GENERATIVE PRE-TRAINING FROM MOLECULES

A PREPRINT

Sanjar Adilov
Department of Biomedical Informatics
Romanovsky Institute of Mathematics
Tashkent, Uzbekistan
s.adilov@mathinst.uz

September 16, 2021

ABSTRACT

SMILES is a line notation for entering and representing molecules. Being inherently a language construct, it allows estimating molecular data in a self-supervised fashion by employing machine learning methods for natural language processing (NLP). The recent success of attention-based neural networks in NLP has made large-corpora transformer pretraining a de facto standard for learning representations and transferring knowledge to downstream tasks. In this work, we attempt to adapt transformer capabilities to a large SMILES corpus by constructing a GPT-2-like language model. We experimentally show that a pretrained causal transformer captures general knowledge that can be successfully transferred to such downstream tasks as focused molecule generation and single-/multi-output molecular-property prediction. For each task, we freeze model parameters and attach trainable lightweight networks between attention blocks—adapters—as alternative to fine-tuning. With a relatively modest setup, our transformer outperforms the recently proposed ChemBERTa transformer and approaches state-of-the-art MoleculeNet and Chemprop results. Overall, transformers pretrained on SMILES corpora are promising alternatives that do not require handcrafted feature engineering, make few assumptions about structure of data, and scale well with the pretraining data size.

Keywords transformers · transfer learning · adapters · de novo molecule generation · molecular property prediction

1 Introduction

Learned self-supervised representations of a general-domain dataset provide reliable initialization for downstream tasks. Transferring domain knowledge is especially useful in cases of low-data tasks, when models trained from scratch cannot adequately recognize patterns in limited data. The recent success of *transformer* models [Vaswani et al., 2017] established transfer learning as a de-facto standard for downstream-task learning in natural language processing [see Radford et al., 2018, Devlin et al., 2019]. Such software packages as HuggingFace Transformers [Wolf et al., 2020], BertViz [Vig, 2019], and AdapterHub [Pfeiffer et al., 2020] became widely popular among NLP practitioners.

As for drug discovery, state-of-the-art property-prediction models are primarily graph- or fingerprint-related [see Wu et al., 2018]; de novo molecular design enjoys promising prospects of graph and recurrent neural networks [see Brown et al., 2019]. Basically, molecules are encoded with *simplified molecular input line entry system* (SMILES) [Weininger, 1988], which is a language construct with its own vocabulary of molecule constituents and grammar rules. To adapt NLP techniques to molecular design, the aforementioned recurrent neural networks were evaluated with autoregressive language modeling objective [see Segler et al., 2018, Brown et al., 2019]. While proven useful, RNNs can be inferior to transformers in capturing sequential dependencies and scaling to larger corpora. In this paper, we make another attempt to train a multi-task transformer for drug discovery. Motivated by challenges and prospects of the recent ChemBERTa masked language model [Chithrananda et al., 2020] and Transformer-CNN [Karpov et al., 2020] method, we present our *causal* (autoregressive) GPT-2-like [Radford et al., 2019] model pretrained on the first 5M compounds from a curated PubChem-10M dataset released by ChemBERTa authors. We hypothesize that a large-scale, pretrained causal transformer is a multi-task learner capable of transferring learned representations with no significant architecture

modifications to such downstream tasks as focused molecule generation and single-/multi-output molecular-property prediction. For every task, we freeze transformer parameters and employ trainable *adapter modules* [Houlsby et al., 2019]—lightweight networks between transformer blocks,—which are alternative to *fine-tuning* requiring separate serialization of the retrained model for every task and tending to "forget" the domain knowledge during transfer.

2 Related work

Gupta et al. [2017] and Segler et al. [2018] made one of the first attempts to pretrain autoregressive language models on SMILES corpora to facilitate focused molecular-library generation. They experimentally proved the ability of character-level RNNs to capture SMILES language patterns. Chithrananda et al. [2020] introduced ChemBERTa, a masked language model based on RoBERTa transformer [Liu et al., 2019]. It was pretrained on 10M compounds from PubChem [Kim et al., 2019] and evaluated on single-output classification tasks from MoleculeNet [Wu et al., 2018].

