Multitask Graph Convolutional Networks for Molecular Property Prediction

Sohom Paul[1], Emily Ye[1]

Abstract
Chemical machine learning has the potential to greatly accelerate the drug discovery process. However, in order for these machine learning techniques to be useful to pharmaceutical researchers, these networks must be able to achieve high accuracies on moderately-sized datasets. While prior research has developed a multitude of different neural architectures, we study the utility of multitask learning in learning quantum mechanical features, as well as compare two of the most promising neural architectures, graph isomorphism networks and edge-conditioned convolutions. We find that while multitask learning does not produce improved accuracies, the edge-conditioned convolutions greatly increases performance while reducing training times, providing a new avenue for future research.

[1] *Thomas Jefferson High School for Science and Technology, Alexandria VA 22312*

Contents

1. Introduction

1.1 Background

Drug research is expensive. While the global pharmaceutical market is projected to exceed $1.5 trillion by 2023 [1], recent estimates show that the median cost to bring a new therapeutic drug to market is $985 million [2]. Part of the reason that drug development is so costly is because only a small proportion of drugs pass each phase of preclinical and clinical trials. Approximately 13.8% of all drugs in the United States that reach phase I clinical trials make it to market; in some fields, like oncology, the success rate are as low as 3.4% [3]. By expanding the throughput and screening at early stages of research and development, we can hopefully increase the efficacy of the drug discovery process.

To this end, computer-aided drug design has seen heavy and continuing research over the past few decades. In particular, density functional theory (DFT), the premier quantum chemical modeling method, is critical in understanding drug-protein interactions [4, 5, 6, 7, 8, 9, 10]. DFT calculations of electronic properties forms an important part of drug research and screening [11, 12, 13, 14, 15, 16].

However, DFT scales poorly. Letting n_e be the number of electrons, the time complexity of DFT is between $O(n_e^2)$ (the cost of simulating pairwise electron-electron interactions) and $O(n_e^3)$ (the cost of diagonalizing the Hamiltonian on a basis set sized proportionally to the number of electrons). In practice, inference using neural networks tends to be more than 10^5 times faster than DFT, even for small molecules [17]. Furthermore, recent neural techniques using variational auto-encoders [18] and generative adversarial networks [19, 20] allow for the creation of large numbers of candidate molecules at the start of the drug design pipeline, so we need improved throughput. Thus, the goal in recent computational chemistry research has been to adapt and create neural methods to learn on graphical data and ultimately provide fast and accurate predictions of chemical properties. For the sake of generalizability of our work, we focus on learning electronic and thermodynamic properties rather than specific drug-ligand or solvent interactions.

1.2 Contributions

We analyze the potential benefits of multitask learning when applied to predict electronic and thermodynamic properties of small organic molecules taken from the QM9 dataset. We study both edge-conditioned convolutional networks and graph isomorphism networks. We find that while multitask learning does not produce increased accuracies for these quantum mechanical properties, the edge-conditioned convolutional

networks outperforms the popularly studied graph isomorphism network and indicates that this architecture should be studied further.

1.3 Related Work

Gilmer et al. in their seminal 2017 paper [17] consolidated the work of a large number of previous researchers by describing how previously-studied architectures were all special cases of their message-passing neural network, using message and vertex update functions to share hidden states of nodes of a graph. Gilmer et al. further extended the gated graph neural network (GG-NN) proposed by Li et al. (2016) [21] and achieved chemical accuracy on 11 of 13 targets in the QM9 dataset.

In order to address the dearth of labelled data in chemical machine learning (as labelling requires either expensive DFT calculations or, at worst, real-world laboratory experiments), Wang et al. (2019) created SMILES-BERT [22]. SMILES-BERT adapts the BERT natural language representation model [23]. Wang et al. pre-trained a semi-supervised model with attention in the transformer layer to recover masked SMILES strings. SMILES-BERT is a promising resource for future chemical machine learning and its application would be an interesting extension of our research.

Complementing our own work in convolutional networks applied to molecular data, Shindo and Matsumoto (2019) [24] extended the results of Li et al. (2016) [21] and Gilmer et al. (2017) [17] to create a gated graph recurrent neural network (GGRNN). Their GGRNN uses 3D coordinates of the atoms as additional input to the network and includes skip connections so the input representations are fed into each level of the network. Shindo and Matsumoto found that their GGRNet exceeded the performance of graph convolutional and MPNN networks on all non-thermodynamic properties on the QM9 dataset. However, their use of geoemtric information as input to the network makes their network difficult to use on molecules outside of the dataset without solved geometries.

