Topological Distance-Based Electron Interaction Tensor: A Novel Molecular Structure Representation to Bridge Convolutional Neural Network Studies in Computer Vision to Drug-Like Compound Datasets

Hyun Kil Shin (hyunkil.shin@kitox.re.kr)
Institute of Toxicology

Research article

Keywords: Electron configuration, Density matrix, Deep neural network, Deep learning, Molecular descriptors

Posted Date: July 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-709747/v1

Abstract

Owing to the success achieved by deep learning, researchers are exploringthe application of deep learning in drug discovery to improve the accuracy of prediction models. Significant performance improvement has been achieved by diverse convolutional neural network (CNN) models in computer vision, and the preparation of an input format suitable for CNN is one of the major questions required to be answered in order to harness the advancements in using CNNs for chemical data. It was reported that the models achieved improvement in prediction accuracy, in deep learning studies on molecular structure data; however, the improvement was insufficient from an industry perspective. Furthermore, a recent study suggested that conventional machine learning models can outperform deep learning models on chemical data. As only a limited number of feature calculation methods are available for molecules in deep learning studies, it is crucial to develop more methods to calculate features appropriate for deep learning model development.

A topological distance-based electron interaction (TDEi) tensor has been introduced in this study to transform a molecular structure into image-like 3D arrays based on electron interactions (Eis) within a molecule. The prediction accuracy of the CNN model with the TDEi tensor was tested with four datasets: MP (275,131), Lipop (4,193), Esol (1,127), and Freesolv (639), and the models achieved desirable prediction accuracy. Ei is the fundamental level of information that determines the chemical properties of a molecule. Feature space variation was visualized by taking outputs from the middle of the CNN architecture as the CNN model exhibited outstanding performance in automatic feature extraction.The correlation between features from the CNN, and target endpoints was strengthened as outputs were extracted from the deeper layer of the CNN.

Introduction

Diverse *in silico* models have been used in drug discovery projects to reduce the time and cost required for drug development [1, 2]. Quantitative structure–activity relationship (QSAR) is a type of computational model that predicts the physicochemical properties, potency, pharmacokinetic properties, and safety of drug candidates from only their molecular structures [3]. Even though QSAR models have been successful in filtering out poor molecules in the early phase of drug discovery, the models have failed to discover good drug candidates based on their prediction outcomes alone, which implies that the prediction accuracy of QSAR models is not satisfactory [4]. Most QSAR models have been developed using machine learning (ML) algorithms; however, deep learning algorithms have recently been used in QSAR model development to improve its prediction accuracy [5]. Typically, the first attempt to apply deep learning in different fields involves the application of identical hyperparameters and architectures studied in other deep learning studies. As convolutional neural networks (CNNs) have achieved outstanding performance, previous CNN studies were followed by diverse studies; however, it was not easy to do so in chemistry because chemical data has a completely different structure from image data.

The application of advanced deep learning algorithms to molecular structures requires the development of novel descriptors [6]. Fully connected architectures of artificial neural networks, such as feedforward neural network (FNN) models are commonly used in conventional QSAR model development. In FNN, molecular descriptors calculated from the molecular structures are the model input. As molecular descriptors were calculated in 1D vector format, FNN was an appropriate architecture for QSAR model development [7]. Thus, when deep learning was applied to molecular structures, the easiest approach was to use an FNN with various hidden layers, on 1D descriptor vectors. Although many complicated neural network architectures have been developed, these architectures cannot be used with a 1D descriptor vector as an input. Thus, more research is required to develop novel feature tensors for molecular structure representation, that are suitable for application of advanced neural network architecture, in order to exploit advancement made by a wide range of deep learning research.

Graph neural networks (GNNs) have been widely used to train a model directly with the molecular structure because molecular geometry is considered as a graph, chemical bonds as edges, and atoms as nodes, [8]. In the GNN, the feature of each atom and their neighbor information are used as descriptors [9]. The feature for each atom represents the character of the atom based on its microenvironment, and neighbor information is described by the distance between the atoms, or connectivity features, such as chemical bond features [10]. In natural language processing, string data are one-hot encoded, and word embeddings are used to encode the tensor, before the model training. The simplified molecular input line entry system (SMILES) code is a string format representation for molecular structures, [11] and is broadly used in public databases. As molecular structures can be represented by SMILES, one-hot encoding on each symbol of SMILES [12], or SMILES-embeddings [13], was used to generate a matrix, representing each molecule as an input for the CNN architecture.

Most deep learning studies claim that the application of deep learning algorithms improved prediction accuracy in molecular property prediction [6, 9, 14–16]. However, Jian et al. experimented with diverse datasets used in deep learning model development studies to compare the prediction accuracy between deep learning, and feature-based ML models. This study showed that the feature-based ML models outperformed the deep learning models in terms of prediction accuracy [17]. Moreover, ML models do not require demanding computation in the training process; therefore, the feature-based ML algorithm is a much more efficient way of developing the model. Unfortunately, the volume of datasets used in deep learning studies on molecular structures is much smaller than that in computer vision. Few datasets contain a large number of compounds; however, their labels are significantly unbalanced [18], which hinders the appropriate training of deep learning models [19]. Given that deep learning models in computer vision achieved significantly improved prediction accuracy owing to the application of deep learning algorithms and a huge amount of data, deep learning models on molecular structures are still not appropriately validated because of the small size of available datasets. This creates

a hindrance as mining of molecular structure data is one of the paramount tasks that is required to be performed for the meaningful application of deep learning in chemistry.

In this study, a topological distance-based electron interaction (TDEi) tensor has been developed as a novel molecular representation for CNN architecture, in order to transform molecular structures into image-like 3D arrays. In the TDEi tensor calculation, each atom was represented by the electron configuration of atoms, and the number of interactions between each atomic orbital was calculated to prepare the TDEi tensor. Molecular properties are calculated in quantum mechanics (QM) based on the interaction of electrons within a molecule according to the distance between atoms; similarly, the TDEi tensor was designed based on the assumption that CNN can extract significant features from Eis, through weights in filters to predict molecular properties. The TDEi tensor was designed to be adjustable as per the size of the data and the complexity of the molecular structure, by changing the electron configuration vector and topological distance channel to avoid the generation of a sparse tensor. The CNN model developed with the TDEi tensor achieved desirable prediction accuracy, and analysis of the features processed by the CNN filters revealed that extracted features achieved higher correlation with target properties, when the features were obtained from the deeper layer.

Methods

TDEi tensor calculation

Definition of an electron configuration vector

TDEi tensor was calculated based on the electron configuration (EC) of atoms in a molecule. The EC vector was defined in a previous study by giving a zero, for each unoccupied atomic orbital (AO) and one, for each occupied AO with two different electron spins marked by positive and negative signs [20]. The EC vector can be varied by combining degenerated AOs or electron spins because these electrons possess identical levels of energy. Given that the size of data in chemistry is generally much smaller than the size of data in computer vision, thus reducing sparse information and condensing the feature size are significant for efficient model training and accurate prediction model development. Therefore, sparse information or invariable features were integrated to condense the information without loss. Such information condensation was successful in the prediction model development with a small dataset [21]. The possible variations of EC vectors are summarized in Fig. 1.

