Efficient Training of Transformers for Molecule Property Prediction on Small-scale Datasets

Shivesh Prakash, *Student Member, IEEE, and Rishit Dagli, Student Member, IEEE,*

Abstract—The blood-brain barrier (BBB) serves as a protective barrier that separates the brain from the circulatory system, regulating the passage of substances into the central nervous system. Assessing the BBB permeability of potential drugs is crucial for effective drug targeting. However, traditional experimental methods for measuring BBB permeability are challenging and impractical for large-scale screening. Consequently, there is a need to develop computational approaches to predict BBB permeability. This paper proposes a GPS Transformer architecture augmented with Self Attention, designed to perform well in the low-data regime. The proposed approach achieved a state-of-the-art performance on the BBB permeability prediction task using the BBBP dataset, surpassing existing models. With a ROC-AUC of 78.8%, the approach sets a state-of-the-art by 5.5%. We demonstrate that standard Self Attention coupled with GPS transformer performs better than other variants of attention coupled with GPS Transformer.

Keywords—Blood-Brain Barrier Permeability, Molecule Property Prediction, Transformers, Geometric Learning

I. INTRODUCTION

The blood-brain barrier acts as a protective shield, separating the brain from the circulatory system of the body. Its primary function is to restrict the passage of solutes from the bloodstream into the central nervous system, where neurons reside [1].

For drugs to effectively act on the brain, they must be capable of crossing the blood-brain barrier. Conversely, drugs intended for peripheral action should exhibit limited ability to penetrate this barrier in order to avoid undesired effects on the central nervous system. Therefore, it is crucial to determine the blood-brain barrier permeability of potential drugs. However, experimental methods for measuring brain-blood permeability are challenging, time-consuming, expensive, and impractical for large-scale screening of chemicals [2].

The blood-brain barrier's highly selective nature enables endothelial cells to tightly regulate the central nervous system's homeostasis [3]. This regulation is vital for proper neuronal function and provides protection against toxins, pathogens, inflammation, injury, and disease. However, the restrictive properties of the blood-brain barrier pose a significant obstacle to drug delivery into the central nervous system. Consequently, extensive research has been conducted to develop methods for modulating or circumventing the blood-brain barrier to enhance therapeutic delivery [1].

The process of brain entry for compounds is multifaceted and influenced by various factors. Lipophilic drugs, for in-

S. Prakash is with the Department of Computer Science, University of Toronto, Canada, ON, M5S 2E4 E-mail: shivesh.prakash@mail.utoronto.ca

R. Dagli is with the Department of Computer Science, University of Toronto, Canada, ON, M5S 2E4

stance, can passively diffuse across the blood-brain barrier, with their ability to form hydrogen bonds playing a role. In contrast, polar molecules typically face difficulties in crossing this barrier, although active transport mechanisms can facilitate their penetration [4]. Local hydrophobicity, ionization profile, molecular size, lipophilicity, and flexibility are among the important parameters influencing blood-brain barrier permeation [5].

For compounds targeting the central nervous system, it is crucial to achieve brain penetration, ensuring the site of action is reached effectively [6]. Conversely, compounds designed for peripheral targets should minimize blood-brain barrier penetration to reduce potential central nervous system-related side effects. Therefore, selecting compounds with appropriate brain penetration properties is a critical consideration during the drug discovery phase [6].

In this paper we focus on producing more powerful results using smaller datasets, thus increasing the computational efficiency.

The contributions of this paper can be summarized as:

- A GPS Transformer architecture similar to GPS [7] and GPS++ [8] paired with Self Attention which is able to perform well in the low-data regime is proposed.
- With this approach a new state-of-the-art for the task of predicting Blood-Brain Barrier permeability using the BBBP dataset by Wu, Ramsundar, Feinberg, *et al.* [9] was established, and to the best of our knowledge this is the first work to use Self Attention coupled with the GPS Transformer to improve models in this domain. A 78.8 ROC-AUC was achieved, beating the current state-of-the-art [10] by 5.5.

II. RELATED WORK

The paper by Miao, Xia, Chen, *et al.* [11] proposes a Deep Learning method for predicting the blood-brain barrier permeability based on clinical phenotypes data. The method aims to overcome the limitations of existing prediction approaches that rely on physical characteristics and chemical structure of drugs, which are typically applicable only to small molecule compounds that passively diffuse through the BBB.

