ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction

Seyone Chithrananda
University of Toronto
seyone.chithrananda@utoronto.ca

Gabriel Grand
Reverie Labs
gabe@reverielabs.com

Bharath Ramsundar
DeepChem
bharath.ramsundar@gmail.com

Abstract

GNNs and chemical fingerprints are the predominant approaches to representing molecules for property prediction. However, in NLP, transformers have become the de-facto standard for representation learning thanks to their strong downstream task transfer. In parallel, the software ecosystem around transformers is maturing rapidly, with libraries like HuggingFace and BertViz enabling streamlined training and introspection. In this work, we make one of the first attempts to systematically evaluate transformers on molecular property prediction tasks via our ChemBERTa model. While not at state-of-the-art, ChemBERTa scales well with pretraining dataset size, offering competitive downstream performance on MoleculeNet and useful attention-based visualization modalities. Our results suggest that transformers offer a promising avenue of future work for molecular representation learning and property prediction. To facilitate these efforts, we release a curated dataset of 77M SMILES from PubChem suitable for large-scale self-supervised pretraining.

1 Motivation

Molecular property prediction has seen a recent resurgence thanks to the success of graph neural networks (GNNs) on various benchmark tasks [1, 2, 3, 4, 5, 6]. However, data scarcity remains a fundamental challenge for supervised learning in a domain in which each new labelled data point requires costly and time-consuming laboratory testing. Determining effective methods to make use of large amounts of unlabeled structure data remains an important unsolved challenge.

Over the past two years, the transformer [7, 8] has emerged as a robust architecture for learning self-supervised representations of text. Transformer pretraining plus task-specific finetuning provides substantial gains over previous approaches to many tasks in natural language processing (NLP) [9, 10, 11]. Meanwhile, software infrastructure for transformers is maturing rapidly: HuggingFace [12] provides streamlined pretraining and finetuning pipelines, while packages like BertViz [13] offer sophisticated interfaces for attention visualization. Given the availability of millions of SMILES strings, transformers offer an interesting alternative to both expert-crafted and GNN-learned fingerprints. In particular, the masked language-modeling (MLM) pretraining task [8] commonly used for BERT-style architectures is analogous to atom masking tasks used in graph settings [14]. Moreover, since modern transformers are engineered to scale to massive NLP corpora, they offer practical advantages over GNNs in terms of efficiency and throughput.

Though simple in concept, the application of transformers to molecular data presents several questions that are severely underexplored. For instance: How does pretraining dataset size affect downstream task performance? What tokenization strategies work best for SMILES? Does replacing SMILES

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

with a more robust string representation like SELFIES [15] improve performance? We aim to address these questions via one of the first systematic evaluations of transformers on molecular property prediction tasks.

2 Related Work

In cheminformatics, there is a long tradition of training language models directly on SMILES to learn continuous latent representations [16, 17, 18]. Typically, these are RNN sequence-to-sequence models and their goal is to facilitate auxiliary lead optimization tasks; e.g., focused library generation [19]. Thus far, discussion of the transformer architecture in chemistry has been largely focused on a particular application to reaction prediction [20].

Some recent work has pretrained transformers for molecular property prediction and reported promising results [21, 22]. However, the datasets used for pretraining have been relatively small (861K compounds from ChEMBL and 2M compounds from ZINC, respectively). Other work has used larger pretraining datasets (18.7M compounds from ZINC) [23] but the effects of pretraining dataset size, tokenizer, and string representation were not explored. In still other work, transformers were used for supervised learning directly without pretraining [24].

Recently, a systematic study of self-supervised pretraining strategies for GNNs helped to clarify the landscape of those methods [14]. Our goal is to undertake a similar investigation for transformers to assess the viability of this architecture for property prediction.