Transformation of molecular structure into tensor shape

In QM calculations, molecular orbitals were calculated through a linear combination of AOs, and coefficients for each AO were estimated during the calculation in the density matrix, which is a diagonal matrix whose rows and columns are all AOs in a molecule. As a molecule can be translated into a matrix format with AO information, the concept was used with adaptations in the TDEi tensor design. First, the Ei matrix was designed with rows and columns with a fixed size of the EC vector such that every input has an identically sized matrix. The size of the density matrix is dependent on the number of AOs in the molecule; however, the input shape must be equal regardless of the size of the molecule in order to input them into the CNN. As the size of the Ei matrix was fixed, molecular geometry differences were lost in the matrix because the EC vector was solely based on the composition of molecules; however, molecules with identical compositions can have different molecular geometries. To consider the difference in the topological structure of a molecule, a matrix was generated based on the topological distance within a molecule.

The matrix for the topological distance 0 is shown in Fig. 2A. Topological distance 0 means the atom itself; thus, the EC vector of a C atom was multiplied by a row and a column in order to calculate the number of interactions between all electrons within the C atom. The topological distance 0 matrix is the sum of the Ei matrices for all atoms within a molecule. The matrix for the topological distance 1D is explained in Fig. 2B. Further, the pairs of atoms within the molecule were considered to calculate the Eis between them. As the Ei matrix was calculated for all atoms in a molecule, the Ei matrices between the C and N atoms were calculated twice in the example. Therefore, they were divided by two, and all Ei matrices for topological distance 1D were added. All Ei matrices with topological distances greater than 1 were calculated, as explained in Fig. 2B. In QM calculations, chemical bond information is not required, thus the physical distance between the atoms was measured based on the coordination of each atom. The precise 3D geometry of a molecule should be prepared to accurately calculate the physical distance between atoms. However, 3D geometry optimization requires an expensive computational cost, and it is not suitably accurate. Hence, 2D structural information alone was used, and the topological distance between atoms was used in the TDEi tensor calculation. The GetDistanceMatrix function implemented in RDKit was used to obtain the topological distance of atoms within a molecule once hydrogen was added to it.

The Ei matrix can be calculated from the topological distance. In the example molecule (Fig. 3), atom pairs existed up to a topological distance of 4D. The Ei matrices from atom pairs with greater topological distance can be calculated if the size of the molecule increases. When Ei matrices were prepared from a range of predetermined topological distances, they were concatenated to form the TDEi tensor (Fig. 4). As the TDEi tensor size can be varied based on the size of the EC vector and the topological distance, it can be flexibly adjusted according to the size of the data or the diversity of chemical space.

Datasets

In this study, four datasets were selected for the regression tasks: melting point (MP), water solubility (Esol), octanol/water distribution coefficient (Lipop), and hydration free energy (Freesolv). There are various publicly available datasets for the classification problem that are used in deep

learning model development studies [9, 10, 13, 17] such as human immunodeficiency virus replication inhibition (HIV), human ß-secretase 1 inhibition (BACE), blood-brain barrier penetration (BBBP), toxicity in clinical trials (ClinTox), drug adverse reactions (SIDER), biological targets screened in Tox21 and ToxCast, and PubChem BioAssay data (MUV); however, they were not used in this study because labels in the datasets were seriously imbalanced, whether they were binary, or multiple classification tasks.

MP was obtained from the study by Igor V. Tetko et al., in which 275,131 compounds were extracted using their normal melting point values by mining patent documents [22]. The dataset was divided into training, validation, and external test sets by a random split, in a ratio of 8:1:1. It is the largest publicly available, labeled chemical dataset. ESOL, Freesolv, and Lipop were obtained from the study by Jian et al. [17]. Because the datasets were already divided into the three given categories by the authors, I used them as such. The number of data and the range of the endpoint are listed in Table 1, and the chemical space of the datasets were plotted to verify the structural diversity in the training, validation, and external test sets (Fig. 5).

Table 1
Datasets for model building

Endpoints	Totalnum.	Trainining set		Validation set		Test set	
		Num.	Range	Num.	Range	Num.	Range
MP	**275,131**	220,104	-199.0 to 517.0	27,513	-157.15 to 420	27,514	-185.18 to 438.65
Lipop	**4,193**	3,354	-1.5 to 4.5	420	-1.42 to 4.49	419	-1.17 to 4.5
Esol	**1,127**	901	-11.6 to 1.58	113	-9.16 to 0.94	113	-8.40 to 1.02
Freesolv	**639**	511	-25.47 to 3.16	64	-9.76 to 3.43	64	-20.52 to 3.12

TDEi parameter search

Because TDEi tensors can have different shapes based on EC vectors and topological distances, experiments were performed to determine which option can provide the best prediction outcome for each dataset. Six different EC vectors and topological distances from zero to four were used in the TDEi tensor calculation. Models were developed with identical hyperparameters, such as Relu for the activation function, RMSprop for the optimizer, and identical CNN architecture. Epoch was applied differently because of differences in the size of data, such as 15 for MP and 100 for other datasets, in this preliminary search.

Model development and validation

CNN was used in this study for model development, through Tensorflow 2 [23], and the network architecture was designed based on VGGNet as a backbone with modifications, such as (1) the size of the initial filter channel was reduced by half from that is 64 to 32, (2) the filter shape was reduced from three by three, to two by two, (3) average pooling was used to minimize information loss, and (4) a convolutional layer was applied once before the pooling layer (Fig. 6). A grid search was performed on the CNN architectures, activation functions, and epoch numbers to obtain the finest hyperparameters for model development. Model training was conducted using the NEURON system of the National Supercomputing Center of South Korea (https://www.ksc.re.kr/eng/resource/neuron).

The prediction accuracy of the model was measured using four metrics: mean absolute error (MAE), normalized mean absolute error (NMAE), R square (R^2), and Spearman's rank correlation coefficient (S_r).

$$\mathrm{MAE} = \left|\frac{y_{pred} - y_{obs}}{n}\right|$$

$$\mathrm{NMAE} = \frac{MAE}{max(y_{obs}) - min(y_{obs})}$$

$$R^2 = 1 - \frac{\sum(y_{pred} - y_{obs})^2}{\sum(y_{obs} - \overline{y_{obs}})^2}$$

$$S_r = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

where, y_{pred} is a model's prediction value, y_{obs} is an observation value, n is the number of compounds, $\overline{y_{obs}}$ is the average of observation values, and d is the difference between the ranks of each compound. The prediction model with R^2 higher than 0.6, on the external test set, is considered as an accurate model. Even though the model did not achieve $R^2 > 0.6$, it was still able to make an accurate prediction of the target value when NMAE was

less than 10%. As the QSAR model was used in the prioritization of compounds, S_r higher than 0.6 implies that the model's prediction is valid and useful in relative comparison of chemicals, even if NMAE is over 10% [24].

Model analysis

CNN models were developed over four datasets; however, the CNN model developed with MP alone was analyzed because this model was trained with the largest dataset. In the QSAR study, the capacity to separate different molecular structures was the most significant point in the descriptor design to facilitate valid predictions using the descriptor. As the CNN extracted features from the TDEi tensor, the performance of these features in distinguishing compounds along the MP was examined. The final model outputs were extracted from the middle of the CNN before the final prediction value was calculated. Principal component analysis (PCA) was used to project extracted features into 2D space, and extracted feature variation was examined to determine how the model correlates with the extracted features for the prediction of target values.

