Abstract  The ability of quantitative molecular prediction is of great significance to drug discovery, human health, and environmental protection. Despite considerable efforts, quantitative prediction of various molecular properties remains a challenge. Although some machine learning models, such as bidirectional encoder from transformer, can incorporate massive unlabeled molecular data into molecular representations via a self-supervised learning strategy, it neglects three-dimensional (3D) stereochemical information. Algebraic graph, specifically, element-specific multiscale weighted colored algebraic graph, embeds complementary 3D molecular information into graph invariants. We propose an algebraic graph-assisted bidirectional transformer (AGBT) model by fusing representations generated by algebraic graph and bidirectional transformer, as well as a variety of machine learning algorithms, including decision trees, multitask learning, and deep neural networks. We validate the proposed AGBT model on five benchmark molecular datasets, involving quantitative toxicity and partition coefficient. Extensive numerical experiments suggest that AGBT outperforms all other existing methods for all these molecular predictions.

Keywords  Algebraic graph, Transformer, Self-supervised Learning, Toxicity, Partition Coefficient, Multi-task Learning

1 Introduction

The fact that there is no specific and effective drug for coronavirus disease 2019 (COVID-19) one year after the outbreak reminds us that drug discovery remains a grand challenge. Rational drug discovery involves a long list of molecular properties, including binding affinity, toxicity, partition coefficient, solubility, pharmacokinetics, pharmacodynamics, etc. [1] Experimental determination of molecular properties is very time consuming and expensive. Additionally, experimental testing involving animals or humans is subject to serious ethical concerns. Therefore, various computer-aided or in silico approaches have become highly attractive because they can produce quick results without seriously sacrificing accuracy in many cases. [2] One of the most popular approaches is the quantitative structure activity relationship (QSAR) analysis. It assumes that similar molecules have similar bioactivities and physicochemical properties. [3]
Recently, machine learning (ML), including deep learning (DL), has emerged as a powerful approach for
data-driven discovery of in molecular science. For example, generative adversarial networks (GANs)\,[4, 5]
graph convolutional networks (GCNs)\,[5, 6, 7], convolutional neural networks (CCNs)\,[8], and recurrent
neural networks (RNNs)\,[9, 10], have become popular for drug discovery and molecular analysis\,[11, 12, 9].
However, DL methods require large datasets to determine their large number of weights and might not be
competitive for small datasets\.[13]

Although DL methods, particularly CNN and GANs, can automatically extract features from simple
data, such as images and/or texts, the performance of ML and DL methods for molecules, particularly macro-
molecules, crucially depends on the molecular descriptors or molecular representations due to their intricate
structural complexity \[14\]. Earlier molecular descriptors are designed as the profiles or fingerprints of interpretable physical properties in a bit string format \[15\]. Various fingerprints have been developed in the past few decades\.[16, 17] There are four main categories of two-dimensional (2D) fingerprints \[17\], namely sub-
structure key-based fingerprints\,[18] topological or path-based fingerprints\,[19] circular fingerprints\,[16] and phar-macophore fingerprints\.[20] However, 2D fingerprints lack three-dimensional (3D) structural information
of molecules, especially stereochemical descriptions.

To deal with the aforementioned problems, 3D-structure based fingerprints have been developed to
capture 3D patterns of molecules. \[21\] However, the molecular structural complexity and high dimensionality
are the major obstacles in designing efficient 3D fingerprints \[14\]. Recently, a variety of 3D molecular
representations based on advanced mathematics, including algebraic topology \[8, 22\], differential geometry
\[23\], and algebraic graph \[24\] have been proposed to simplify the structural complexity and dimensionality
of molecules and biomolecules \[14\]. These methods have had tremendous success in protein classification,
virtual screening, and the predictions of solubility, solvation free energies, toxicity, partition coefficients,
protein folding stability changes upon mutation, and Drug Design Data Resource (D3R) Grand Challenges
\[25, 14\], a worldwide competition series in computer-aided drug design. However, this approach depends on
the availability of reliable 3D molecular structures.

Alternatively, a self-supervised learning (SSL) strategy can be used to pre-train an autoencoder model
that can produce latent space vectors as molecular representations without 3D molecular structures. The
initial development of SSL was due to the need for natural language processing (NLP) \[26, 27\]. For example,
bidirectional encoder representations from transformers (BERT) is designed to pre-train deep bidirectional
transformer representations from unlabeled texts. \[26\] The techniques developed in understanding sequential
words and sentences in NLP have been used for understanding the fundamental constitutional principles of
molecules expressed as a simplified molecular-input line-entry system (SMILES) \[28\]. Unlabeled SMILES
strings can be considered as text-based chemical sentences and are used as inputs for SSL pre-training.
\[27, 29\] It is worth noting that the availability of large public chemical databases such as ZINC \[30\] and
ChEMBL \[31\] makes SSL a viable option for molecular representation generation. However, latent-space
representations ignore much stereochemical information, such as the dihedral angle \[32\] and chirality \[33\].
Additionally, latent-space representations lack specific physical and chemical knowledge about task-specific
properties. For example, van der Waals interactions can play a greater role than the covalent interactions in
many drug-related properties \[34\], and need to be considered in the description of these properties.