3 Methods

ChemBERTa is based on the RoBERTa [25] transformer implementation in HuggingFace [12]. Our implementation of RoBERTa uses 12 attention heads and 6 layers, resulting in 72 distinct attention mechanisms. So far, we have released 15 pre-trained ChemBERTa models on the Huggingface's model hub; these models have collectively received over 30,000 Inference API calls to date.[1]

We used the popular Chemprop library for all baselines [6]. We trained the directed Message Passing Neural Network (D-MPNN) with default hyperparameters as well as the sklearn-based [26] Random Forest (RF) and Support Vector Machine (SVM) models from Chemprop, which use 2048-bit Morgan fingerprints from RDKit [27, 28].

3.1 PreTraining on PubChem 77M

We adopted our pretraining procedure from RoBERTa, which masks 15% of the tokens in each input string. We used a max. vocab size of 52K tokens and max. sequence length of 512 tokens. We trained for 10 epochs on all PubChem subsets except for the 10M subset, on which we trained for 3 epochs to avoid observed overfitting. Our hypothesis is that, in learning to recover masked tokens, the model forms a representational topology of chemical space that should generalize to property prediction tasks.

For pretraining, we curated a dataset of 77M unique SMILES from PubChem [29], the world's largest open-source collection of chemical structures. The SMILES were canonicalized and globally shuffled to facilitate large-scale pretraining. We divided this dataset into subsets of 100K, 250K, 1M, and 10M. Pretraining on the largest subset took approx. 48 hours on a single NVIDIA V100 GPU. We make this dataset publicly available as a part of MoleculeNet, and leave pretraining on the full 77M set to future work.

3.2 Finetuning on MoleculeNet

We evaluated our models on several classification tasks from MoleculeNet [30] selected to cover a range of dataset sizes (1.5K - 41.1K examples) and medicinal chemistry applications (brain penetrability, toxicity, and on-target inhibition). These included the BBBP, ClinTox, HIV, and Tox21 datasets. For datasets with multiple tasks, we selected a single representative task: the clinical toxicity

[1] The main model directory can be viewed here. Each model includes the specific tokenizer (BPE, SMILES-tokenized), representation (SMILES, SELFIES) and number of training steps ('150k') appended in its name.

	BBBP 2,039		**ClinTox (CT_TOX)** 1,478		**HIV** 41,127		**Tox21 (SR-p53)** 7,831	
	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC
ChemBERTa 10M	0.643	0.620	0.733	0.975	0.622	0.119	**0.728**	0.207
D-MPNN	**0.708**	0.697	**0.906**	**0.993**	0.752	0.152	0.688	**0.429**
RF	0.681	0.692	0.693	0.968	**0.780**	**0.383**	0.724	0.335
SVM	0.702	**0.724**	0.833	0.986	0.763	0.364	0.708	0.345

Table 1: Comparison of ChemBERTa pretrained on 10M PubChem compounds and Chemprop baselines on selected MoleculeNet tasks. We report both ROC-AUC and PRC-AUC to give a full picture of performance on class-imbalanced tasks.

(CT_TOX) task from ClinTox and the p53 stress-response pathway activation (SR-p53) task from Tox21. For each dataset, we generated an 80/10/10 train/valid/test split using the scaffold splitter from DeepChem [31]. During finetuning, we appended a linear classification layer and backpropagated through the base model. We finetuned models for up to 25 epochs with early stopping based on evaluation loss. We release a set of tutorials in DeepChem which allows users to go through loading two types of pre-trained ChemBERTa models (with differing tokenizers), running masked token inference, visualizing the attention of the model on several molecules, and fine-tuning the model on the ClinTox dataset.

4 Results

On the MoleculeNet tasks that we evaluated, ChemBERTa approaches, but does not beat, the strong baselines from Chemprop (Table 1).[2] Nevertheless, downstream performance of ChemBERTa scales well with more pretraining data (Fig. 1). On average, scaling from 100K to 10M resulted in ΔROC-AUC $= +0.110$ and ΔPRC-AUC $= +0.059$. (HIV was omitted from this analysis due to resource constraints.) These results suggest that ChemBERTa learns more robust representations with additional data and is able to leverage this information when learning downstream tasks.