Results And Discussion

TDEi parameter search

The TDEi parameter search results are presented in the supplementary tables: MP (Table S1), Lipop (Table S2), Esol (Table S3), and Freesolv (Table S4). As the TDEi tensor can be varied by changing the EC vectors and topological distances, the influence of different options in the TDEi tensor on prediction accuracy was analyzed. In the Lipop, Esol, and Freesolv datasets, a dramatic decrease in prediction accuracy was observed regardless of topological distance, when the EC vector size was reduced from full, and full bit strings were condensed to EC vectors without degenerated AOs and spin numbers, whereas the MP model showed a mild decrease in prediction accuracy. The full EC vector achieved the highest accuracies in MP, Lipop, and Esol, whose data size was greater than 1,000, and the condensed full EC vector in Freesolv, whose data size was less than 1,000. Thus, it appeared that the EC vector size in the TDEi tensor reduced if the models were trained with a smaller size of data. According to this experiment, full or condensed full EC vectors should be used in the development of a CNN model for drug-like compounds.

Desirable prediction accuracy was achieved in MP when the TDEi tensor with full EC vector and topological distance 3D was used, and a further increase in topological distance did not lead to a significant improvement in the accuracy. In the other three datasets, the TDEi tensor with topological distance 2D achieved the highest accuracy. In the MP model, prediction accuracy gradually increased as the topological distance increased until 3D, whereas it fluctuated in other datasets. Instability in prediction accuracy in the three datasets implied that the training process of the deep learning model could be stabilized if larger datasets were used. Because the prediction accuracy of the model varied significantly based on the topological distance of the TDEi tensor in each dataset, a preliminary search was required to select the most suitable topological distance for the datasets and the target endpoint.

Based on this preliminary study, the most suitable TDEi options were selected for each dataset, such as topological distance 3D with full EC vector for MP, 2D with full EC vector for Lipop and Esol, and 2D with condensed full EC vector for Freesolv. In this study, experiments were performed on small drug-like compounds. Given that models developed for drug-like compounds showed poor prediction accuracy for molecules whose structural diversity was dissimilar to the drug chemical space [25, 26], TDEi tensor options should be examined before model development if the structural diversity of datasets is different from that of drug-like molecules.

Model prediction accuracy

In CNN model development, TDEi tensors with the leading results in the preliminary search have been used for each dataset (Table 2), and the goodness-of-fit of each model is shown in Fig. 7. R^2 of the MP model was 0.565 for the external data set. Prediction errors between 0 and 400°C were relatively high, as data points were widely distributed across the best-fit line (Fig. 7A). However, NMAE = 5.27% indicates that prediction values were accurate on average, and S_r = 0.729 implies that the model correctly ordered the molecules as per normal melting point values. It was arduous to make precise predictions for LogP, as the Lipop model achieved an R^2 of 0.516 on the external test set, and NMAE was 10.93%. Even though most of the data points were close to the best-fit line, some of the compounds that were located away from the best-fit line were predicted inaccurately (Fig. 7B). The models developed by Esol and Freesolv achieved high R^2, and Figs. 7C and 7D display the fact that the model achieved goodness-of-fit. CNN models were developed with more convolutional layers to examine whether the prediction accuracy would improve significantly. However, adding more convolutional layers or increasing the number of nodes within fully connected layers did not lead to a meaningful improvement in prediction accuracy. Thus, CNN models with deeper layers, such as ResNet and Inception, were not applied. This was similar to the previous study where increasing the weights within the neural network architecture did not always improve prediction accuracy [20]. Moreover, it is important to search for a model architecture with the minimum number of weights and the highest prediction accuracy, because the use of an excessive number of weights in the model could induce false positives in prediction outcomes [17].

Table 2
Best prediction results for each endpoint

Endpoints	TD	EC vector	Training set				Validation set				Test set			
			MAE	NMAE	R^2	S_r	MAE	NMAE	R^2	S_r	MAE	NMAE	R^2	S_r
MP	3D	full	31.959	4.46%	0.584	0.742	33.352	5.78%	0.553	0.720	32.874	5.27%	0.565	0.729
Lipop	2D	full	0.450	7.50%	0.726	0.867	0.654	11.07%	0.525	0.740	0.620	10.93%	0.516	0.724
Esol	2D	full	0.346	2.63%	0.948	0.976	0.557	5.52%	0.872	0.922	0.465	4.94%	0.896	0.951
Freesolv	2D	full (cond.*)	0.425	1.48%	0.968	0.989	0.729	5.53%	0.875	0.942	0.563	2.38%	0.961	0.979

*cond.: condensed

In QSAR modeling, datasets were collected from a wide range of studies in which experimental values were measured using different experimental protocols. This difference is a source of experimental error in the target dataset [27]. As the model aims to predict the endpoints with their experimental noises, understanding the experimental errors of the dataset is of great aid in determining whether the prediction accuracy of the model is meaningful. In particular, deep learning studies in chemistry have attempted to utilize large volumes of datasets; thus, it is inevitable to integrate datasets measured by different protocols to increase the size of data, which deteriorates data quality and increases inherent experimental errors. If the prediction errors of the model were lower than the experimental errors, then there is a possibility that such accuracy was not a meaningful achievement, even though prediction accuracy was improved as compared to other methods [3]. In deep learning model studies, the authors compared the prediction accuracy of their models with others to prove that their own methods achieved improvement in prediction accuracy. However, it is challenging to find studies that have compared the prediction accuracy of their model with the experimental errors of the target endpoint. It may be attributed to dataset curation being done without an understanding of their inherent experimental errors; however, it is critical to verify the prediction accuracy of the model based on experimental errors to test the validity of the improvement in prediction accuracy. Among the four datasets used in this study, MP data analyzed experimental errors based on 18,058 duplicated compounds and estimated that the inherent experimental error of the dataset was 35°C [22], which was larger than the MAE of the MP model in this study. The higher inherent experimental errors in the MP dataset than the MAE of the MP model suggest that the actual prediction accuracy of the model might be higher than that measured by the external test set when the model was used for the prediction of unseen compounds. Unless the prediction errors of deep learning models in chemistry are analyzed based on experimental errors in the dataset, a simple comparison between the prediction accuracy of deep learning models may not be adequate to provide decisive evidence of significant improvement in prediction accuracy. As models in computer vision predict unambiguous and invariable labels, a higher prediction accuracy implies a better model., If the problem of mislabeling was excluded from the discussion, prediction models in computer vision achieved great success because of certainty in the dataset. It is practically impossible to obtain experimental noise-free datasets in chemistry. To make a successful case in chemistry, inherent experimental errors in the dataset must be understood precisely, such that the models are trained and validated reliably.

Model analysis

In QSAR, descriptors aim to distinguish compounds based on their structural similarity. As its purpose is comparison, differences in feature values between different compounds are significant in predicting the target endpoint. To examine the performance of the extracted features on clustering molecules, PCA was performed to exhibit how the feature space was varied as the TDEi tensor was processed within the CNN architecture. In Fig. 8, the brightness of colors implies the value of the melting point; dots are brighter if the value is higher and darker if they are lower. Initially, the original TDEi tensor's feature space established a low correlation with the normal melting point (Fig. 8A). Once the TDEi tensor was processed up to the last convolutional layer that is the eighth layer, compounds were prioritized as per the normal melting point (Fig. 8B). An additional pooling layer strengthened the trend in data distribution by separating compounds with a low melting point to the upper left side and a high melting point to the lower right side in the projected space (Fig. 8C). When the extracted features from the convolutional layer and pooling layer were processed using a fully connected layer, most of the compounds were arranged with a stronger correlation with their normal melting point values (Fig. 8D).