The work of Yang et al. (2019) [25] evaluates the relative effectivenesses of convolution- and descriptor-based methods to learn molecular representations in a molecular property prediction pipeline. They conclude that while convolutions are prone to overfit the training data, convolutional methods are superior on large datasets ($>$1000 training molecules).

Our work is most similar to Capela et al. (2019) [26], which also studies the multitask paradigm applied to various graph neural network models to predict physical chemical properties. The main distinction between this and our work is the problem domain; we are are studying quantum mechanical rather than physical chemical properties, potentially forcing successful neural model to learn representations of additional electronic and geometric interactions. Also, as outlined above, much current drug research requires the computation of QM properties of drug candidates in order to understand protein binding [4]. Finally, as a practical consideration, the datasets used by [26] (which can be found at [27]) for physical chemical properties are much smaller than that for quantum chemical properties (~1000 versus ~100000). We sought to investigate if the positive inductive bias from multitask learning still improves accuracy even with substantially larger training sets.

2. Methods

2.1 Graph Convolutional Networks

Convolutional neural networks (CNNs) are completely ubiquitous in computer vision and other applications. CNNs are interesting theoretically because convolutional layers impose shift-invariance, meaning a translation of the input image produces the same translation of the output of the convolution. This allows CNNs to learn generalizable features that can apply across the image domain. This property makes the application of convolutional neural networks to molecular data promising; for example, a CNN ought to be able to learn functional groups that can appear in various locations across the molecule. Modifying the classic CNN architecture to admit graphical data has been extensively studied [28, 29, 30, 31, 17]. In order to indulge the graphical structure of the data, these networks involve various forms of feature aggregation between a node and its neighbors, similarly to how a convolutional kernel mixes the state of a pixel and its surrounding pixels. In this work, we focus on two promising proposed architectures, the edge-conditioned convolutional network and graph isomorphism networks.

2.1.1 Edge-Conditioned Convolutions

Simonovsky and Komodakis (2017) [31] argued that feature aggregation steps in prior architectures, such as summing adjacent node features in Duvenaud et al. (2015) [29], unnecessarily destroyed information about graphical structure by sharing the same weights across all edges. The edge-conditioned convolution aggregates the neighboring nodes' hidden states using learned weight matrices unique to each edge in the graph. Formally, the propagation rule is given as

$$X^{(l)}(i) = \frac{1}{|N(i)|} \sum_{j \in N(i)} \Theta^{(l)}_{ji} X^{(l-1)}(j) + b^{(l)} \tag{1}$$

where i, j are vertices, $X^{(l)}$ is the output of the lth layer in the feed-forward network, $N(i)$ is the neighborhood of vertex i, $\Theta^{(l)}_{ji}$ is a learned matrix for the lth layer and ji edge, and $b^{(l)}$ is the bias vector for the lth layer. Conditioning the weight matrix on the edges in this manner in practice leads to the network identifying important graph structures, improving empirical performance.

2.1.2 Graph Isomorphism Network

Xu et al. (2019) [32] drew inspiration between the similarities between various graph neural network architectures and the Weisfeiler-Lehman test for graph isomorphim. The authors prove that their proposed architecture, the graph isomorphism network (GIN), is as expressive as the Weisfeiler-Lehman test.

The GIN has as its update step

$$X^{(l)}(i) = \text{MLP}^{(l)}\left(\sum_{j\in N(i)} X_j^{(l-1)} + (1+\varepsilon^{(l)})X^{(l-1)}(i)\right) \quad (2)$$

using the same variable naming conventions as Equation (1) with $\varepsilon^{(l)}$ being some parameter for each layer and MLP a learned multilayer perceptron. Xu et al. perform the final readout by concatentating the sum of node features for each layer.