The Deep Learning method leverages clinical phenotypes data to train a predictive model. Clinical phenotypes refer to observable characteristics or traits associated with drugs, such as drug side effects and drug indications. By incorporating this information into the prediction model, the method aims to capture more comprehensive and complex mechanisms of BBB penetration beyond passive diffusion.

Miao, Xia, Chen, *et al.* [11] validate their Deep Learning method using three datasets. The validation results demonstrate that their approach outperforms existing methods in terms of prediction accuracy. Specifically, the average accuracy achieved by their method is 0.97, the area under the receiver operating characteristic curve (AUC) is 0.98, and the F1 score is 0.92. These performance metrics indicate a high level of accuracy in predicting BBB permeability. By incorporating clinical phenotypes data into the predictive model, the method demonstrates superior performance compared to existing physical, chemical, and supervised learning approaches [11].

Jain and Shanmuganathan [12] in their paper address the limitations of existing models used to predict blood-brain barrier (BBB) permeability by exploring the role of inflammation in influencing BBB permeability. Inflammation, measured through acute phase C-reactive protein (CRP) levels, has been found to have a direct correlation with BBB permeability. A threshold of 2.5 μg/ml of CRP is established, above which BBB impairment is expected.

To overcome the challenges of existing models, a custom-built dataset of 281 molecules is created, and a machine learning approach is employed. Initially, different models such as multi-layer perceptron regression (MPR) and ensemble models are tested, but they struggle with outliers and lack the ability to predict descriptors for future drugs. Finally, a fully connected neural network (FCNN) is chosen as the predictive logBB model, which performs well in terms of learning the distribution and achieving lower error levels.

The research also incorporates a neuroinflammation model, which considers CRP levels alongside the logBB model. Although CRP levels do not directly correlate with logBB values, the model explores the second and third-order feature derivatives that show significant correlations. A quadratic polynomial regression model is used to determine this correlation. The software developed for this research includes preprocessing, the predictive logBB model, and the neuroinflammation model, providing a continuous logBB value that represents BBB permeability based on the patient's inflammation level. This approach aims to improve the accuracy of predicting BBB permeability for drugs and better understand the impact of inflammation on BBB function.

The study by Saber, Mhanna, and Rihana [13] focuses on predicting drug permeability across the blood-brain barrier (BBB) using in silico models. The researchers compare the performance of sequential feature selection (SFS) and genetic algorithms (GA) in selecting relevant molecular descriptors to enhance the accuracy of permeability prediction. Five different classifiers are trained initially using eight molecular descriptors, and then SFS and GA are separately applied to choose the descriptors for each algorithm.

The results show that both SFS and GA improve the accuracy of the classifiers, but GA outperforms SFS. The highest accuracy of 96.23% is achieved with GA, specifically with a fitness function based on the performance of a support vector machine. The study highlights the significance of the polar surface area (PSA) of drugs in crossing the BBB. GA consistently selects the PSA and the number of hydrogen bond donors as the most relevant descriptors, providing better results compared to using other features.

The findings suggest that GA is a more robust approach for selecting relevant descriptors in predicting BBB permeability. The selected classifiers demonstrate a good balance in predicting BBB+ and BBB- drugs. Accurate in silico BBB models are crucial for early-phase drug discovery, reducing the need for extensive in vitro testing and saving time and resources.

The paper by Yuan, Zheng, and Zhan [14] describes the importance of predicting the permeability of compounds through the blood-brain barrier (BBB) for drug discovery targeting the brain. It discusses the use of computational methods, particularly support vector machine (SVM), in predicting BBB permeability. Different types of descriptors, such as molecular property-based descriptors (1D, 2D, and 3D descriptors) and fragment-based descriptors (fingerprints), have been utilized in SVM models. The selection of descriptors greatly influences the performance of the SVM model.