To examine the change in feature extraction by changing the CNN architecture, an identical analysis was performed using the CNN model with an increased number of convolutional layers. In Fig. 6, the convolutional layer is applied once before the pooling layer. Here, convolutional layers were applied twice with identical hyperparameters before the pooling layer. Features from the last pooling layer and the second fully connected layer were extracted and visualized (Fig. 9). PCA showed that the features extracted after additional convolutional layers were strongly correlated with the melting point. Although an additional convolutional layer did not improve the prediction accuracy, this analysis established that the CNN architecture can be modified for novel feature extraction.

In the CNN model trained with image data, initially, the fundamental level of features was extracted, and a higher level of features was found as the layer went deeper. EC is fundamental level information as compared to atom-level features; thus, the use of EC in CNN was expected to fully harness the CNN's automatic feature extraction capacity through filters establishing significant Eis for prediction of the target endpoint. Feature space

variation in PCA supported this idea because the extracted features were rearranged with a stronger correlation toward the melting point values as the layer went deeper.

Conclusions

The TDEi tensor was introduced in this study as a novel way to represent the molecular structure. Since electron interactions in a molecule determine molecular properties, the TDEi tensor was designed with electron interactions between each atom in a molecule, based on topological distance. Given that the data size is much smaller in chemistry as compared to computer vision studies, and sparsity in features can significantly deteriorate the performance of the model, the TDEi tensor was devised to be robust to the size of the dataset and structural diversity by changing the EC vector and topological distances considered in Ei calculations. Because the chemical space difference significantly influences the model prediction accuracy, a preliminary search on TDEi tensor options may be required if the structural diversity of the target dataset is different from the target chemical space in this study, namely small-molecule drugs.

The TDEi tensor was used in the CNN model, whose architecture was developed based on VGG net, with reduced weights because an increase in the number of layers did not improve the prediction accuracy. Desirable prediction accuracy was achieved for the four datasets. Unlike image data in computer vision, data in chemistry contains experimental noise in the target endpoint; therefore, deep learning studies in chemistry require a comparison between the prediction errors of the models and the inherent experimental errors in the dataset. As CNN was suitable for extracting relevant features automatically to predict the target endpoint, feature space changes were traced, through PCA on features obtained from the middle of the CNN architecture. A stronger correlation was found between the extracted features from the deeper layer and the target endpoint, implying that the CNN correctly modeled Eis that are significant for the prediction of the property.

Abbreviations

TDEi: topological distance-based electron interactionCNN: convolutional neural network

QSAR: quantitative structure–activityrelationship

ML: machine learning

FNN: fully connected architecture of artificial neural network

GNN: graph neural networks

SMILES: simplified molecular input line entry system

QM: quantum mechanics

EC: electron configuration

AO: atomic orbital

MP: melting point

Lipop: octanol/water distribution coefficient

Esol: water solubility

Freesolv: hydration free energy

MAE: mean absolute error

NMAE: normalized MAE

Sr: Spearman's rank correlation coefficient

PCA: principal component analysis

Declarations

Acknowledgements

I would like to thank Editage (www.editage.co.kr) for editing the manuscript.

Availability of data and materials

The MP dataset was obtained from the work ofIgor V. Tetko et al.[22], and others were obtained from Dejun Jian et al.[17].

Competing interests

The author declares that I have no competing interests.

Funding

This work was financially supported by a National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. NRF-2019R1F1A1061955)

Author contributions

Not applicable.

References

1. Piñero J, Furlong LI, Sanz F (2018) *In silico* models in drug development: Where we are. Curr Opin Pharmacol 42:111–121. doi: 10.1016/j.coph.2018.08.007.
2. Shin HK, Kang Y-M, No KT (2017) Predicting ADME properties of chemicals. In: Leszczynski J, Kaczmarek-Kedziera A, Puzyn T, Papadopoulos MG (ed) Handbook of computational chemistry, Springer, Cham, pp 2265–2301. doi: 10.1007/978-3-319-27282-5_59.
3. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A (2014) QSAR modeling: Where have you been? Where are you going to? J Med Chem 57:4977–5010. doi: 10.1021/jm4004285.
4. Huang J, Fan X (2011) Why QSAR fails: An empirical evaluation using conventional computational approach. Mol Pharm 8:600–608. doi: 10.1021/mp100423u.
5. Kim H, Kim E, Lee I, Bae B, Park M, Nam H (2020) Artificial intelligence in drug discovery: A comprehensive review of data-driven and machine learning approaches. Biotechnol Bioprocess Eng 25:895–930. doi: 10.1007/s12257-020-0049-y.
6. Jiménez-Luna J, Grisoni F, Schneider G (2020) Drug discovery with explainable artificial intelligence. Nat Mach Intell 2:573-584. doi: 10.1038/s42256-020-00236-4.
7. Baskin II, Palyulin VA, Zefirov NS (2006) Neural networks in building QSAR models. In: Livingstone DJ (ed) Artificial neural networks. Methods in Molecular Biology. vol 458, Humana Press. doi: 10.1007/978-1-60327-101-1_8.
8. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. Drug Discov Today 23:1241-1250. doi: 10.1016/j.drudis.2018.01.039.
9. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V (2017) MoleculeNet: a benchmark for molecular machine learning. arXiv 1703.00564. doi: 10.1039/c7sc02664a.
10. Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, Li Z, Luo X, Chen K, Jiang H, Zheng M (2020) Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem 63:8749-8760. doi: 10.1021/acs.jmedchem.9b00959.
11. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31-36. doi: 10.1021/ci00057a005.
12. Hirohara M, Saito Y, Koda Y, Sato K, Sakakibara Y (2018) Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. BMC Bioinformatics 19:526. doi: 10.1186/s12859-018-2523-5.
13. Karpov P, Godin G, Tetko IV (2020) Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J Cheminform 12:17. doi: 10.1186/s13321-020-00423-w.
14. Cova TFGG, Pais AACC (2019) Deep learning for deep chemistry: Optimizing the prediction of chemical patterns. Front Chem 7:809. doi: 10.3389/fchem.2019.00809.
15. Cui Q, Lu S, Ni B, Zeng X, Tan Y, Chen YD, Zhao H (2020) Improved prediction of aqueous solubility of novel compounds by going deeper with deep learning. Front Oncol 10:121. doi: 10.3389/fonc.2020.00121.
16. Xu Y, Dai Z, Chen F, Gao S, Pei J, Lai L (2015) Deep learning for drug-induced liver injury. J Chem Inf Model 55:2085-2093. doi: 10.1021/acs.jcim.5b00238.
17. Jiang D, Wu Z, Hsieh C-Y, Chen G, Liao B, Wang Z, Shen C, Cao D, Wu J, Hou T (2021) Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J Cheminform 13:12. doi: 10.1186/s13321-020-00479-8.
18. Zakharov AV, Peach ML, Sitzmann M, Nicklaus MC (2014) QSAR Modeling of imbalanced high-throughput screening data in PubChem. J Chem Inf Model 54:705-712. doi: 10.1021/ci400737s.

