Geometric Deep Learning For Molecule Similarity Prediction

Candidate number: 1071068

Abstract

Molecule similarity prediction has applications in numerous domains such as drug discovery, materials science, environmental science, and bioinformatics, enabling the identification of structurally similar molecules. Geometric Deep Learning models, especially Graph Neural Networks, have proven to be successful in capturing graph structures and complex topological relationships in the data. In this study, we explore the influence of various fingerprint techniques and molecule similarity measures on the task of molecule similarity prediction. We compare different neural network architectures to identify the important factors affecting model performance. Our investigation seeks to provide insights into the significance of factors such as model complexity (i.e., the number of layers and parameters), expressive ability of a model, robustness to transformations (i.e., invariance, equivariance), and the impact of dataset generation choices on the model's predictions. Please see the GitHub repository.

1. Introduction

1.1. Background

Molecule similarity prediction allows scientists to identify molecules which share similar structures - this has a wide range of applications in drug discovery, materials science, environmental science and bioinformatics. Predicting molecule similarity provides us with information about biological activity, allows to identify potential harmful compounds which share similar properties as known toxic compounds, discover new materials with specific desired properties or help us to understand functional and structural similarities between molecules (Shaw & Schneider, 2021).

Geometric Deep Learning models such as Graph Neural Networks (GNNs) are able to capture complex relationships between atoms and edges of molecules as well as the structural and topological features of molecular graphs. One of the ways we can represent molecules is using graphs where nodes model atoms and chemical bonds are represented as edges. It is plausible to think we could use GNNs and try to learn the structural information about each molecule as well as the chemical properties of atoms and bonds to predict molecule similarity. Another promising aspect is that some models are invariant or equivariant to graph isomorphisms (Gilmer et al., 2017), ensuring that the model's similarity predictions are consistent regardless of the molecular rotation and translation. On the other hand, Convolutional Neural Networks (CNNs) are known for their capability to extract most relevant features and although they are not well-suited for graph-based data, they can process molecules in their fingerprint-based representations, which capture the global and local structure of each molecule (Goh et al., 2017). Transformer-based neural networks are well-suited for capturing long-range dependencies and due to their attention mechanism suffer less from oversmoothing (Li et al., 2018) that could arise if the networks need to capture properties of large and complex molecules.

Some of the previous approaches which tackle molecule similarity prediction include traditional machine learning or kernel-based algorithms. The downside of these techniques is that they often require advanced feature-engineering and extensive data pre-processing. Some of the recent deep learning approaches for molecule similarity prediction include Convolutional Neural Networks (CNN) for graphs (Duvenaud et al., 2015) and Graph Convolutional Networks (GCN) for molecular property prediction (Kearnes et al., 2016). Since then, several more powerful GNNs have emerged, including models such as Graph Attention Networks (GAT) (Veličković et al., 2018), Graph Isomorphism Networks (GIN) (Xu et al., 2019), ChebNet (Defferrard et al.) and the recent rise in popularity of applying message passing architectures with invariant and equivariant layers (Kondor & Trivedi, 2018).

1.2. Research question

We experiment with different fingerprint techniques, molecule similarity measures and neural network architectures to answer the following question: What are the most important factors affecting the model's performance on the task of molecule similarity prediction? Through experimental study of different architectures we hope to understand in what way dataset generation (by using different fingerprint and similarity techniques) and the model complexity changes the way how the model learns and generalizes on

unseen data. Our experiments attempt to provide an insight into whether GNNs outperform other models and whether factors such as expressive power of a model, complexity of a model and robustness to transformations such as rotations or translations have an impact on the test performance.

2. Methodology

This section describes our approach of creating a dataset of pairs of similar and dissimilar molecules from the QM9 dataset (Ramakrishnan et al., 2014). We sample a pair of molecules, compute the fingerprint score for each molecule (using selected fingerprint technique), compute similarity of fingerprints (using selected similarity measure) and create a dataset. We control the split of positive and negative pairs by the positivity ratio parameter (ensuring 50 / 50 split of label 0 and 1 pairs) and the positivity threshold parameter (which specifies the minimum similarity score for a pair of molecules to be considered similar and labelled 1).

