Few-shot classification of drug target activity augmented with pretrained protein embeddings

David Kuo
Stanford University
dekuo

Tom Pritsky
Stanford University
tom5

Davey Huang
Stanford University
huangdh

Abstract

The process of drug discovery is extremely long and expensive. A key component of this process is hit identification and lead optimization, which involves identifying compounds that bind to a biological target of interest and optimizing target affinity. Due to the paucity and heterogeneity of high-quality experimental data, we propose a novel few-shot meta-learning methodology that efficiently brings together chemical structure information and biological protein information, leveraging recent advances in large language models for chemical and biological sequence data.

We created a few-shot protein-ligand binding classification dataset of 1047 target protein tasks from 17818 protein-ligand affinity measurements in the `BindingDB` dataset and leveraged the pre-trained embeddings from ChemBERTa and ESM-2 transformer models to represent each small molecule ligand and target protein sequence, respectively. We additionally trained a variational autoencoder (VAE) to reduce the dimensionality of the generated ligand and protein embeddings for improved computational efficiency.

We implemented and evaluated 3 different meta-learning models: a LSTM-based black-box meta-learner, a BERT-based black-box meta-learner, and a non-parametric prototypical network. For each of our model classes, we experimented with using only SMILES embeddings as inputs, concatenating SMILES and protein embeddings as inputs, and concatenating SMILES and protein embeddings with reduced dimensionality ($d = 50$) generated by a variational autoencoder (VAE) as inputs. For the black-box meta-learners, we additionally explored introducing protein embeddings at an intermediate step rather than immediately prior to input to the model, in order to assess the optimal combination point.

We trained and evaluated our models on 5-shot 2-way classification of binding affinity. Our best performing model was a LSTM with protein embeddings concatenated at the beginning of the model, which attained a test set query accuracy of 70.7%. For all our models, we found that introducing target protein information led to improved few-shot performance. We also found that the optimal point of protein information introduction was at the beginning of the model. We then showed that significantly reducing the dimensionality of the protein embeddings with a VAE maintained reasonable accuracy while significantly improving runtime. These results suggest promising future directions for few-shot protein-ligand binding prediction.

1 Introduction

1.1 Background

The process of drug discovery is often long and complex, with a very high failure rate. A key step in this process is hit identification and lead optimization, which involves finding a small molecule that binds to a certain protein target. This is a challenging problem due to the paucity and heterogeneity of experimental data measuring the activity of small molecules against protein targets, but one that fits well within the few-shot learning framework, and where meta-learning methods may lead to significant improvements.

1.2 Related Work

There is currently limited literature addressing few-shot learning for protein-ligand binding prediction (also called drug-target interaction). The closest analog to our current work, and the dataset and benchmark that we originally planned to build from, is the Few-shot Learning Dataset of Molecules (`FS-MOL`) [1]. `FS-MOL` provides a few-shot learning dataset of 5120 protein tasks with a median of 46 ligand compounds per task, with options for representing each ligand as a SMILES string, extended connectivity fingerprint, or molecular graph. `FS-MOL` furthermore implemented several baseline models including:

1. A random forest and graph neural network trained from scratch to represent single task methods
2. A GNN-based multitask model to represent multi-task pretraining
3. A Molecular Attention Transformer to represent self-supervised pretraining
4. A GNN meta-trained with MAML and a GNN meta-trained with PN to represent standard meta-learning approaches

Among these models, GNN-PN achieved the best overall results followed by GNN-MAML; however, performance was highly task-dependent as demonstrated by the wide range in performance between tasks. Notably, no target protein information was provided for each task, and all of the baseline methods used either extended connectivity fingerprints or molecular graph representations of the small molecules.

Nguyen et al. and Pappu et al. explored gradient-based meta-learning vs. multi-task approaches with extended connectivity molecular fingerprints and molecular graph representations of small molecule ligands respectively, and found that MAML and its variants FO-MAML and ANIL generally outperformed multi-task pre-training baselines (especially on out-of-distribution tasks), and consistently produced the best-performing models across fine-tuning sets across all support set sizes [2][3].

Finally, Lee et al. developed an Attentive Neural Process with extended connectivity molecular fingerprint representations of small molecule ligands and raw protein amino acid sequences as inputs which demonstrated superior few-shot classification performance compared to conventional supervised single-task approaches [4].