In this work, we introduce algebraic graph-assisted bidirectional transformer (AGBT) to construct new
molecular representations via combining the advantages of 3D element-specific weighted colored algebraic
graphs and deep bidirectional transformers. The element-specific weighted colored algebraic graphs generate
intrinsically low-dimensional molecular representations, called algebraic graph-based fingerprints (AG-FPs),
that significantly reduce the molecular structural complexity while retaining essentially physical/chemical
information and physical insight. \[24\] Deep bidirectional transformer (DBT) utilizes an SSL-based pre-
training process to learn fundamental constitutional principles from massive unlabeled SMILES data and
a fine-tuning procedure to further train the model with task-specific data. The resulting molecular fingerprint,
called bidirectional transformer-based fingerprints (BT-FPs), are latent-space vectors of the DBT.
The proposed AGBT model is applied to five benchmark molecular datasets involving quantitative toxicity.
and partition coefficient. [2, 13, 35, 36] Extensive validation and comparison suggest that the proposed AGBT model gives rise to some of the best predictions of molecular properties.

2 Results

Figure 1: Illustration of AGBT model. For a given molecular structure and its SMILES strings, AG-FPs are generated from element-specific algebraic subgraphs module and BT-FPs are generated from a deep bidirectional transformer module, as shown inside the dashed rectangle, which contains the pre-training and fine-tuning processes, and then finally completes the feature extraction using task-specific SMILES as input. Then the random forest algorithm is used to fuse, rank, and select optimal fingerprints (AGBT-FPs) for machine learning.

In this section, we present the proposed AGBT model and its results for molecular prediction on five datasets, i.e., LD50 dataset, IGC50 dataset, LC50 dataset, LC50DM dataset, and partition coefficient dataset. Table S1 lists the basic information of these five datasets and the CheMBL [31] dataset we used in the pre-training dataset. More descriptions of the datasets can be found in Section S1 of the Supplementary Information.

Algebraic graph-assisted deep bidirectional transformers (AGBT) As shown in Figure 1, the proposed AGBT consists of four major modules: AG-FP generator (i.e., the blue rectangles), BT-FP generator (i.e., the orange rectangles), random forest (RF)-based feature-fusion module (i.e., the green rectangle), and downstream machine learning module (i.e., the pink rectangle). For the graph fingerprint generation, we use element-specific multiscale weighted colored algebraic graphs to encode the chemical and physical interactions into graph invariants and capture 3D molecular structural information. The BT-FPs have created in two steps: an SSL-based pre-training step with massive unlabeled input data and a task-specific fine-tuning step. The task-specific fine-tuning step can be executed in two ways. The first way is merely to adopt the same SSL procedure to fine-tune the model with task-specific data and generate their BT-FPs. The other way is to utilize labels in task-specific data via a supervised learning (SL) procedure to fine-tune model and generate latent-space vectors of task-specific data, denoted as BTs-FPs (i.e, the orange vector). The random forest algorithm is used to rank the importance of fused AG-FP and BT-PF features and select an optimal set of AGBT-FPs of a fixed number of components. The downstream machine learning algorithms are fed with optimal features to achieve the best performance on four benchmark toxicity datasets.
We carry out our final predictions by using some standard machine learning algorithms, namely, gradient boosted decision tree (GBDT) and deep neural networks (DNNs), including single-task DNN (ST-DNN, Figure S7a) and multitask DNN (MT-DNN, Figure S7b). Our training follows the traditional pipeline.\cite{37} To evaluate the variance from machine learning predictions, we repeat our calculations 20 times on each set of parameters and use the average result as the final prediction. In this work, squared Pearson correlation coefficient ($R^2$) and root-mean-square error (RMSE) are used to assess the accuracy of predictions. Further details on our AGBT model are given in the Section 5 and Section S2 of the Supplementary Information.

**Toxicity prediction** Toxicity, a critical issue to consider in drug lead optimization, measures the degree to which a chemical compound can affect an organism adversely.\cite{2} Indeed, toxicity and side effect are responsible for more than half of drug candidate failures on their path to the market.\cite{38} The LC50DM set refers to the concentration of test chemicals in the water in milligrams per liter that cause 50% Daphnia Magna to die after 48 h. Its size is the smallest among the four datasets. Among its 353 molecules, 283 are used as a training set and the rest 70 as a test set.\cite{2} The small size leads to difficulties in building a robust prediction model. The overfitting issue poses a challenge to traditional machine learning methods if a large number of descriptors is used. MT-DNN is a method to extract information from data sets that share certain statistical distributions, which can effectively improve the predictive ability of models on small data sets\cite{2,13}. Based on the AGBT framework, we fuse AG-FPs and BT-FPs, i.e., BT-FPs with a supervised fine-tuning procedure for task-specific data. We applied MT-DNN in the downstream task and obtained $R^2=0.830$ and RMSE=0.743. As shown in Figure 2b, our model yields the best result, which is over 13% better than the previous best score of $R^2 = 0.733$. This result indicates the power of our method.

