SMILES natural language encodings can accurately predict JAK bioactivity

Michael Shteyn

Abstract

The relationship between chemical structure and bioactivity is complex. Chemical structures of similar atomic compositions, but consisting of distinct bond formations in distinct sequences, can have differing impacts upon exposure to living tissue. Importantly, the influence on these differences on biological activity can be difficult to predict. Here, I fine-tuned BERT, a transformer-based neural network model, to predict bioactivity among four Janus kinases (JAKs) based on a string-level encoding of molecular structure alone. The fine-tuned transformer was successful at predicting the activity of each of four JAKs with a hidden-test accuracy of >74%.

Background

Janus kinases (JAKs) are a family of enzymes that play an important role in intracellular signaling. All mammals have four JAKs, which are activated by receptors in extracellular space and participate in phosphorylation pathways that can lead to activating DNA transcription with a cell's the nucleus. Disruption in the JAK family can lead to a variety of diseases including disordered immunity and cancers (Aronson, 2002).

Drug interactions with biological tissue are complex at the molecular level. Being able to predict the effects, and side-effects, of exposing living cells to chemical compounds remains a challenge in drug development and design. It is therefore essential to better understand the effect of molecular structure and composition on bioactivity within cells to develop effective therapies with predictable outcomes on intracellular processes. Predicting the effect of chemical compounds on JAK activity is especially important, as disorders in JAK signaling are known to have serious medical consequences.

Advances in natural language processing (NLP) techniques have opened new opportunities to better understand the effect of molecular structure on bioactivity. Importantly, the simplified molecular-input line-entry system (SMILES), which catalogues the structure of hundreds of thousands of chemical species in ASCII format, has enabled researchers to explore hidden relationships between structure and function. Here, I fine-tune a state-of-the-art natural language processing model to generate a learned representation of chemical species based on their SMILES identities. The learned representation is then utilized to predict the effect molecular species may have on the bioactivity of JAKs.

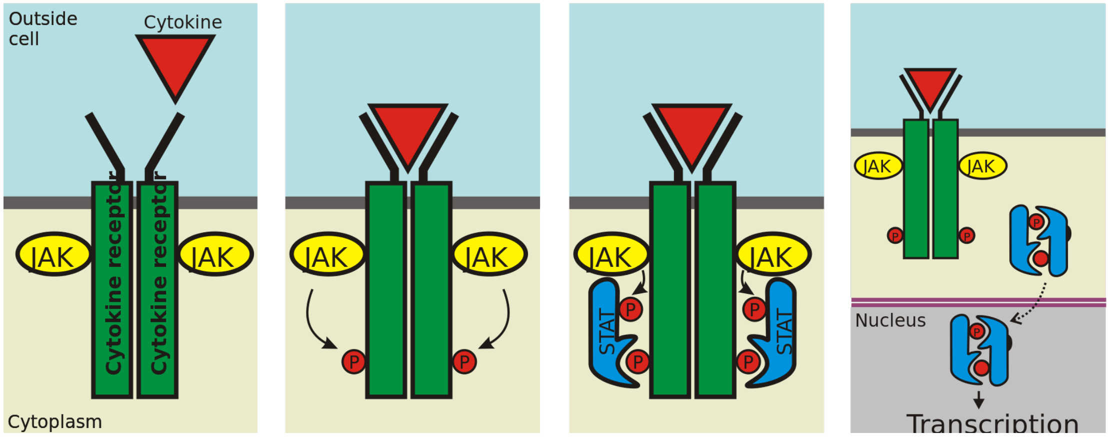

Figure 1. JAKs respond to extracellular activity to activate intracellular signaling cascades (from Peter Znamenkiy, published on wikipedia)

Methods

The training database consisted of 92921 SMILES entries representing distinct chemical structures, along with labels for whether each of four mammalian JAKs were active as a result of exposure to them in a laboratory setting. To successfully predict the bioactivity of JAKs based on SMILES encoding alone, an appropriate neural network architecture was required. The Bidirectional Encoder Representations from Transformers (BERT) is a family of language models introduced by Google (Devlin et al., 2018). The key innovation of BERT is applying the bidirectional training of a popular attention model, the Transformer, to the language domain. This architecture is in contrast to models that train only unidirectionally, or sequentially in both directions from the cursor. A masked language model (MLM) is used for bidirectional training. Using MLM, BERT is able to learn the relationship between natural language tokens and their context. BERT contains 768 hidden layers and a total of 110 million parameters.

