TDiMS : A Topological Distance based Intra-Molecular Substructure Descriptor for Improved Machine Learning Predictions

Lisa Hamada, Indra Priyadarsini, Seiji Takeda, Onur Boyar

IBM Research - Tokyo

Lisa.Hamada@ibm.com, indra.ipd@ibm.com, seijitkd@jp.ibm.com, Onur.Boyar@ibm.com,

Abstract

Various molecular descriptors have been developed due to their diverse roles and importance in material informatics. However, they still have challenges in accurately capturing the global relationship of intra-molecular substructures, which significantly influence on the physical property. In this paper, we introduced a novel molecular descriptor which can extract topological distance between each pair of substructures within a molecule. Our evaluations reveal that the proposed descriptor outperformed existing baselines in downstream tasks, including neural-network-based models. Moreover, this descriptor enables to acquire important chemical insight into what substructure pairs need to be considered with topological distance, which is crucial for advanced tasks such as molecular generation.

Introduction

Machine learning (ML) has played a significant role in accelerating material discovery aiming to reduce the time/-cost and increase variability (Wei et al. 2019). ML models, specifically designed for predicting properties, are trained using features that encapsulate the characteristics of molecules, including molecular descriptors which capture different facets of these molecules. Consequently, the efficiency with which structural features are extracted plays a crucial role. Various molecular descriptors have been developed, ranging from Quantitative Structure-Property Relationships (QSPR) based descriptors (D and M 2010; Carhart and Venkataraghavan 1985; Capecchi, Probst, and Reymond 2020; Moriwaki et al. 2018), which basically enumerate constituent elements, to neural-network-based descriptors (Duvenaud et al. 2015; Ross et al. 2022a; Wang et al. 2022b; Ahmad et al. 2022). However, they still have limitations in accurately capturing global relationship of intra-molecular substructures. In addition, the interpretability of prediction model is crucial, as it contributes to subsequent tasks in material science, such as molecular generation.

Herein, we introduce a new molecular descriptor - Topological Distance of intra-Molecular Substructures (TDiMS), which can extract topological distance between each pair of substructures within a molecule. A topological distance between a substructure pair is approximately defined as the total mean of the shortest bond distances between atoms constituting each substructure, enabling long-distance interactions capture and flexible fragment targeting. As the combination of substructure pairs tends to significantly increase the dimensionality of the feature vector, TDiMS employs duplicate feature handling and feature selection to reduce it to a manageable size. Moreover, since the feature values represent structural pairwise distances, the TDiMS descriptor preserves the interpretability typical of QSPR-based methods. In this study, we not only demonstrated the strong effectiveness of TDiMS in prediction tasks but also confirmed that valuable chemical insights into key substructure pairs, where distance plays a crucial role, can be obtained according to the specific task. This study also provides an important direction for descriptor development including neural-network models that combining topological distance of intra-molecular substructures information can lead to further improvement.

Related Works

Mordred (Moriwaki et al. 2018) is an advanced descriptor calculation open-source software that primarily focuses on counting substructures based on physical chemistry knowledge, allowing the calculation of over 1800 types of two- and three-dimensional descriptors. However, the substructure counting method lacks global molecule information, such as intra-molecular positional relations, which can significantly influence the physicochemical properties of the molecule. Thus, the Atom-Pair (Carhart and Venkataraghavan 1985) descriptor is focused on capturing global molecular information, which captures the atomic environments and the shortest path separations between all pairs of atoms within a molecule. Despite this improvement, focusing solely on individual atoms brings its own challenges. To overcome these issues, the MinHashed Atom-Pair fingerprint up to four bonds (MAP4) (Capecchi, Probst, and Reymond 2020) was proposed. MAP4 encodes pairs of atoms and their bond distances, similar to the Atom-Pair fingerprint, but replaces atomic characteristics with the circular substructures surrounding each atom. Given the large number of possible substructure pairs, MAP4 employs MinHash values derived from Locality Sensitive Hashing (LSH) to efficiently represent a molecule. While MinHash values enable fast similarity searches in very large databases, they results in a trade-off with interpretability. More recently, latent vec-