2.2 Multitask Learning

Multitask learning has seen usage across diverse areas of machine learning as a form of implicit data augmentation [33]. Simply put, given some input data for which we want to predict multiple descriptors, multitask learning trains several networks with shared layers. For example, in Figure 1 depicting our final model architecture, we see how three tasks (corresponding to the 3 dense layers at the bottom of the diagram) are learned simultaneously while sharing the same convolutions and inputs. Multitask learning is useful because it can create a positive inductive bias among tasks [34]. Because the same convolutional layers must be learned from all 3 tasks below, the network is forced to construct some molecular representation that can generalize well to different applications. However, multitask learning with dissimilar tasks runs the risk of failing to find a good representation for any of the tasks.

2.2.1 Task Clustering

However, in order to leverage the potential benefits of multitask learning, we have to ensure that our task clusters contain similar tasks. Otherwise, training the tasks simultaneously can lead to negative inductive bias and worse training and test losses. Rather than hand-select task clusters, we chose to use a fully automated solution. We used our single-task trained networks to compute cross-task transfer scores and then performed spectral clustering on the resultant similarity matrix.

Cross-Task Transfer Scores Our method of constructing a task similarity matrix has been adapted from Yu et al. (2017) [35]. Because we wanted to compare the performances of single-task and multitask networks, we already had trained single-task networks on all the tasks. Thus, we could use transfer learning in order to estimate the similarity between the molecular representations learned by the convolutional layers in each network.

Conceptually, our single-task networks can be considered encoder-decoder pairs where the training was performed end-to-end. Thus, to compute the similarity S_{ij} between tasks i and j, we merely compute $\mathscr{L}(M_j^{\text{dec}}(M_i^{\text{enc}}(x_j)), y_j)$, where (x_j, y_j) is a pair of vectors containing the testing data for task j. labeled test set for task j and $\mathscr{L}$ is some loss function. We chose mean squared error as our loss for this step. Computing S_{ij} for all pairs i, j gives us our similarly matrix S. Note that S_{ii} merely represents the testing loss. Then, we used spectral clustering with our similarity matrix to create task clusters.

Spectral Clustering Using the similarity matrix S described in the previous section, we constructed 6 task clusters using the spectral clustering implementation contained in the `scikit-learn` library [36, 37]. The algorithm is described in Figure 1 below, adapted from [38]. Essentially, we cluster based on components of the first k eigenvalues of the similarity matrix in a low-dimensional embedding of the matrix, where k is our desired number of clusters. Note that in order to guarantee that our matrix has real eigenvalues, we must first symmetrize our similarity matrix: $S' = (S + S^\top)/2$.

- compute graph Laplacian $L = nI - S'$, where n is the number of tasks
- compute the first k eigenvectors $v_1, \ldots, v_k$ of L, sorted in decreasing order by size of the eigenvalue
- define U to be the matrix containing $v_1, \ldots, v_k$ as columns
- define $(y_i)_{i=1}^n$ as the rows of U
- apply k-means clustering to (y_i)

Figure 1. Spectral Clustring Algorithm

2.3 Dataset and Training Details

We used the QM9 dataset. QM9, compiled by researchers in [39, 40] and available for download at [41], is a dataset of 133,885 small organic molecules with at most nine heavy atoms (carbon, oxygen, nitrogen, fluorine [CONF]). The properties included in the dataset is summarized in Table 1. The labels in the dataset were computed using DFT with the B3LYP/6-31G(2df,p) functional. Even though QM9 does not contain experimentally-verified labels, we choose to use this dataset because experimental data is extremely sparse and prohibitively difficult to compile and train on. Instead, we make the standard assumption that any architecture that can be trained to high accuracy on DFT-calculated values can later be retrained to high accuracy on true values. Furthermore, QM9 is a standard dataset, making is easier for our results to be compared to other methods in the literature.

Even though the QM9 dataset provides geometric data about our molecules, giving the locations of the constituent atoms in 3D space, we fed our network only the information that could be inferred from a SMILES string. We opted not to use the solved 3D geometry because calculating that geometry requires a DFT call almost as expensive as solving for the property directly. Expecting geometry-labeled data is not helpful in practice. Our only node feature was a one-hot encoding of the atomic number and our only edge feature was a one-hot encoding of bond type (single, double, or triple).