19. Soufan O, Ba-alawi W, Magana-Mora A, Essack M, Bajic VB (2018) DPubChem: a web tool for QSAR modeling and high-throughput virtual screening. Sci Rep 8:9110. doi: 10.1038/s41598-018-27495-x.
20. Shin HK (2020) Electron configuration-based neural network model to predict physicochemical properties of inorganic compounds. RSC Adv 10:33268-33278. doi: 10.1039/D0RA05873D.
21. Shin HK, Kim S, Yoon S (2021) Use of size-dependent electron configuration fingerprint to develop general prediction models for nanomaterials. NanoImpact 21:100298. doi: 10.1016/j.impact.2021.100298.
22. Tetko IV, M. Lowe D, Williams AJ (2016) The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS. J Cheminform 8:2. doi: 10.1186/s13321-016-0113-y.
23. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jozefowicz R, Jia Y, Kaiser L, Kudlur M, Levenberg J, Mané D, Schuster M, Monga R, Moore S, Murray D, Olah C, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X: TensorFlow: Large-scale machine learning on heterogeneous systems. In. Google; 2015: Software available from tensorflow.org.
24. Alexander DLJ, Tropsha A, Winkler DA (2015) Beware of R^2: Simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. J Chem Inf Model 55:1316-1322. doi: 10.1021/acs.jcim.5b00206.
25. Seo M, Shin HK, Myung Y, Hwang S, No KT (2020) Development of natural compound molecular fingerprint (NC-MFP) with the dictionary of natural products (DNP) for natural product-based drug development. J Cheminform 12:6. doi: 10.1186/s13321-020-0410-3.
26. Shin HK, Lee S, Oh HN, Yoo D, Park S, Kim WK, Kang MG (2021) Development of blood brain barrier permeation prediction models for organic and inorganic biocidal active substances. Chemosphere 277:130330. doi: 10.1016/j.chemosphere.2021.130330.
27. Zhao L, Wang W, Sedykh A, Zhu H (2017) Experimental errors in QSAR modeling sets: What we can do and what we cannot do. ACS Omega 2:2805-2812. doi: 10.1021/acsomega.7b00274.

Figures

Figure 1

Calculation of electronic configuration (EC) vectors of each atom. EC vector of each atom was used in order to calculate Eis within a molecule. (A) Indices of full EC vector are atomic orbitals with different spin, and EC vector was designed to be reduced by integrating (B) atomic orbitals in an identical energy level (degenerate orbitals) and (C) in different spins. Each EC vector was condensed in order to reduce sparsity in feature space (D–F).

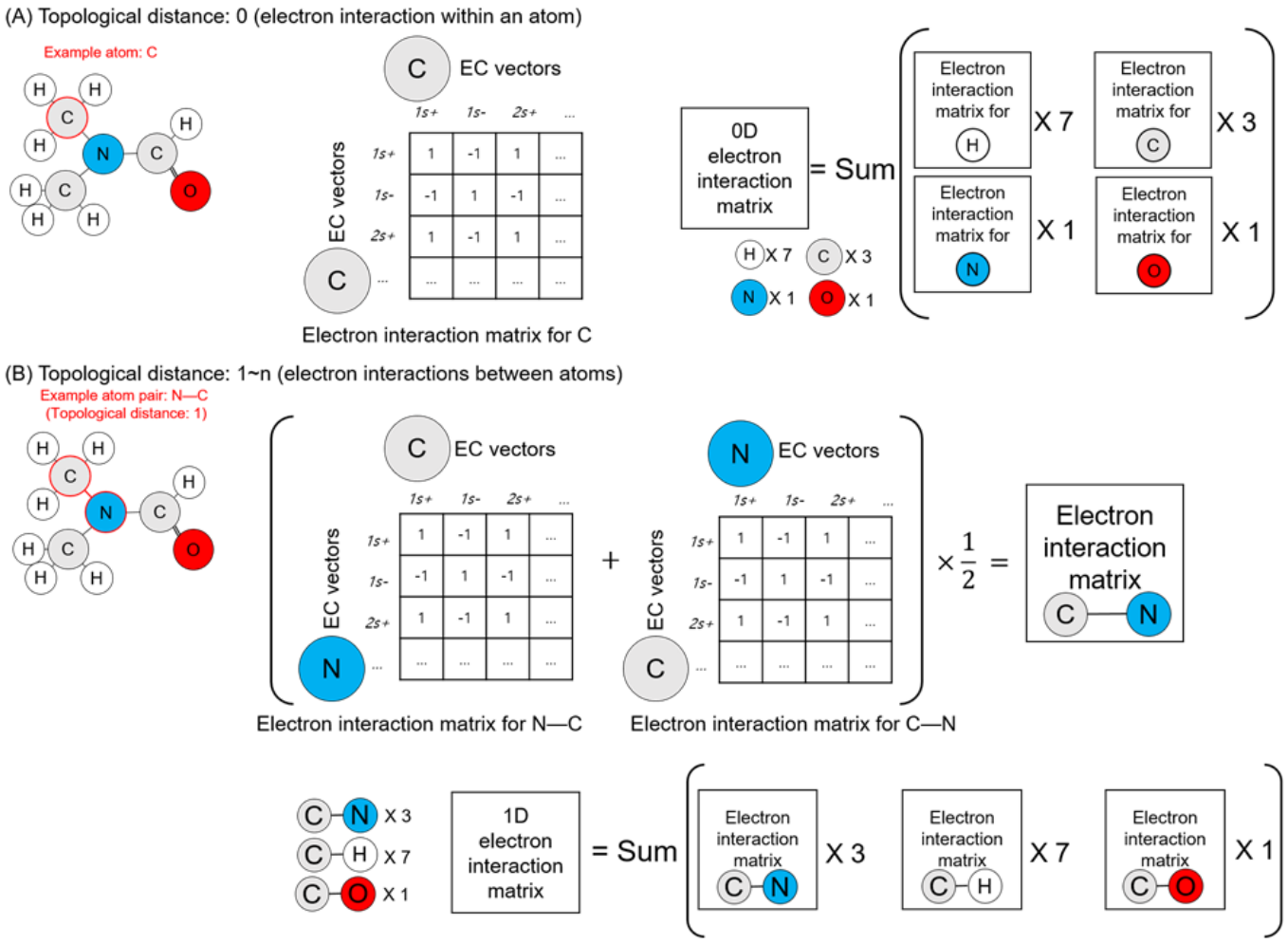

Figure 2

Electron interaction (Ei) matrices in each specified topological distance calculated based on EC vector of each atom in a molecule. (A) Eis within topological distance 0D are electron interactions within an atom, and (B) Eis within topological distance longer than 0D are Eis between pair of atoms according to the distance. After calculation of Ei matrices, they were summed according to the topological distance.