As mentioned before, RNNs are autoregressive models—i.e., SMILES sequences are processed from left to right, so the current hiden representations are updated using only the previous ones. ChemBERTa relies on the encoder part of the original transformer [Vaswani et al., 2017] and during pretraining, a given percentage of tokens is masked and the model learns to recover the original inputs. Our goal is to investigate the viability of causal transformers, which rely on the decoder part of the original transformer, for molecular-property prediction and low-data focused molecule generation. Inspired by OpenAI GPT models [Radford et al., 2018, 2019], we aim to learn a comprehensive language model capable of adapting to downstream tasks with no significant data and architecture preprocessing. Another goal we seek to accomplish is to assess the ability of adapters [Houlsby et al., 2019, Pfeiffer et al., 2020] to solve downstream molecular tasks. Adapters are increasingly becoming popular among NLP practitioners as they are better alternatives in terms of the source-target knowledge tradeoff and model serialization.

3 Methods

Our transformer decoder replicates GPT-2 [Radford et al., 2019] (Figure 1) except that during tokenization, we used the character-level *byte-pair encoding* [Sennrich et al., 2016] instead of byte-level. We reserved 72 characters from the SMILES alphabet as an initial vocabulary and supplemented the vocabulary with up to 1000 most frequent merges. The model uses parameterized token and position embeddings, 8 attention heads, and 4 attention blocks. With the embedding/hidden dimension of 512, it has 13.4M parameters. See our github repository for implementation details:

`http://github.com/sanjaradylov/smiles-gpt`

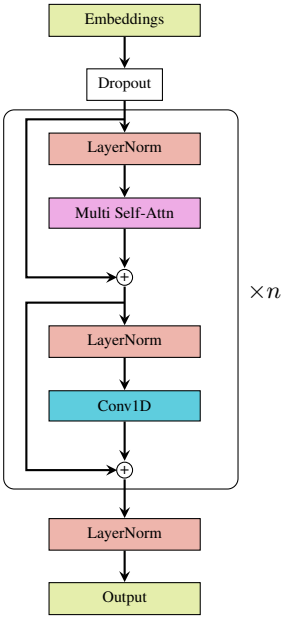


Figure 1: Causal transformer for generative pretraining from SMILES. Token-embedding parameters are shared with linear output; position embeddings are learned. Dropouts are also applied to attention weights and projections.

3.1 Pretraining on PubChem

We trained our model for 6 epochs on the first 5M SMILES strings of ChemBERTa's PubChem-10M. We used *AdamW* [Loshchilov and Hutter, 2017] for optimization and *cosine annealing* [Loshchilov and Hutter, 2016] for learning-rate scheduling. The initial and final learning rates were set to $5e-4$ and $5e-8$, respectively. We kept the default Adam hyperparameters and optimized the batch size (128) and maximum sequence length (512).

After training, we generated a set of 10,000 SMILES sequences using top-96% sampling [see Holtzman et al., 2019] and evaluated the ability of the model to produce diverse and valid molecules with the following distribution-learning metrics, each of which ranging from 0 to 1 (higher is better):

1. *Validity* – the rate of valid molecules.
2. *Uniqueness* – the rate of unique sequences.
3. *Novelty* – the rate of sequences not presented in the baseline (training) dataset.
4. *Internal Diversity* [Benhenda, 2017] – the average pairwise dissimilarity measure based on Tanimoto distance between Morgan fingerprints (ECFP) [Rogers and Hahn, 2010] of two molecules in a set $\boldsymbol{M}$:

$$\mathrm{IntDiv}(\boldsymbol{M}) = 1 - \frac{1}{|\boldsymbol{M}|^2} \sum_{\boldsymbol{m},\boldsymbol{m}' \in \boldsymbol{M}} T(\boldsymbol{m}, \boldsymbol{m}'),$$

$$T(\boldsymbol{m}, \boldsymbol{m}') = \frac{|\boldsymbol{m} \cap \boldsymbol{m}'|}{|\boldsymbol{m} \cup \boldsymbol{m}'|}.$$