Finally, we noticed that many of our tasks had extremely skewed data. For example, while more than 95% of the data for `A` was in the range 0 to 6 GHz, there was one data point at 619867 GHz. Rather than corrupt the dataset by excluding

Table 1. Properties in QM9

Property	Unit	Description
A	GHz	rotational constant A
B	GHz	rotational constant B
C	GHz	rotational constant C
mu	Debye	dipole moment
alpha	Bohr^3	isotropic polarizability
homo	Hartree	energy of highest occupied molecular orbital
lumo	Hartree	energy of lowest unoccupied molecular orbital
r2	Bohr^2	electronic spatial extent
zpve	Hartree	zero point vibrational energy
u0	Hartree	internal energy at 0 K
u298	Hartree	internal energy at 298.15 K
h298	Hartree	enthalpy at 298.15 K
g298	Hartree	Gibbs free energy at 298.15 K
cv	cal/(mol K)	constant volume heat capacity at 298.15 K

outliers, we chose to preprocess our data by normalizing and applying a Yeo-Johnson transformation [42]. Yeo-Johnson tranformations are a generalization of the power and boxcox transforms to nonpositive inputs. Yeo-Johnson transformations, with formula given below, are characterized by a parameter λ computed on the data to minimize the skew of the output. Note that while our training losses measure the difference between the transformed predictions and transformed labels, we report our final errors in terms of the untransformed predictions and labels in order to see how useful our network would be for real-world uses.

$$f_\lambda(y) = \begin{cases} ((y+1)^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y+1) & \text{if } \lambda = 0, y \geq 0 \\ -\frac{(-y+1)^{2-\lambda}-1}{2-\lambda} & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y+1) & \text{if } \lambda = 2, y < 0 \end{cases} \tag{3}$$

All of our code is publicly hosted on Github[1]. We used the Spektral graphical deep learning library [43], written on top of Tensorflow [44].

3. Results and Discussion

The clusters used for multitask learning are shown in Table 2, and the hand-selected hyperparameters used for training are shown in Table 3. The layer widths used are shown in the model architecture visualization in Figure 4. A comparison of the number of learned parameters in a single-task network for each type of convolution is shown in Table 4. Finally, to evaluate our networks, we computed the mean absolute error (MAE) for each testing set and reported this error in comparison to the thresholds for chemical accuracy used by prior researchers in Gilmer et al. (2017) [17] and Faber et al. (2017) [45]. All of this information can be found in Table 5. Bolded entries are the best of the row; italicized entries are those that meet chemical accuracy (e.g. estimated error is lower than target error).

Table 2. Training Clusters

[A, lumo, homo]
[B, r2, cv]
[alpha, zpve]
[C, u0, u298, mu]
[g298, h298]

Table 3. Hyperparameters

Learning rate	0.001
Batch size	32
Training samples	30000
Epochs	40
Loss	Mean Absolute Error
Optimizer	Adam

Table 4. Number of Learned Parameters

Conv	Params
ECC	693121
GIN	559619

We see that in all but two cases (for h298 and g298) multitask learning does not perform as well as single-task learning. Our results are at odds with the findings of Capela et al. (2019) [26], who performed a similar experiment on physical chemical data and found multitask learning outperformed single-task learning. There are two possible conclusions for this. The first is that the underlying tasks in the QM9 dataset are too dissimilar for positive inductive bias through multitask learning. However, given the intuitive similarity between tasks like learning internal energy at 0 K and internal energy at 298 K, this seems unlikely. The more plausible explanation was that our datasets were too large to benefit from multitask learning. Most of the datasets in Capela et al. are smaller than 5000 molecules, and they found that multitask learning only outperformed single-task learning when using less than 50% of the dataset for learning LogP, a task with 14000 samples. Considering that our training set samples 30000 entries from the dataset, it stands to reason that the data augmentation offered by multitask learning did not aid our networks because there was already enough data to learn generalizable trends.

The other thing of note was that our single-task networks achieved chemical accuracy (score less than 1) on only 3 of the tasks (homo, lumo, and zpve), as opposed to Gilmer et al., who achieved chemical accuracy on all but two of the tasks they studied. The most likely reason for this disparity