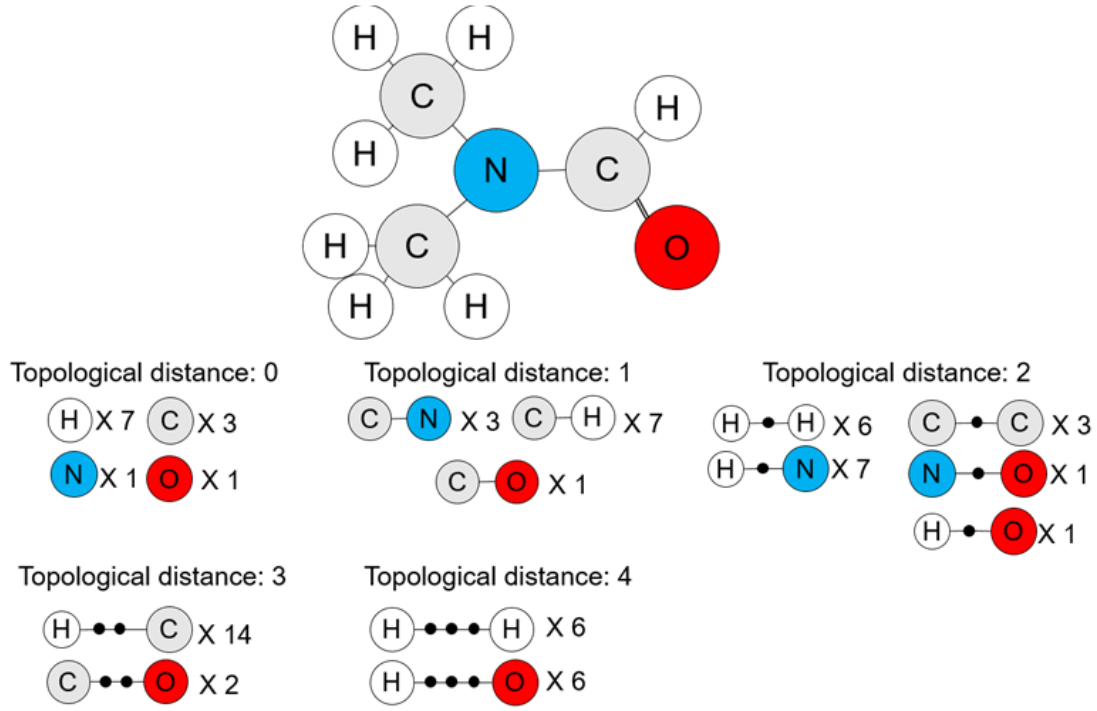

Figure 3

Possible atom pairs in the example molecule. In this molecule, the longest topological distance was 4D; however, further atom pairs can be found if size of the molecule increases.

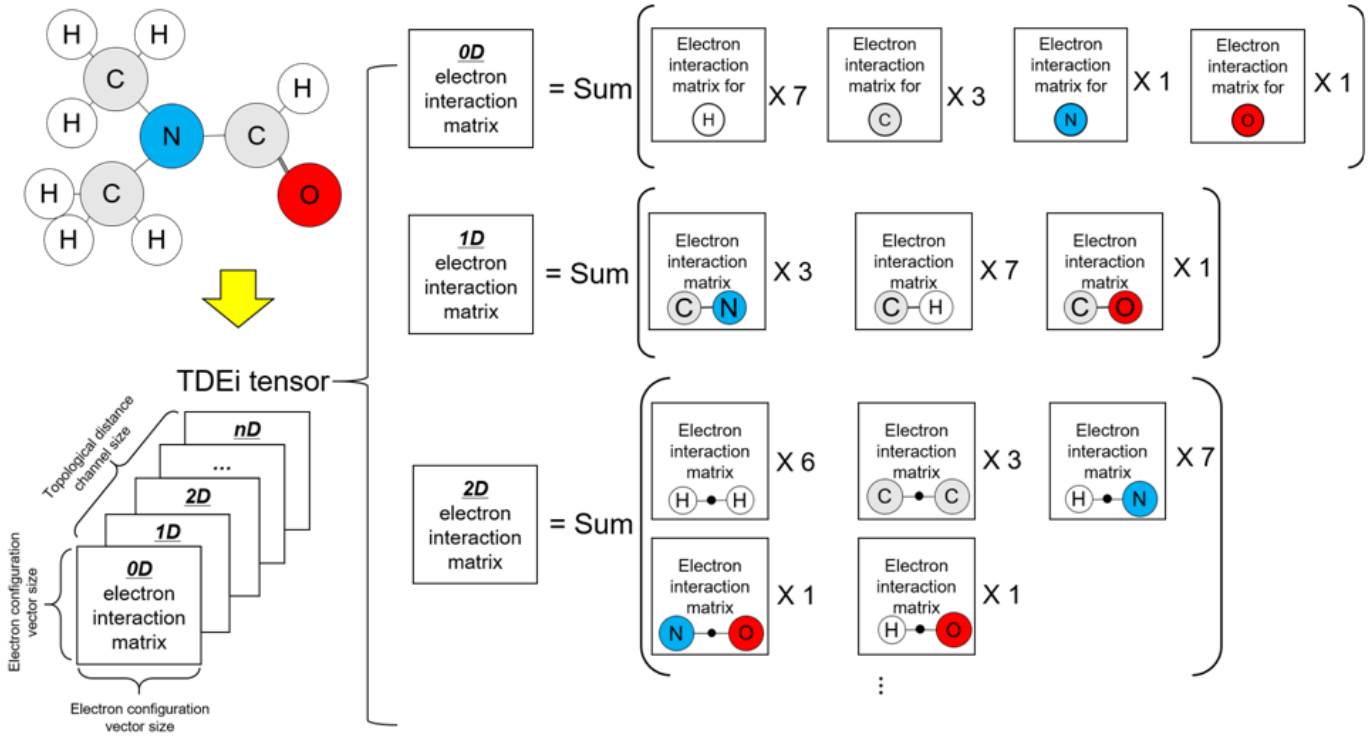

Figure 4

Preparation of TDEi tensor. TDEi tensor was prepared by concatenating Ei matrices in each topological distance and designed to be adaptable according to structural diversity and the size of datasets by adjusting EC vectors and topological distances.

