where long-distance interactions are more likely to influence target properties. Furthermore, we show that incorporating custom-defined substructures tailored to the dataset leads to additional improvements in predictive performance. This study also provides an important direction for descriptor development, including neural-network-based models, by showing that capturing long-range substructure distances can further improve predictive performance.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Materials Science (AI4Mat) Workshop at NeurIPS 2025.

## 2 Related Works

Mordred [18] is an advanced descriptor open-source tool that calculates over 1800 two- and three-dimensional descriptors by counting substructures based on physical chemistry knowledge. However, this approach lacks global molecule information, such as intra-molecular positional relationships, which can critically affect molecular properties. To address this, the Atom-Pair descriptor [3] encodes global information by capturing atomic environments and shortest path distances between all atom pairs. Still, relying solely on atoms presents limitations. MAP4 [2] extends Atom-Pair by replacing atomic features with circular substructures around each atom and encoding their distances. To handle the combinatorial explosion of substructure-pairs, MAP4 uses MinHash values from Locality Sensitive Hashing (LSH) for efficient representation. While this enables fast similarity search in large databases, it comes at the cost of reduced interpretability.

Latent vectors from neural-network models, including Transformer-based chemical language models (CLMs) and graph neural networks (GNNs), are increasingly used as molecular descriptors [1, 15]. These models are pretrained on large datasets such as PubChem [13] and ZINC [10]. For example, MolCLR [22] is a contrastive-learning-based GNN that learns molecular graph embeddings via self-supervised pretraining. MolFormer [20], on the other hand, is a Transformer-based CLM that directly encodes SMILES strings to generate context-aware molecular representations.

GNNs capture atomic and bond-level information but are limited by the number of message-passing steps, restricting the range of bond-path distances [14]. CLMs can, in principle, model long-range relationships via attention, but are constrained by input representations like SMILES, which the sequence of characters does not necessarily reflect the actual spatial arrangement of atoms in the molecular structure [25, 11]. Moreover, the feature vectors derived from neural-network models, commonly referred to as latent vectors, often lack interpretability, as individual features typically have no clear chemical meaning.

## 3 Method

### 3.1 TDiMS algorithm

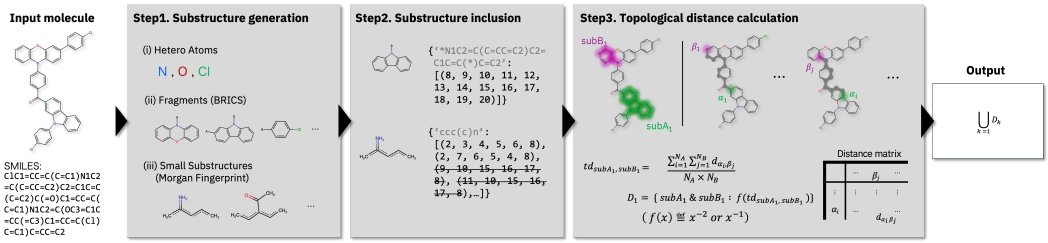


Figure 1: Workflow of TDiMS for a target molecule in dataset.

Figure 1 outlines the TDiMS workflow. Canonical SMILES are used as input, and all substructure-pairs within a molecule are exhaustively enumerated. We consider three types of substructures: (i) Hetero atoms, (ii) fragments, and (iii) circular substructures from Morgan fingerprints [17]. When substructures are extracted using Morgan fingerprints or fragments, smaller substructures are often entirely encompassed by larger ones. To avoid redundant representation of the same structural effect, TDiMS excludes smaller substructures from a pair if they are fully contained within larger substructures (Step 2 in Fig. 1). The topological distance (TD) between two substructures is computed as the mean shortest bond distance between their heavy atoms using the Floyd-Warshall algorithm [7, 12, 23]:

$$td_{subA,subB} = \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N_B} bd_{\alpha_i\beta_j}}{N_A \times N_B},$$

where $td_{subA,subB}$ denotes the average shortest bond distance between all pairs of heavy atoms in substructures $subA$ and $subB$, with $N_A$ and $N_B$ being the number of heavy atoms, and $bd_{\alpha_i\beta_j}$ the shortest bond distance between atom $\alpha_i$ in $subA$ and $\beta_j$ in $subB$. This formulation captures the spatial spread between substructures, making it robust to variations in their internal geometry. It also allows flexible targeting of arbitrary substructure definitions. The feature values are computed as the inverse or inverse square of the TD, emphasizing short-range effects and reflecting physical principles such as Coulomb's law (Step3 in Fig. 1). To handle identical substructure-pairs appearing