[1] https://github.com/tjresearch/research-sohom_emily

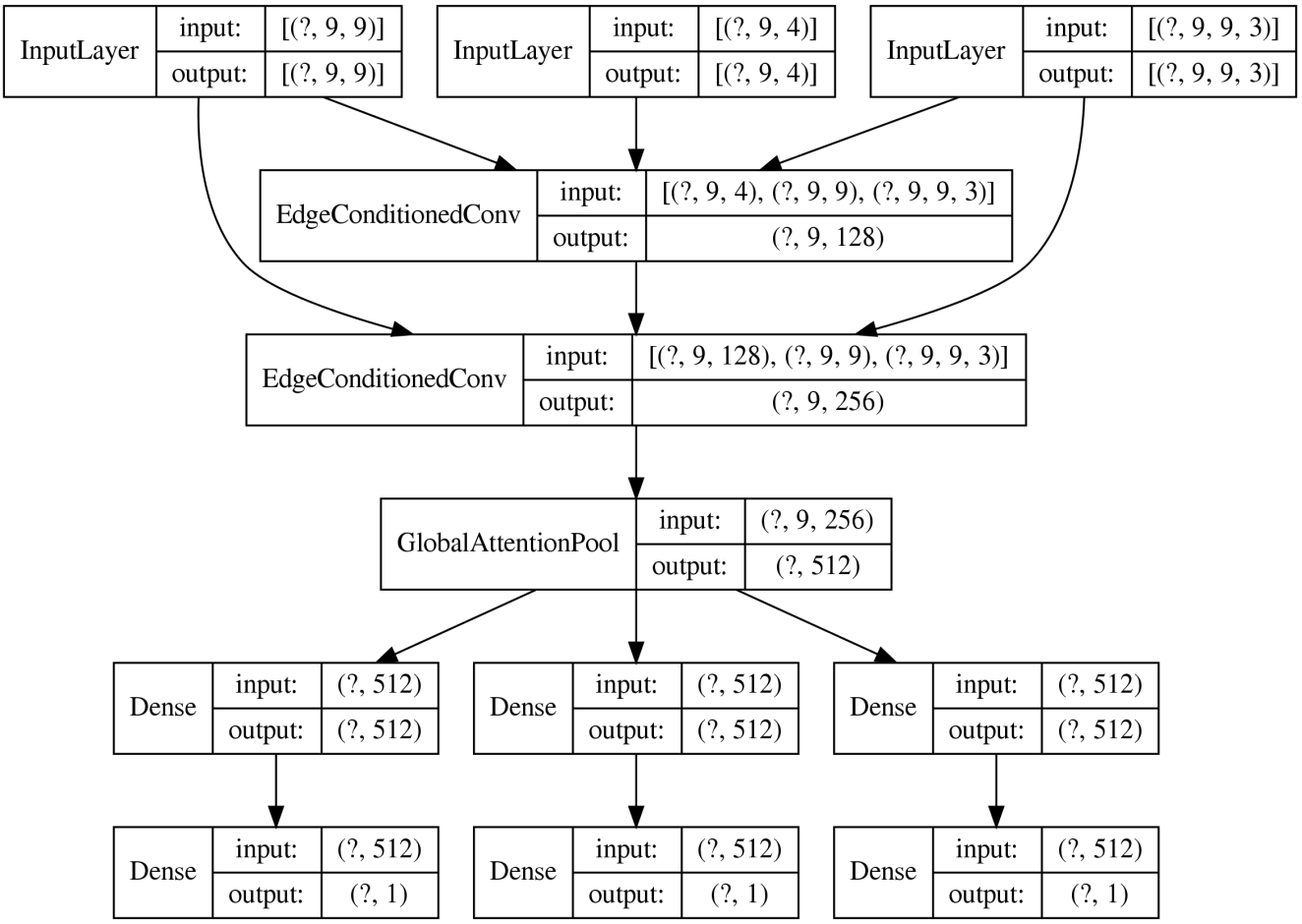

Figure 2. Model Architecture with Input Shapes

is that Gilmer et al. used geoemetric data when training their networks, meaning that they fed in the 3D coordinates of each of the atoms into their network. While this use of geometric data leads to stunning state-of-the-art results, these networks are difficult to apply to the real world, where solving for the geometry of molecules is incredibly expensive and negates the point of having a fast neural predictor in the first place. Thus, even though we did not achieve chemical accuracy on most of the tasks, our results (within an order of magnitude for most tasks) are promising for real-world application.

4. Conclusions and Open Questions

We find that while multitask learning did not produce performance gains for learning quantum mechanical properties, reaffirming the need for training separate networks for separate chemical tasks. However, we also find that the edge-conditioned convolutional network is a powerful tool for learning chemical properties, out-performing the graph isomorphism network that has been the subject of much past research [26, 17, 46, 47]. Future studies should look further to study the learning capacity of these networks.