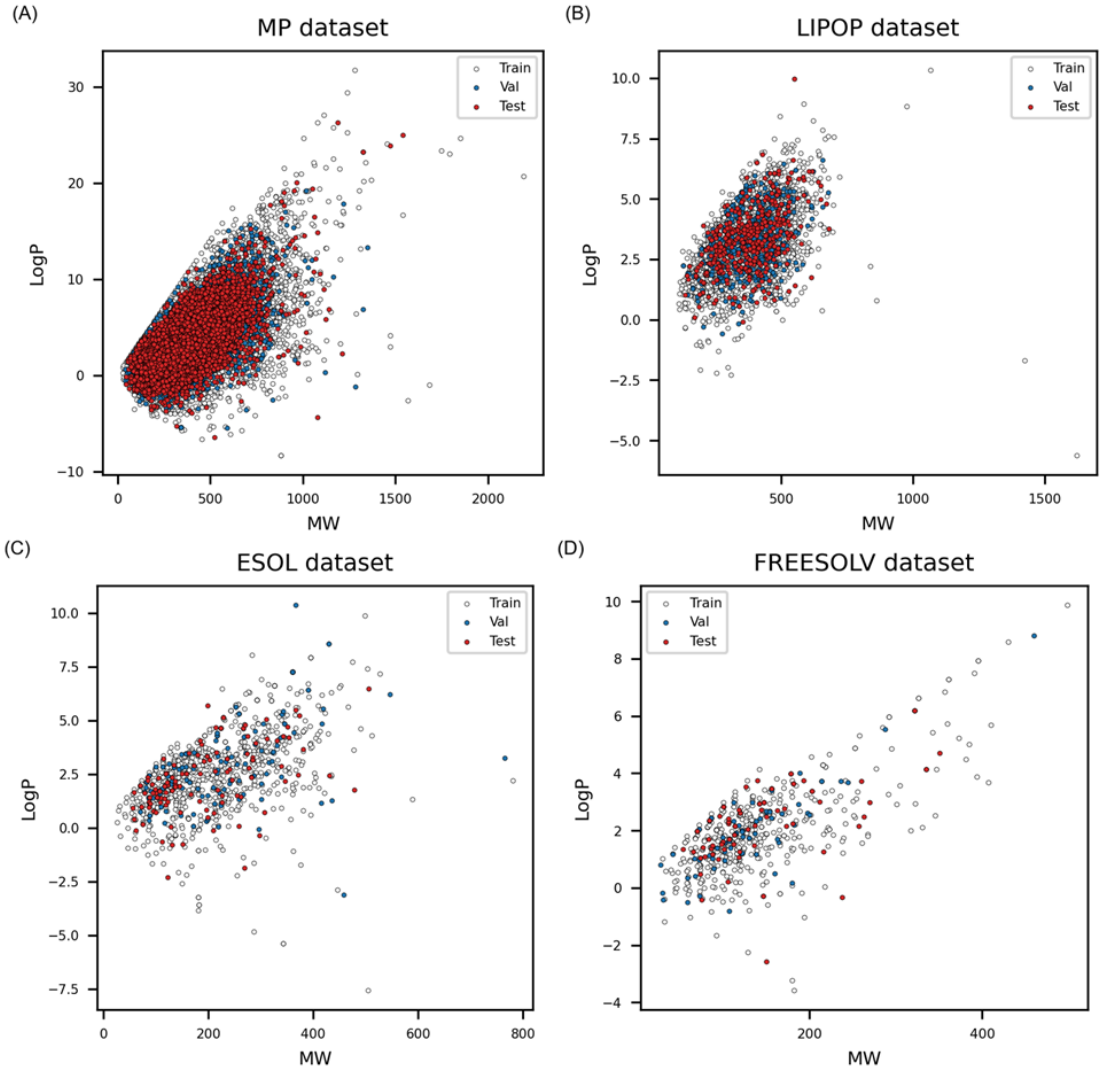

Figure 5

Structure diversities between training, validation, and external test set. (A) Normal melting point (MP) data was the largest data set. Datasets for Lipop (B), Esol (C), and Freesolv (D) were octanol/water partition coefficient, water solubility, and hydration free energy, respectively.

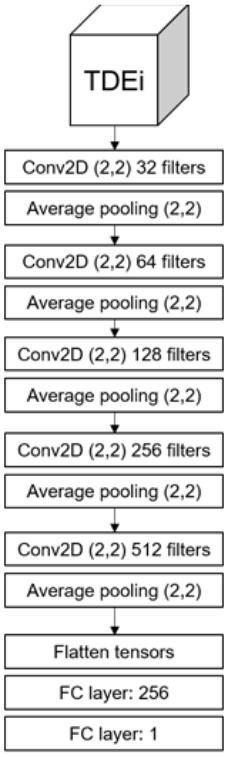

Figure 6

CNN architecture used in this study. VGG Net was modified by decreasing channel size of filter, filter size in convolutional layer to two by two, and the number of convolutional layers before pooling layer. In grid search, diverse CNN architectures were tested; however, increasing thenumber of layers did notimprove prediction accuracy.

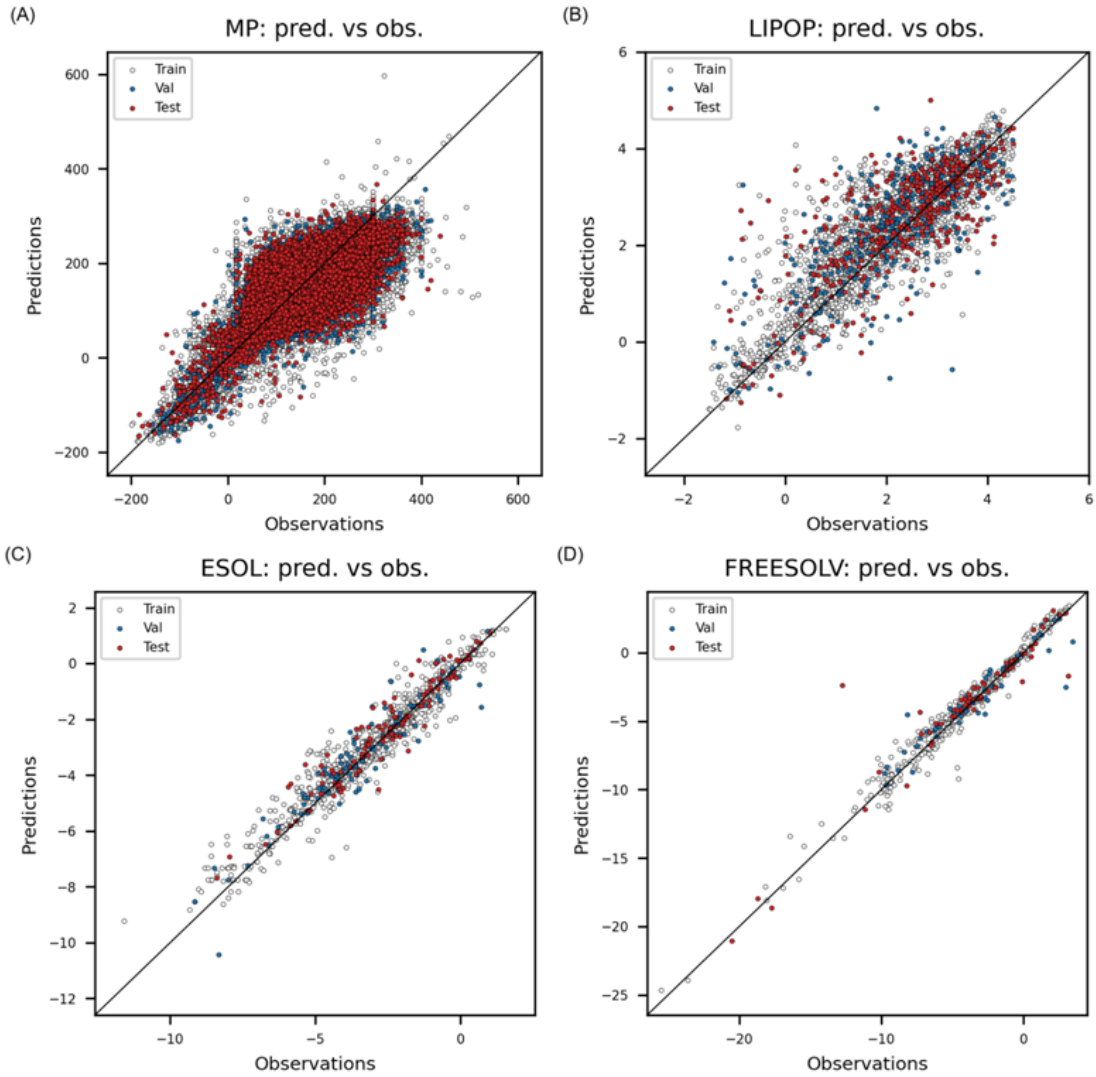

Figure 7

Examination of goodness-of-fit on four endpoints.