While the current literature in few-shot protein-ligand binding prediction focuses heavily on connectivity-based representations of small molecule ligands, previous work has shown that pre-training a RoBERTa-based model on SMILES codes (string-based representations of chemical structures) generates information-rich molecular embeddings that can outperform graph-based methods [5][6]. Furthermore, recent development of large protein language models such as Evolutionary Scale Modeling (`ESM`) have enabled the generation of information-rich protein embeddings able to predict 3D protein structures from amino acid sequence with accuracy comparable to that of AlphaFold2 without requiring multiple sequence alignment [7]. These advances suggest that leveraging large language models with SMILES and protein sequence data may be a powerful and currently under-explored approach for few-shot protein-ligand binding prediction.

1.3 Objective

The objective of our project is to explore the use of large language models pretrained on large corpora of protein and small molecule sequence information for few-shot protein-ligand binding classification. The development of increasingly large transformer-based language models trained in a self-supervised fashion has resulted in substantial gains in performance on natural language processing tasks such as text summarization, question answering, and translation, as well as in unexpected emergent behaviors such as the ability to perform in-context learning. In the biomedical sciences, transformer-based models have demonstrated the ability to learn salient representations of biology from sequence data and effectively predict properties such as solubility, toxicity, and even the 3D structure of proteins [5] [7]. Our intuition is that the rich representations of biological information captured by these large language models should result in more efficient few-shot learning.

2 Methods

2.1 Dataset

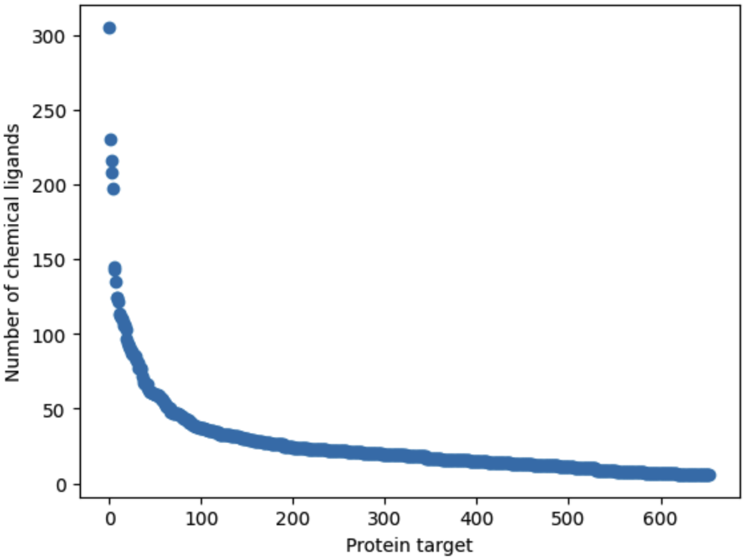

Figure 1: Distribution of the number of chemical ligands per protein target in our dataset shows a long-tailed distribution. Proteins with less than $K = 5$ chemical ligands were excluded since they were not sampled.

Our initial plan was to build upon the `FS-MOL` dataset and benchmark; however, due to challenges setting up dependencies for `FS-MOL`, we created our own few-shot protein-ligand binding classification meta-learning dataset modeled after `FS-MOL`. To do this, we extracted the target protein sequence, ligand SMILES string, and binding affinity (pK_d) for each protein-ligand binding affinity in the `TorchProtein BindingDB` dataset and grouped them by target protein. In this way we treated binding each target protein as a separate task and framed our meta-learning task as learning to learn to predict protein-ligand binding. Emulating `FS-MOL`, we also converted binding affinities (pK_d) to binary binding classification by determining the median binding affinity for each target protein and classifying all ligands with $pK_d \geq$ this median binding affinity as binding and all other ligands as non-binding. In total, we created 1047 target protein tasks with a median of 18 ligands per target protein (interquartile range 11-28) from 17818 protein-ligand affinities in the `TorchProtein BindingDB` dataset.