The IGC50 set is the second-largest toxicity set and its toxicity values range from 0.334 $-\log_{10}$ mol/L to 6.36 $-\log_{10}$ mol/L.\cite{2} As shown in Figure 2a, the $R^2$s from different methods fluctuate from 0.274 to 0.810. Karim et al.\ also studied IGC50 dataset, but their training set and test set are different from those of others\cite{2} and thus their results cannot be included in the present comparison. For our method, the $R^2$ of MT-DNN with AGBT-FP is 0.842, which exceeds that of all existing methods. It suggests that our AGDMT framework not only can overcome overfitting problem, but also is not sensitive to dataset size.

The oral rat LD50 set measures the number of chemicals that can kill half of the rats when orally ingested.\cite{40,35,36} This dataset is the largest set among the four sets with as many as 7413 compounds. However, a large range of values in this set makes it relatively difficult to predict.\cite{41} Gao et al.\ studied this problem using many 2D molecular fingerprints and various machine learning methods, which includes GBDT, ST-DNN, and MT-DNN. However, the prediction accuracy of the LD50 data set did not improve much. As shown in Table 1, the $R^2$ values for all existing methods range from 0.392 to 0.643. In our case, our method can achieve $R^2$ 0.671 and RMSE 0.554 log(mol/L), which are better than those from other existing methods.

LC50 dataset reports the concentration of test chemicals in water by milligrams per liter that cause 50% of fathead minnows to die after 96 hours.\cite{41,2} Wu et al.\ used physical information including energy, surface energy, electric charge and so on to construct molecular descriptors. These physical properties are related to molecular toxicity, achieving the prediction accuracy of $R^2$ 0.771. In this work, our AGBT-FPs with MT-DNN deliver the best $R^2$ of 0.776. We also test the performance of our BT-FPs, which achieve $R^2$ 0.783 with MT-DNN. As listed in Table 1, our model outperforms all other existing methods.

**Partition coefficient prediction** Partition coefficient, denoted $P$, derived from the ratio of the concentration of a mixture of two mutually insoluble solvents at equilibrium, measures the drug relevance of the compound as well as its hydrophobicity to the human bodies. The logarithm of this coefficient is denoted as $\log P$.\cite{41} The training set used for partition coefficient prediction includes 8199 molecules.\cite{41} A set of 406 molecules approved by the Food and Drug Administration (FDA) is used as organic drugs were used as the test set for $\log P$ prediction\cite{41} and its $\log P$ values range from -3.1 to 7.57. The comparison of different prediction methods for FDA molecular data set are listed in Table 1. It should be mentioned that the ALOGPS
Table 1: Comparison of the $R^2$ of various predicting methods on the LD50, LC50, and FDA Approved Small-Molecule data sets.

<table>
<thead>
<tr>
<th>Method</th>
<th>LD50 $R^2$</th>
<th>LC50 $R^2$</th>
<th>FDA $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGBT-FP</td>
<td>0.671</td>
<td>0.776/0.783</td>
<td>0.905</td>
</tr>
<tr>
<td>MACCS[17]</td>
<td>0.643</td>
<td>0.750</td>
<td>0.893</td>
</tr>
<tr>
<td>FP2[17]</td>
<td>0.631</td>
<td>0.745</td>
<td>0.893</td>
</tr>
<tr>
<td>HybridModel[39]</td>
<td>0.629</td>
<td>0.724</td>
<td>0.872</td>
</tr>
<tr>
<td>Daylight[17]</td>
<td>0.624</td>
<td>0.710</td>
<td>0.870</td>
</tr>
<tr>
<td>BESTox[42]</td>
<td>0.619</td>
<td>0.704</td>
<td>0.867</td>
</tr>
<tr>
<td>BTAMDL1[13]</td>
<td>0.605</td>
<td>0.694</td>
<td>0.857</td>
</tr>
<tr>
<td>Estate1[17]</td>
<td>0.605</td>
<td>0.686</td>
<td>0.848</td>
</tr>
<tr>
<td>Estate2[17]</td>
<td>0.589</td>
<td>0.678</td>
<td>0.847</td>
</tr>
<tr>
<td>ECFP[17]</td>
<td>0.586</td>
<td>0.662</td>
<td>0.838</td>
</tr>
<tr>
<td>Hierarchical[43]</td>
<td>0.578</td>
<td>0.626</td>
<td>0.819</td>
</tr>
<tr>
<td>Nearest neighbor[43]</td>
<td>0.557</td>
<td>0.609</td>
<td>0.815</td>
</tr>
<tr>
<td>FDA[43]</td>
<td>0.557</td>
<td>0.608</td>
<td>0.800</td>
</tr>
<tr>
<td>Pharm2D[17]</td>
<td>0.443</td>
<td>0.573</td>
<td>0.777</td>
</tr>
<tr>
<td>ERG[17]</td>
<td>0.392</td>
<td>0.528</td>
<td>0.771</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.348</td>
<td>0.491</td>
</tr>
</tbody>
</table>

* only BT-FP is used as input;

model established by Tetko et al.[45] can also be used in logP prediction, however, there is no guarantee that the training set of ALOGPS are independent of the test set and thus its result is not included in the comparison. As we can see from the Table 1, our AGBT$_s$-FPs with STDNN model produce the best $R^2$ of 0.905. The predicted result of AGBT$_s$-FPs with STDNN model for FDA data set are shown in Figure S8c. Notice that the neural network structure of the downstream STDNN model is still the same as the one in toxicity prediction. It proves robust of our AGBT$_s$-FPs.