The paper aims to develop a new SVM model by combining molecular property-based descriptors and fingerprints to improve the accuracy of BBB permeability prediction. The results indicate that the proposed SVM model outperforms existing models in predicting BBB permeability. The paper also provides an overview of the blood-brain barrier and the complex factors that influence compound penetration into the brain. Additionally, it discusses the classification of quantitative and qualitative BBB permeability prediction models and highlights SVM as a superior method for qualitative prediction. The paper concludes that the combined use of property-based descriptors and fingerprints improves the accuracy of SVM models and suggests that similar approaches may enhance computational predictions for other molecular activities in the future.

However, our work explores how we find a simple modification to a General, Powerful, Scalable Graph Transformer (GPS) Rampášek, Galkin, Dwivedi, *et al.* [7] and improve upon the performance for the task of blood-brain barrier permeability while showing the effectiveness of simple methods within the graph setting.

III. DATA

We perform our experiments on the dataset 'BBBP: Binary labels of blood-brain barrier penetration (permeability)' which is a classification dataset available as a part of MoleculeNet [9] which contains 2,050 molecules and each molecule comes with a name, label, and SMILES string. The label is a boolean integer which denotes if a given compound can pass through the blood-brain barrier or not.

The Blood–brain barrier penetration (BBBP) dataset comes from a study by Martins, Teixeira, Pinheiro, *et al.* [15] on the modeling and prediction of the barrier permeability using a Bayesian approach. For this well-defined target scaffold splitting is also recommended by Wu, Ramsundar, Feinberg, *et al.* [9].

In the paper by Martins, Teixeira, Pinheiro, *et al.* [15], a unique approach based on Bayesian statistics, in conjunction with state-of-the-art machine learning techniques, was employed to create a robust model suitable for real-world drug research applications. The dataset used for the Bayesian analysis

consisted of 1970 carefully curated molecules, making it one of the largest, but still small, datasets used in similar studies. Various configurations of Random Forests and Support Vector Machines, coupled with different combinations of chemical descriptors, were evaluated. To assess the performance of the models, a 5-fold cross-validation process was employed, and the best-performing model was further tested on an independent validation set.

The model proposed by Martins, Teixeira, Pinheiro, *et al.* [15] achieved an impressive overall accuracy of 95%, as measured by a mean square contingency coefficient (ϕ) of 0.74. Furthermore, this model exhibited a high capacity for predicting blood-brain barrier (BBB) positives, with an accuracy of 83%, and a remarkable accuracy of 96% in determining BBB negatives. These findings highlight the efficacy of the developed model in accurately predicting the permeability of molecules across the blood-brain barrier, based on experiments by Martins, Teixeira, Pinheiro, *et al.* [15].

IV. METHODS

The proposed approach closely follows the framework of General, Powerful, Scalable Graph Transformer (GPS) by Rampášek, Galkin, Dwivedi, *et al.* [7] and that of GPS++ by Masters, Dean, Klaser, *et al.* [8]. This paper explores how these frameworks could be used to learn from low amounts of data. Thus, the BBBP dataset was used which contains only 2050 molecules.

A. GPS Block

Masters, Dean, Klaser, *et al.* [8] proposed that at each layer, the features are updated by aggregating the output of an MPNN layer with that of a global attention layer which is described by the equations below. Note that the edge features are only passed to the MPNN layer, and that residual connections with batch normalization [16] are omitted for clarity. Both the MPNN and GlobalAttn layers are modular, i.e., MPNN can be any function that acts on a local neighborhood and GlobalAttn can be any fully-connected layer.

$$\mathbf{X}^{\ell+1}, \mathbb{E}^{\ell+1} = \text{GPS}^\ell \left(\mathbf{X}^\ell, \mathbb{E}^\ell, \mathbf{A} \right), \quad (1)$$

computed as
$$\mathbf{X}_M^{\ell+1}, \ \mathbb{E}^{\ell+1} = \text{MPNN} \ _e^\ell \left(\mathbf{X}^\ell, \mathbb{E}^\ell, \mathbf{A} \right), \quad (2)$$

$$\mathbf{X}_T^{\ell+1} = \text{GlobalAttn} \ ^\ell \left(\mathbf{X}^\ell \right), \quad (3)$$

$$\mathbf{X}^{\ell+1} = \text{MLP}^\ell \left(\mathbf{X}_M^{\ell+1} + \mathbf{X}_T^{\ell+1} \right) \quad (4)$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix of a graph with N nodes and E edges; $\mathbf{X}^\ell \in \mathbb{R}^{N \times d_\ell}, \mathbb{E}^\ell \in \mathbb{R}^{E \times d_\ell}$ are the d_ℓ-dimensional node and edge features, respectively; MPNN $_e^\ell$ and GlobalAttn $^\ell$ are instances of an MPNN with edge features and of a global attention mechanism at the ℓ-th layer with their corresponding learnable parameters, respectively; MLP$^\ell$ is a 2-layer MLP block.