5. *PhysChem KL Divergence* [Brown et al., 2019] – the average descriptor similarity between $\boldsymbol{M}$ and $\boldsymbol{R}$ based on the Kullback-Leibler divergence:

$$D_{KL}(\boldsymbol{p} \parallel \boldsymbol{q}) = \sum_i p_i \log \frac{p_i}{q_i},$$

$$\mathrm{KL}(\boldsymbol{M}, \boldsymbol{R}) = \frac{1}{|\mathfrak{D}|} \sum_{\mathfrak{d} \in \mathfrak{D}} \exp\Big(- D_{KL}\big(\mathfrak{d}(\boldsymbol{M}) \parallel \mathfrak{d}(\boldsymbol{R})\big)\Big),$$

where $\mathfrak{D}$ is the set of target descriptors (e.g., *physicochemical* [see Brown et al., 2019]) and $\mathfrak{d}$ is mapping from a molecule set into a descriptor distribution.

3.2 Adapter training on downstream tasks

For every task, we employed a lightweight adapter network (Figure 2). It learns task-specific patterns by projecting the hidden representations of a transformer down to a significantly lower dimension and back to the original dimension. The pretrained model parameters were fixed. Thus, with the reduction factor of 16 and one- or two-layer feed-forward output, we learned and serialized only task-specific and output parameters—1-5% of the number of the pretrained model parameters. This way we alleviated the cumbersome process of (layer-wise) learning-rate optimization, which could potentially lead to re-fitting of the domain model to the given task.

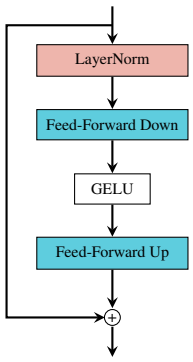


Figure 2: Adapter network architecture for downstream tasks. Applied after every attention block—i.e., after the second residual connection in Figure 1. Can be stacked multiple times.

For focused molecule generation, we used the TRPM8 dataset [Gupta et al., 2017], from which the authors clustered 5 most active molecules for training and left the rest 472 for comparison. After training, we generated 200 distinct

Table 1: Evaluation of GPT-2 pretrained on 5M PubChem compounds on distribution-learning metrics. ECFPs were computed with radius=2 and bits=1024; distributions of discrete and continuous physicochemical descriptors were computed via histograms and gaussian KDE, respectively.

Validity	Uniqueness	Novelty	Internal Diversity	PhysChem KL
0.9884	0.9999	0.9802	0.8832	0.9895

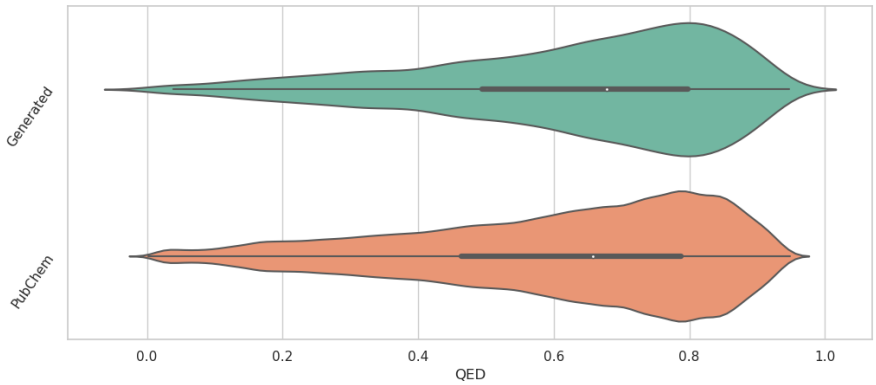


Figure 3: Kernel density estimations of QEDs of generated set and 1M subset of PubChem.

and valid molecules and measured the distributions of both datasets by calculating 9 physicochemcial descriptors [see Brown et al., 2019] and running *t-SNE* [van der Maaten and Hinton, 2008] on the obtained descriptor space for visual comparison.