at multiple locations within a molecule, TDiMS applies an aggregation function to their TDs. This function, such as sum, max, or min, is selected based on the property of interest and how repeated occurrences should be weighted. This reduces redundancy while preserving meaningful recurring interactions. Then TDiMS calculates the set of the feature values for all substructure-pairs $\bigcup_{k=1} D_k$. TDiMS computes feature values for all observed substructure-pairs $\bigcup_{k=1} D_k$, and constructs a unified feature vector for each molecule based on the full set across the dataset. Missing pairs are filled with zero. Feature vectors are normalized across all molecules. We tested all combinations of substructure types (Hetero atoms, fragments, circular substructures), feature functions (raw, inverse, inverse square), and aggregation methods for duplicate substructure-pairs (min, max, sum), and selected the best-performing configuration for each task.

### 3.2 Evaluation Tasks

To investigate the molecular size range where TDiMS is most effective, we conducted a dipole moment prediction task using subsets of the PubChemQC dataset [19], grouped by heavy atom count (HAC). Each subset consists of 1,000 randomly selected molecules per HAC group, real-world data scarcity commonly observed in materials discovery datasets. Distributions of HAC and dipole moments are shown in Figure 2 (a) and (b).

As described in the Method section, TDiMS features were generated under various configurations, and the best-performing setup was selected based on initial validation. Hetero atoms were always included as a target substructure. Since no established fragment database exists for dipole moment prediction, we used the MacFrag method [5] to extract candidate substructures from the PubChemQC dataset. We compared TDiMS against five representative baseline descriptors covering a diverse range of molecular representation strategies: one knowledge-driven descriptor (Mordred), two neural network-based descriptors (MolFormer and MolCLR), and two enumerative descriptors based on intramolecular TDs (Atom-Pair and MAP4).

For prediction, we adopted the elastic net (including Lasso and Ridge) and random forest, following prior studies [21]. Hyperparameters were tuned via grid search with 3-fold cross-validation and 10 repeats using the RepeatedKFold class. To interpret the contribution of each substructure-pair feature, we computed feature importance scores using SHAP [16]. This allowed us to quantify the impact of individual features on the model's predictions.

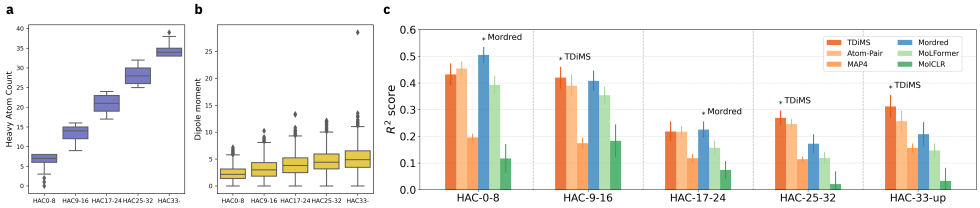


Figure 2: **Data characteristics and performance comparison.** (a) Heavy atom count. (b) Dipole moment. (c) $R^2$ scores for each descriptor. Each subset is labeled as HAC $X$–$Y$, indicating the HAC range. "HAC 33–" denotes molecules with 33 or more heavy atoms.

Table 1: **$R^2$ scores of TDiMS under different configurations.** $r$ indicates the radius of circular substructures; w/ and w/o refer to the inclusion or exclusion of custom fragments. Bold values indicate the best-performing configurations.

| Method | $HAC \leq 8$ | $9 \leq HAC \leq 16$ | $17 \leq HAC \leq 24$ | $25 \leq HAC \leq 32$ | $33 \leq HAC$ |
|---|---|---|---|---|---|
| $r = 1$, w/o | $\mathbf{0.432 \pm 0.041}$ | $\mathbf{0.420 \pm 0.041}$ | $0.200 \pm 0.029$ | $0.242 \pm 0.028$ | $0.293 \pm 0.035$ |
| $r = 1$, w/ | $0.423 \pm 0.040$ | $0.417 \pm 0.028$ | $0.200 \pm 0.025$ | $0.244 \pm 0.024$ | $0.299 \pm 0.044$ |
| $r = 2$, w/o | $0.310 \pm 0.051$ | $0.307 \pm 0.037$ | $0.209 \pm 0.029$ | $0.255 \pm 0.020$ | $0.310 \pm 0.049$ |
| $r = 2$, w/ | $0.298 \pm 0.053$ | $0.392 \pm 0.043$ | $\mathbf{0.218 \pm 0.038}$ | $\mathbf{0.269 \pm 0.028}$ | $\mathbf{0.312 \pm 0.043}$ |

## 4 Results and Discussion

Figure 2 (c) compares TDiMS with other descriptors. TDiMS outperformed others for molecules with HAC $\geq$ 25, while Mordred showed stronger performance for smaller molecules, particularly in the 0–8 HAC range. In intermediate HAC groups (9–24), the two were competitive. Table 1 summarizes

$R^2$ scores across TDiMS configurations. Radius 1 was effective for small molecules, while radius 2 yielded better results for larger ones, validating the effectiveness of our configuration selection. Custom fragments had limited impact on small molecules but consistently improved performance for those with HAC $\geq 17$, highlighting the value of task-specific substructure extraction in TDiMS.