B. GPS++ Block

Rampášek, Galkin, Dwivedi, *et al.* [7] proposed the GPS++ block to be defined as follows for layers $\ell > 0$:

$$\mathbf{X}^{\ell+1}, \mathbf{E}^{\ell+1}, \mathbf{g}^{\ell+1} = \text{GPS++} \left(\mathbf{X}^\ell, \mathbf{E}^\ell, \mathbf{g}^\ell, \boldsymbol{B} \right) \quad (5)$$

$$\mathbf{Y}^\ell, \ \mathbf{E}^{\ell+1}, \ \mathbf{g}^{\ell+1} = \text{MPNN} \ \left(\mathbf{X}^\ell, \mathbf{E}^\ell, \mathbf{g}^\ell \right), \quad (6)$$

$$\mathbf{Z}^\ell = \text{BiasedAttn} \left(\mathbf{X}^\ell, \boldsymbol{B} \right), \quad (7)$$

$$\forall i : \quad \mathbf{x}_i^{\ell+1} = \text{FFN} \left(\mathbf{y}_i^\ell + \mathbf{z}_i^\ell \right) \quad (8)$$

In GPS++ [8] which is analogous to MPNN modules as shown by Gilmer, Schoenholz, Riley, *et al.* [17] and Battaglia, Hamrick, Bapst, *et al.* [18], and Bronstein, Bruna, Cohen, *et al.* [19]. They indicate that their MPNN variant maximizes the expressivity of the model and increases the generalizability of the model.

C. Our Architecture

The proposed block is defined as follows which is similar to the construction of GPS++ [8].

$$\mathbf{X}^{\ell+1}, \mathbf{E}^{\ell+1}, \mathbf{g}^{\ell+1} = \text{GPS++} \left(\mathbf{X}^\ell, \mathbf{E}^\ell, \mathbf{g}^\ell, \boldsymbol{B} \right) \quad (9)$$

$$\mathbf{Y}^\ell, \ \mathbf{E}^{\ell+1}, \ \mathbf{g}^{\ell+1} = \text{MPNN} \ \left(\mathbf{X}^\ell, \mathbf{E}^\ell, \mathbf{g}^\ell \right), \quad (10)$$

$$\mathbf{Z}^\ell = \text{Attn} \left(\mathbf{X}^\ell, \boldsymbol{B} \right), \quad (11)$$

$$\forall i : \quad \mathbf{x}_i^{\ell+1} = \text{FFN} \left(\mathbf{y}_i^\ell + \mathbf{z}_i^\ell \right) \quad (12)$$

where $\mathbf{X}^0$, $\mathbf{E}^0$, $\mathbf{g}^0$, and $\boldsymbol{B}$ defined in Equation 13.

We do not use the MPNN variant proposed by GPS++ [8] in our architecture which increases the overfitting considering the sizes of the datasets we work with. Our approach also incorporates Attn which is defined as the standard attention block from Vaswani, Shazeer, Parmar, *et al.* [20]. We also find that for smaller graph datasets standard Attention, modules work better than Attention variants like Biased self-attention [21] which was used by GPS++ [8]. The Feed Forward Network is defined simply as a stack of MLP layers and also uses Drop GNN [22]. These architectural choices are also summarized in Figure 1.

With these changes in the Architecture of GPS++, our approach outperforms GPS++ on smaller datasets and sets a new state-of-the-art on the BBBP dataset achieving 78.8 ROC-AUC, beating the current state-of-the-art [10] by 5.5.

V. EXPERIMENTAL SETUP

This section explains the implementation details of the experiments and our proposed model.

