JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2020, 60, 22−28

pubs.acs.org/jcim

Graph Convolutional Neural Networks as "General-Purpose" Property Predictors: The Universality and Limits of Applicability

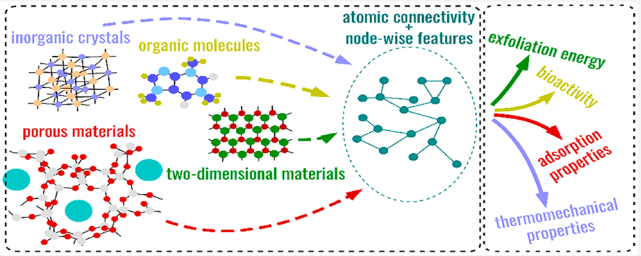
Vadim Korolev,*[,†,‡] Artem Mitrofanov,[†,‡] Alexandru Korotcov,[†] and Valery Tkachenko[†]

[†]Science Data Software, LLC, 14909 Forest Landing Circle, Rockville, Maryland 20850, United States

[‡]Department of Chemistry, Lomonosov Moscow State University, Leninskie gory, 1 bld. 3, Moscow 119991, Russia

Supporting Information

ABSTRACT: Nowadays the development of new functional materials/chemical compounds using machine learning (ML) techniques is a hot topic and includes several crucial steps, one of which is the choice of chemical structure representation. The classical approach of rigorous feature engineering in ML typically improves the performance of the predictive model, but at the same time, it narrows down the scope of applicability and decreases the physical interpretability of predicted results. In this study, we present graph convolutional neural networks (GCNNs) as an architecture that allows for successfully predicting the properties of compounds from diverse domains of chemical space, using a minimal set of meaningful descriptors. The applicability of GCNN models has been demonstrated by a wide range of chemical domain-specific properties. Their performance is comparable to state-of-the-art techniques; however, this architecture exempts from the need to carry out precise feature engineering.



INTRODUCTION

The design of new functional materials is associated with significant computational costs. The use of different approximations substantially based on density functional theory[1,2] (DFT) makes it possible to solve the Schrödinger equation numerically without critical loss of accuracy for calculation of ground-state properties. However, DFT calculations are still disappointingly time-consuming for materials and molecular design, especially if we consider their extensive use for large databases.[3−5] An alternative approach combines the quantitative structure−property relationship (QSPR) modeling and ML techniques for materials property predictions based on their structure. This method accelerates the calculation of properties by several orders of magnitude. Nevertheless, it remains unclear how widely this approach can be applied and replaces the DFT calculations.

It is well-known that appropriate representation of considered structures plays a crucial role in property prediction using machine learning.[6−9] Design of input data, and so-called feature engineering, represents a challenge. In particular, a specific feature vector has to be constructed for each use case; the more specific the target property, the more specific the feature vector should be. All these issues prevent the use of one unified approach for different domains of chemical space.

Nevertheless, several attempts have been made to create universal models that do not require precise feature engineering. The Gaussian process regression[10] and deep tensor neural networks[11] were used as basic algorithms. However, it should be noted that priority attention in these studies was paid to the prediction of the thermodynamic stability of molecules/materials, i.e., their energy. State-of-the-art results have been achieved using graph-based neural networks[12] on large data sets generated in a high-throughput manner (e.g., QM9 for molecules, Materials Project database for inorganic crystals). At the same time, many physicochemical quantities cannot be obtained as a result of DFT calculations. Data scarcity makes it very difficult to build highly efficient predictive models (including those based on graph-based neural networks) in the case of small experimentally obtained data sets. Another issue of particular interest is a generalization of such universal models to porous or low-dimensional materials, which represent an intermediate case between molecular and crystalline systems from an atomic connectivity perspective. However, the above difficulties have not been considered in detail.

This study is devoted to the development of a universal approach that would allow predicting the properties of both molecules and materials with varied atomic connectivity, using simple, physically interpretable descriptors. A graph-based convolutional neural network is considered here as a base architecture for universal property prediction. Similar architectures were successfully used to predict the properties of structures from distinct domains of chemical space.[13−15] To the best of our knowledge, this work also represents the first attempt to apply graph-based NNs for the prediction of two-dimensional and porous materials' properties. The limits of

Received: July 18, 2019
Published: December 20, 2019

DOI: 10.1021/acs.jcim.9b00587
J. Chem. Inf. Model. 2020, 60, 22−28

applicability of graph-based NNs, as a flip side of their universality, are also clearly demonstrated.