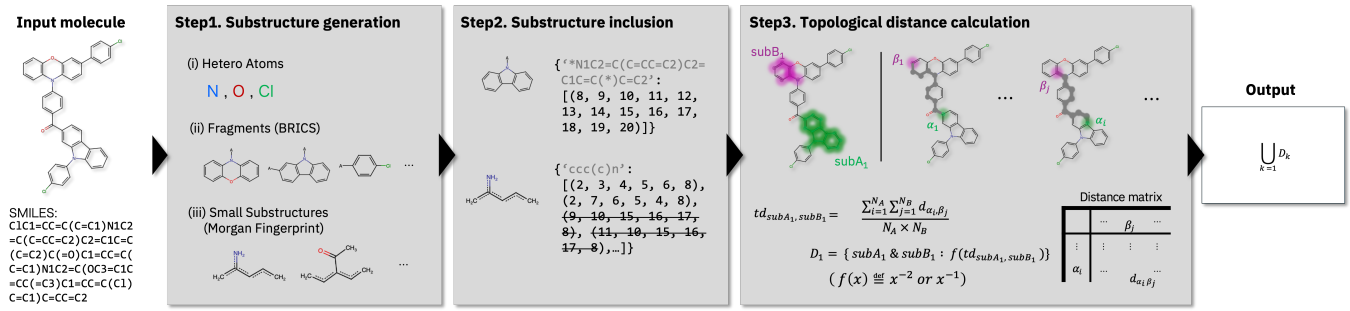

Figure 1: Workflow of TDiMS for a target molecule in dataset.

tors derived from various neural-network models, including Transformer-based chemical language models (CLMs) and graph neural-network (GNN) models, have been employed as feature vectors for downstream tasks (Ross et al. 2022a; Wang et al. 2022b; Ahmad et al. 2022; Liu, Demirel, and Liang 2019). These models are pretrained on molecular structures from large standard datasets, for example, PubChem (Kim et al. 2023) and ZINC (Irwin et al. 2012), which contain hundreds of millions or even billions of molecules. CLMs and GNN models are capable of learning information about atoms and their bonds. However, in GNN models, the technical limitation on the number of steps imposes a restriction on the range of bond-path distances that can be captured (Li, Han, and Wu 2018). In contrast, CLM models, with their attention mechanism, can potentially learn relationships between atoms and bonds even at long distances. Nevertheless, they are directly impacted by the limitations of molecule representations, for example, in the case of SMILES strings, the sequence of characters does not necessarily reflect the actual spatial arrangement of atoms in the molecular structure (Xia et al. 2023; Jin, Barzilay, and Jaakkola 2018). Furthermore, the feature vectors derived from neural-network models, commonly referred to as latent vectors, generally have no inherent meaning in each feature value, leading to a lower level of interpretability.

Method

TDiMS algorithm

Figure 1 shows an algorithm workflow of TDiMS for an input molecule. Canonical SMILES representations were used as input. Substructure pairs within the target molecule are first comprehensively explored. We targeted three types of substructures in this study; (i) Hetero atoms, (ii) fragments from BRICS method (Degen et al. 2008), and (iii) circular substructures from Morgan Fingerprint (Morgan 1965). When substructures are extracted by Morgan Fingerprint or fragments, a smaller substructure is often entirely encompassed by a larger superset substructure. Considering both small and large substructures cause a duplicate count of the same effect stemming from these substructures, TDiMS eliminates pairs of smaller substructures if they are completely included in larger ones (Step2 in Fig. 1). The topological distance between substructure pair, is approximately defined as the total mean of the shortest bond distances between atoms constituting each substructure using Floyd-Warshall Algorithm (Floyd 1962; Johnson 1977; Warshall 1962):