None of all other existing methods can provide the best prediction simultaneously for all LD50, IGC50, LC50, and LC50DM. However, our AGBT framework delivers state-of-the-art performance on all four toxicity data sets as well as in the FDA partition coefficient data set. This indicates the power of our AGBT framework and its stable performance on datasets with different sizes and different molecular properties.

3 Discussion

In this section, we discuss how the AGBT model brings new insights to quantitative molecular property predictions, as well as the enhancement that algebraic graph-based fingerprints and deep bidirectional transformers-based fingerprints give rise to our proposed AGBT method.

Impact of algebraic graph descriptor Pre-trained on a large number of molecules, deep SSL-based molecular fingerprints could achieve high accuracy. Many deep learning-based molecular fingerprints have shown a better performance than conventional fingerprints. However, deep learning fingerprints, including our BT-FPs, are prone to the loss of molecular stereochemical information. Therefore, we propose the use of algebraic graph theory in association with our AGBT framework to retain stereochemical and physical information and enhance the performance of original BT-FPs. Moreover, in this work, we set the total number of molecular fingerprints after feature fusion to 512, and thus we only need to optimize one neural network architecture. Our AGBT model is an efficient framework for molecular property predictions.
Figure 2: a and b illustrate the comparison of the $R^2$ by various methods for IGC50 set and LC50DM set, respectively. AGBT$_s$-FP means a fine-tuning process is applied to AGBT-FP. The other results were taken from Refs. 2, 17, 39, 13, 43, 23. c, The bar charts illustrate the average $R^2$ of AGBT-FPs and BT-FPs with three machine learning algorithm for the IGC50 dataset. All points in the figure show the $R^2$ of the prediction from the 20 repeated experiments. d, The bar charts illustrate the average $R^2$ of AGBT-FPs and AGBT$_s$-FPs with three machine learning algorithm for the LC50DM dataset. All points in the figure show the $R^2$ of the prediction from the 20 repeated experiments. e, Visualization of LD50 set. The axes are the top three important features of AGBT-FPs. f, Predicted results of AGBT-FPs with MT-DNN model for IGC50 and LC50DM set, respectively. The box plots in each figure statistic the $R^2$ for 20 experiments. g, Variance ratios in the first two components from the principal component analysis (PCA) are used to visualize the four toxicity data sets. h, Variance ratios in the first two components from the principal component analysis (PCA) are used to visualize the four toxicity data sets.

Figure 2f shows the best prediction performance on the IGC50 and LC50DM datasets using the AGBT framework, namely, $R^2 = 0.842$ on IGC50 and $R^2 = 0.830$ on LC50DM. The orange bar at each point is the deviation of predicted toxicity with 20 experiments. For each experiment, $R^2$ was calculated and the distribution of $R^2$ is shown in subfigures. The performance on the LD50 and LC50 datasets are shown in Figure S8b. For LD50, IGC50, and LC50DM datasets, the best prediction results are obtained by the algebraic graph-assisted fingerprints. For the LC50 dataset, the AGBT-FP prediction of 0.776 is very close to the best performance of 0.783 obtained by BT-FPs. This indicates that the AGBT model can produce stable and robust prediction performance on various datasets. Moreover, For the IGC50 dataset, the $R^2$ of the toxicity predictions from the three machine learning algorithms, i.e., GBDT, ST-DNN, and MT-DNN, are shown in the bar plot of Figure 2c. All points in the figure show the $R^2$ of the prediction from the 20 repeated experiments. It is obvious that for the IGC50 dataset, AGBT-FP performs better than BT-FP with GBDT and MT-DNN. However, when ST-DNN is used as a predictor, AGBT-FP has higher fluctuations in its predictions, resulting in a worse average of $R^2$. AG-FPs and BT-FPs are produced from two different molecular fingerprint generators and have the dimensions of 1800 and 512, respectively. The fused molecular
Table 2: Comparison of Prediction Results of AGBT-FP and AGBTₙ-FP for the toxicity datasets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>AGBT-FP R²</th>
<th>RMSE</th>
<th>AGBTₙ-FP R²</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD50</td>
<td>0.671</td>
<td>0.554</td>
<td>0.612</td>
<td>0.606</td>
</tr>
<tr>
<td>IGC50</td>
<td>0.842</td>
<td>0.391</td>
<td>0.805</td>
<td>0.437</td>
</tr>
<tr>
<td>LC50</td>
<td>0.776</td>
<td>0.703</td>
<td>0.750</td>
<td>0.734</td>
</tr>
<tr>
<td>LC50DM</td>
<td>0.781</td>
<td>0.824</td>
<td>0.830</td>
<td>0.743</td>
</tr>
</tbody>
</table>

fingerprints, AGBT-FPs, return 512 components with heterogeneous information from AG-FPs and BT-FPs and require a longer training time.

For the IGC50 dataset, only 1434 molecular structures were used to train the AGBT model, leading to a high fluctuation in prediction. Similar situations are found in the LD50 dataset and LC50DM dataset, as shown in Figure S9. For the LC50 dataset, the best result is obtained with BT-FPs, but the result of AGBT-FPs also reaches $R^2$ 0.776, exceeding all the other existing methods. Therefore, the fusion of AG-FPs and BT-FPs improves the accuracy of predictions for most datasets. Molecular descriptors based on mathematics complement to data-driven latent space descriptors.