MATERIALS AND METHODS

Convolutional neural networks (CNNs) are one of the most promising and fast-developing classes of machine learning algorithms due to their exceptional performance in image, video, and speech recognition tasks.[16]

Ordinary CNNs (strictly speaking, convolutional kernels) require input data to be on the regular grid. Unfortunately, most of the real use cases deal with highly unordered data, which can be sometimes represented as a graph. Extension of convolutional kernels to an irregular domain is a nontrivial task, thus in recent years the efforts of several scientific groups have focused on its solution.[17,18] Graph convolutional neural networks (GCNNs) as a subclass of neural networks applied to graphs[19,20] are strongly inspired by previously developed categories: graph neural networks and recurrent graph neural networks. Thus, some GCNNs (spatial-based ones) inherited the idea of information propagation/message passing. Information is passed from one node to another along edges. On the other hand, GCNNs differ sharply from all other graph-based architectures operating directly on graphs (graph neural networks, recursive and recurrent graph neural networks, etc.) since they perform through two fundamental operations taken from "vanilla" CNNs and generalized to an irregular domain — convolution and pooling. The comparison of standard 2D operations and its graph counterparts is provided in Figure 1.

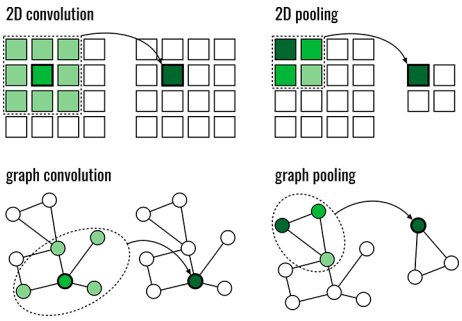


Figure 1. Scheme representation of two essential operations for (graph) convolutional neural networks: convolution and pooling. Original 2D convolution and pooling operate on a regular two-dimensional grid. A set of nodes that form new pixel/node representation (receptive field) is marked with color and restricted by a dashed line. Kernel size is defined by height × width (3 × 3 in the present case) of the convolution window for 2D convolution; the kernel size of graph convolution is determined by the node's neighborhood (a first neighborhood in the present case).

Both convolution operations are responsible for extracting high-level node representations by taking the weighted average of node feature values of the considered node along with its neighbors. A set of neighboring nodes and corresponding weights determines the convolutional filter, and a set of convolutional filters applied to all nodes forms the convolutional layer. In general, (G)CNNs contain multiple convolutional layers; each of them contains several filters. Therefore, the final node representation is formed not only by its nearest neighborhood but also each added convolutional layer extends the set of nodes that form final node representation, i.e., its receptive field. In the simplest case (one convolutional layer), the receptive field equals the convolution kernel size.

Both pooling operations aim to reduce the dimensionality of representations. Sum/max/mean strategy, i.e., calculating the sum/max/mean value for the target node and its neighborhood (pooling window), is the most popular choice for 2D pooling. However, as in the case of convolution operation, the notion of pooling on a 2D grid cannot be directly applied to graphs; a more efficient pooling strategy should be implied (see the Supporting Information).

The attractive idea is to adapt and apply graph convolutional neural networks for materials property prediction since chemical structures can be represented in the form of graphs with atoms for nodes and bonds for edges. Previously, this concept was used mainly for organic molecules (starting from the seminal work of Baskin et al.[21]), and so far only a few studies have been devoted to their application for periodic structures (inorganic crystals) property prediction.[12,15,22] We propose to use unified representation for (non)periodic structures. Each structure is represented as graph $G = (V,E)$, where V is the set of nodes (atoms), and E is the set of edges (bonds). Binary adjacency matrix $A^{N\times N}$ (where N is the number of atoms) is used to describe connectivity between atoms. To define the graph for periodic structures, only atoms in one unit cell were considered. Two atoms are supposed to be connected if the interatomic distance is shorter than the sum of the corresponding covalent radius and tolerance distance (0.25 Å). A similar approach has been successfully used to establish connectivity between atoms for universal fragment descriptors.[23]

We based our architecture on the GCNN proposed by Such et al.[24] The exact architecture used in this study and schematic representation of convolution and pooling operations are provided in Figure 2. After spatial-based graph convolutions and down-sampling by graph embed pooling, graph-level representation is aggregated from node-level representations with read-out operation. For our model, this operation is a special case of graph embed pooling that reduces graph-level representation to a one-dimensional vector. For the sake of universality, only the adjacency matrices and the set of the