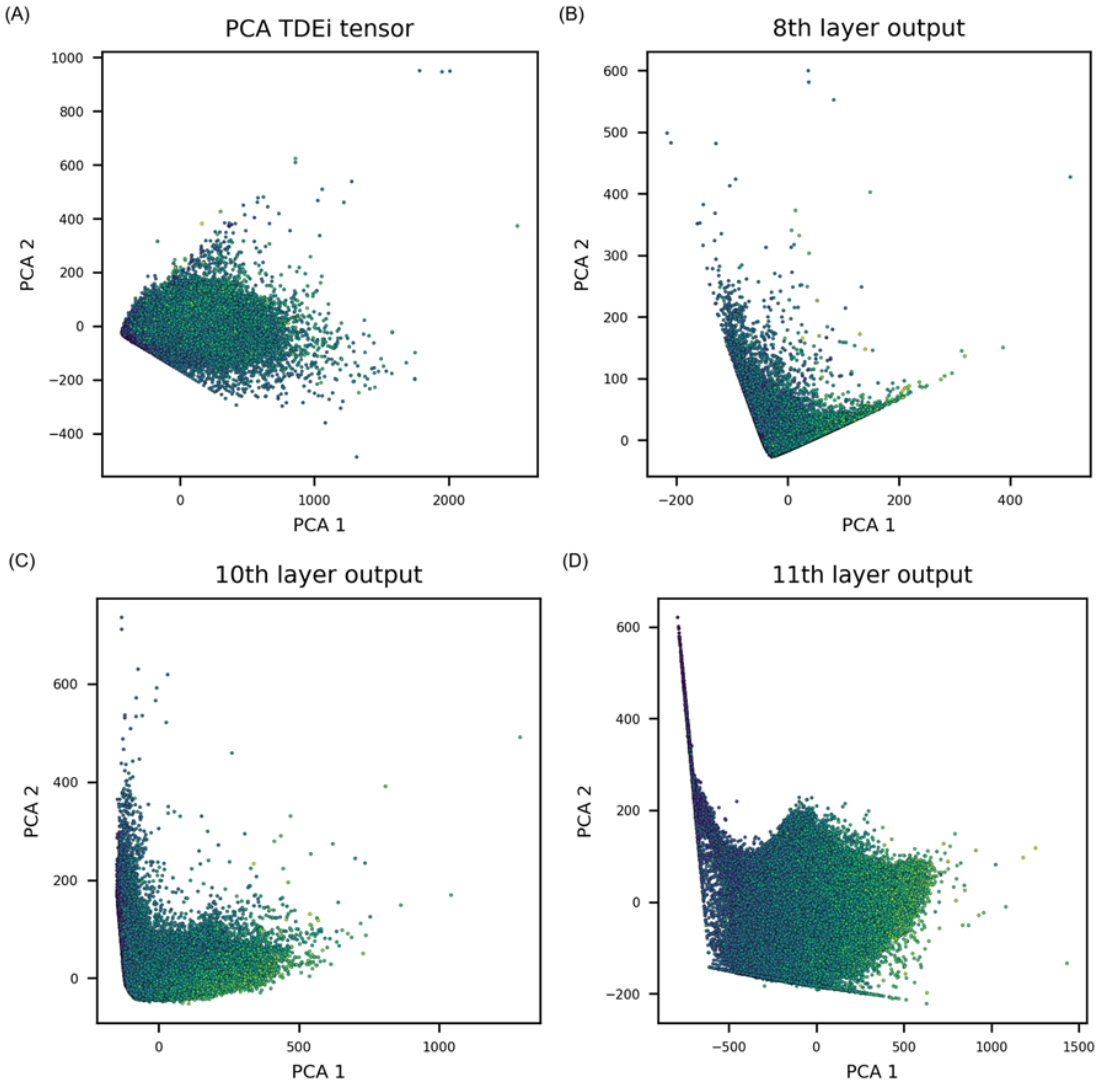

Figure 8

Results of principal component analysis (PCA) performed to examine feature space variation throughout CNN model. MP prediction model was analyzed since it was developed with the largest dataset. Each point was colored based on MP value: a brighter color implies higher MP whereas darker color indicates lower MP. (A) Initially, TDEi tensor itself could not sufficiently prioritize compounds according to MP. (B) Features extracted after the last convolutional layer showed that data points were arranged along the trend of MP values. (C) The trend was strengthened after the last average pooling layer, and (D) stronger correlation was established after the fully connected layer.

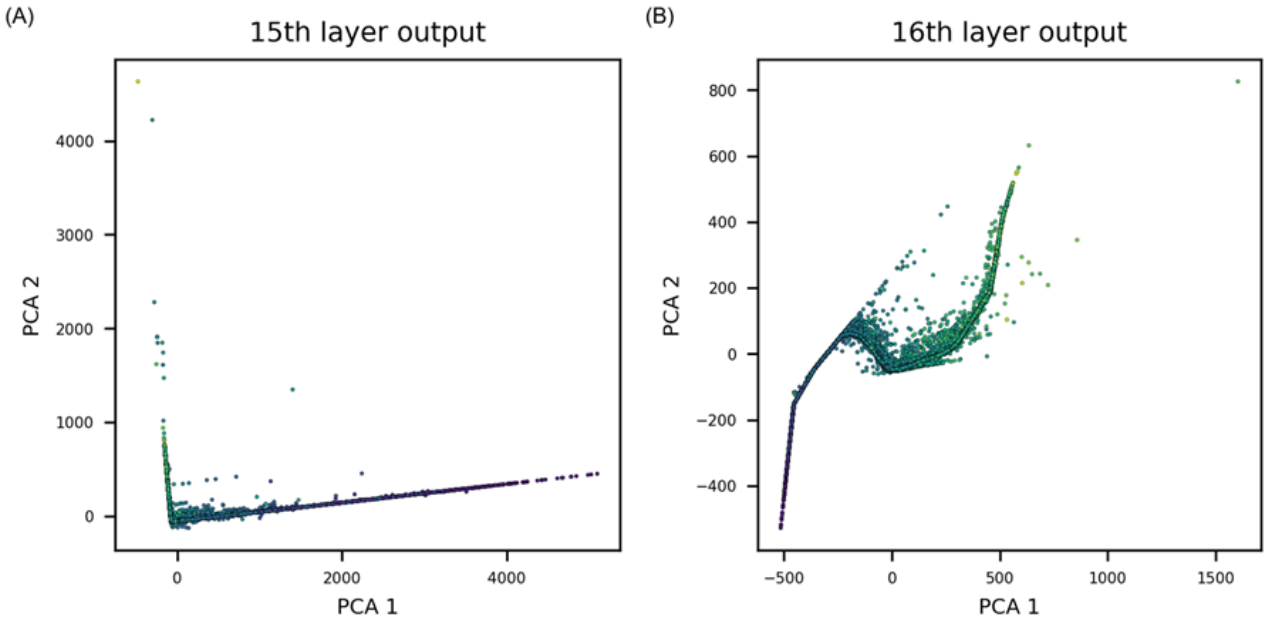

Figure 9

Results of using deeper layer of CNN in feature space variation.Results shown in this figure were obtained by slightly modifying the CNN shown in Figure 6 by applying convolutional layer twice before the average pooling layer. (A) Features extracted from the last pooling layer prioritize compounds accurately. (B) After fully connected layer, features were in stronger correlation with the target endpoint.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Graphicalabstract.png
- TDEicalculationcode.zip
- supplementarytables.xlsx