Capturing Long-Range Intramolecular Interactions with TDiMS for Interpretable Property Prediction

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

Long-range intra-molecular interactions are not well represented by existing molecular descriptors, which limits the accuracy of machine learning models for molecular property prediction. We introduce TDiMS, a descriptor that encodes topological distances between substructure pairs, enabling explicit handling of long-range distances while retaining chemical meaning. Applied to molecular datasets, TDiMS shows particular advantages for larger molecules, where long-range interactions strongly influence target properties. We further demonstrate that choosing appropriate substructure definitions, such as tailored fragments, enhances predictive performance. Beyond accuracy, TDiMS provides interpretable features essential for material discovery, offering insights into structural motifs driving predictions. These results highlight distance-based, interpretable descriptors as a promising route for machine learning in the materials discovery.

1 Introduction

Machine learning (ML) is increasingly applied across the physical sciences, enabling efficient prediction, simulation, and design. Among these domains, materials discovery has emerged as a key area, where accurate and interpretable molecular descriptors are essential for property prediction and molecular design [1]. Conventional Quantitative Structure–Property Relationship (QSPR) descriptors [2–5] enumerate substructures or physicochemical properties, while neural-network–based approaches [6–9] learn data-driven embeddings from molecular graphs or Simplified Molecular Input Line Entry System (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 introduce the Topological Distance of intra-Molecular Substructures (TDiMS), a descriptor that captures long-range topological relationships between substructure-pairs while maintaining interpretability, even with a potentially vast number of features. Its design also allows flexible substructure definitions, ranging from generic fragments to task-specific motifs, making it adaptable to diverse prediction tasks. Prior work showed that TDiMS outperforms conventional and neural-network-based descriptors on datasets such as Chromophore [10] and MoleculeNet [11], particularly where long-range interactions are critical.

In this study, we analyze TDiMS with a particular focus on the molecular size range where it shows clear advantages over existing descriptors We find that TDiMS demonstrates particularly strong performance for larger molecules, where long-distance interactions are more likely to influence target properties. We show that tailored substructures enhance predictive performance compared with generic schemes. Attribution analysis further identifies the substructure pairs that dominate predictions in each dataset, highlighting which features are relevant to the target property. This study 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: Machine Learning and the Physical Sciences (ML4PS) Workshop at NeurIPS 2025.

2 Related Works

Mordred [5] 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 [4] 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 [9, 12]. These models are pretrained on large datasets such as PubChem [13] and ZINC [14]. For example, MolCLR [8] is a contrastive-learning-based GNN that learns molecular graph embeddings via self-supervised pretraining. MolFormer [7], 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 [15]. 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 [16, 17]. 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.

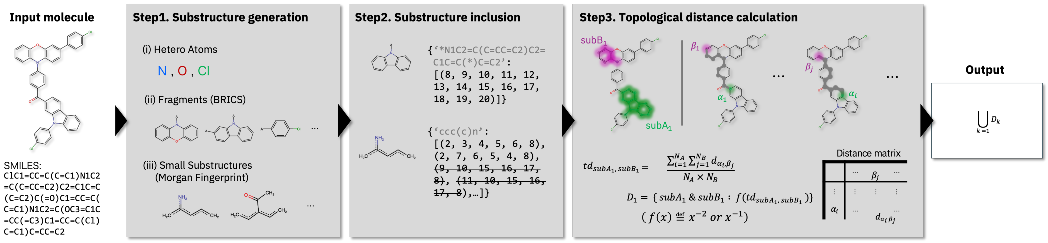
Input molecule
SMILES:
ClC1=CC=C(C=C1)N1C2=C(C=CC=C2)C2=C1C=C(C=C2)C(=O)C1=CC=C(C=C1)N1C2=C(OC3=C1C=CC(=C3)C1=CC=C(Cl)C=C1)C=CC=C2
Step1. Substructure generation
(i) Hetero Atoms
N , O , Cl
(ii) Fragments (BRICS)
(iii) Small Substructures (Morgan Fingerprint)
Step2. Substructure inclusion
{'*N1C2=C(C=CC=C2)C2=C1C=C(*)C=C2': [(8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20)]}
{'ccc(c)n': [(2, 3, 4, 5, 6, 8), (2, 7, 6, 5, 4, 8), ~~(9, 10, 15, 16, 17, 8)~~, ~~(11, 10, 15, 16, 17, 8)~~,…]}
Step3. Topological distance calculation
subB$_1$
subA$_1$
β_1
α_1
β_j
α_i
$td_{subA_1,subB_1} = \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N_B} d_{\alpha_i,\beta_j}}{N_A \times N_B}$
$D_1 = \{ subA_1 \,\&\, subB_1 : f(td_{subA_1,subB_1}) \}$
$(f(x) \overset{\text{def}}{=} x^{-2} \text{ or } x^{-1})$
Distance matrix
β_j
α_i
$d_{\alpha_i\beta_j}$
Output
$\bigcup_{k=1} D_k$