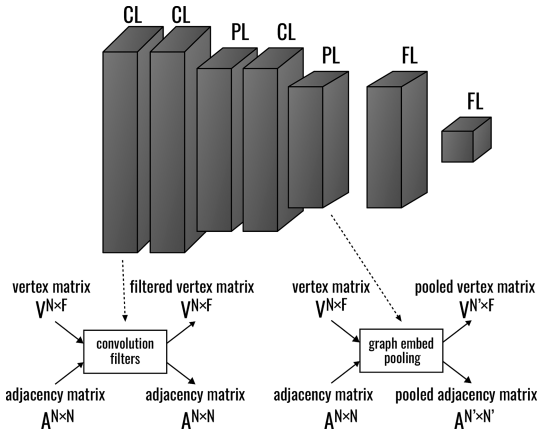


Figure 2. GCNN architecture presented in this study. It includes the following layers: two convolutional layers (CL), pooling layer (PL), convolutional layer, pooling layer, fully connected layer (FL), and output one (two)-neuron layer for regression (classification) task. Convolutional layers update vertex matrix (node representation via a 2D matrix where each row contains a feature vector of the corresponding node); the adjacency matrix remains unchanged. Graph embed pooling operation reduces the dimensionality (number of nodes) and also updates the vertex matrix for a new set of nodes.

DOI: 10.1021/acs.jcim.9b00587
J. Chem. Inf. Model. 2020, 60, 22−28

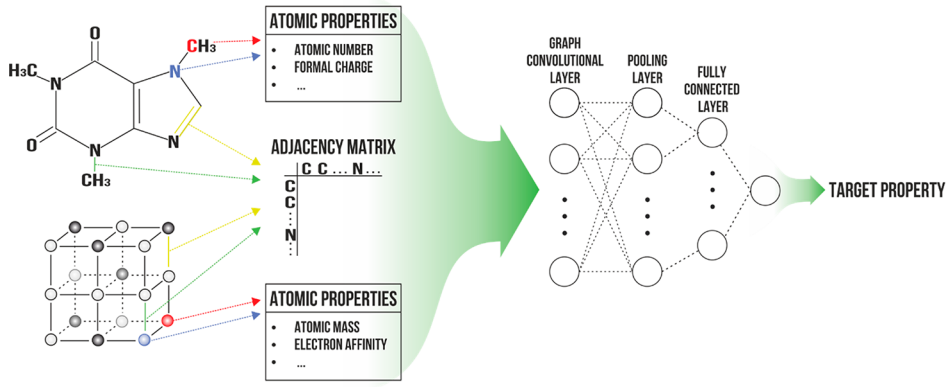


Figure 3. General workflow of training process.

most common properties of the corresponding element/lattice site are used as input data for the GCNN model. More detailed information on GCNN architecture is provided in the Supporting Information.

Two slightly different sets of descriptors for nonperiodic and periodic structures were used to represent structures, and each of them reflects the specificity of the above systems. For nonperiodic structures (organic molecules), only node-wise descriptors were used, and structural features were excluded entirely from consideration, which is opposite from the case of periodic structures where those features were included (for more details see the Supporting Information). The general scheme is presented in Figure 3.

It should be noted that we purposely excluded from consideration combinations (multiplication, ratio, etc.) of the mentioned descriptors. Such artificially generated quantities are widely used in conjunction with random forest/gradient boosting models, with the subsequent selection of the most valuable features. In fact, this methodology is one of the variations of feature engineering. However, as stated earlier, our goal was to build the simplest, most universal model, without the need to use extensive sets of domain-specific features and their combinations.

To demonstrate the universality and limits of applicability of GCNN models, we train models on multiple diverse data sets, which cover a wide range of various parameters:

- Dimensionality − 0D (organic molecules), 2D (layered materials), 3D (all others).
- Structural diversity − high (porous materials), medium/low (all others).
- Compositional diversity − low (porous materials), medium/high (all others).

Molecules. Previously, similar graph-based neural networks demonstrated high accuracy of prediction of multiple molecular properties related to different levels, including the following: 1) quantum mechanical, 2) physicochemical, 3) biophysical, and 4) macroscopic physiological properies.[25] Only a few data sets from the last three levels were used to confirm the performance of GCNN models.