**Predictive power of fine-tuning strategies**  In this work, we develop two strategies in the fine-tuning stage: SSL and SL with task-specific data. It is found that SSL strategy (See Figure S3) performs better on LD50, IGC50, and LC50 data sets, as shown in Figure 2f and Figure S8, while SL strategy with task-specific data (See Figure S4) is the best for LC50DM dataset. The LC50DM dataset is the smallest set with only 283 molecules in its training set. Conventional methods cannot capture enough information from such a small dataset to achieve satisfactory results. In the AGBT model, the pre-training strategy with bidirectional transformer enables the model to acquire a general knowledge of molecules. During the fine-tuning phase, we further feed the model with four toxicity datasets with labels, and the labeled data guide the model to specifically extract toxin-related information from all the training data. Then we complement fine-tuning fingerprints with algebraic graph descriptors to ultimately enhance the robustness of the AGBT model and improve the performance on the LC50DM set ($r^2$=0.830, RMSE=0.743).

Figure 2d shows the performance of AGBT-FPs and AGBTₙ-FPs on the LC50DM dataset using three advanced machine learning methods. The bar charts show the $R^2$ of prediction results with three machine learning algorithms and the subplot displays the distribution of $R^2$ in 20 experiments. This figure shows that AGBTₙ-FPs have an excellent performance with all three machine learning algorithms, with $R^2$ values being 0.822 (GBDT), 0.815 (ST-DNN), and 0.830 (MT-DNN), respectively. This indicates that AGBTₙ-FPs can capture general toxin-related information during the sequential fine-tuning process. There is no significant difference among the three predictions based on GBDT, ST-DNN, and MT-DNN. In contrast, AGBT-FPs are derived from the model after self-supervised training. Their pre-training and fine-tuning processes do not involve any labeled data. The resulting prediction accuracies with GBDT and ST-DNN are quite low with $R^2$ being 0.587 and 0.659 respectively. Through the MT-DNN model, the performance of AGBT-FPs can be significantly improved from $R^2$ 0.587 to 0.781.

The above discussion indicates that SSL can acquire general molecular information without the guidance of labels. In downstream tasks, the MT-DNN model can also help to extract the task-specific information from related data. As for extremely small datasets, such as the LC50DM dataset (300 samples), the subsequent fine-tuning with a SL strategy is much more promising. The results of all four datasets using AGBT-FPs and AGBTₙ-FPs with MT-DNN model are shown in Table 2. For partition coefficient prediction, a major challenge of the test set (FDA) is that its structures are much more complex than that of the training set. So that the supervised learning in the fine-tuning procedure can give a better result, the value of $R^2$ for AGBT-FP and AGBTₙ-FP are 0.901 and 0.885 separately.
**Molecular representations and structural genes** In chemistry, the properties of molecules, such as toxicity, are often determined by some specific functional groups or fragments. Similar to biological genes, molecules have some determinants of their properties, which are called structural genes in this work. For some path-based fingerprints, such as FP2, a molecule is represented by 256 length vectors, each corresponding to a specific fragment. However, for molecular toxicity, it is difficult to achieve the best results from such a fingerprint, as shown in Figure 2a and b. The proposed AGBT-FP is a 512-dimensional fingerprint, with each dimension being a projection of various physical information about the molecule. In this section, we hope to characterize the key dimensions of AGBT-FPs to identify the structural genes.

Using a random forest algorithm, we performed a feature importance analysis of AGBT-FPs. As shown in Figure S10, for the LD50, IGC50, and LC50 datasets, the top three features in the feature importance ranking are all from algebraic graph-based descriptors. For the LC50DM dataset, the most important feature is from BT-FPs and the 2nd and 3rd important features are from AG-FPs. This implies that the multiscale weighted colored algebraic graph-based molecular descriptors contribute the most critical molecular features, which are derived from embedding specific physical and chemical information into graph invariants. The top three important features of the LD50 set are illustrated in Figure 2e, where each point represents a molecule and the toxicity is represented by the color. It is easy to find that the top three important dimensions in AGBT-FP, denoted as Feature 1, Feature 2, and Feature 3, divide the molecules into two groups: one can be distinguished by Feature 3 and the other is a linear combination of Feature 1 and Feature 2. This means the molecule can be classified by just three key dimensions (features), indicating that these three features, or structural genes, dominate the intrinsic characteristics of molecules. However, since predicting molecular toxicity is complex, it is difficult to directly distinguish the toxicity of each molecule in AGBT-FPs through the first three dimensions. Similarly, the visualizations for the IGC50, LC50, and LC50DM datasets can be seen in Figure S11.