A. Sampling

There is a class imbalance in the dataset, meaning that not all classes have a similar number of images. For this reason, we follow a stratified sampling strategy during data loading to ensure each batch contains $50 \pm 5\%$ instances of each label class.

$$\mathbf{X}^0 = \texttt{Dense}\big([\mathbf{X}^{\text{atom}} \mid \mathbf{X}^{\text{LapVec}} \mid \mathbf{X}^{\text{LapVal}} \mid \mathbf{X}^{\text{RW}} \mid \mathbf{X}^{\text{Cent}} \mid \mathbf{X}^{\text{3D}}]\big) \qquad \in \mathbb{R}^{N \times d_{\text{node}}}$$
$$\mathbf{E}^0 = \texttt{Dense}\big([\mathbf{E}^{\text{bond}} \mid \mathbf{E}^{\text{3D}}]\big) \qquad \in \mathbb{R}^{M \times d_{\text{edge}}} \qquad (13)$$
$$\mathbf{g}^0 = \texttt{Embed}_{d_{\text{global}}}(0) \qquad \in \mathbb{R}^{d_{\text{global}}}$$
$$\boldsymbol{B} = \boldsymbol{B}^{\text{SPD}} + \boldsymbol{B}^{\text{3D}} \qquad \in \mathbb{R}^{N \times N}$$
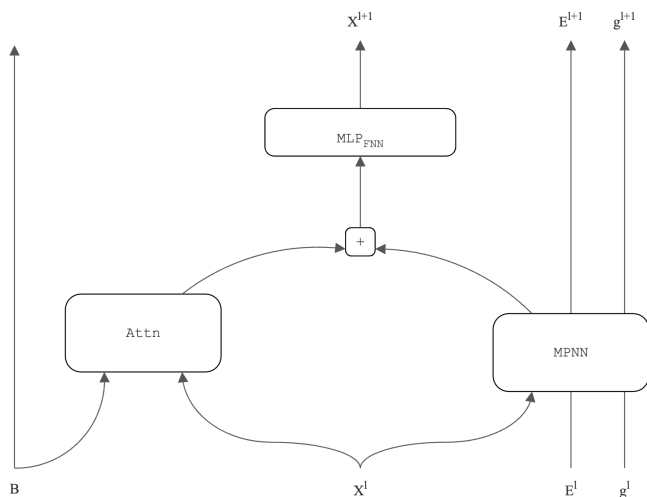

Fig. 1: Overview of the proposed model, notice the architectural changes made from GPS++.

B. Model backbone

Throughout this work GPS [7] and GPS++ [8] is used as the architecture backbone due to their success in using MPNNs and Transformers. Our model backbone is standard self-attention [20] which we demonstrate works well for the task of molecule property prediction. Since the BBBP dataset does not contain a large amount of data, other transformer-based models did not show great results for this task whereas standard self-attention model with regularization and augmentation techniques, was able to generalize well and achieved better results.

C. Baseline model

We established a naive baseline with random guesses. The baseline model we chose was training GPS++ [8] without any design modifications. This gets to a ROC-AUC of 72.7 on the BBBP dataset. To train this model, we employ standard augmentations using AugLiChem [23] and train the network for 300 epochs with a batch size of 256.

D. Loss function

The final model prediction is formed by global sum-pooling of all node representations and then passing it through a MLP. The regression loss is the mean absolute error (L1 loss) between a scalar prediction and the ground truth.

E. Code

Our code is in PyTorch 1.10 [24] and we also provide TensorFlow code. We use a number of open-source packages to develop our training workflows. Most of our experiments and models were trained with PyTorch Geometric [25]. Our hardware setup for the experiments included either four NVIDIA Tesla V100 GPUs or a TPUv3-8 cluster. We utilized mixed-precision training with PyTorch's native AMP (through `torch.cuda.amp`) for mixed-precision training and a distributed training setup (through `torch.distributed.launch`) which allowed us to obtain significant boosts in the overall model training time. Our code to reproduce the results along with Tensorboard logs of the training runs are available at [1].

VI. Results

We report all results and compare them against previous models and a random baseline (equivalent to making a guess) in Table I. The performance of models is calculated using the metrics that are typical for a molecular property prediction problem, ROC-AUC. Additionally, it is found that using any of the other variants of GPS++ is detrimental for this task and the model starts heavily overfitting even with our design changes, augmentation, and regularization. We also observe that our architectural changes over GPS++ [8] improves GPS++ by 6.1%.