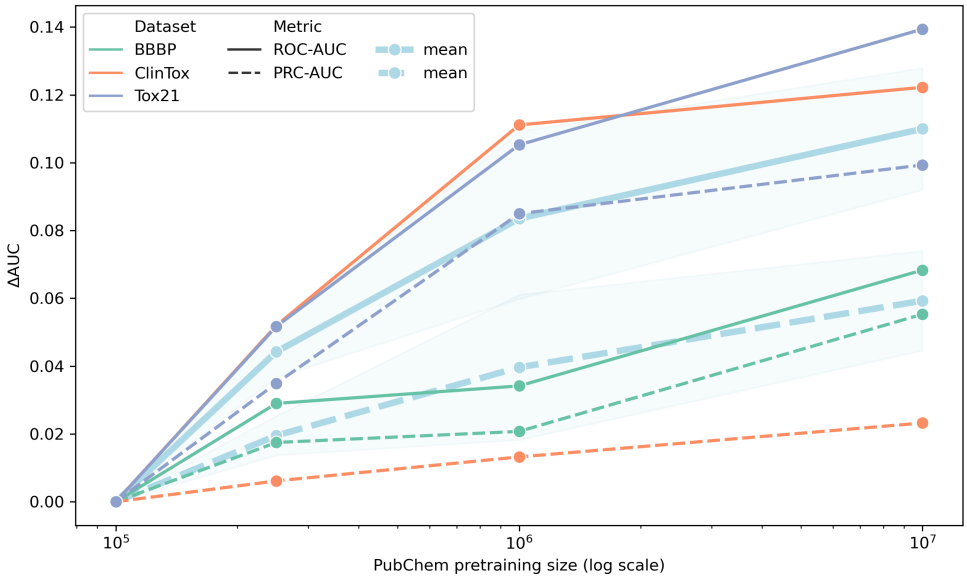


Figure 1: Scaling the pretraining size (100K, 250K, 1M, 10M) produces consistent improvements in downstream task performance on BBBP, ClinTox, and Tox21. Mean ΔAUC across all three tasks with a 68% confidence interval is shown in light blue.

[2]While Tox21 ROC-AUC is better than the baselines, PR-AUC is considerably lower.

4.1 Tokenizers

Our default tokenization strategy uses a Byte-Pair Encoder (BPE) from the HuggingFace tokenizers library [12]. BPE is a hybrid between character and word-level representations, which allows for the handling of large vocabularies in natural language corpora. Motivated by the intuition that rare and unknown words can often be decomposed into multiple known subwords, BPE finds the best word segmentation by iteratively and greedily merging frequent pairs of characters [32]. We compare this tokenization algorithm with a custom SmilesTokenizer based on a regex from work applying transformers to reaction prediction [20], which we have released as part of DeepChem [31].[3]

To compare tokenizers, we pretrained two identical models on the PubChem-1M set, one utilizing the SmilesTokenizer, and one utilizing the default BPE tokenization algorithm. The pretrained models were evaluated on the Tox21 SR-p53 task, the ClinTox dataset, and BBBP, utilizing 80-10-10 scaffold splits for the train, evaluation, and test sets. We found that SmilesTokenizer narrowly outperformed BPE by ΔPRC-AUC $= +0.015$, with no boost to ΔROC-AUC on the Tox21 set. Interestingly, SmilesTokenizer also observes strong boosts against BPE the BBBP dataset, achieving a boost of ΔPRC-AUC $= +0.021$, ΔROC-AUC $= +0.029$. However, we observe no boosts on the Clintox dataset. Though this result suggests that a more semantically-relevant tokenization may provide performance benefits, further benchmarking on the remainder of the MoleculeNet suite is needed to validate this finding. Additionally, ChemBERTa utilizes a default vocabulary size of 52,0000, as used in the original RoBERTa model with BPE tokenization [25]. However, the SMILES molecular string representation has a much more narrower vocabulary, as seen in the succinctness of the regex pattern used by SmilesTokenizer. This large vocabulary size used by BPE may increase the difficulty of learning effective contextual representations for each token, and so a similar pre-training procedure with a vocabulary size of 500-1000 may improve downstream benchmarking results.