Also, this study only looked at simple properties of small molecules. There are two axes in which to extend this work. The first is seeking to understand how these networks may learn features that may help us better understand specific drug-ligand interactions; for example, we would like to be able to label molecules with regions that are likely to interact with a specific protein. The second axis is to expand the size of the molecules, seeing if our networks our capable of learning geometries of more complicated molecules. While our results show that feeding in only bond connectivities results in less accuracy than feeding in geoemetric data, it remains open whether neural networks are able to meaningfully learn geometry as well as molecular properties, which would be an exciting step towards a better understanding of chemistry and molecular biology.

5. Acknowledgments

We thank Dr. Patrick White and Dr. Peter Gabor for advising and overseeing this work, and we are grateful to TJCSL for providing us with access to computational resources.

Table 5. Accuracies

Task	Target	Single ECC	Single GIN	Multi ECC	Multi GIN
A	-	**0.3807**	0.5988	0.8262	0.9116
B	-	**0.1096**	0.1912	0.5207	0.4821
C	-	**0.0596**	0.1091	0.2658	0.2615
mu	0.1	**0.5986**	0.7512	1.8025	1.7889
alpha	0.1	**0.6219**	1.4663	2.7315	1.8966
homo	0.043	***0.0053***	*0.0097*	*0.0125*	*0.0130*
lumo	0.043	***0.0063***	*0.0273*	*0.0263*	*0.0328*
r2	1.2	**42.7545**	90.8275	103.6453	121.0475
zpve	0.0012	***0.0010***	0.0138	0.0070	0.0166
u0	0.043	**0.2375**	0.8871	0.8205	2.2232
u298	0.043	**0.3653**	1.1732	1.1946	2.1242
h298	0.043	0.3654	0.8069	**0.3532**	0.8076
g298	0.043	0.5945	1.0879	**0.3024**	0.8347
cv	0.05	**0.3393**	1.2018	1.0241	1.6607

References

[1] IQVIA Institute. The global use of medicine in 2019 and outlook to 2023. Technical report, Jan 2019.

[2] Olivier J. Wouters, Martin McKee, and Jeroen Luyten. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *JAMA*, 323(9):844–853, 03 2020.

[3] Chi Heem Wong, Kien Wei Siah, and Andrew W. Lo. Estimation of clinical trial success rates and releated parameters. *Biostatistics*, 20(2):273–286, 2018.

[4] Olayide A. Arodola and Mahmoud E. S. Soliman. Quantum mechanics implementation in drug-design workflows: does it really help? *Drug Design, Development and Therapy*, 11:2551–2564, 2017.

[5] Jeffrey Augen. The evolving role of information technology in the drug discovery process. *Drug Discovery Today*, 7:315–323, Mar 2002.

[6] Hiteshi Tandon, Tanmoy Chakraborty, and Vandana Suhag. A brief review on importance of DFT in drug design. 7, Feb 2019.

[7] Kenny Lipkowitz. Abuses of molecular mechanics: Pitfalls to avoid. *Journal of Chemical Education*, 72, Dec 1995.

[8] Alejandro Crespo, Agustina Rodriguez-Granillo, and Victoria T. Lim. Quantum-mechanics methodologies in drug discovery: Applications of docking and scoring in lead optimization. *Current Topics in Medicinal Chemistry*, 17, 2017.

[9] Steven Volney Jerome. *Methods Development in Quantum Mechanics and Molecular Mechanics for Drug Discovery*. PhD thesis, Columbia University, 2015.

[10] Claudio Cavasotto, María Gabriela Aucar, and Natalia S. Adler. Computational chemistry in drug lead discovery and design. *International Journal of Quantum Chemistry*, 119, Sep 2018.

[11] Raya Ahmadi, Mohammad Reza Jalali Sarvestani, and Babak Sadeghi. Computational study of the fullerene effects on the properties of 16 different drugs: A review. *International Journal of Nano Dimension*, 9:325–335, 2018.

[12] Ankush W. Wakode, Archana S. Burghate, and Shrikant A. Wadhal. Studies of electric dipole moment and magnetic dipole moment of substituted benzothiazolyl and benzimidazolyl derivatives. *Oriental Journal of Chemistry*, 33, 2017.