Water solubility models were trained using a data set[26] containing 1299 molecules with logS = −5 taken as a cutoff value for classification. We also trained a number of models for biological end points, including human ether-a-go-go-related gene (hERG) inhibitors data set[27] with 373 active and 433 inactive molecules; 175 active, 19 604 inactive molecules from CDD Public data sets[28−30] (malaria); and a data set with 253 active, 69 inactive molecules that have been scored for probe-likeness by medical chemists.[31]

Inorganic Crystals. Experimental determination of inorganic crystal properties is a complex task due to difficulties with a monocrystalline synthesis. In this study, we used data sets with DFT calculated properties to train models. Two sources of data were used: AFLOWlib[32] and JARVIS-DFT[33] databases. To train and validate models for prediction of thermomechanical properties, data sets with 2748 and 770 compounds, respectively, were taken from Isayev et al.[23] Six properties include the bulk modulus, shear modulus, Debye temperature, heat capacity at constant pressure, heat capacity at constant volume, and the thermal expansion coefficient which were calculated with the AEL−AGL integrated framework.[34] Metal/insulator and magnetic/nonmagnetic classification tasks were performed on 25468 and 25131 compounds, extracted from AFLOWlib and JARVIS-DFT databases, respectively. Optimized two-dimensional structures and corresponding exfoliation energies for 601 compounds were extracted from the JARVIS-DFT[35] database. All calculations were provided with PBE and optB88 functionals.

Porous Crystalline Materials. To demonstrate the applicability of GCNN models for porous materials' property predictions, we consider two domain-specific properties: the bulk and shear modulus of pure-silica zeolites from the IZA database and Xe/Kr infinite dilution selectivity of Computation-Ready, Experimental (CoRE) Metal−Organic Frameworks (MOF).

To predict the bulk and shear modulus of pure silica zeolites, we use a subset of the Database of Zeolite Structures presented by Coudert.[36] Seven of 122 zeolite frameworks were excluded from consideration since errors occurred during the generation of the corresponding structural descriptors (see the Supporting Information). The B3LYP hybrid exchange−correlation functional was used to obtain the elastic data; isotropic values of bulk and shear moduli were calculated using Voigt−Ruess−Hill averages.

Xe/Kr adsorption data for all MOF structures (for which density derived electrostatic and chemical charges have been obtained) was modeled via a classical force field (FF), namely, a universal force field.[37] To demonstrate the influence of intrinsic flexibility on the Henry regime adsorption properties in CoRE MOF structures, all calculations were carried out both in a rigid and in a flexible approximation.

DOI: 10.1021/acs.jcim.9b00587
J. Chem. Inf. Model. 2020, 60, 22−28

RESULTS AND DISCUSSION

Molecules. Like many other breakthroughs in chemoinformatics, the deep learning revolution in the field was mainly caused by the needs of drug design.[38−41] Graph-related neural networks are no exception. To date, they are mainly used to predict molecular properties.[14,42] Recently, several studies present the large-scale comparison of various ML techniques, including graph-based neural networks, proving their high performance.[25,43] Therefore, we take into consideration only a few molecular data sets to prove the concept.

We compare the prediction performance of GCNN models against the prediction performance of the support vector machine (SVM) and the feed-forward neural networks (FNN) models on four molecular data sets with ECFP6[45] fingerprints. Following the original methodology,[44] we test our models on an external test set (20% compounds from initial set) with a 5-fold cross-validation. The results are presented in Table 1. These two algorithms have been chosen as the most efficient—5-layer FNN and SVM (among all classic machine learning algorithms) rank above all other presented ML methods.[44]

Table 1. Summary of Performances (Molecules-Related Tasks): GCNN Models versus SVM and FNN Models

	GCNN (this study)		SVM[44]		FNN[44]	
	ROC AUC	accuracy	ROC AUC	accuracy	ROC AUC	accuracy
solubility	0.97	0.96	0.93	0.90	0.93	0.93
hERG	0.85	0.78	0.86	0.80	0.84	0.80
malaria	0.93	0.99	0.98	0.97	0.97	0.99
probe-like	0.62	0.76	0.66	0.76	0.56	0.77

The most obvious tendency deals with data scarcity and class imbalance. GCNN models outperform or perform similarly as other algorithms on well-balanced data sets (solubility and hERG). Nevertheless, GCNN models as a particular case of graph-based models are not robust enough[25,43] to perform well on highly imbalanced (malaria, active/inactive ratio is 0.0089) and small data sets (probe-like, 322 samples in total).

As an intermediate conclusion, we may denote GCNN as a convenient tool for the purposes of molecule QSPR modeling. The accuracy of the approach is comparable with well-known and reliable methods, though it cannot be called a breakthrough in regards to small organic molecules.