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

where $td_{subA,subB}$ is a defined topological distance between $subA$ and $subB$, N_A and N_B are the number of heavy atoms in $subA$ and $subB$, α_i and β_j denote the i-th and j-th heavy atoms in $subA$ and $subB$, and $bd_{\alpha_i\beta_j}$ is the shortest bond distance between α_i and β_j. We aim to capture the distance with spread in order to be independent of the shape of particular substructures. Additionally, using this calculation method enables to freely target any desired fragment. The feature values are calculated as the inverse or inverse square of this topological distance value to emphasize the influence of shorter distances and to account for chemical phenomenona such as Coulomb's law (Step3 in Fig.1). If a target substructure pair exists at multiple locations within a molecule, the total of each feature value is calculated in order to constrain feature number by grouping similar phenomena. Then TDiMS calculates the set of the feature values for all substructure pairs $\bigcup_{k=1} D_k$. Finally, feature vectors are generated corresponding to the integrated set of substructure pairs among whole molecules in the dataset. The number 0 is stored if the corresponding substructure pairs are not in the target molecules. Normalization was applied to the feature vectors across the whole molecules in the dataset. As mentioned above, there are a number of possible combinations of the features: substructure type combination, Hetero atom, fragments, circular substructures, and feature value calculation, inverse or inverse square. We evaluated with every combination of conditions and choose the best score in each task.

Feature Selection

We have performed grid search for optimizing hyperparameters while selecting features using the SelectFromModel class in Scikit-learn (Pedregosa et al. 2011). Lasso Regression and RandomForestClassifier were used as estimators for the prediction and classification tasks, respectively. The hyperparameter α in Lasso Regression was searched within the discrete set $10^k | -5 \leq k \leq 1$, where k is an integer, while the hyperparameter min_samples_split in RandomForestClassifier was searched in the range [2, 3].

Evaluation Tasks

To demonstrate the effectiveness of TDiMS, we benchmark the performance on multiple challenging regression and classification tasks from MeleculeNet (Wu et al. 2018). This allowed for a comprehensive comparison with baseline models, especially for neural-network models, across various tasks. We compared the performance of our proposed method with prior works such as D-MPNN (Yang et al. 2019), (Hu et al. 2019), MGCN (Lu et al. 2019), GEM (Fang et al. 2022), SchNet (Schütt et al. 2017), KPGT (Li, Zhao, and Zeng 2022), GraphMVP-C (Liu et al. 2021), GCN (Kipf and Welling 2016), GIN (Xu et al. 2018), MolCLR (Wang et al. 2022a), ChemBERTa-2 (Ahmad et al. 2022), MolFormer (Ross et al. 2022b), RF (Ross et al. 2022b), SVM (Ross et al. 2022b), N-Gram (Liu, Demirel, and Liang 2019), Galatica (Taylor et al. 2022), and Uni-Mol (Zhou et al. 2023). For QSPR-based descriptors, Morgan Fingerprint, Mordred, Atom-Pair, and MAP4, the corresponding packages were used to generate feature vectors (Moriwaki et al. 2018; Capecchi, Probst, and Reymond 2020; RDKit-Community 2024).

XGBoost (Chen and Guestrin 2016) was employed for the tasks, with hyperparameter tuning performed using Optuna (Akiba et al. 2019). The results based on the optimal hyperparameters are reported, with performance evaluated using the RMSE and ROC-AUC metrics. Feature importance was analyzed using XGBoost implemented method, identifying the key features contributing to the prediction models. We compared the performance of TDiMS with QSPR-based descriptors as well as neural-network models, including CLMs and GNN models. A feature vector derived from TDiMS was evaluated both independently and in concatenation with Mordred, aiming to comprehensively capture global and local molecular information.

The details of the benchmarks used are illustrated in Table 1. We evaluate 3 datasets for regression tasks and 4 datasets for classification tasks. To ensure a robustness and unbiased evaluation, we followed the MoleculeNet benchmark by using the same train/validation/test splits across all tasks.