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

3 Method

3.1 TDiMS algorithm

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) heteroatoms, (ii) fragments, and (iii) circular substructures from Morgan fingerprints [18]. To avoid double-counting structural effects, smaller substructures fully contained within larger ones are excluded (Step 2). The topological distance (TD) between two substructures is computed as the mean shortest bond distance between their heavy atoms using the Floyd–Warshall algorithm [19–21], which is executed once per molecule to obtain the all-pairs shortest path matrix that can be reused across all substructure pairs:

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

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 α_i in $subA$ and β_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, a choice that is physically motivated by distance-dependent interactions such as Coulomb's law (Step 3).

Because exhaustive enumeration of substructure pairs can generate an extremely large number of features, especially for larger molecules, TDiMS introduces additional mechanisms to control feature dimensionality beyond the avoidance of double-counting described in Step 2. To handle identical substructure-pairs appearing at multiple locations within a molecule, an aggregation function, such as *sum*, *max*, or *min*, is applied to combine their TDs. This reduces redundancy while preserving meaningful recurring interactions that are chemically relevant. Furthermore, to further limit the dimensionality of the resulting feature set, feature selection is performed using the `SelectFromModel` class in Scikit-learn [22], which automatically prunes low-importance features based on model weights. These three procedures, including the structural pruning in Step 2, collectively maintain the interpretability of TDiMS while ensuring computational feasibility.

The resulting features across all observed pairs are combined into a unified feature vector for each molecule, with missing entries filled by zero and values normalized across the dataset. We tested combinations of substructure types, feature functions, and aggregation methods, selecting 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 [23], grouped by heavy atom count (HAC). Each subset consists of 1,000 randomly selected molecules per HAC group, reflecting the data scarcity commonly observed in materials datasets. Distributions of HAC and dipole moments are shown in Fig. 2 (a) and (b).

Heteroatoms and circular substructures (radius 1 or 2) were always included as target substructures. Since no established fragment database exists for dipole moment prediction, we primarily used the MacFrag method [24] to extract custom fragments from the PubChemQC dataset (Fig. 2 (c)). Custom fragments can enhance predictive performance when chemically relevant motifs exist, while generic definitions such as BRICS [25] or circular substructures remain applicable for general-purpose modeling. This flexibility enables users to balance interpretability, generality, and domain specificity depending on the application. For the largest molecules (HAC $\geq$ 33), we also tested fragment definitions derived from BRICS and from the Harvard Clean Energy Project (CEP) [26] to examine their impact on predictive performance.

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 intra-molecular TDs (Atom-Pair and MAP4). For prediction, we adopted the elastic net (including Lasso and Ridge) and random forest, following prior studies [27]. 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 Shapley additive explanations (SHAP) [28], allowing us to identify which substructure pair features dominate predictions. In each trained prediction model, SHAP additively partitions the prediction value into *SHAP values*, which represent the contributions of individual features.