2.1. Fingerprints

Fingerprints are compact representations of molecules which encode specific molecular structural features (Gobbi & Poppinger, 2008). We use them when creating datasets of pairs of similar and dissimilar molecules to encode the structure and chemical properties of each molecule in a form of numerical vectors, which are passed to a similarity measure function to determine similarity of 2 molecules. In our experiments we explore the following fingerprints based on review of (O'Boyle & Sayle, 2016):

Morgan: Encode local structural information around each atom and atom neighbourhoods at increasing radii.
Topological (daylight): Encode molecular graph's connectivity, topology and the presence of specific substructures.
Avalon: Encode the substructures within a certain radius of each atom in a molecule, as well as their connectivity and the bond types between them.
Path-based: Encode the frequency of paths of a certain length between atoms and the environment surrounding each atom.
Torsion: Encode the torsion angles between connected atoms in a molecule and their frequencies .

2.2. Similarity measures

We consider the following similarity measures to determine which similarity measure is best suited for our dataset and the task of molecule similarity prediction:

Tanimoto similarity: Proportion of shared bits in a fingerprint between two molecules, takes values between 0 (no shared features) and 1 (identical features).
Dice coefficient: Total number of shared features divided by the average number of features in both molecules.
Euclidean distance: L2 distance between two fingerprints.
Pearson rank correlation: measures the strength and direction of a linear relationship between two fingerprints, takes values between -1 (negative correlation) to 1 (positive correlation).

2.3. Dataset generation

We generate training, validation and testing datasets of similar and dissimilar pairs from the QM9 benchmark dataset (using different fingerprints and different similarity functions). Each generated dataset provides pairs of molecular graph representations along with a label indicating whether they are similar or dissimilar. We first compute molecular fingerprint for each molecule, compute the similarity score between corresponding fingerprints and create a dataset (with specified similarity threshold, in our case 0.7) of 50 / 50 split of pairs of similar (label 1) and dissimilar (label 0) molecules.

3. Models

The architecture of all models can be decomposed into the following:

Base Model: A neural network consisting of a sequence of layers, activation functions, dropout and pooling layers.

Siamese Network: A network which combines the embeddings of 2 molecules produced as output from the base model which are concatenated and passed through a fully connected layer to compute similarity score between 2 molecules. Approach inspired by (Altalib & Salim, 2022).

3.1. Model architectures

GCN, GAT, GIN, ChebNet, Transformer: The base model consists of 2 layers (varying layer specific to each model) and ReLU activation function between the layers. The features are pooled using global mean pooling to aggregate the node features into a single vector and create a fixed-size graph embedding. The siamese network takes two graph inputs and computes their embeddings using the base model. These embeddings are then concatenated and passed through a fully connected layer to compute the similarity score between the two graphs. The similarity score is returned as a value between 0 and 1 (using sigmoid activation function), with 1 representing high similarity and 0 representing no similarity. Specific to GIN: Each layer uses 2-layer sequential neural network with ReLU as its message-passing function. Specific to ChebNet: We set the parameter K (the order of the Chebyshev polynomial approximation) to 4 which produces smooth filters but allows the model to learn more complex patterns. Specific to Transformer: The layers use custom message passing which applies a transformer-based convolution using multi-head self-attention.

I-MPNN: The base model uses linear layer to transform input features followed by a sequence of 2 invariant message passing layers and a global mean pooling layer. Each invariant layer uses 2 multi-layer perceptrons for message computation (which takes into account distance between node positions and concatenates it with the node features) and the node update. The siamese network combines 2 produced graph embeddings as described in section above. The main difference between this model and models above is the use of the distance between nodes in the message computation of the layers.