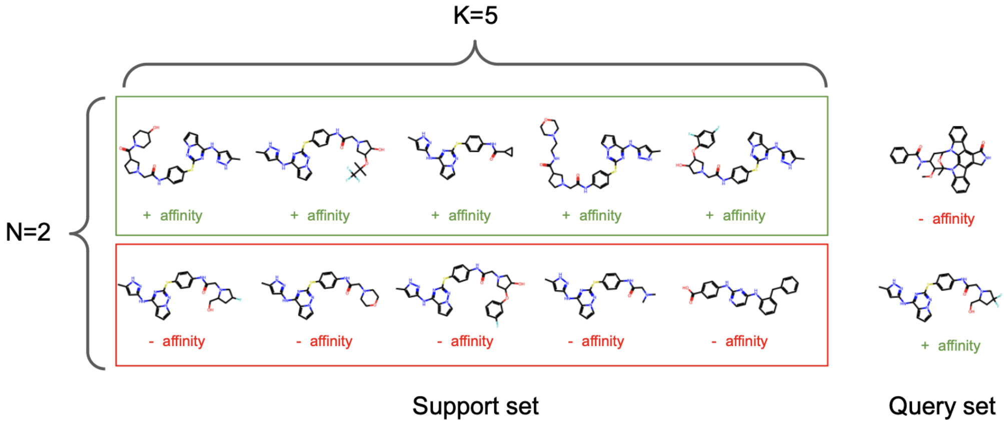

Figure 2: Representative few-shot classification task with K=5 support examples and 1 query example for each of the N=2 classes (binding or not binding).

2.2 Task

Our learning task is 5-shot, 2-way classification of ligand binding activity against a given target protein. Each task is then defined by a randomly selected target protein, five positive support set examples of ligands classified as having positive binding affinity (i.e. having a $pK_d \geq$ the median binding affinity of all ligands for the target protein), five negative support set examples of ligands classified as having negative binding affinity (i.e. having a $pK_d \leq$ the median binding affinity of all ligands for the target protein), one positive query set example, and one negative query set example. During meta-train time, the model is trained on batches of tasks from a random sample of 80% of the protein targets and validated on batches of tasks from a random sample of 10% of the protein targets set aside for this purpose. During meta-test time, the model is evaluated on a completely unseen held-out set of 10% of the protein targets from our dataset. In other words, there is no overlap between the proteins seen during meta-train and meta-test time.

2.3 SMILES Embeddings

In order to represent our small molecule ligands, we utilized prior work from ChemBERTa [5], in which the authors pre-train a RoBERTa-like model on 77 million unique SMILES strings using masked language modeling, in which spans of the chemical sequence are masked out and predicted by the model. We use their pretrained model weights in order to generate 767-dimensional embeddings from the SMILES strings for our ligands. Due to the computational demand of running ChemBERTa, we decided to pre-compute and cache the embeddings of molecules in our dataset, which is equivalent to freezing the weights of ChemBERTa. We use these pretrained SMILES embeddings as a starting point for the development of our models.

2.4 Protein Embeddings

Similarly, we used pretrained protein embeddings from the 150 million parameter ESM-2 model [7] trained via masked language modeling on roughly 138 million protein sequences from the UniRef protein sequence database. ESM and other large transformer models have been demonstrated to be highly expressive and able to predict 3D protein folding structure at atomic scale at the level of AlphaFold2 while also not requiring multiple sequence alignment as an initial preprocessing step and being orders of magnitude faster, which was particularly important given our compute constraints. Similar to our SMILES embeddings, we cached the 640-dimensional protein embeddings for our target proteins in order to significantly reduce training time, which is equivalent to freezing the layers in ESM-2.

2.5 Variational Autoencoder

Due to the high dimensionality of the protein (and SMILES) embeddings generated, we explored the use of dimensionality reduction techniques to improve computational efficiency and reduce runtime. We chose to use a variational auto-encoder for dimensionality reduction due to it's simplicity and a training paradigm which enables validation of compression performance during training. The compression performance can be tracked during training by recording the decoder loss on the validation set.

Our VAE is comprised of a two-layer fully-connected encoder (and associated decoder) trained with an 80/20 train/validation split. Various training parameters were explored, including the number of model layers, the size of the bottleneck embedding vector, the weight of the Kullback-Leibler divergence loss term, and the number of training epochs. After training, the encoder was used to reduce the dimensionality of our input protein embeddings.