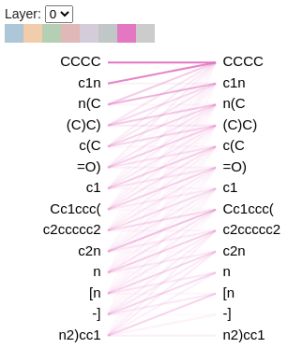
For molecular property prediction, we selected 4 classification datasets from MoleculeNet [Wu et al., 2018]: BBBP (2K compounds, 1 output), HIV (42K compounds, 1 output), Tox21 (8K compounds, 12 outputs, contains missing labels), and Clintox (1.5K compounds, 2 outputs). For multi-output datasets, we replaced the linear output layer with *bypass multi-task network* [Ramsundar et al., 2017]. It consists of n_{task} independent blocks mapping inputs to the corresponding tasks, and a shared block, whose outputs are summed with the task-specific representations to obtain the final n_{task} predictions. The single-output blocks consist of two feed-forward layers and an intermediate dropout. The reduction factor was chosen between 12 and 16. Parameters were optimized with AdamW and initial learning rates were in $[25e-5, 1e-3]$. In general, models were trained for 10-25 epochs with early stopping on ROC-AUC.

4 Results

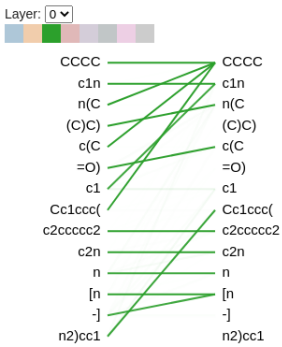
Table 1 lists the values for the distribution-learning evaluation metrics. Figure 3 shows kernel density estimations of *quantitative estimation of drug-likeness* (QED) [Bickerton et al., 2012] of training and generated sets. Training with different parameter initialization and/or generating another set of molecules produce similar results. Overall, the model is capable of producing valid, diverse, and novel sets of molecules close to the baseline set in various molecular properties.

One of the key distinctions of attention-based models is interpretability (see Figure 4). Attention heads tend to exhibit expository patterns in linguistic data, including positional, syntactic, and token-specific patterns. Although causal transformers' heads evidence rather inexplicable patterns compared to encoding transformers', there are still visible, distinct heads capturing specific relationships between tokens in the first attention block, and the long-term-dependency heads in the remaining blocks.

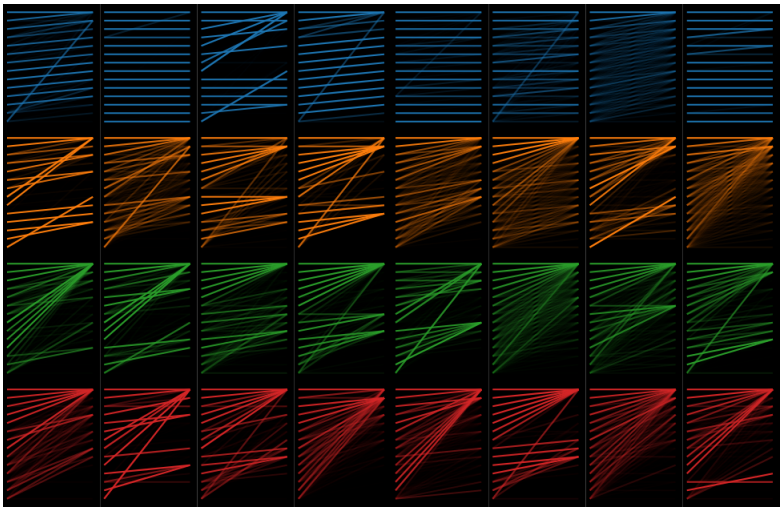
Figure 5 shows t-SNE of physicochemical descriptors of TRPM8-active molecules. The descriptors were standardized beforehand. Adjusting perplexity (from 2 to 20), learning rate, and early stopping parameters produce similar topology. We also calculated ECFPs (radius=2, length=1024) of molecules to find closest neighbors to 5 active TRMP8 inhibitors (Figure 6).



(a) "Bag-of-token" head attending 2-4 previous tokens.



(b) Syntactic head that closes brackets.



(c) Heads across all of the layers. The first layer typically captures short-term relationships, higher layers attend farther tokens or concentrate on some specific tokens.

Figure 4: Attention head and model views produced via BertViz.