[13] Mohamed Hagar, Hoda A. Ahmed, Ghadah Aljohani, and Omaima A. Alhaddad. Investigation of some antiviral N-heterocycles as COVID 19 drug: molecular docking and DFT calculations. *International Journal of Molecular Science*, 21, 2020.

[14] Enrico Redenti, Lajos Szente, and Josef Szejtli. Cyclodextrin complexes of salts of acidic drugs. thermodynamic properties, structural features, and pharmaceutical applications. *Journal of Pharmaceutical Sciences*, 90:979–986, Aug 2001.

[15] Etratsadat Dadfar and Fatemeh Shafiei. Prediction of some thermodynamic properties of sulfonamide drugs using genetic algorithm-multiple linear regressions. *Journal of the Chinese Chemical Society*, 67, Sep 2019.

[16] Shoji Hirokawa, Tomoko Imasaka, and Totaro Imasaka. Chlorine substitution pattern, molecular electronic properties, and the nature of the ligand-receptor interaction: Quantitative property-activity relationships of polychlorinated dibenzofurans. *Chemical Research in Toxicology*, 18:232–238, Jan 2005.

[17] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the*

34th International Conference on Machine Learning - Volume 70, ICML'17, page 1263–1272. JMLR.org, 2017.

[18] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4:268–276, Jan 2018.

[19] Artur Kadurin, Sergey Nikolenko, Kuzma Khrabrov, Alex Aliper, and Alex Zhavoronkov. druGAN: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Molecular Pharmaceutics*, 14:3098–3104, Jul 2017.

[20] Thomas Blaschke, Marcus Olivecrona, Ola Engkvist, J urgen Bajorath, and Hongming Chen. Application of generative autoencoder in de novo molecular design. *Molecular Informatics*, 37, Dec 2017.

[21] Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. Gated graph sequence neural networks. In *Proceedings of ICLR'16*, April 2016.

[22] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, page 429–436, New York, NY, USA, 2019. Association for Computing Machinery.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding.

[24] Hiroyuki Shindo and Yuji Matsumoto. Gated graph recursive neural networks for molecular property prediction.

[25] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Tim Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59, 07 2019.

[26] Fabio Capela, Vincent Nouchi, Ruud van Deursen, Igor V. Tetko, and Guillaume Godin. Multitask learning on graph neural networks applied to molecular property predictions. *arXiv preprint arXiv:1910.13124*, 2019.

[27] MoleculeNet. Dataset collection. `http://moleculenet.ai/datasets-1`.

[28] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann Lecun. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, 2014.

[29] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2224–2232. Curran Associates, Inc., 2015.

[30] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3844–3852, Red Hook, NY, USA, 2016. Curran Associates Inc.

[31] Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. pages 29–38, 07 2017.

[32] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

[33] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *ArXiv*, abs/1706.05098, 2017.

[34] Rich Caruana. Multitask learning. *Machine Learning*, 28, 1997.

[35] Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. Robust task clustering for deep many-task learning, 2017.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[37] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[38] Ulrike Von Luxburg. A tutorial on spectral clustering, 2007.

[39] Lars Ruddigkei, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52:2864–2875, 2012.

[40] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry

structures and properties of 134 kilo molecules. *Scientific Data*, Aug 2014.

[41] Quantum Machine homepage. `http://quantum-machine.org/datasets/`. Accessed: 2020-06-27.

[42] In-Kwon Yeo and Richard A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 12 2000.

[43] Daniele Grattarola. Spektral. `https://github.com/danielegrattarola/spektral`, May 2020.

[44] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[45] Felix A. Faber, Luke Hutchinson, Bing Huang, Justin Gilmer, Samuel S. Schoenholz, George E. Dahl, Oriol Vinyals, Steven Kearnes, Patrick F. Riley, and O. Anatole von Lilienfeld. Prediction errors for molecular machine learning models lower than hybrid DFT error. *Journal of Chemical Theory and Computation*, 13:5255–5264, Nov 2017.

[46] Yu Xie, Maoguo Gong, Yuan Gao, A. Kai Qin, and Xiaolong Fan. A multi-task representation learning architecture for enhanced graph classification. *Frontiers in Neuroscience*, 13, 2019.

[47] Federico Errica, Marco Podda, Davide Bacciu, and Alessio Micheli. A fair comparison of graph neural networks for graph classification. *ArXiv*, abs/1912.09893, 2020.