Inorganic Crystals. AFLOWlib[32] and JARVIS-DFT[33] databases have already been used to develop ML models for prediction of thermomechanical properties. In contrast to this study, the available machine-learning frameworks are based on precise feature engineering and algorithms with high interpretability, such as gradient boosting decision trees (GBDT). Following the original methodology,[23] we test our models on an external test set (∼20% compounds from initial set) with a 5-fold cross-validation. Table 2 contains the performance metrics for six regression models that predict thermomechanical properties of bulk materials and exfoliation energy of potentially exfoliable 2D-layered materials. In most cases, GBDT models slightly outperform GCNN models. Furthermore, we also develop two predictive models for metal/insulator and magnetic/nonmagnetic classification tasks. The area under the ROC curve and accuracy for the two classification tasks obtained with GCNN/GBDT models were used to evaluate and compare models. These models show similar accuracy with the area under the curve at 0.97/0.98[23] for metal/insulator and 0.94/0.96[46] for magnetic/nonmagnetic classification tasks, respectively.

Surprisingly, GCNN models demonstrate sufficient accuracy not only for bulk materials property prediction but also for initial screening of potentially exfoliable materials. Following the original methodology,[46] we test our models on an external test set (10% compounds from initial set) with 5-fold cross-validation. Our GCCN model has MAE for exfoliation energy (37 meV/atom) that is comparable with an MAE of the GBDT model and is significantly less than the threshold value for potentially exfoliable 2D-layered materials (200 meV/atom). Strictly speaking, only a few 2D materials are true monolayers. Most of them contain several layers in the direction perpendicular to the free surface, thus, the corresponding chemical graph used to predict exfoliation energy should be *weighted*, since the edges have an unequal contribution to the surface/exfoliation energy. Nevertheless, it should be concluded that elemental descriptors are sufficient for prediction of exfoliation energy with appropriate accuracy (3D Voronoi descriptors were excluded from consideration).

Due to the high interpretability of GBDT models, it is possible to rank the features of importance for the predictive model. According to Isavev et al.,[23] the most important features are combinations of element properties, while within the framework of our GGNNs model, linear combinations of properties are not taken into account for the sake of model simplicity. Additionally, we do not use some of the specific properties also included in the descriptor list (effective atomic charge, chemical hardness, van der Waals radius, second and third ionization potentials). Moreover, structural representation used by Choudhary[46] contains very specific features—charge-based and classical force-field inspired descriptors, bond-angle and dihedral-angle distributions, etc.

Porous Crystalline Materials. Recently, bulk and shear moduli of pure-silica zeolites were calculated using five classical

Table 2. Summary of Performances (Materials-Related Tasks): GCNN Models versus GBRT Models

	GCNN (this study)			GBDT[23]		
	MAE	RMSE	R^2	MAE	RMSE	R^2
bulk modulus, GPa	15.21	24.61	0.91	12.00	21.13	0.93
shear modulus, GPa	14.88	21.01	0.87	13.31	18.94	0.90
Debye temperature, K	50.97	71.21	0.91	42.92	64.04	0.93
heat capacity at constant pressure, k_b/atom	0.07	0.11	0.91	0.06	0.10	0.92
heat capacity at constant volume, k_b/atom	0.05	0.08	0.93	0.05	0.07	0.95
thermal expansion coefficient, K^{-1}	9.63×10^{-6}	1.56×10^{-5}	0.81	5.77×10^{-6}	1.95×10^{-5}	0.76
exfoliation energy, meV/atom	36.5	69.4	0.22	37.3		

DOI: 10.1021/acs.jcim.9b00587
J. Chem. Inf. Model. 2020, 60, 22−28

interatomic potentials[47] (BKS,[48] Catlow,[49] Gale,[50] Matsui,[51] Sastre[52]). These results can serve as a ground-state level to probe the accuracy of other approaches (using machine learning, in our case), due to the wide applicability of classical model potentials for calculation of mechanical properties.[53]

Table 3 contains performance metrics for force field models as mentioned above, based on five state-of-the-art interatomic potentials. Besides, the gradient boosting regressor (GBR) model was also used to predict the mechanical properties of zeolites from the same data set.[54] Following the original methodology,[54] we test our models using 3-fold internal cross-validation. The values of the bulk K and shear G moduli obtained with GBR and GCNN models are also provided in Table 3.

Table 3. Summary of Performances (Porous Materials-Related Tasks): GCNN Models versus GBR Model and Conventional Model Potentials

	RMSE, K	RMSE, G
GCNN	13.14	6.40
GBR[54]	10.00	4.74
BKS[47]	22.7	36.1
Catlow[47]	18.8	11.7
Gale[47]	20.0	12.6
Matsui[47]	16.8	29.4
Sastre[47]	18.1	14.1
	RMSE, S_f	RMSE, S_r
GCNN	3.98	5.32

The GCNN model significantly outperforms all force field models, but the accuracy of the GBR model is slightly better. According to Evans et al.,[54] the local descriptors (in particular, Si−O−Si angles and parameters related to the Si−O bonds) are the most crucial features for the prediction of mechanical properties. Due to the peculiarities of graph-based structure representation, it is not a trivial task to implement the descriptors associated with the statistical distribution for bonds and angles between atoms as node-wise features. Also, as an alternative approach, the secondary building blocks can be used as the vertices of the chemical graph instead of atoms. This insight is to be addressed in coming studies.