Dataset	Description	#Mol.	Metric
Esol	Water solubility prediction of small molecules	1,128	RMSE
Freesolv	Hydration free energy of small molecules in water	642	RMSE
Lipo.	Prediction of octanol-water partition coefficient (logD)	4,200	RMSE
SIDER	Drug side effect classification for 27 types of adverse effects	1,427	ROC-AUC
ClinTox	Binary labels on clinical toxicity data on FDA-approved drugs	1,478	ROC-AUC
BACE	Binary labels on β-secretase 1 (BACE1) binding properties	1,513	ROC-AUC
BBBP	Binary labels on blood–brain barrier permeability	2,039	ROC-AUC

Table 1: Description of the benchmark datasets used in the evaluation of the proposed model. #Mol. and Lipo. stands for Molecule numbers and Lipophilicity, respectively.

Results and Discussion

Table 2 shows the TDiMS performance compared to other molecule descriptors derived from QSPR-based as well as neural-network models. The TDiMS and Mordred combined descriptors, which effectively capture both local and global molecule information, outperformed in all regression tasks. We also evaluated the TDiMS performance on classification tasks, as shown in Table 3. The descriptors including TDiMS achieved the best scores in all classification tasks.

Model	ESOL	FreeSolv	Lipophilicity
D-MPNN	1.050	2.082	0.683
Hu et al.	1.220	2.830	0.740
MGCN	1.270	3.350	1.110
GEM	0.798	1.877	0.660
SchNet	1.050	3.220	0.910
KPGT	0.803	2.121	**0.600**
GraphMVP-C	1.029	-	0.681
GCN	1.430	2.870	0.850
GIN	1.450	2.760	0.850
MolCLR	1.110	2.200	0.650
ChemBERTa-2	-	-	0.986
MolFormer	0.755	2.022	0.840
Morgan Fingerprint	0.769	1.756	0.691
Mordred	**0.311**	**1.307**	0.659
Atom-Pair	0.471	1.411	0.665
MAP4	0.962	2.595	0.837
TDiMS	0.423	1.325	0.649
TDiMS+Mordred	**0.252**	**1.076**	**0.545**

Table 2: RMSE scores (↓) of the evaluation on regression tasks. **Red** and **Blue** indicates best and second-best performing model, respectively.

Model	SIDER	ClinTox	BACE	BBBP
RF	68.4	71.3	86.7	71.4
SVM	68.2	66.9	86.2	72.9
MGCN	55.2	63.4	73.4	85.0
Hu, et al.	65.2	78.9	85.9	70.8
N-Gram	63.2	85.5	87.6	91.2
MolCLR	68.0	93.2	89.0	73.6
GEM	67.2	90.1	85.6	72.4
ChemBerta-2	-	90.7	85.1	71.94
Galatica 120B	63.2	82.6	61.7	66.1
Uni-Mol	65.9	91.9	85.7	72.9
MoLFormer-XL	69.0	94.8	88.2	93.7
Morgan Fingerprint	68.2	82.8	88.5	93.0
Mordred	82.2	57.3	90.6	**96.5**
Atom-Pair	77.1	89.4	**91.7**	94.7
MAP4	75.4	92.9	91.4	92.9
TDiMS	**84.4**	**99.0**	**92.0**	96.2
TDiMS+Mordred	**83.7**	**99.7**	91.4	**96.6**

Table 3: ROC-AUC scores (↑) of the evaluation on classification tasks. **Red** and **Blue** indicates best and second-best performing model, respectively.

Interestingly, individual TDiMS outperformed Atom-Pair and MAP4, both of which also target the topological distance of intra-molecule structure, across most regression and classification tasks in this study. The best score of TDiMS was mainly achieved when targeting heavy atoms and circular small substructures derived from Morgan Fingerprint