Interestingly, TDiMS outperformed both Atom-Pair and MAP4, which also target TDs between substructures. As shown in Table 1, it achieved higher prediction scores than MAP4 even without custom fragments, namely when targeting equivalent substructures. Figure 3 shows that for larger molecules (HAC 17+), features involving substructures with more than two heavy atoms contributed more significantly, explaining TDiMS's growing advantage over Atom-Pair. While Atom-Pair is limited to small substructures and MAP4 sacrifices interpretability through MinHash compression, TDiMS retains interpretability by removing redundancy and grouping similar substructure-pairs. These results highlight the advantage of explicitly modeling distances between meaningful substructures, rather than compressing or simplifying them.

TDiMS consistently outperformed neural-network models, likely due to the limitations of GNNs and CLMs discussed in Related Works. As shown in Figure 3 (b), an increase in heavy atom count (HAC) leads to a greater contribution from features with bond-path distances of five or more. In the HAC 33+ dataset, where TDiMS achieved the largest performance gains, these long-range features accounted for a substantial portion of the model's output. When they were removed, over 36% of TDiMS's predictive performance was lost. These findings highlight the importance of modeling long-range structural interactions, which are difficult to capture with conventional GNNs constrained by message-passing depth [14].

Next, we further analyze the substructure-pairs that contributed most to the prediction of dipole moment. Figure 3 (c) shows the top 20 substructure-pairs with the highest total SHAP values across all HAC groups. While detailed interpretation is beyond this study's scope, further analysis may yield deeper insights into their chemical relevance. Such interpretability is made possible by TDiMS's explicit and structured design.

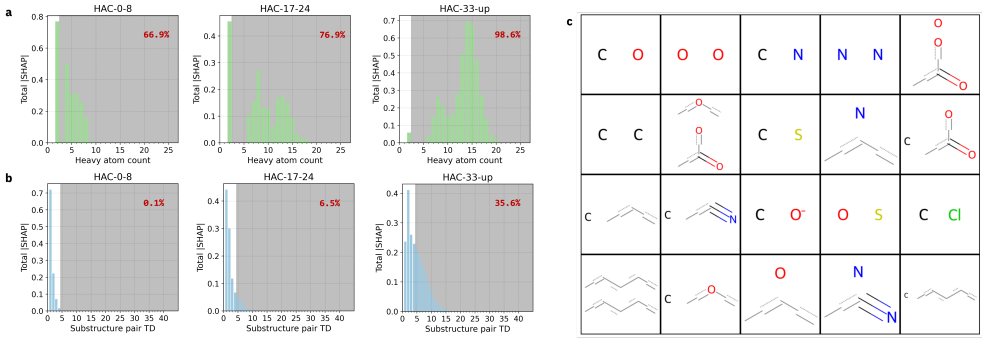


Figure 3: **SHAP analysis of TDiMS features.** SHAP values by (a) HAC and (b) TD of substructures(red numbers: % in shaded region). (c) Top 20 substructure-pairs.

## 5 Conclusion and Future Work

In this study, we demonstrated that the TD of intra-Molecular Substructures (TDiMS) descriptor provides a significant advantage in predicting molecular properties, as demonstrated on a dipole moment prediction task using the PubChemQC dataset. TDiMS consistently outperformed competitive descriptors, particularly for larger molecules with more than 24 heavy atoms, where modeling long-range substructure interactions becomes essential. From the perspective of AI-guided materials design, TDiMS offers a compelling alternative to end-to-end deep learning models. Its ability to capture long-range intra-molecular interactions while preserving interpretability provides a valuable tool for the design and screening of molecules with targeted properties. Additionally, the flexible design of TDiMS allows it to incorporate task-specific substructures, further enhancing its adaptability to different material property prediction tasks. Future work will focus on expanding TDiMS to support multi-property prediction and integration with generative models. Another promising direction is to integrate TDiMS with graph transformers or virtual-node GNNs to enhance their ability to represent long-range substructure interactions while preserving interpretability. We are also exploring its application to broader materials datasets beyond organic molecules, such as polymers or inorganic compounds.