E-MPNN: Same setup as I-MPNN except it uses equivariant layers in which the message function takes into account the distance between nodes but does not use the positional information in the update function. The equivariant layer is less sensitive to the distances between nodes and more sensitive to relative order of nodes.

CNN: The base model takes as input 1D molecular fingerprints and passes them through a series of 2 blocks of a convolutional layer (kernel size=3, padding=1), ReLU activation and max-pooling layer (kernel size=2, stride=2) and a fully connected layer. The siamese architecture remains same but instead of embeddings of molecules it generates embeddings of fingerprints. This configuration has significantly more parameters than other models so we apply dropout 0.5 to reduce over-fitting.

3.2. Training and hyperparameter tuning

We pefromed extensive hyperparameter search. We trained the models in Google Colab environment (using GPUs), used Adam optimizer (best learning rate specific for each model), BCE loss for classification tasks and MSE loss for prediction tasks. We performed hyperparameter tuning on the following: learning rate, dropout rate, number of layers, number of hidden units, batch size, number of attention heads (for GAT and Transformer), and other parameters specific to each network (i.e. value of K in ChebNet, kernel size in CNN). We trained the models on the training dataset and tuned hyperparameters on the validation dataset and kept the architectures which achieved the highest validation accuracy. Overall, there was no advantage of adding more than 2 layers for all models except for the CNN. We used dropout=0 for all models except for the CNN and applied early stopping with patience of 10. The learning rate used for GCN, GAT, ChebNet, GIN is 0.005, for Transformer 0.003, for I-MPNN and E-MPNN 0.01 and for CNN 0.001. We used batch size 64 for all models and hidden sizes 64 or 128, depending on the model.

4. Experiments

For reporting statistics, we repeat experiments 10 times (where relevant) and show the average of measurements.

Ex1 - Fingerprints & similarity measure comparison: We investigate the performance of fingerprint techniques and similarity measures for creating the datasets and their effect on the model's performance. We report test accuracy and provide insights into how the fingerprint and similarity measure selection affect the training procedure.

Ex2 - Model benchmarking on molecule classification: We compare the performance of the models on the task of molecule classification where the task is to predict a discrete label whether the molecules are similar (1) or dissimilar (0). The experiment uses our created dataset of similar and dissimilar pairs using Morgan fingerprints and Tanimoto similarity measure. We report test accuracy, number of model parameters and training time to observe how performance varies with model architecture and complexity.

Ex3 - Model benchmarking on molecule similarity prediction: We compare performance on the task of similarity score prediction where the task is to predict the similarity score of 2 molecules. We report test MSE and use Morgan fingerprints and Tanimoto similarity measure.

5. Results

Ex1: Table 1 provides insights into created datasets after we generate 2 classes of molecule pairs, using differing fingerprint techniques. Morgan, Topological, and Avalon fingerprints are suitable because they distinguish between similar and dissimilar molecules, as shown by the larger differences between the mean positive and negative fingerprint similarities. In contrast, Torsion and Path-based fingerprints are not suitable for this task because they fail to differentiate between similar and dissimilar molecules, as shown by the close values of their mean positive and negative similarities. Morgan fingerprint is the most suitable option because there is greatest distance between mean positive similarity 0.8732 and mean negative similarity 0.5204. Similarly, Topological fingerprint has mean positive similarity 0.8623, which is higher than the mean negative similarity 0.5139, and low standard deviations within classes. Avalon fingerprint has mean positive similarity 0.8681, which is higher than the mean negative similarity 0.4734, but higher standard deviations of fingerprints in each group, hence there is greater range of fingerprint values in each group which increases the difficulty of learning on this type of dataset.