4.2 SMILES vs. SELFIES

In addition to SMILES, we pretrained ChemBERTA using the SELFIES (SELF-referencing Embedded Strings) string representation [15]. SELFIES is an alternate molecular string representation designed for machine learning. Because every valid SELFIES corresponds to a valid molecule, we hypothesized that SELFIES would lead to a more robust model. After pre-training ChemBERTa on the PubChem 1M dataset which was converted to SELFIES, we conducted a similar set of benchmarks as done in the previous tokenizer study, using the Tox21 SR-p53, BBBP, and ClinTox tasks. Using the BPE tokenizer for comparison, we find no significant boosts to performance in the Tox21 and ClinTox prediction tasks, and a marginal improvement of ΔROC-AUC $= +0.010$ and ΔPRC-AUC $= +0.007$. We believe a custom SELFIES tokenizer in parallel to SmilesTokenizer may help demonstrate stronger performance, as BPE often splits sequences into tokens which are not semantically valuable in respect to the SELFIES/SMILES grammar. Further benchmarking on the entire MoleculeNet suite is needed to validate this idea.

4.3 Attention Visualization

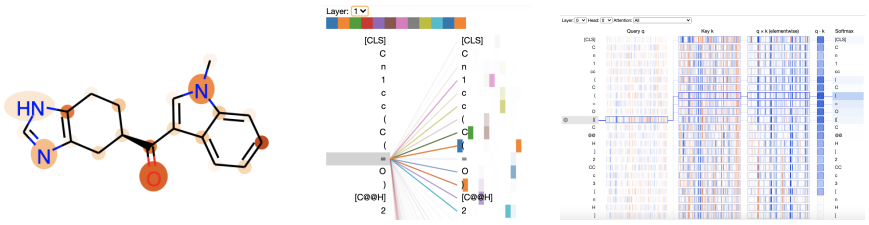
We used BertViz [13] to inspect the attention heads of ChemBERTa (SmilesTokenizer version) on Tox21, and contrast them to the molecular graph visualization of an attention-based GNN. We found certain neurons that were selective for chemically-relevant functional groups, and aromatic rings. We also observed other neurons that tracked bracket closures – a finding in keeping with results on attention-based RNNs showing the ability to track nested parentheses [33, 34].

5 Discussion

In this work, we introduce ChemBERTa, a transformer architecture for molecular property prediction. Initial results show that MLM pretraining provides a boost in predictive power for models on selected downstream tasks from MoleculeNet. However, with the possible exception of Tox21, ChemBERTa still performs below state-of-the-art on these tasks.

Our current analysis covers only a small portion of the hypothesis space we hope to explore. We plan to expand our evaluations to all of MoleculeNet, undertake more systematic hyperparameter

[3] https://deepchem.readthedocs.io/en/latest/tokenizers.html#smilestokenizer



(a) (b) (c)

Figure 2: (a) Attention in GNNs highlights a problematic ketone in a Tox21 compound. (b) Attention over SMILES tokens in ChemBERTa provides a close analogue to graph attention. (c) Neural stack trace enables fine-grained introspection of neuron behavior. (b - c) produced via BertViz [13].

tuning, experiment with larger masking rates, and explore multitask finetuning. In parallel, we aim to scale up pretraining, first to the full PubChem 77M dataset, then to even larger sets like ZINC-15 (with 270 million compounds). This work will require us to improve our engineering infrastructure considerably.

As we scale up, we are also actively investigating methods to improve sample efficiency. Alternative text-based pretraining methods like ELECTRA may be useful [10]. Separately, there is little question that graph representations provide useful inductive biases for learning molecular structures. Recent hybrid graph transformer models [22, 35] may provide better sample efficiency while retaining the scalability of attention-based architectures.