## References


[1] Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.

[2] Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(43), June 2020. URL https://doi.org/10.1186/s13321-020-00445-4.

[3] Raymond E. Carhart and Dennis H. Smith R. Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, 1985.

[4] Rogers D and Hahn M. Extended-connectivity fingerprints. *J Chem Inf*, 50(5):742–754, April 2010. URL https://doi.org/10.1021/ci100050t.

[5] Yanyan Diao, Feng Hu, Zihao Shen, and Honglin Li. MacFrag: Segmenting large-scale molecules to obtain diverse fragments with high qualities. *Bioinformatics*, 39(1), 2023.

[6] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R.l Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems*, pages 2224–2232, 2015.

[7] R.W. Floyd. Algorithm 97: Shortest path. *Communications of the ACM*, 5(6):345, June 1962. doi: 10.1145/367766.368168. URL https://doi.org/10.1145/367766.368168.

[8] Lisa Hamada, Akihiro Kishimoto, Kohei Miyaguchi, Masataka Hirose, Junta Fuchiwaki, Indra Priyadarsini, and Seiji Takeda. Revisiting molecular descriptors with TDiMS for interpretable intramolecular interactions based on substructure pairs, 2025. URL https://doi.org/10.21203/rs.3.rs-6141459/v1. Preprint, Research Square.

[9] Lisa Hamada, Indra Priyadarsini, Seiji Takeda, and Onur Boyar. Tdims: A topological distance based intra-molecular substructure descriptor for improved machine learning predictions. In *Proceedings of the 4th Annual AAAI Workshop on AI to Accelerate Science and Engineering (AI2ASE)*, Philadelphia, PA, USA, March 2025. URL https://ai-2-ase.github.io/papers/21_241220_TDiMS_AI2ASE_Camera-ready.pdf.

[10] John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, , and Ryan G. Coleman. ZINC: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, May 2012. URL https://doi.org/10.1021/ci3001277.

[11] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

[12] D.B. Johnson. Efficient algorithms for shortest paths in sparse networks. *Journal of the ACM*, 24(1):1–13, January 1977. doi: 10.1145/321992.321993. URL https://doi.org/10.1145/321992.321993.

[13] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2023 update. *Nucleic acids research*, 51(D1):D1373–D1380, January 2023. URL https://doi.org/10.1093/nar/gkac956.

[14] Li Q. Li, Z. Han, and X. Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 3538–3545, 2018.

[15] Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *Advances in neural information processing systems*, 32, 2019.

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

[17] H. L. Morgan. The generation of a unique machine description for chemical structures – a technique developed at chemical abstracts service. *J. Chem. Doc.*, 5(2):107–113, May 1965. URL https://doi.org/10.1021/c160017a018.

[18] H. Moriwaki, Y. S. Tian, N. Kawashita, and T. Takagi. Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, 10(4), February 2018. URL https://doi.org/10.1186/s13321-018-0258-y.

[19] Mutsumi Nakata and Toshihiko Shimazaki. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *Journal of Chemical Information and Modeling*, 57(6):1300–1308, 2017. doi: 10.1021/acs.jcim.7b00083.

[20] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4:1256–1264, December 2022. URL https://doi.org/10.1038/s42256-022-00580-7.

[21] S. Takeda, T. Hama, H.-H. Hsu, V. A. Piunova, D. Zubarev, D. P. Sanders, J. W. Pitera, M. Kogoh, T. Hongo, Y. Cheng, W. Bocanett, H. Nakashika, A. Fujita, Y. Tsuchiya, K. Hino, K. Yano, S. Hirose, H. Toda, Y. Orii, and D. Nakano:. Molecular inverse-design platform for material industries. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2961–2969, 2020.

[22] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4:279–287, March 2022. URL https://doi.org/10.1038/s42256-022-00447-x.

[23] S. Warshall. A theorem on boolean matrices. *Journal of the ACM*, 9(1):11–12, January 1962. doi: 10.1145/321105.321107. URL https://doi.org/10.1145/321105.321107.

[24] Jing Wei, Xuan Chu, Xiang-Yu Sun, Kun Xu, Hui-Xiong Deng, Jigen Chen, Zhongming Wei, and Ming Lei. Machine learning in materials science. *InfoMat*, 1(3):338–358, September 2019. URL https://doi.org/10.1002/inf2.12028.

[25] Jun Xia, Yanqiao Zhu, Yuanqi Du, and Stan Z. Li. A systematic survey of chemical pre-trained models. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, pages 6787–6795, 2023.