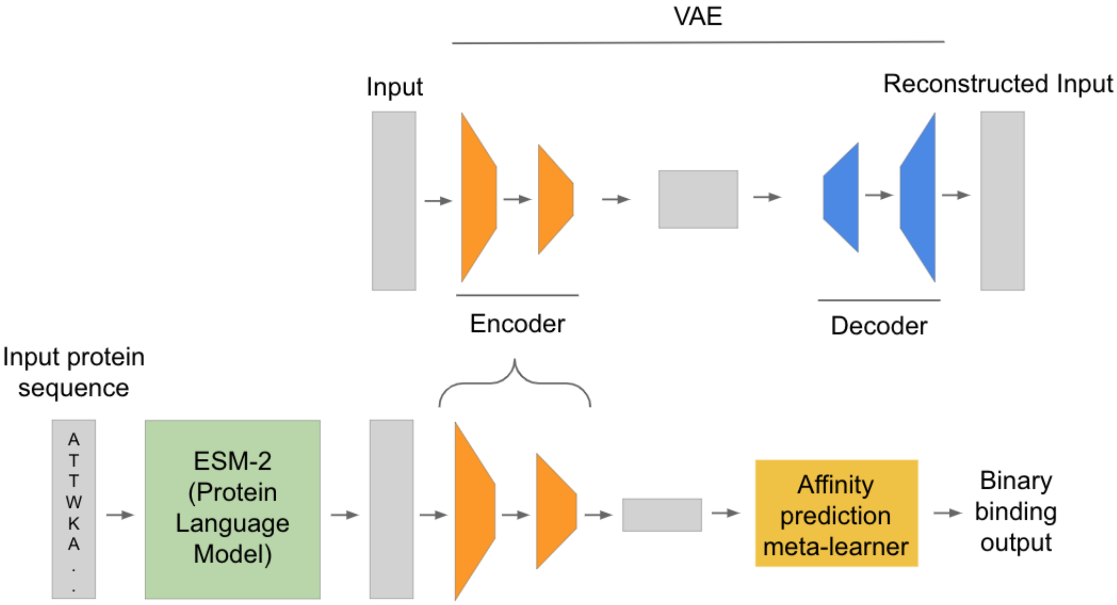

Figure 3: Diagram of variational autoencoder architecture.

2.6 Black Box LSTM

The first black-box meta-learner we experimented with was a 2-layer LSTM model. We used a hidden dimensionality of 128, a meta batch size of 128, learning rate of 1e-5, dropout probability of 0.35, and trained for 150,000 steps. We experimented with concatenating the ESM-2 protein embedding for the respective target protein to the input ChemBERTA SMILES sequence embedding. We also experimented with inserting the ESM-2 protein embedding at an intermediate step of black-box meta-learning, namely after the first LSTM module but before the second LSTM module. We replicated both experiments again using reduced dimensionality ($d = 50$) protein embeddings generated by a variational autoencoder. For all experiments, we saved the model with the highest validation query accuracy, which we then evaluated on the held-out meta-test set.