Furthermore, Database of Zeolite Structures is a typical example of small materials data set, and, in a sense, it is similar to the previously discussed probe-like data set. Data scarcity imposes principal limits on the level of accuracy of ML models regardless of structure representation. Only advanced techniques, such as incorporating the crude estimation of property (CEP) in the feature vector,[55] can significantly improve the performance of implemented models.

Table 3 also contains performance metrics of GCNN models predicted infinite dilution selectivity of Xe/Kr in rigid S_r and flexible S_f approximations. As in the previous case, GCNN models have low accuracy in prediction of porous materials' properties as opposed to other subdomains of chemical space. The models for prediction of infinite dilution selectivity in flexible mode show even better performance, but due to the low accuracy of both models, the result is not significant. Nevertheless, considering the ratio between the range of selectivity values and corresponding RMSE, GCNN models enable at least a meaningful qualitative comparison of most promising candidates for adsorption-related applications. Furthermore, graph-based representation is suitable for modeling of strong interatomic interactions, and at the same time it is well-known that van der Waals interactions play a significant role in MOFs and similar classes of porous materials. Also, the interaction of atoms only with the first coordination sphere was taken into account for GCNN (see the Supporting Information). This restriction reduces the time required for training the model but also makes it challenging to consider long-range forces. This approach can be compared to using a short cutoff radius in the Bayesian GPR framework.[10]

In contrast to the properties of crystalline materials, a network of pores primarily determines the properties of porous materials associated with adsorption. As has been shown, even only structural descriptors that ignore entirely chemical diversity are suitable for the clustering of porous materials.[56] Given the specifics of these properties, pore-centered descriptors, instead of atom-centered ones, used successfully to predict the properties of crystalline materials, seem to be a more appropriate choice.

CONCLUSION

In this study, we have shown that GCNN architecture with a minimal set of interpretable descriptors could be a universal tool for fast initial screening and search of perspective materials from the various subdomains of chemical space. Its performance was tested with a broad set of chemical domain-specific properties, including biological activity for organic molecules, thermomechanical properties for inorganic crystals, exfoliation energy for potentially exfoliable materials, elastic moduli, and infinite dilution selectivity of Xe/Kr for porous materials. Except for the last subdomain, the GCNN models demonstrate excellent accuracy, which is comparable with the best known approaches. However, even for the domain of porous materials GCNN models can be applied for an initial search of advanced materials for specific applications. GCNN models are still vulnerable to the effects of data scarcity and class imbalance data, but at the same time, graph-based NNs demonstrate state-of-the-art performance on well-balanced data sets with a sufficient number of samples.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b00587.

Lists of descriptors for periodic/nonperiodic structures, description of GCNN architecture and model training (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: korolev@colloid.chem.msu.ru.

ORCID

Vadim Korolev: 0000-0001-6117-5662

Artem Mitrofanov: 0000-0001-8891-6862

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Russian Science Foundation (project no. 19-73-20115).