Table 1. Molecule Similarity Statistics (varying fingerprints)

Fingerprint	Mean Similarity	Similarity Std	Mean Pos	Mean Neg	Std Pos	Std Neg
Morgan	0.6837	0.2199	0.8732	0.5204	0.1250	0.1073
Topological	0.6968	0.2224	0.8623	0.5139	0.1244	0.1165
Avalon	0.6902	0.2372	0.8681	0.4734	0.1463	0.1495
Torsion	0.7024	0.1054	0.7124	0.6924	0.0959	0.0954
Path-based	0.6899	0.1123	0.7010	0.6889	0.0825	0.0811

Ex1: Figures 1-4 display the UMAP (McInnes et al., 2018) plots of molecule clustering using different fingerprint techniques projected onto 2D space. Morgan, Topological and Avalon fingerprints cluster the molecules into groups which are desirable for creating a dataset for molecule similarity prediction. Morgan and Topological fingerprints cluster molecules in a similar way whereas Avalon fingerprints produce greater number of smaller clusters. Path-based and Torsion fingerprints are not suitable candidates for further investigation as they are not capable of separating molecules and instead produce one large cluster - this makes training a model on such datasets impossible as the model just learns to predict all molecules uniformly as similar (with label 1).

We found that the choice of similarity measure, unlike the choice of fingerprint technique, has no effect on clustering - the plots and dataset statistics show insignificant differences. We show selected 2D projections to point out the differences between fingerprints (plots use Tanimoto similarity).

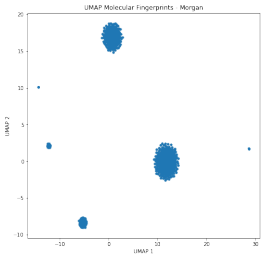

Figure 1. Morgan

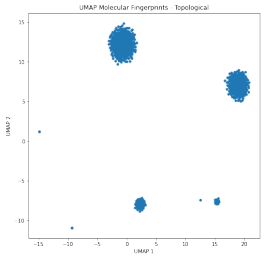

Figure 2. Topological

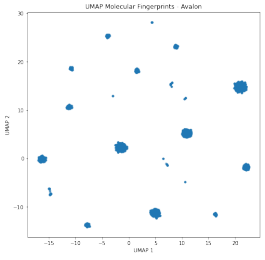

Figure 3. Avalon

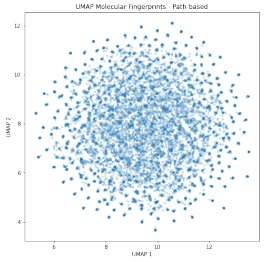

Figure 4. Path-based

Ex1: Figures 5 and 6 show the differences in generated datasets. Topological (and Morgan) fingerprints produce fewer large clusters, therefore most molecules are either very close to each other (high similarity indicated by strong red) or very far apart (low similarity indicated by dark blue). Training a model on such dataset is easier because the boundaries are well-defined. In contrast, Avalon fingerprints produce more clusters with similar molecules so there is varying degree of similarity between molecules (more tones of red) because some of the clusters are close together - this makes a dataset more challenging for the models to perform well.