combinations in this study, which indicates targeting the same substructure as MAP4. In general, enumerative descriptors suffer from the number of structural patterns that grows exponentially with respect to the maximum size of the patterns. This issue is mitigated by either considering the small patterns only (Atom Pair) or compressing the information (MAP4). MAP4 represents a molecule as a set of so-called MinHash values derived from LSH; therefore, it probabilistically represents molecules and lacks accuracy and interpretability. In contrast, the feature set in TDiMS was saved by handling substructure inclusion relationships and integrating multiple pairs existence. Additionally, feature selection was applied to refine the key features. As shown in Fig. 2, the number of substructure pair combinations increases with the sample size and molecular diversity within the dataset, leading to a large feature set. However, by applying feature selection, the dimensionality is reduced to a manageable level for XGBoost. These TDiMS approaches effectively reduced the number of features while preserving the information and the interpretability of the feature vector. Indeed, our results indicate the importance of targeting substructures composed of multiple atoms and reveal that the method employed by TDiMS for capturing intra-molecular substructure distances is more suitable than that of MAP4.

The fact that TDiMS outperformed neural-network models across all tasks is likely attributed to the limitations that still exist in GNN and CLM models, as discussed in the Related Works section. TDiMS achieved significantly better performance compared to the baselines in SIDER and ClinTox datasets, both of the top two were derived from TDiMS. SIDER, followed by ClinTox, are the datasets that include molecules with large structures in terms of the number of atoms and aromatic rings compared to other datasets. By analyzing the longest atom-pair band-path-based topological distance within the target substructure pair feature and the total absolute feature importance values of each feature, we reveal that without the features with the bond-path-based distance of five or greater, 64.2% and 65.7% of the predictive capability of TDiMS were lost for SIDER and ClinTox tasks, respectively. This result shows the strength of TDiMS, which has no restrictions on the distances it can handle, over current GNNs that are limited by layer size (Li, Han, and Wu 2018).

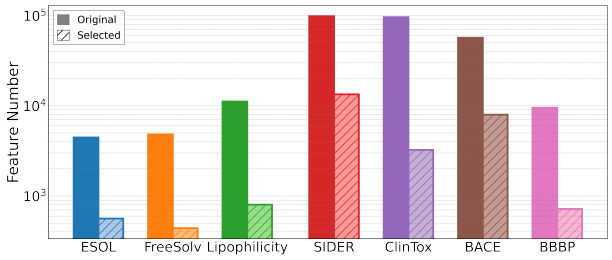

Figure 2: Original and selected feature vector dimension of TDiMS in each task.

Next, we further analyze the features that played important roles in the regression task. As a representative example, we demonstrate the results of concatenating feature vectors derived from TDiMS and Mordred for the Lipophilicity task in this study. The top 9 substructure pairs derived from TDiMS features based on the Feature Importance scores are illustrated in Figure 3. The Lipophilicity task targets the prediction of the Log D value, which measures the molecule's distribution between octanol and water. The substructure pairs in this figure appear to be closely related to hydrophilicity, hydrophobicity, or polarity, indicating that TDiMS successfully captured the structural features relevant to the target property. Further analysis could lead to a deeper and more precise understanding of the relationship between these substructure pairs and the target property, providing important chemical insights. Furthermore, when summing the absolute values of the Feature Importance for the features derived from TDiMS and Mordred, respectively, the ratio was found to be 9:1, indicating that TDiMS features played a more dominant role compared to those from Mordred.

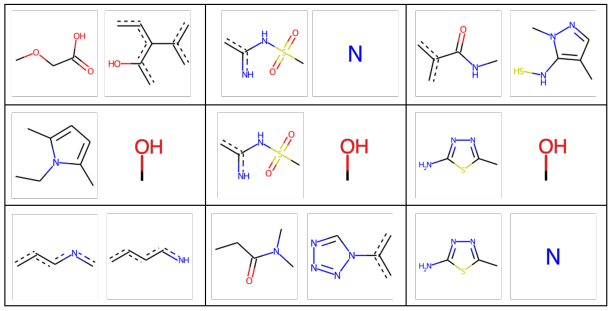

Figure 3: Substructure pairs of TDiMS features with high feature importance in Lipophilicity task.