DOI: 10.1021/acs.jcim.9b00587
J. Chem. Inf. Model. 2020, 60, 22−28

REFERENCES

(1) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864−B871.
(2) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133−A1138.
(3) Curtarolo, S.; Setyawan, W.; Hart, G. L. W.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; Mehl, M. J.; Stokes, H. T.; Demchenko, D. O.; Morgan, D. AFLOW: An Automatic Framework for High-Throughput Materials Discovery. *Comput. Mater. Sci.* **2012**, *58*, 218−226.
(4) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65*, 1501−1509.
(5) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 011002.
(6) Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 468−481.
(7) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.
(8) Seko, A.; Hayashi, H.; Nakayama, K.; Takahashi, A.; Tanaka, I. Representation of Compounds for Machine-Learning Prediction of Physical Properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, 144110.
(9) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3*, 54.
(10) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, No. e1701816.
(11) Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R. SchNet - A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(12) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564−3572.
(13) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 13890.
(14) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; Von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255−5264.
(15) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.
(16) Lecun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436−444.
(17) Defferrard, M.; Bresson, X.; Vandergheynst, P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. *Adv. Neural Inf. Process. Syst.* **2016**, 3844−3852.
(18) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 2224−2232.
(19) Zhang, Z.; Cui, P.; Zhu, W. Deep Learning on Graphs: A Survey. 2018, *arXiv Prepr. arXiv1812.04202*. https://arxiv.org/abs/1812.04202 (accessed Dec 27, 2019).
(20) Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A Comprehensive Survey on Graph Neural Networks. 2019, *arXiv Prepr. arXiv1901.00596*. https://arxiv.org/abs/1901.00596 (accessed Dec 27, 2019).
(21) Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. A Neural Device for Searching Direct Correlations between Structures and Properties of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 715−721.
(22) Fedorov, A. V.; Shamanaev, I. V. Crystal Structure Representation for Neural Networks Using Topological Approach. *Mol. Inf.* **2017**, *36*, 1600162.
(23) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nat. Commun.* **2017**, *8*, 15679.
(24) Such, F. P.; Sah, S.; Dominguez, M. A.; Pillai, S.; Zhang, C.; Michael, A.; Cahill, N. D.; Ptucha, R. Robust Spatial Filtering with Graph Convolutional Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 884−896.
(25) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513−530.
(26) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(27) Wang, S.; Li, Y.; Wang, J.; Chen, L.; Zhang, L.; Yu, H.; Hou, T. ADMET Evaluation in Drug Discovery. 12. Development of Binary Classification Models for Prediction of HERG Potassium Channel Blockage. *Mol. Pharmaceutics* **2012**, *9*, 996−1010.
(28) Guiguemde, W. A.; Shelat, A. A.; Bouck, D.; Duffy, S.; Crowther, G. J.; Davis, P. H.; Smithson, D. C.; Connelly, M.; Clark, J.; Zhu, F.; Jiménez-Díaz, M. B.; Martinez, M. S.; Wilson, E. B.; Tripathi, A. K.; Gut, J.; Sharlow, E. R.; Bathurst, I.; Mazouni, F. E.; Fowble, J. W.; Forquer, I.; McGinley, P. L.; Castro, S.; Angulo-Barturen, I.; Ferrer, S.; Rosenthal, P. J.; DeRisi, J. L.; Sullivan, D. J., Jr; Lazo, J. S.; Roos, D. S.; Riscoe, M. K.; Phillips, M. A.; Rathod, P. K.; Van Voorhis, W. C.; Avery, V. M.; Guy, R. K. Chemical Genetics of Plasmodium Falciparum. *Nature* **2010**, *465*, 311−315.
(29) Gamo, F. J.; Sanz, L. M.; Vidal, J.; De Cozar, C.; Alvarez, E.; Lavandera, J. L.; Vanderwall, D. E.; Green, D. V. S.; Kumar, V.; Hasan, S.; Brown, J. R.; Peishoff, C. E.; Cardon, L. R.; Garcia-Bustos, J. F. Thousands of Chemical Starting Points for Antimalarial Lead Identification. *Nature* **2010**, *465*, 305−310.
(30) Gagaring, K.; Borboa, R.; Francek, C.; Chen, Z.; Buenviaje, J.; Plouffe, D.; Winzeler, E.; Brinker, A.; Diagena, T.; Taylor, J.; Glynne, R.; Chatterjee, A.; Kuhen, K. Novartis-GNF Malaria Box. *ChEMBL-NTD*. https://chembl.gitbook.io/chembl-ntd/ (accessed Dec 27, 2019).
(31) Litterman, N. K.; Lipinski, C. A.; Bunin, B. A.; Ekins, S. Computational Prediction and Validation of an Experts Evaluation of Chemical Probes. *J. Chem. Inf. Model.* **2014**, *54*, 2996−3004.
(32) Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L. W.; Sanvito, S.; Buongiorno-Nardelli, M.; et al. AFLOWLIB.ORG: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations. *Comput. Mater. Sci.* **2012**, *58*, 227−235.
(33) Choudhary, K.; Zhang, Q.; Reid, A. C. E.; Chowdhury, S.; Van Nguyen, N.; Trautt, Z.; Newrock, M. W.; Congo, F. Y.; Tavazza, F. Computational Screening of High-Performance Optoelectronic Materials Using OptB88vdW and TB-MBJ Formalisms. *Sci. Data* **2018**, *5*, 180082.
(34) Toher, C.; Plata, J. J.; Levy, O.; De Jong, M.; Asta, M.; Nardelli, M. B.; Curtarolo, S. High-Throughput Computational Screening of Thermal Conductivity, Debye Temperature, and Grüneisen Parameter Using a Quasiharmonic Debye Model. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *90*, 174107.
(35) Choudhary, K.; Kalish, I.; Beams, R.; Tavazza, F. High-Throughput Identification and Characterization of Two-Dimensional Materials Using Density Functional Theory. *Sci. Rep.* **2017**, *7*, 5179.
(36) Coudert, F.-X. Systematic Investigation of the Mechanical Properties of Pure Silica Zeolites: Stiffness, Anisotropy, and Negative Linear Compressibility. *Phys. Chem. Chem. Phys.* **2013**, *15*, 16012−16018.
(37) Witman, M.; Ling, S.; Jawahery, S.; Boyd, P. G.; Haranczyk, M.; Slater, B.; Smit, B. The Influence of Intrinsic Framework Flexibility on