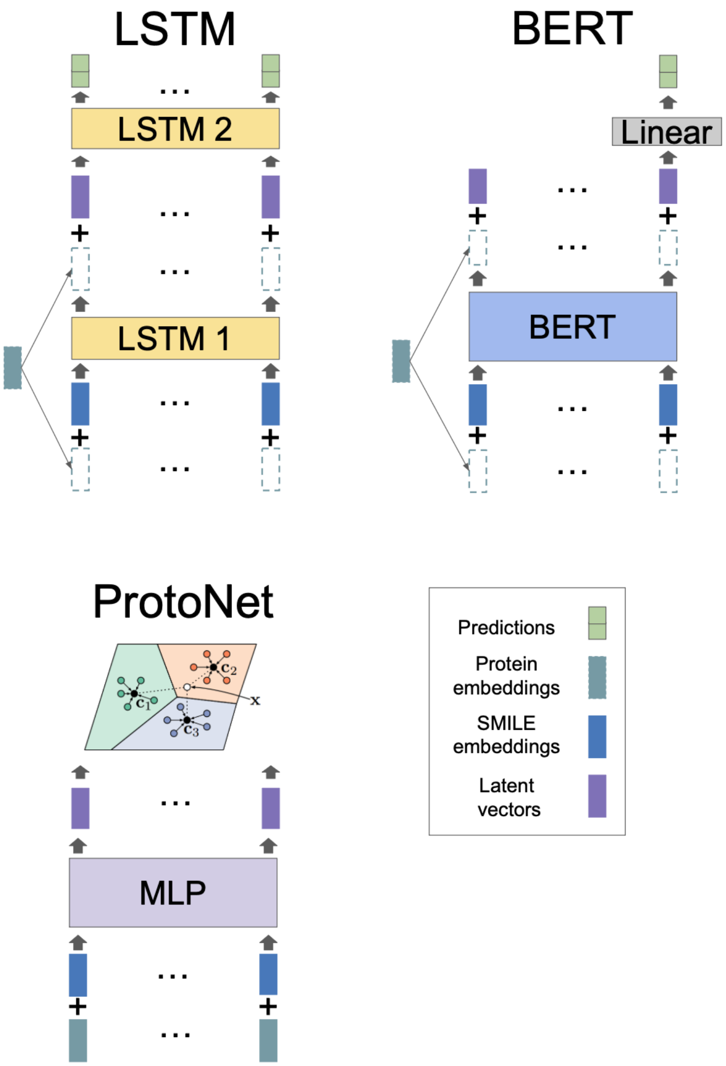

Figure 4: Diagrams of LSTM black-box, BERT black-box, and ProtoNet non-parametric meta-learning models. Experiments also included different methods of feeding protein embeddings into the models.

2.7 Black Box BERT

Given the significant improvements across natural language processing tasks spurred by the adoption of transformer models, we also developed a BERT-based memory-augmented neural network as a more expressive black-box meta-learner. Specifically, we constructed a BERT encoder with 4 self-attention blocks, 1 attention head, an intermediate dimensionality of 64, and a fixed input size of $N(K + 1)$, where N is the number of classes and K is the number of shots. Therefore, our BERT-based model had a fixed input sequence length of 12. We used a meta batch size of 128, learning rate of 5e-5, dropout probability of 0.35, and trained for 25k steps. Similar to the LSTM experiments, we also concatenated the ESM-2 protein embeddings for the respective target protein to either the beginning of the model or after the BERT encoder and before the final linear layer. We also ran the same experiments using the protein embeddings with reduced dimensionality ($d = 50$) generated by a variational autoencoder, which we will further discuss in the VAE section below.

2.8 Prototypical Network

We additionally experimented with a prototypical network since our few-shot classification problem lends itself well to a non-parametric approach, and furthermore, because prototypical networks achieved the best overall performance in the `FS-MOL` dataset and benchmark. To limit computational cost and mitigate the high dimensionality of our ligand SMILES and protein embeddings with respect to our input space, we opted to use a simple 2-layer MLP with a hidden layer of size 128 as our encoder and restricted the dimensionality of our latent embedding space to 64. We used a meta batch size of 100, learning rate of 1e-5, and trained for 500 outer-loop iterations. As with black-box meta-learning, we also experimented with using only using SMILES embeddings as inputs, concatenating SMILES and protein embeddings as inputs, and concatenating SMILES and protein embeddings with reduced dimensionality ($d = 50$) generated by a variational autoencoder as inputs to the prototypical network.

3 Results

	LSTM	BERT	ProtoNet
Without protein embeddings	0.647	0.653	0.662
With protein embeddings			
Concatenated before	**0.707**	**0.701**	**0.693**
Concatenated after	0.667	0.663	N/A
With VAE reduced protein embeddings			
Concatenated before	0.690	0.651	0.688
Concatenated after	0.669	0.632	N/A

Table 1: 5-shot 2-way accuracy on a held-out meta-test set.

	Hidden dimensionality		
	d=50	d=100	d=600
SMILES decoder	97.2	121.3	198.5
Protein decoder	22.2	19.8	19.0

Table 2: VAE mean squared error (MSE) loss at various hidden dimensionalities.

In table 1, we show 5-shot, 2-way accuracy results for the meta-test query set. Our baseline performance (without protein embeddings) remained consistent at approximately 65-66% for our LSTM, BERT, and ProtoNet models.

By incorporating the full 640-dimensional ESM-2 embeddings, we achieved a notable performance improvement, with accuracies of approximately 69-71% for each of our models. Performance improvements were less significant for our LSTM and BERT models when concatenating the protein embeddings between LSTM layers or after the BERT encoder block.

Performance was also roughly maintained by the LSTM and ProtoNet models when we substituted the full 640-dimensional protein embeddings for the reduced 50-dimensional embeddings.

As seen in table 2, the VAE performed significantly better on protein compared to SMILES embeddings. The encoder dimension size also impacted decoder performance, though the trend varied between SMILES and protein embeddings.