TABLE I: Model Performance on the BBBP dataset for predicting blood-brain membrane permeability.

Method Description	ROC-AUC (↑)
GAL 120B [26]	66.1
PretrainGNN [27]	68.7
N-GramRF [28]	69.7
GROVER [29]	70.0
D-MPNN [30]	71.0
ChemRL-GEM [31]	72.4
ChemBERTa-2 [32]	72.8
Uni-Mol [26]	72.9
SPMM [10]	73.3
GPS	72.34
GPS++	72.73
Ours (GPS Transformer with Self Attention)	**78.80**

VII. Conclusion

This research has provided valuable insights into the challenges and opportunities associated with drug delivery across the blood-brain barrier. By understanding the intricate mechanisms of brain entry and the parameters influencing blood-brain barrier permeation, this paper have paved the way for the development of novel strategies in the field of therapeutic

[1]https://github.com/Rishit-dagli/GraphBRAIN

delivery. The introduction of GPS Transformer architecture, combined with Self Attention, has demonstrated significant advancements in predicting blood-brain barrier permeability using limited data. Notably, the approach has surpassed the current state-of-the-art performance, as evidenced by achieving an impressive ROC-AUC of 78.8. This breakthrough opens up new possibilities for efficient and effective drug discovery and development, ultimately aiming to improve treatments for central nervous system disorders. The findings contribute to the bigger vision of enhancing patient care by enabling targeted drug delivery to the brain, thus revolutionizing common practices and offering hope for transformative advancements in neurological medicine.

References

[1] R. Daneman and A. Prat, "The blood–brain barrier," *Cold Spring Harbor perspectives in biology*, vol. 7, no. 1, a020412, 2015.

[2] T. NQ, T. PE, and L. MS, "Fast estimation of the blood-brain barrier permeability by pulling a ligand through a lipid membrane," *Journal of chemical information and modeling*, 2020.

[3] S. Lee, T. T. Chen, C. L. Barber, *et al.*, "Autocrine vegf signaling is required for vascular homeostasis," *Cell*, vol. 130, no. 4, pp. 691–703, 2007, ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2007.06.054. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867407009051.

[4] M. H. Abraham, H. S. Chadha, and R. C. Mitchell, "Hydrogen bonding. 33. factors that influence the distribution of solutes between blood and brain," *Journal of Pharmaceutical Sciences*, vol. 83, no. 9, pp. 1257–1268, 1994, ISSN: 0022-3549. DOI: https://doi.org/10.1002/jps.2600830915. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022354915495929.

[5] P. Crivori, G. Cruciani, P.-A. Carrupt, and B. Testa, "Predicting blood- brain barrier permeation from three-dimensional molecular structure," *Journal of medicinal chemistry*, vol. 43, no. 11, pp. 2204–2216, 2000.

[6] X. Liu, M. Tu, R. S. Kelly, C. Chen, and B. J. Smith, "Development of a computational approach to predict blood-brain barrier permeability," *Drug metabolism and disposition*, vol. 32, no. 1, pp. 132–139, 2004.

[7] L. Rampášek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, and D. Beaini, "Recipe for a general, powerful, scalable graph transformer," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14501–14515, 2022.

[8] D. Masters, J. Dean, K. Klaser, *et al.*, *Gps++: An optimised hybrid mpnn/transformer for molecular property prediction*, 2022. arXiv: 2212.02229 [q-bio.QM].

[9] Z. Wu, B. Ramsundar, E. N. Feinberg, *et al.*, "Moleculenet: A benchmark for molecular machine learning," *Chemical science*, vol. 9, no. 2, pp. 513–530, 2018.

[10] J. Chang and J. C. Ye, *Bidirectional generation of structure and properties through a single molecular foundation model*, 2023. arXiv: 2211.10590 [cs.LG].

[11] R. Miao, L.-Y. Xia, H.-H. Chen, H.-H. Huang, and Y. Liang, "Improved classification of blood-brain-barrier drugs using deep learning," *Scientific Reports*, vol. 9, Jun. 2019. DOI: 10.1038/s41598-019-44773-4.