We projected both AGBT-FPs and AGBT<br>**s**-FPs into an orthogonal subspace by principal component analysis (PCA). As shown in Figure 2g, the first two principal components of AGBT-FPs can roughly divide the data into two clusters and the molecules in the same cluster have similar toxicity. Similarly, the top two components of AGBT<br>**s**-FPs are given in Figure 2h. Along the direction of the first principal components, the molecular data can be well clustered according to the toxicity, with low toxic molecules on the left (green) and higher toxic molecules on the right (red). It indicates these two molecular fingerprints contain very different information. As shown in Figure S12, for AGBT-FPs we need 112 components to explain 90% of the variance, while for AGBT<br>**s**-FPs we only need 48 components. The top two principal components of AGBT-FPs are just explaining 9% and 8% of the variance, which indicates that, since there is no labeled data to train the model, the generated AGBT-FPs represent general information about the molecular constitution rather than specific molecular properties. The first two components for AGBT<br>**s**-FPs can explain 40% and 13% of the variance respectively, which indicates that by using SL-based fine-tuning training, the model can effectively capture task-specific information.

The AGBT<br>**s**-FP model performs better in predicting specific properties because the labeled data are used to train the model during fine-tuning. It should be noted that some molecular information irrelevant to that particular property might be lost in this way. This strategy leads to better results for some datasets with minimal data, such as LC50DM, whose small amount of data is not enough to effectively obtain property-specific information in downstream tasks. However, if more downstream data are available, such as LD50, IGC50, and LC50, downstream machine learning methods can also derive property-specific information from general molecular information. For example, AGBT-FPs perform better on LD50, IGC50, and LC50 datasets.

**4 Conclusion**

Despite many efforts in the past decade, accurate and reliable prediction of numerous molecular properties remains a challenge. Recently, deep bidirectional transformers have become a popular new approach
in molecular science for their ability to extract fundamental constitutional information of molecules from
massive self-supervised learning (SSL). However, they neglect crucial stereochemical information. The alge-
braic graph is effective in simplifying molecular structural complexity but relies on the availability of three-
dimensional structures. We propose an algebraic graph-assisted bidirectional transformers (AGBT) model
for the prediction of toxicity and partition coefficient. Specifically, element-specific multiscale weighted col-
ored algebraic subgraphs are introduced to characterize crucial physical/chemical interactions. Moreover, for
small datasets, we introduce a supervised fine-tuning procedure to the standard pre-trained SSL to focus on
task-specific information. These approaches are paired with random forest, gradient boosted decision trees,
multitask deep learning, and deep neural network algorithms in AGBT. We demonstrate that the proposed
AGBT achieves $R^2$ values of 0.671, 0.842, 0.783, 0.830, and 0.905 on LD50, IGC50, LC50, LC50DM, and
FDA log$P$ dataset, respectively, which are the best available predictions. Our model can be easily extended
to the prediction of other molecular properties. Our results show that the proposed AGBT is a new and
powerful framework for studying various properties of small molecules in drug discovery and environmental
sciences.

5 Methods

![Figure 3: Illustration of weighted colored element-specific algebraic graphs.](image)

a. The molecular structure of 2-Trifluoroacetyl. b and c give a traditional graph representation and a colored graph representation respectively. d. Illustration of the process of decomposing a colored graph into element specific CC, FO, and CH subgroups. e. Illustration of weighted colored element-specific subgraph $G_{SH}$, its adjacency matrix, and Laplacian matrix.

Algebraic graph-based molecular fingerprints (AG-FPs) Graph theory can encode the molecular
structures from a high-dimensional space into a low-dimensional representation. The connections between
atoms in a molecule can be represented by graph theory, as shown in Figure 3a and b. However, ignoring the
quantitative distances between atoms and the different atomic types in traditional graphs will result in the
loss of critical chemical and physical information about the molecule. Element-specific multiscale weighted
colored graph representations can quantitatively capture the patterns of different chemical aspects, such as
van der Waals interactions and hydrogen bonds between different atoms [24]. Figure 3c illustrates a colored
graph representation, which captures the element information by using colored vertices and different edges
are corresponding to different pairwise interaction in the molecule. Moreover, the algebraic graph features
are easily obtained from the statistics of the eigenvalues of appropriated graph Laplacians and/or adjacency
matrices [24].
As shown in Figure 3d, for a given molecule, we first construct element-specific colored subgraphs using selected subsets of atomic coordinates as vertices,

\[ V = \{ (\mathbf{r}_i, \alpha_i) | \mathbf{r}_i \in \mathbb{R}^3; \alpha_i \in \mathcal{E}; i = 1, 2, ..., N \} \]  

(1)

where \( \mathcal{E} = \{ \text{H, C, N, O, S, P, F, Cl, Br,...} \} \) is a set of commonly occurring element types for a given dataset.

And \( i \)th atom in a \( N \)-atom subset is labeled both by its element type \( \alpha_i \) and its position \( \mathbf{r}_i \). We denote all the pairwise interactions between element types \( \mathcal{E}_{k_1} \) and \( \mathcal{E}_{k_2} \) in a molecule by fast-decay radial basis functions

\[ W = \{ \Psi(||\mathbf{r}_i - \mathbf{r}_j||; \eta_{k_1}, \eta_{k_2})|\alpha_i = \mathcal{E}_{k_1}, \alpha_j = \mathcal{E}_{k_2}; i, j = 1, 2, ..., N; ||\mathbf{r}_i - \mathbf{r}_j|| > r_i + r_j + \sigma \} \]  

(2)

where \( ||\mathbf{r}_i - \mathbf{r}_j|| \) is the Euclidean distance between \( i \)th and \( j \)th atoms in a molecule, \( r_i \) and \( r_j \) are the atomic radii of \( i \)th and \( j \)th atoms, respectively, and \( \sigma \) is the mean standard deviation of \( r_i \) and \( r_j \) in the dataset.