4 Analysis/Discussion

In nearly all scenarios, we found that using a combination of information about both the chemical drug molecules as well as the biological protein target led to improved performance for few-shot prediction of target binding affinity. We showed here that the placement of the pre-trained ESM-2 protein embedding within the model was important to performance. Concatenating the protein embedding at the beginning of the model led to substantially better performance than concatenating the embeddings between LSTM layers or after the BERT decoder. This evidence suggests that our models are learning complex relationships between protein embeddings and the few-shot affinity prediction task. For instance, different proteins may often share similar structural motifs. These structural motifs can range from simple alpha-helices to complex hinge-like mechanisms that could suggest new binding pockets for the target drug.

Since ESM-2 embeddings are able to accurately predict 3D structure of the protein, we can assume that the embeddings themselves are suggestive of the entire protein structure. Using structural information in order to predict binding affinity has been a common approach with molecular dynamics and molecular docking simulations. However, the use of structural information in deep learning methods for this particular task has not been well explored in prior literature. Therefore, we are among the first to show that protein structure could notably augment the predictive capabilities of meta-learning methodologies for few-shot drug target affinity prediction.

We also found that we were able to significantly reduce the compute time by reducing the original dimensionality of the pre-trained ESM-2 protein embeddings, with a minimal loss in task performance. This could be particularly important for being able to run these models in a reasonable amount of time. In our study, we used very low hidden dimensionalities (e.g. 64, 128) and a small value of $K = 5$, and training was still highly compute intensive. In production scenarios, we may want to significantly scale up the number of support examples to match the amount of data we have available for particular targets, which could lead to significantly longer training times. Combined with the iterative process of active learning approaches, in which compound activities are measured in vitro in order to guide the exploration of the chemical search space, it is advantageous to be able to run the model efficiently. Our results suggest that a variational autoencoder can help maintain computational efficency while preserving accuracy in the cases of our LSTM and ProtoNet models.

Our best performing model was a black-box LSTM meta-learner, which achieved a 5-shot 2-way accuracy of 70.7% on a held out meta-test set. Our BERT-based black-box meta-learner achieved similar performance. However, BERT was trained for fewer iterations than LSTM due to time and compute constraints, which might indicate promising future experiments using a BERT-based model trained for a longer duration. In general, future work could explore significantly increasing the dimensionality and the size of the support set in order to achieve better performance. Additionally, our performance results here suggest promising future directions for the use of multiple input data types for the chemical affinity prediction task. For example, future work could also incorporate additional information about the protein target from either the Gene Ontology database or from experimental data. Furthermore, many previous works have used graph neural networks (GNNs) in order to build molecular representations. Therefore, further experiments may confirm whether our findings also translate to the molecular representations learned by GNNs.

5 Contributions

All authors contributed equally to the development and written portions of the project.

David Kuo: Helped extract data from `BindingDB` and build the few-shot classification dataset. Helped develop and debug LSTM and BERT-based black-box meta-learner. Implemented ProtoNet meta-learner. Contributed to model training and hyperparameter tuning. Helped make and record poster, and write the report.

Tom Pritsky: Generated the ESM-2 embeddings and trained a variational autoencoder on SMILES and protein embedding information. Helped develop the BERT black box meta-learner. Helped make and record poster, and contributed to writing the report.

Davey Huang: Helped developed the dataloader, LSTM black-box meta-learner, and BERT black-box meta-learner. Contributed to model training, hyperparmeter tuning, and analysis. Helped make and record poster, write the report, and generate the figures.

References

[1] Megan Stanley, John F Bronskill, Krzysztof Maziarz, Hubert Misztela, Jessica Lanini, Marwin Segler, Nadine Schneider, and Marc Brockschmidt. FS-mol: A few-shot learning dataset of molecules. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[2] Cuong Q. Nguyen, Constantine Kreatsoulas, and Kim M. Branson. Meta-learning gnn initializations for low-resource molecular property prediction, 2020.

[3] Aneesh Pappu and Brooks Paige. Making graph neural networks worth it for low-data molecular machine learning, 2020.

[4] Eunjoo Lee, Jiho Yoo, Huisun Lee, and Seunghoon Hong. MetaDTA: Meta-learning-based drug-target binding affinity prediction. In *ICLR2022 Machine Learning for Drug Discovery*, 2022.

[5] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020.

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

[7] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.