Acknowledgments and Disclosure of Funding

We would like to thank the Tyler Cowen and the Emergent Ventures fellowship for providing the research grant to S.C. for cloud computing and various research expenses, alongside the Thiel Foundation for funding the grant. Thanks to Mario Krenn, Alston Lo, Akshat Nigam, Professor Alan Aspuru-Guzik and the entire Aspuru-Guzik group for early discussions and mentorship regarding the potiential for applying large-scale transformers on molecular strings, as well as in motivating the utilization of SELFIES in this work.

We would also like to thank the entire DeepChem team for their support and early discussions on fostering the ChemBERTa concept, and helping with designing and hosting the Tokenizers API and ChemBERTa tutorial. Thanks to the Reverie team for authorizing our usage of the PubChem 77M dataset, which was processed, filtered and split by them.

References

[1] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.

[2] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8):595–608, 2016.

[3] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[4] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212*, 2017.

[5] Connor W Coley, Regina Barzilay, William H Green, Tommi S Jaakkola, and Klavs F Jensen. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of chemical information and modeling*, 57(8):1757–1772, 2017.

[6] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molec-

ular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[10] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910, 2019.

[13] Jesse Vig. A multiscale visualization of attention in the transformer model. *CoRR*, abs/1906.05714, 2019.

[14] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

[15] Mario Krenn, Florian Hase, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 2020.

[16] Zheng Xu, Sheng Wang, Feiyun Zhu, and Junzhou Huang. Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. In *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics*, pages 285–294, 2017.

[17] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. *arXiv preprint arXiv:1703.01925*, 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(2):268–276, 2018.

[19] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1):120–131, 2018.

[20] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

[21] Shion Honda, Shoi Shi, and Hiroki R Ueda. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*, 2019.

[22] Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. Molecule attention transformer. *arXiv preprint arXiv:2002.08264*, 2020.

[23] 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*, pages 429–436, 2019.

[24] Benson Chen, Regina Barzilay, and Tommi Jaakkola. Path-augmented graph transformer network. *arXiv preprint arXiv:1905.12712*, 2019.

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[26] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[27] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[28] Greg Landrum et al. Rdkit: Open-source cheminformatics. 2006.

[29] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109, 2019.

[30] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[31] B Ramsundar, P Eastman, E Feinberg, J Gomes, K Leswing, A Pappu, M Wu, and V Pande. Deepchem: Democratizing deep-learning for drug discovery, quantum chemistry, materials science and biology, 2016.

[32] Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. Byte pair encoding: A text compression scheme that accelerates pattern matching. Technical report, Technical Report DOI-TR-161, Department of Informatics, Kyushu University, 1999.

[33] Mirac Suzgun, Sebastian Gehrmann, Yonatan Belinkov, and Stuart M Shieber. Lstm networks can perform dynamic counting. *arXiv preprint arXiv:1906.03648*, 2019.

[34] Xiang Yu, Ngoc Thang Vu, and Jonas Kuhn. Learning the dyck language with attention-based seq2seq models. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 138–146, 2019.

[35] Anonymous. Modelling drug-target binding affinity using a bert based graph neural network. *Submitted to International Conference on Learning Representations*, 2021.

[36] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Broader Impact

A core goal of AI for drug discovery is to accelerate the development of new and potentially life-saving medicines. Research to improve the accuracy and generalizability of molecular property prediction methods contributes directly to these aims. Nevertheless, machine learning—and particularly large-scale pretraining of the form we undertake here—is a resource-intensive process that has a growing carbon footprint [36]. According to the Machine Learning Emissions Calculator (`https://mlco2.github.io/impact`), we estimate that our pretraining generated roughly 17.1 kg CO_2eq (carbon-dioxide equivalent) of emissions. Fortunately, Google Cloud Platform, which we used for this work, is certified carbon-neutral and offsets 100% of emissions (`https://cloud.google.com/sustainability`). Even as we advocate for further exploration of large-scale pretraining for property prediction, we also encourage other researchers to be mindful of the environmental impact of these efforts and opt for sustainable cloud compute solutions where possible.