Figure 3e gives the illustration of Laplace and adjacency matrices based on the weighted colored subgraph.

For the prediction of toxicity, van der Waals interactions are much more critical than covalent interactions and thus the distance constraint \( (||\mathbf{r}_i - \mathbf{r}_j|| > r_i + r_j + \sigma) \) is used to exclude covalent interactions. In biomolecules, we usually choose generalized exponential functions or generalized Lorentz functions as \( \Psi \), which are weights between graph edges [46]. Here, \( \eta_{k_1}, \eta_{k_2} \) in the function is a characteristic distance between the atoms and thus is a scale parameter. Therefore, we generate a weighted colored subgraph \( G(V, W) \).

In order to construct element-specific molecular descriptors, the multiscale weighted colored subgraph rigidity is defined as

\[ \text{RI}^G(\eta_{k_1}, \eta_{k_2}) = \sum_i \mu_i^G(\eta_{k_1}, \eta_{k_2}) = \sum_i \sum_j \Psi(||\mathbf{r}_i - \mathbf{r}_j||; \eta_{k_1}, \eta_{k_2}) \]  

\[ \alpha_i = \mathcal{E}_{k_1}, \alpha_j = \mathcal{E}_{k_2}; ||\mathbf{r}_i - \mathbf{r}_j|| > r_i + r_j + \sigma \]  

(3)

where \( \mu_i^G(\eta_{k_1}, \eta_{k_2}) \) is a geometric subgraph centrality for the \( i \)th atom.[47] The summation over \( \sum_j \mu_i^G(\eta_{k_1}, \eta_{k_2}) \) represents the total interaction strength for the selected pair of element types \( \mathcal{E}_{k_1} \) and \( \mathcal{E}_{k_2} \), which provide the element-specific coarse-grained description of molecular properties. By choosing appropriate element combinations \( k_1 \) and \( k_2 \), the characteristic distance \( \eta_{k_1}, \eta_{k_2} \), and subgraph weight \( \Psi \), we finally construct a family of element-specific, scalable (i.e., molecular size independent), multiscale geometric graph-based molecular descriptors [24].

To generate associated algebraic graph fingerprints, we construct corresponding graph Laplacians and/or adjacency matrices. For a given subgraph, its matrix representation can provide a straightforward description of the interaction between subgraph elements. To construct a Laplacian matrix, we consider a subgraph \( G_{k_1, k_2} \) for each pair of element types \( \mathcal{E}_{k_1} \) and \( \mathcal{E}_{k_2} \) and define an element-specific weighted colored Laplacian matrix \( L(\eta_{k_1}, \eta_{k_2}) \) as [24]

\[ L_{ij}(\eta_{k_1}, \eta_{k_2}) = \begin{cases} -\Psi(||\mathbf{r}_i - \mathbf{r}_j||) & \text{if } i \neq j, \alpha_i = \mathcal{E}_{k_1}, \alpha_j = \mathcal{E}_{k_2} \text{ and } ||\mathbf{r}_i - \mathbf{r}_j|| > r_i + r_j + \sigma; \\ -\sum_j L_{ij} & \text{if } i = j \end{cases} \]  

(4)

Mathematically, the element-specific weighted Laplacian matrix is symmetric, diagonally dominant, and positive semi-definite, and thus all the eigenvalues are non-negative. The first eigenvalue of the Laplacian matrix is zero because the summation of every row or every column of the matrix is zero. The first non-zero eigenvalue of \( L_{ij}(\eta_{k_1}, \eta_{k_2}) \) is the algebraic connectivity (i.e., Fiedler value). Furthermore, the rank of the zero-dimensional topological invariant, which represents the number of the connected components in the graph, is equal to the number of zero eigenvalues of \( L_{ij}(\eta_{k_1}, \eta_{k_2}) \). A certain connection between geometric graph formulation and algebraic graph matrix can be defined by:

\[ \text{RI}^G(\eta_{k_1}, \eta_{k_2}) = \text{Tr}L(\eta_{k_1}, \eta_{k_2}), \]  

(5)

where \( \text{Tr} \) is the trace. Therefore, we can directly construct a set of element-specific weighted colored Laplacian matrix-based molecular descriptors by the statistics of nontrivial eigenvalues \( \{\lambda_i^L\}_{i=1,2,3,...} \), i.e., summation,
minimum, maximum, average, and standard deviation of nontrivial eigenvalues. Note that the Fiedler value is included as the minimum.

Similarly, an element-specific weighted adjacency matrix can be defined by

\[
A_{ij}(\eta_{k_1k_2}) = \begin{cases} 
\Psi(\|r_i - r_j\|) & \text{if } i \neq j, \alpha_i = \mathcal{E}_{k_1}, \alpha_j = \mathcal{E}_{k_2} \text{ and } \|r_i - r_j\| > r_i + r_j + \sigma; \\
0 & \text{if } i = j
\end{cases}
\]

Mathematically, adjacency matrix \(A_{ij}(\eta_{k_1k_2})\) is symmetrical non-negative matrix. The spectrum of the proposed element-specific weighted colored adjacency matrix is real. A set of element-specific weighted labeled adjacency matrix-based molecular descriptors can be obtained by the statistics of \(\{\lambda^A_k\}_{k=1,2,3,...}\), i.e., summation, minimum, maximum, average, and standard deviation of all positive eigenvalues. To predict the properties of a molecule, graph invariants, such as the eigenvalue statistics of the above matrix, can capture topological and physical information about the molecules, which is named algebraic graph fingerprints (AG-FPs). Detailed parameters of the proposed algebraic graph model can be found in Section S2.3 of the Supplementary Information.