DOI: 10.1021/acs.jcim.9b00587
J. Chem. Inf. Model. 2020, 60, 22−28

Adsorption in Nanoporous Materials. *J. Am. Chem. Soc.* **2017**, *139*, 5547−5557.
(38) Gawehn, E.; Hiss, J. A.; Schneider, G. Deep Learning in Drug Discovery. *Mol. Inf.* **2016**, *35*, 3−14.
(39) Mamoshina, P.; Vieira, A.; Putin, E.; Zhavoronkov, A. Applications of Deep Learning in Biomedicine. *Mol. Pharmaceutics* **2016**, *13*, 1445−1454.
(40) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *J. Comput. Chem.* **2017**, *38*, 1291−1307.
(41) Aspuru-Guzik, A.; Lindh, R.; Reiher, M. The Matter Simulation (R)Evolution. *ACS Cent. Sci.* **2018**, *4*, 144−152.
(42) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.
(43) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D. A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **2018**, *9*, 5441−5451.
(44) Korotcov, A.; Tkachenko, V.; Russo, D. P.; Ekins, S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharmaceutics* **2017**, *14*, 4462−4475.
(45) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(46) Choudhary, K.; DeCost, B.; Tavazza, F. Machine Learning with Force-Field Inspired Descriptors for Materials: Fast Screening and Mapping Energy Landscape. *Phys. Rev. Mater.* **2018**, *2*, 083801.
(47) Siddorn, M.; Coudert, F. X.; Evans, K. E.; Marmier, A. A Systematic Typology for Negative Poisson's Ratio Materials and the Prediction of Complete Auxeticity in Pure Silica Zeolite JST. *Phys. Chem. Chem. Phys.* **2015**, *17*, 17927−17933.
(48) Van Beest, B. W. H.; Kramer, G. J.; Van Santen, R. A. Force Fields for Silicas and Aluminophosphates Based on Ab Initio Calculations. *Phys. Rev. Lett.* **1990**, *64*, 1955−1958.
(49) Sanders, M. J.; Leslie, M.; Catlow, C. R. A. Interatomic Potentials for SiO2. *J. Chem. Soc., Chem. Commun.* **1984**, No. 19, 1271−1273.
(50) Gale, J. D. Analytical Free Energy Minimization of Silica Polymorphs. *J. Phys. Chem. B* **1998**, *102*, 5423−5431.
(51) Tsuneyuki, S.; Tsukada, M.; Aoki, H.; Matsui, Y. First-Principles Interatomic Potential of Silica Applied to Molecular Dynamics. *Phys. Rev. Lett.* **1988**, *61*, 869−872.
(52) Sastre, G.; Gale, J. D. Derivation of an Interatomic Potential for Germanium- and Silicon-Containing Zeolites and Its Application to the Study of the Structures of Octadecasil, ASU-7, and ASU-9 Materials. *Chem. Mater.* **2003**, *15*, 1788−1796.
(53) Combariza, A. F.; Gomez, D. A.; Sastre, G. Simulating the Properties of Small Pore Silica Zeolites Using Interatomic Potentials. *Chem. Soc. Rev.* **2013**, *42*, 114−127.
(54) Evans, J. D.; Coudert, F. X. Predicting the Mechanical Properties of Zeolite Frameworks by Machine Learning. *Chem. Mater.* **2017**, *29*, 7833−7839.
(55) Zhang, Y.; Ling, C. A Strategy to Apply Machine Learning to Small Datasets in Materials Science. *npj Comput. Mater.* **2018**, *4*, 25.
(56) Lee, Y.; Barthel, S. D.; Dłotko, P.; Moosavi, S. M.; Hess, K.; Smit, B. Quantifying Similarity of Pore-Geometry in Nanoporous Materials. *Nat. Commun.* **2017**, *8*, 15396.

DOI: 10.1021/acs.jcim.9b00587
J. Chem. Inf. Model. 2020, 60, 22−28