Table 2 shows average evaluation results on MoleculeNet data. Datasets were divided into 80/10/10 train/valid/test sets using scaffold split. Experiments were run 3 times with different adapter and output parameter initializations. We approach Chemprop [Yang et al., 2019] and MoleculeNet [see Chithrananda et al., 2020] baselines on BBBP and HIV tasks, and with much lighter model, outperform ChemBERTa on every task except Clintox on AUC-PRC. Note that we report only excerpts from ChemBERTa paper since we follow the same evaluation criteria—scaffold split and both AUC-ROC and AUC-PRC calculation; MoleculeNet and Chemprop choose criteria specific for a particular dataset.

Table 2: Evaluation of GPT-2 pretrained on 5M PubChem compounds on selected MoleculeNet tasks.

	BBBP		**HIV**		**Clintox (CT_TOX)**		**Tox21 (SR-p53)**		**Tox21 (All)**	
	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC
GPT-2	0.730	0.745	0.770	0.302	0.988	0.947	0.737	0.325	0.734	0.361
ChemBERTa	0.643	0.620	0.622	0.119	0.733	0.975	0.728	0.207	N/A	N/A

5 Conclusion and further work

In this work, we present a causal transformer for drug discovery. Pretrained on a large SMILES corpus, it learns molecular representations that can be successfully transferred to a variety of downstream tasks. For transfer learning, one can employ adapter modules, which accept hidden representations from the transformer block and process them to learn task-specific patterns.

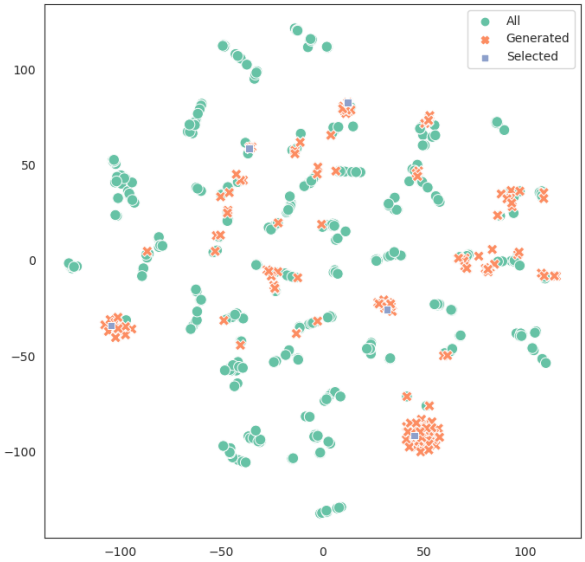


Figure 5: t-SNE of physicochemical descriptors of TRPM8 data and generated adaptations. Every selected inhibitor is surrounded by a cluster of close generated neighbors, albeit of different size. PhysChemKL between the baseline TRPM8 and generated sets is 0.811.

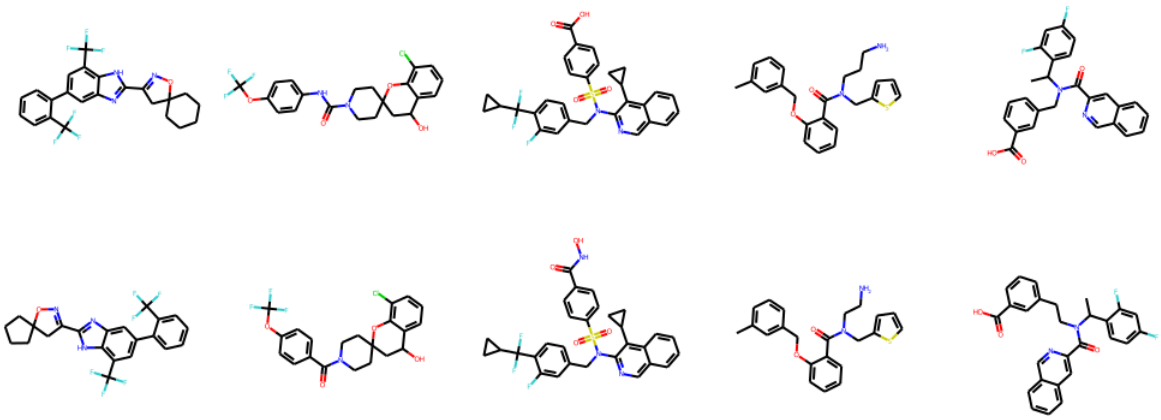


Figure 6: Nearest ECFP-neighbors (2nd row) of selected 5 TRMP8 inhibitors (1st row).