**Bidirectional transformer fingerprints (BT-FPs)** Unlike RNN-based models, deep bidirectional transformer (DBT) is based on the attention mechanism and it is more parallelizable to reduce the training time with massive data. Based on the DBT architecture, Devlin et al. introduced a new representation model called bidirectional encoder representations from transformer (BERT) for natural language processing. There are two tasks involved in BERT, masked language learning and consecutive sentences classification. Masked language learning uses a partially masked sentence (i.e., words) as input and employs other words to predict the masked words. The consecutive sentences classification is to classify if two sentences are consecutive. In the present work, the inputs of the deep bidirectional transformer are molecular SMILES strings. Unlike the sentences in traditional BERT for natural language processing, the SMILES strings of different molecules are not logically connected. However, we train the bidirectional encoder from the transformer to recover the masked atoms or functional groups.

Because a molecule could have multiple SMILES representations, we first convert all the input data into canonical SMILES strings, which provide a unique representation of each molecular structure. Then, a SMILES string is split into symbols, e.g., C, H, N, O, =, Br, etc., which generally represent the atoms, chemical bonds, and connectivity, see Table S2 for more detail. In the pre-training stage, we first select a certain percentage of the input symbols randomly for three types of operations: mask, random changing, and no changing. The purpose of the pre-training is to learn fundamental constitutional principles of molecules in a SSL manner with massive unlabeled data. A loss function is built to improve the rate of correctly predicted masked symbols during the training. For each SMILES string, we add two special symbols, <s> and </s>. Here, <s> means the beginning of a SMILES string and </s> is a special terminating symbol. All symbols are embedded into input data of a fixed length. A position embedding is added to every symbol to indicate the order of the symbol. The embedded SMILES strings are fed into the BERT framework for further operation. Figure S2 shows the detailed process of pre-training procedure. In our work, more than 1.9 million unlabeled SMILES data from CheMBL are used for the pre-training so that the model learns basic “syntactic information” about SMILES strings and captures global information of molecules.

Both BT-FPs and BT-FPs are created in the fine-tuning training step, which further learns the characteristics of task-specific data. Two types of fine-tuning procedures are used in our task-specific fine-tuning. The first type is still based on the self-supervised learning (SSL) strategy, where the task-specific SMILES strings are used as the training inputs, as shown in Figure S3. To accurately identify these task-specific data, only the ‘mask’ and ‘no changing’ operations are allowed in this fine-tuning. The resulting latent-space representations are called BT-FPs.

The second fine-tuning procedure is based on a supervised learning (SL) strategy with labeled task-specific data. As shown in Figure S4, when dealing with multiple datasets with cross-dataset correlations,
such as four toxicity datasets in the present study (Table S4), we make use of all the labels of four datasets
to tune the model weights via supervised learning before generating the latent-space representations (i.e.,
BT-s-FPs), which significantly strengthens the predictive power of the model on the smallest dataset.

In our DBT, an input SMILES string has a maximal allowed length of 256 symbols. During the training,
each of the 256 symbols is embedded into a 512-dimensional vector that contains the information of the whole
SMILES string. In this extended 256×512 representation, one can, in principle, select one or multiple 512-
dimensional vectors to represent the original molecule. In our work, we choose the corresponding vector of the
leading symbol <s> of a molecular SMILES string as the bidirectional transformer fingerprints (BT-FPs or
BT-s-FPs) of the molecule. In the downstream tasks, BT-FPs or BT-s-FPs are used for molecular property
prediction. Detailed model parameters can be found in Section S2.5 of the Supplementary Information.

Data and model availability

The pre-training dataset used in this work is CheMBL26, which is available at chembl.gitbook.io/
chembl-interface-documentation/downloads. To make sure the reproducibility of this work, the toxicity
datasets and partition coefficient used in this work are also available at weilab.math.msu.edu/Database/.
The overall models and related code have been released as an open-source code in the Github repository:
github.com/ChenDdon/AGBTcode.

Acknowledgment

The research was financially supported by the National Key R&D Program of China (2016YFB0700600).
The work of Gao and Wei was supported in partial by NSF Grants DMS1721024, DMS1761320, IIS1900473,
NIH grants GM126189 and GM129004, Bristol-Myers Squibb, and Pfizer.

Supporting Information

The Supporting Information is available on the website at xxxxxx

References

[1] Li Di and Edward H Kerns. Drug-like properties: concepts, structure design and methods from ADME


[4] Evgeny Putin, Arip Asadulaev, Quentin Vanhaezen, Yan Ivanenkov, Anastasia V Aladinskaya, Alex
Aliper, and Alex Zhavoronkov. Adversarial threshold neural computer for molecular de novo design.