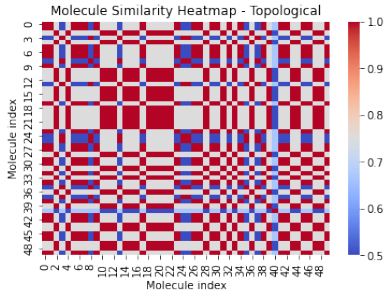

Figure 5. Topological

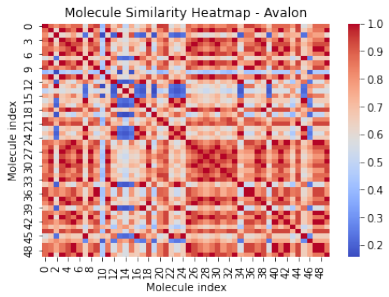

Figure 6. Avalon

Ex2: Table 2 shows classification accuracy of the models on the dataset created using Morgan fingerprints and Tanimoto similarity measure. GAT slightly outperforms GCN and has shorter training time. GIN has more parameters than GCN and GAT, but takes similar time to train and achieves higher test accuracy of 98.23%. ChebNet achieves a test accuracy of 98.32% - it has the highest number of parameters among the 'vanilla' graph-based models and longer training time. Despite having the highest number of parameters and longest training time, CNN has the lowest test accuracy among all the models, suggesting that it is the least suitable model. Transformer achieves high test accuracy and despite having 105 000 parameters, the training only takes 122 seconds due to its optimisations for utilizing GPU resources (Vaswani et al., 2017). The I-MPNN has 236 000 parameters, takes 149 seconds to train, and has test accuracy of 98.65% which makes it the best-performing model. The E-MPNN has the same number of parameters as I-MPNN, but takes longer to train and achieves lower test accuracy.

Table 2. Classification Test Accuracies - Morgan

Model	Params	Training time/s	Test acc
GCN	31,000	271	0.9723
GAT	32,000	216	0.9760
GIN	52,000	223	0.9823
ChebNet	63,000	319	0.9832
CNN	8,000,000	399	0.9052
Transformer	105,000	122	0.9843
I-MPNN	236,000	149	0.9865
E-MPNN	236,000	211	0.9803

Ex1,Ex2: Table 3 shows classification accuracies of the models across datasets created using different fingerprint techniques, which provides insight into how the models handle different dataset complexities. I-MPNN demonstrates the highest classification accuracy. Transformer, E-MPNN and ChebNet also show strong performance, with accuracy scores above 0.9 on all datasets. The CNN model shows the lowest classification accuracy across all datasets due to its limitations in capturing graph inputs.

Table 3. Classification Accuracies (fingerprint comparison)

MODEL	MORGAN	TOPOLOGICAL	AVALON
GCN	0.9723	0.9580	0.7958
GAT	0.9760	0.9639	0.8243
GIN	0.9823	0.9698	0.8322
CHEBNET	0.9832	0.9800	0.9344
CNN	0.9052	0.8182	0.7574
TRANSFORMER	0.9843	0.9803	0.9185
I-MPNN	0.9865	0.9850	0.9542
E-MPNN	0.9803	0.9735	0.9347

Ex3: Test MSE values for the molecule similarity prediction task in Table 4 show that I-MPNN outperforms all other models with the lowest test MSE of 0.00826. E-MPNN follows, achieving the test MSE of 0.01003. This demonstrates that incorporation of invariance and equivariance properties in the neural network architectures improves model's test performance on the molecule similarity prediction task. The worst performing models in this experiment were GCN and GAT, which could have been caused by their lower expressive power compared to other models and greater impact of oversmoothing.

Table 4. Molecule Similarity Prediction Test MSE

MODEL	TEST MSE
GCN	0.03517
GAT	0.03555
GIN	0.01675
CHEBNET	0.01676
CNN	0.01768
TRANSFORMER	0.01192
I-MPNN	0.00826
E-MPNN	0.01003

Ex2: Figures 7-10 show data collected during training. We observe that when we use Morgan fingerprints, all models, when tuned correctly, can achieve high validation accuracy on the task of predicting whether 2 molecules are similar. This indicates that the generated dataset is likely not challenging enough for these models and to make more conclusive statements, we would require larger differences in performance across models. Avalon dataset is more challenging, displaying larger variance in performance, but all models are able to achieve high validation accuracy in 70 epochs. We observe that the curves are not smooth but rather noisy, which suggests that despite achieving high training, validation and test accuracy, the training procedure is unstable. We tried to increase batch size, lower the learning rate further and train for longer period of time. The curves slowly become more stable but it requires significant amount of time and does not further improve test accuracy.