[12] T. Jain and P. K. P. Shanmuganathan, *In silico prediction of blood-brain barrier permeability of chemical compounds through molecular feature modeling*, 2022. arXiv: 2208.09484 [q-bio.QM].

[13] R. Saber, R. Mhanna, and S. Rihana, *A machine learning model for the prediction of drug permeability across the blood-brain barrier: A comparative approach*, May 2020. DOI: 10.21203/rs.3.rs-29117/v1.

[14] Y. Yuan, F. Zheng, and C.-G. Zhan, "Improved prediction of blood-brain barrier permeability through machine learning with combined use of molecular property-based descriptors and fingerprints," *The AAPS journal*, vol. 20, p. 54, Mar. 2018. DOI: 10.1208/s12248-018-0215-8.

[15] I. Martins, A. Teixeira, L. Pinheiro, and A. Falcão, "A bayesian approach to in silico blood-brain barrier penetration modeling," *Journal of chemical information and modeling*, vol. 52, pp. 1686–97, May 2012. DOI: 10.1021/ci300124c.

[16] S. Ioffe and C. Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, 2015. arXiv: 1502.03167 [cs.LG].

[17] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*, PMLR, 2017, pp. 1263–1272.

[18] P. W. Battaglia, J. B. Hamrick, V. Bapst, *et al.*, "Relational inductive biases, deep learning, and graph networks," *arXiv preprint arXiv:1806.01261*, 2018.

[19] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *arXiv preprint arXiv:2104.13478*, 2021.

[20] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2017. arXiv: 1706.03762 [cs.CL].

[21] C. Ying, T. Cai, S. Luo, *et al.*, "Do transformers really perform badly for graph representation?" In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 28877–28888. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/f1c1592588411002af340cbaedd6fc33-Paper.pdf.

[22] P. A. Papp, K. Martinkus, L. Faber, and R. Wattenhofer, *Dropgnn: Random dropouts increase the expressiveness of graph neural networks*, 2021. arXiv: 2111.06283 [cs.LG].

[23] R. Magar, Y. Wang, C. Lorsung, *et al.*, "Auglichem: Data augmentation library of chemical structures for machine learning," *Machine Learning: Science and Technology*, vol. 3, no. 4, p. 045015, 2022. DOI: 10.1088/2632-2153/ac9c84. [Online]. Available: https://dx.doi.org/10.1088/2632-2153/ac9c84.

[24] A. Paszke, S. Gross, F. Massa, *et al.*, *Pytorch: An imperative style, high-performance deep learning library*, 2019. arXiv: 1912.01703 `[cs.LG]`.

[25] M. Fey and J. E. Lenssen, *Fast Graph Representation Learning with PyTorch Geometric*, May 2019. [Online]. Available: https://github.com/pyg-team/pytorch_geometric.

[26] R. Taylor, M. Kardas, G. Cucurull, *et al.*, *Galactica: A large language model for science*, 2022. arXiv: 2211.09085 `[cs.CL]`.

[27] W. Hu, B. Liu, J. Gomes, *et al.*, *Strategies for pretraining graph neural networks*, 2020. arXiv: 1905.12265 `[cs.LG]`.

[28] S. Liu, M. F. Demirel, and Y. Liang, *N-gram graph: Simple unsupervised representation for graphs, with applications to molecules*, 2019. arXiv: 1806.09206 `[cs.LG]`.

[29] Y. Rong, Y. Bian, T. Xu, *et al.*, *Self-supervised graph transformer on large-scale molecular data*, 2020. arXiv: 2007.02835 `[q-bio.BM]`.

[30] K. Yang, K. Swanson, W. Jin, *et al.*, *Analyzing learned molecular representations for property prediction*, 2019. arXiv: 1904.01561 `[cs.LG]`.

[31] X. Fang, L. Liu, J. Lei, *et al.*, "Geometry-enhanced molecular representation learning for property prediction," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 127–134, 2022. DOI: 10.1038/s42256-021-00438-4. [Online]. Available: https://doi.org/10.1038%2Fs42256-021-00438-4.

[32] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, and B. Ramsundar, *Chemberta-2: Towards chemical foundation models*, 2022. arXiv: 2209.01712 `[cs.LG]`.