Conclusion and Feature Work

This paper presents TDiMS, a novel descriptor that effectively captures and directly handles substructure-pair topological distances related to intra-molecular interactions. Experiments show that TDiMS achieves significant improvements across various benchmarks. Moreover, TDiMS successfully captured reasonable features aligned with the characteristics of each task. Further investigation into the relationships between these substructure pairs and their properties could yield deeper chemical insights. This study offers a key perspective for the future development of neural-network models, suggesting that combining topological distance information of intra-molecular substructures can drive further advancements. In addition, the introduction of task-specific fragments in TDiMS has potential to enhance performance further and provide critical insights into molecular design. Extending validation to datasets consisting of molecules with larger structure, which are currently recognized as challenging tasks, is expected to highlight the strengths of TDiMS more effectively. In parallel, we aim to further explore the characteristics of TDiMS through an empirical analysis of the computational costs of the Floyd-Warshall algorithm, focusing on its scaling behavior with respect to molecular size.

References

Ahmad, W.; Simon, E.; Chithrananda, S.; Grand, G.; and Ramsundar, B. 2022. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*.

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Capecchi, A.; Probst, D.; and Reymond, J.-L. 2020. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(43).

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

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794. New York, NY, USA: ACM. ISBN 978-1-4503-4232-2.

D, R.; and M, H. 2010. Extended-connectivity fingerprints. *J Chem Inf*, 50(5): 742–754.

Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; and Rarey, M. 2008. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3(10): 1503–1507.

Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems*, 2224–2232.

Fang, X.; Liu, L.; Lei, J.; He, D.; Zhang, S.; Zhou, J.; Wang, F.; Wu, H.; and Wang, H. 2022. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2): 127–134.

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

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; and Leskovec, J. 2019. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*.

Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; ; and Coleman, R. G. 2012. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling*, 52(7): 1757–1768.

Jin, W.; Barzilay, R.; and Jaakkola, T. 2018. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, 2323–2332. PMLR.

Johnson, D. 1977. Efficient Algorithms for Shortest Paths in Sparse Networks. *Journal of the ACM*, 24(1): 1–13.

Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; and Bolton, E. E. 2023. PubChem 2023 update. *Nucleic acids research*, 51(D1): D1373–D1380.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Li, H.; Zhao, D.; and Zeng, J. 2022. KPGT: knowledge-guided pre-training of graph transformer for molecular property prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 857–867.

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

Liu, S.; Demirel, M. F.; and Liang, Y. 2019. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *Advances in neural information processing systems*, 32.

Liu, S.; Wang, H.; Liu, W.; Lasenby, J.; Guo, H.; and Tang, J. 2021. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*.

Lu, C.; Liu, Q.; Wang, C.; Huang, Z.; Lin, P.; and He, L. 2019. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 1052–1060.

Morgan, H. L. 1965. The generation of a unique machine description for chemical structures – a technique developed at chemical abstracts service. *J. Chem. Doc.*, 5(2): 107–113.

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

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

RDKit-Community. 2024. RDKit: Open-source cheminformatics software, Version 2024.09.1. Accessed: September 2024.

Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; and Das, P. 2022a. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4: 1256–1264.

Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; and Das, P. 2022b. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12): 1256–1264.

Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; and Müller, K.-R. 2017. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30.

Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; and Stojnic, R. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Wang, Y.; Wang, J.; Cao, Z.; and Barati Farimani, A. 2022a. Molecular contrastive learning of representations via graph

neural networks. *Nature Machine Intelligence*, 4(3): 279–287.

Wang, Y.; Wang, J.; Cao, Z.; and Farimani, A. B. 2022b. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4: 279–287.

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

Wei, J.; Chu, X.; Sun, X.-Y.; Xu, K.; Deng, H.-X.; Chen, J.; Wei, Z.; and Lei, M. 2019. Machine learning in materials science. *InfoMat*, 1(3): 338–358.

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

Xia, J.; Zhu, Y.; Du, Y.; and Li, S. Z. 2023. A Systematic Survey of Chemical Pre-trained Models. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, 6787–6795.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. 2019. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8): 3370–3388.

Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; and Ke, G. 2023. Uni-Mol: a universal 3D molecular representation learning framework.