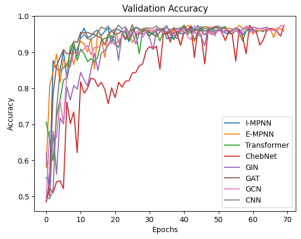

Figure 7. M: Validation Acc

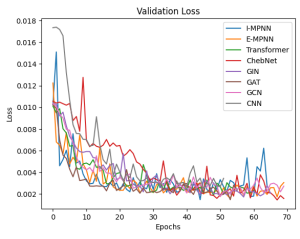

Figure 8. M: Validation loss

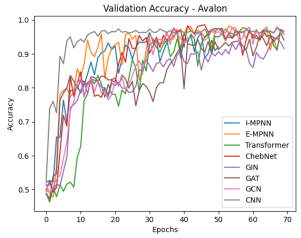

Figure 9. A: Validation Acc

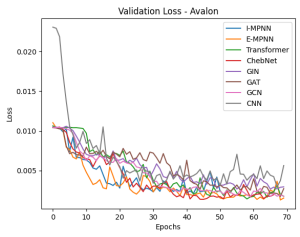

Figure 10. A: Validation loss

6. Discussion

This section links our results back to the research question. In summary, selection of the right fingerprint technique and suitable model are most important factors affecting performance. Choice of similarity measure did not make any difference. Complexity of a model (in terms of trainable parameters) is not as important as its suitability for the required task - as results show, Graph Neural Networks are the most appropriate model based on achieved test performance.

Selection of fingerprints: As shown by the results in Ex1, the choice of fingerprint technique impacts the model's performance. Morgan fingerprints produce the easiest dataset to learn, as demonstrated by the short training times and high test accuracies. This is due to the ability of Morgan fingerprints to capture local chemical environments and patterns (Rogers & Hahn, 2010), which are important for similarity prediction tasks, and also the fact that the dataset has fewer clusters, distanced far apart from one another. Topological fingerprints generate slightly more challenging dataset, due to their focus on the connectivity information between atoms which adds complexity (Systems, 2000). Avalon fingerprints produce the most challenging dataset for the models, with the lowest test performance and, on average, five times longer training times since the boundaries

between similar and dissimilar pairs are more difficult to learn. We omit Path-based and Torsion fingerprints from the model performance analysis as we have shown that they are unable to distinguish molecules from the QM9 dataset.

Selection of similarity measures: Results suggest that the choice of similarity measure has minimal impact on the performance of models. This could be due to small sizes and simple structures of molecules in the data.

GNNs more powerful than CNN: Our results confirm that selecting suitable model architecture for a given task, tuning the hyperparameters and components of the model are more important than model complexity in terms of the number of parameters. Using more complex model (with more layers / parameters) does not translate into better performance or generalization capability, even if regularization is used. GNNs are better choice for graph-based inputs than CNN, which works better on grids. CNN is limited to local dependencies due to the use of convolutional layers which are a bottleneck in molecule similarity prediction (LeCun et al., 2015). GNNs outperfromed CNN due to their ability to capture graph structures, while CNN used fingerprints, which likely caused information loss (Goodfellow et al., 2016). In addition, CNNs require inputs of fixed size while GNNs can process varying sizes and topologies of graphs.

GNNs comparison: GIN and ChebNet models achieved higher performance due to their greater capability to distinguish molecule structures compared to GAT and GCN. GIN captures graph isomorphism and ChebNet uses Chebyshev polynomials, allowing it to capture frequencies in the graph spectrum. Their aggregation mechanisms make them less vulnerable to oversmoothing and oversquashing. GIN uses an injective aggregation function that preserves the original features while aggregating neighbourhood information and ChebNet uses spectral filtering to limit the diffusion process. GIN's injectivity during aggregation helps to avoid oversquashing by maximizing expressive power (Xu et al., 2019) and ChebNet's spectral filtering approach preserves more information by selectively aggregating features based on the graph spectrum (Defferrard et al.).