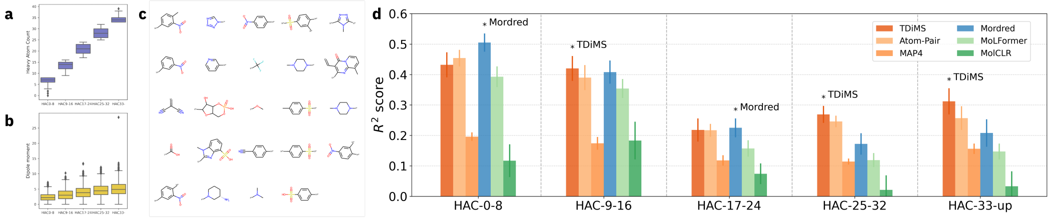

Figure 2: Data characteristics and performance comparison. (a) Heavy atom count. (b) Dipole moment. (c) Custom fragment. (d) R^2 scores for each descriptor. Each subset is labeled as HAC X–Y, indicating the HAC range. "HAC 33–" denotes molecules with 33 or more heavy atoms.

4 Results and Discussion

Figure 2 (d) compares TDiMS with other descriptors across HAC groups. TDiMS outperformed others for molecules with HAC ≥ 25, while Mordred showed stronger performance for smaller molecules, particularly in the 0–8 HAC range. In intermediate ranges (9–24), the two were competitive. These results confirm that long-range intra-molecular interactions become increasingly important as molecular size grows. Table 1 summarizes results for HAC 33+ molecules using different fragment definitions. Custom fragments provided the highest accuracy, surpassing CEP fragments and BRICS-based fragmentation. This demonstrates that tailoring fragment definitions to the dataset and/or target property can enhance the representation power of TDiMS, as dataset-specific motifs are better suited to capture relevant interactions.

Table 1: Comparison of fragment definitions for HAC 33+ molecules in R^2 scores. Best in bold.

	w/o Fragment	w/ Custom fragment	w/ CEP	w/ BRICS
Circular substructure (r=1)	0.293 ± 0.035	0.299 ± 0.044	0.297 ± 0.033	0.293 ± 0.052
Circular substructure (r=2)	0.310 ± 0.049	**0.312** ± **0.043**	0.308 ± 0.048	0.279 ± 0.037

Interestingly, TDiMS outperformed both Atom-Pair and MAP4, even without custom fragments (Table 1). Figure 3 (a) further shows that larger HAC groups rely more on features involving substructures with more than two heavy atoms, explaining TDiMS's advantage over Atom-Pair. Unlike Atom-Pair, which is limited to small substructures, or MAP4, which loses interpretability through MinHash compression, TDiMS explicitly models distances between meaningful substructures while retaining interpretability.

TDiMS consistently outperformed neural-network models, likely due to the limitations of GNNs and CLMs discussed in Related Works. As shown in Fig. 3 (b), which shows the distribution of SHAP absolute values with respect to the longest bond-path distance between atoms in each substructure-pair feature, an increase in 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, features with bond-path distances of five or more accounted for 98.4% of the predictor's total feature contribution. These findings highlight the importance of modeling long-range structural interactions, which are challenging to capture with conventional GNNs constrained by message-passing depth [15].

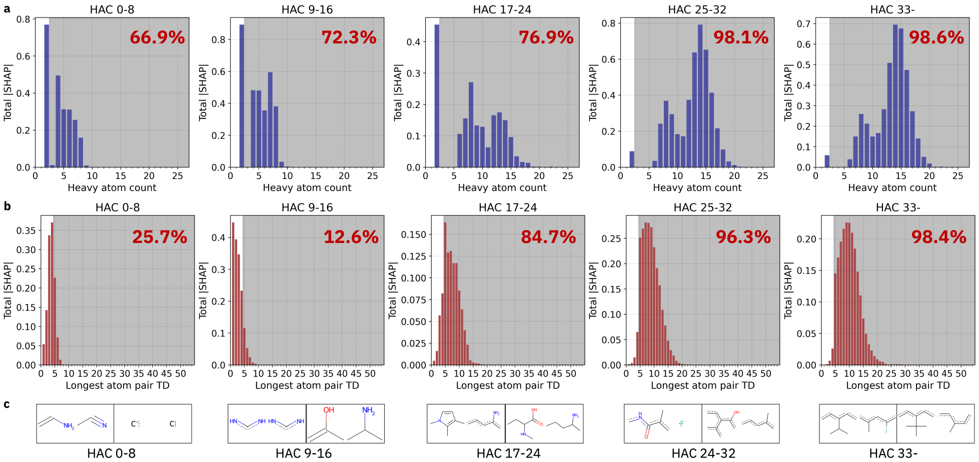

Figure 3: Distribution of SHAP values for TDiMS features: (a) HAC, (b) by longest TD in each substructure pair feature. (red numbers: % in shaded region), (c) top two substructure pairs per HAC group.