Performance on MoleculeNet property-prediction tasks is challenging but not superior. First, we should note that because of resource limitations, we restricted ourselves to lighter modules and experimenting only on a 5M subset, while 10M was initially made available and 77M is on the way. Running experiments on smaller subsets, we conclude that transformers scale well with the pretraining data size, as was previously demonstrated by Chithrananda et al. [2020] on ChemBERTa. Another consideration is tokenization strategy: we use a rather straightforward approach, whereas sophisticated, linguistically-consistent tokenization might lead to learning clearer representations of molecular constituents. Finally, there is a variety of other transformer architectures performing better on many NLP tasks [see Wolf et al., 2020]. While generally being inferior to GNNs in terms of resources, they might boost overall performance on many molecular downstream tasks with even more data.

References

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. 2018.

J. Devlin, M. Chang, K. Lee, and K. 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)*, page 4171–4186. Association for Computational Linguistics, 2019.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A.M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, oct 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

J. Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-3007. URL `https://www.aclweb.org/anthology/P19-3007`.

J. Pfeiffer, A. Rücklé, C. Poth, A. Kamath, I. Vulić, S. Ruder, K. Cho, and I. Gurevych. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, 2020.

Z. Wu, B. Ramsundar, E.N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, K. Leswing, and V. Pande. Moleculenet: a benchmark for molecular machine learning. In *Chemical science, 9(2)*, pages 513—530, 2018.

N. Brown, M. Fiscato, M.H.S. Segler, and A.C. Vaucher. Guacamol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108, 2019. doi:10.1021/acs.jcim.8b00839. URL `https://doi.org/10.1021/acs.jcim.8b00839`.

D. Weininger. Smiles, a chemical language and information system. introduction to methodology and encoding rules. In *J.Chem. Inf. Comput. Sci*, pages 28–31, 1988.

M.H.S. Segler, T. Kogej, C. Tyrchan, and M.P. Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, 4(1):120–131, 2018. doi:10.1021/acscentsci.7b00512. URL `https://doi.org/10.1021/acscentsci.7b00512`. PMID: 29392184.

S. Chithrananda, G. Grand, and B. Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

P. Karpov, G. Godin, and I.V. Tetko. Transformer-cnn: Swiss knife for qsar modeling and interpretation. In *Journal of Cheminformatics 12*, 2020. doi:10.1186/s13321-020-00423-w.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp. *arXiv preprint arXiv:1902.00751*, 2019.

A. Gupta, A. Müller, B. Huisman, J. Fuchs, P. Schneider, and G. Schneider. Generative recurrent networks for de novo drug design. *Molecular Informatics*, 37, 11 2017. doi:10.1002/minf.201700111.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and Stoyanov V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692v1*, 2019.

S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E.E. Bolton. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47 (D1):D1102–D1109, 2019. doi:10.1093/nar/gky1033.

R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi:10.18653/v1/P16-1162. URL `https://aclanthology.org/P16-1162`.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text *de*generation. *arXiv preprint arXiv:1904.09751*, 2019.

B. Benhenda. Chemgan challenge for drug discovery: can ai reproduce natural chemical diversity? *arXiv preprint arXiv:1708.08227*, 2017.

D. Rogers and M. Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5): 742–754, 2010. doi:10.1021/ci100050t.

L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

B. Ramsundar, B. Liu, Z. Wu, A. Verras, M. Tudor, R.P. Sheridan, and V. Pande. Is multitask deep learning practical for pharma? *Journal of Chemical Information and Modeling*, 57(8):2068–2076, 2017. doi:10.1021/acs.jcim.7b00146. URL `https://doi.org/10.1021/acs.jcim.7b00146`. PMID: 28692267.

G.R. Bickerton, G.V. Paolini, J. Besnard, S. Muresan, and A.L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4:90–98, 2012. doi:10.1038/nchem.1243. URL `https://doi.org/10.1038/nchem.1243`.

K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, and R. Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Machine Learning Research*, 59(8):3370–3388, 2019.