Invariance improves performance: We showed that using molecular representations in the message-passing process and ensuring invariance under rotation and reflection has contributed towards better performance across all experiments as the model could capture the underlying properties of molecules without being affected by their specific arrangements. This ensures that the output of the model remains unchanged regardless of different orientations of the same molecule (Gilmer et al., 2017) - this is desirable for molecule similarity prediction because the goal is to identify structurally similar molecules regardless of their orientation. The E-MPNN model with equivariant layers, which depended on relative spatial information between nodes, did not perform as well as the I-MPNN model, because relative orientation of the atoms in the molecule is less relevant for similarity prediction and it is likely that the equivariant layers increased the complexity and variability during training. Since the QM9 dataset contains smaller and simple molecules, there might not have neccessarily been complex spatial relationships that would benefit from the use of equivariant layers. (Ruddigkeit et al., 2012).

7. Conclusion

We have explored Siamese architectures of different families of models (Graph Neural Network, Transformer, CNN) and evaluated their suitability for the task of molecule similarity prediction in various experiments on the datasets which we created using the QM9 dataset, different fingerprint techniques and varying similarity measures.

We demonstrated that creating a dataset and pre-processing data (in particular fingerprint selection) is equally as important as selecting and tuning the right model architecture. Models such as Transformer which can learn long-range dependencies and complex patterns in the structure of molecules perform better than models which focus on local features such as CNNs. In the Graph Neural Networks family, GIN and ChebNet, which have more sophisticated neighbour aggregation functions, perform better than simpler GCN and GAT. Invariant and equivariant Graph Neural Networks were able to achieve the best performance from all models due to their increased robustness to handle transformations such as rotation and translation of molecules, which benefited the molecule prediction similarity task.

Due to high computational requirements to train models to the highest performance (a significant time and resources were spent on tuning hyperparameters), the experimental setup where we tested combinations of fingerprints, similarity measures and model configurations, and the fact that we had to repeat experiments, we only explored molecules from the QM9 dataset. Extensions of this project could be to assess the models on a wider range of molecular structures by using greater variety of datasets like ChEMBL, ZINC, or PubChem - this would help to reach more conclusive results. In addition, more complex and varied synthetic datasets need to be generated such that the task of similarity prediction will become more challenging and not all models will be able to achieve high test accuracies - this will discriminate the performance of models by a larger margin. We could explore other representations and instead of graph-based structure and fingerprints use SMILES strings (Weininger, 1988), encode domain knowledge about molecules, or use unsupervised clustering approach to generate embeddings of molecules. Finally, we could explore the approach of ensembling the best performing models and combine predictions from the models to achieve better performance.

Code: GitHub repository.

References

Altalib, M. K. and Salim, N. Similarity-based virtual screen using enhanced siamese deep learning methods. *Journal of Chemical Information and Modeling*, 62(1):118–132, 2022.

Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems*, pp. 2224–2232, 2015.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1263–1272. JMLR. org, 2017.

Gobbi, A. and Poppinger, D. Performance evaluation of a novel fingerprint for similarity searching. *Journal of chemical information and modeling*, 48(10):2083–2090, 2008.

Goh, G. B., Hodas, N. O., and Vishnu, A. Deep learning for computational chemistry. *Journal of computational chemistry*, 38(16):1291–1307, 2017.

Goodfellow, I., Bengio, Y., and Courville, A. Deep learning. 2016.

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8):595–608, 2016.

Kondor, R. and Trivedi, S. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2747–2755. JMLR. org, 2018.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015.

Li, Q., Han, Z., and Wu, X.-M. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

O'Boyle, N. M. and Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of Chemical Information and Modeling*, 56(7): 1267–1272, 2016.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50 (5):742–754, 2010.

Ruddigkeit, L., Deursen, R. v., Blum, L. C., and Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012.

Shaw, M. L. and Schneider, N. Deep learning for molecular similarity. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 11(2):e1478, 2021.

Systems, D. C. I. Daylight theory manual. 2000. URL http://www.daylight.com/dayhtml/doc/theory/theory.finger.html.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.