Next, we analyze the top two substructure-pairs that contributed most to the prediction of dipole moment. Across HAC groups, the top-ranked SHAP pairs reveal that as molecular size increases, features involving substructures with more than two heavy atoms and longer bond-path distances occur more frequently, as shown in Fig. 3 (c). This trend is consistent with the stronger performance of TDiMS on larger molecules. These attributions indicate what the model relies on under the chosen fragmentation scheme, while deeper mechanistic interpretation is left for future work. Further analysis may yield additional insights into their chemical relevance.

While this study focuses on dipole moments as a prototypical property governed by charge separation, the same framework can be readily applied to other property prediction tasks such as solubility, band gap, or polarizability. The fragment flexibility of TDiMS allows easy adaptation to these different contexts.

5 Conclusion and Future Work

This work has shown that the Topological Distance of intra-Molecular Substructures (TDiMS) descriptor offers clear advantages for molecular property prediction, demonstrated on a dipole moment task with the PubChemQC dataset. TDiMS outperformed conventional descriptors and neural-network–based embedding, with the largest gains for molecules with higher heavy atom counts where long-range intra-molecular interactions are critical. From the perspective of machine learning for materials discovery, TDiMS demonstrates the value of explicitly encoding distance-dependent structural relationships while retaining interpretability. This enables both improved predictive performance and interpretable analysis of dominant substructure pairs, offering insights often less accessible to end-to-end models. Its flexible design, including customizable fragments and feature calculation schemes, further enhances adaptability across datasets and properties, and allows targeted exploration of substructures of interest.

Future work will focus on improving computational scalability through sparse or approximate pair enumeration, and on extending TDiMS to integrate 3D geometrical distances for properties governed by through-space interactions. Although TDiMS currently operates on two-dimensional substructures and supports limited stereochemistry, it can be extended to include 3D fragment representations. For example, MolBar [29] provides a stereochemically aware molecular identifier using three-dimensional fragments to represent global molecular topology. Combining TDiMS with MolBar offers a complementary extension that supports diverse molecular systems while retaining interpretability. In parallel, hybrid modeling that combines TDiMS features with long-range GNN architectures will be explored to bridge interpretable descriptors and learned representations. We also plan to extend applications beyond organic molecules, such as polymers and inorganic compounds.

References

[1] 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.

[2] Rogers D and Hahn M. Extended-connectivity fingerprints. *J Chem Inf*, 50(5):742–754, April 2010.

[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] 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.

[5] H. Moriwaki, Y. S. Tian, N. Kawashita, and T. Takagi. Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, 10(4), February 2018.

[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] 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.

[8] 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.

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

[10] 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. Preprint, Research Square.

[11] 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.

[12] 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.

[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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] R.W. Floyd. Algorithm 97: Shortest path. *Communications of the ACM*, 5(6):345, June 1962.

[20] D.B. Johnson. Efficient algorithms for shortest paths in sparse networks. *Journal of the ACM*, 24(1):1–13, January 1977.

[21] S. Warshall. A theorem on boolean matrices. *Journal of the ACM*, 9(1):11–12, January 1962.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and É. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, October 2011.

[23] 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.

[24] 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.

[25] Jçrg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507, October 2008.

[26] Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S. Sánchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M. Brockway, , and Alán Aspuru-Guzik. The Harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett. 2011*, 1(17):2241–2251, 2011.

[27] 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.

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

[29] Nils van Staalduinen and Christoph Bannwarth. MolBar: A molecular identifier for inorganic and organic molecules with full support of stereoisomerism. *Digital Discovery*, 3:2298–2319, 2024.