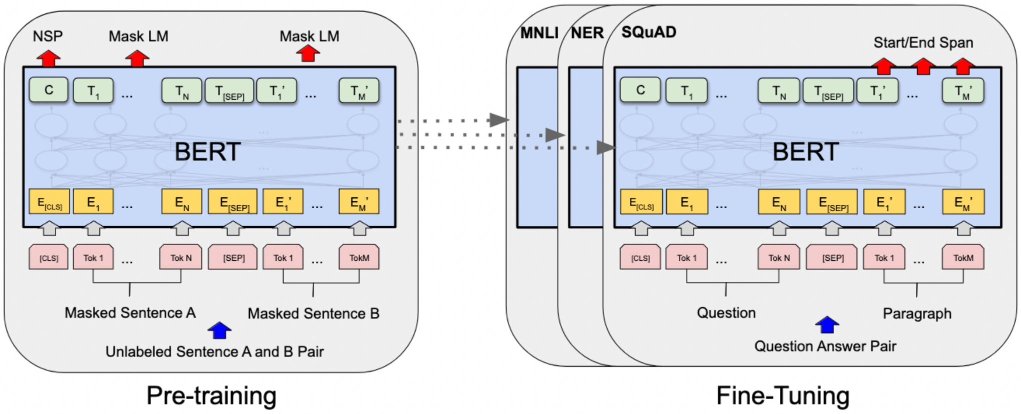

Figure 2. A visual representation of BERT architecture, from Devlin et al.

I applied the standard version of BERT, in the form it is published on *huggingface*, to the SMILES classification task. BERT applies self-attention to learn contextual information about textual inputs by developing a representation of which characters are likely to appear to the left and right of a given string (Fig. 2). To make BERT useful for classifying SMILES, finetuning was required. To this end, SMILES strings were tokenized. I tested two separate tokenizers – the standard BERT tokenizer, as well as a SMILES-specific character-level tokenizer developed by Chitarandra et al. available here. Interestingly, I found that the basic BERT tokenizer enabled the BERT model to converge toward a lower loss faster than the SMILES-specific tokenizer (Fig. 3).

In addition to the tokenizer, I tested how several hyperparameters may affect the ability of the model to train on the SMILES dataset. Originally, I began training the model with a large learning rate (lr = 0.01), under the assumption that SMILES encodings may not follow the typical character structure of natural language that BERT was pre-trained on. This assumption suggested that a significant degree of transfer learning may be required. However, I found a small learning rate (lr = 0.0001) was much more effective for achieving a lower loss most quickly. I also tested a larger batch size (n=64). This batch size exceeded the memory for a single GPU during fine-tuning, so I reduced the batch size to 48 for the final fine-tuning procedure. The loss was computed as the sum of the Cross—Entropy across the binary bioactivities (0 = inactive; 1 = active) predicated over four classes of JAKs.

The final hyperparameters involved a medium batch size (n=48), small learning rate (lr = 0.0001), and relied on the BERT tokenizer. BERT converged quickly on a summed Cross Entropy Loss ~ 25 across four classes. This is an expectedly larger loss than may be anticipated for single-task classification problems. However, after training on just 30,000 samples (less than one full epoch), the network was sufficient to achieve over 70% accuracy across classifying the bioactivity of the four JAKs in a hold-out test set. The network's highest accuracy on the test set was 74.8% correct.

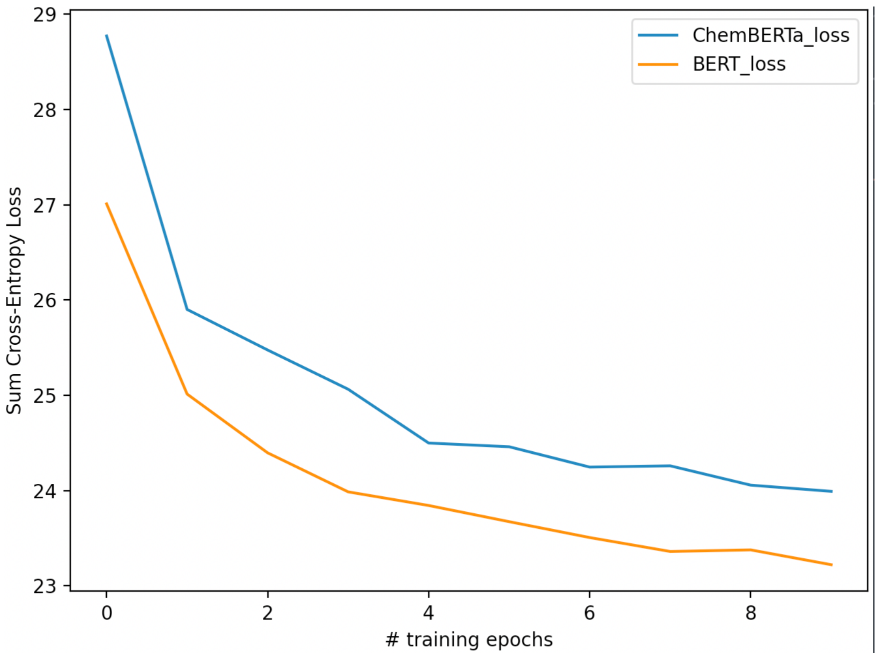

Figure 3. A comparison of loss between the BERT and ChemBERTA tokenizers over ten training epochs

Conclusion

I found that BERT was effective at learning chemical representations through transforming natural language inputs with attention-based processes, despite not having been pre-trained on or being known to have sensitivity to domain-specific encodings in SMILES format. The representation the BERT model developed of molecular structure was sufficient to achieve a robust ability in predicting the impact of different molecular formations on the bioactivity of four mammalian JAKs, with a modest degree of fine-tuning.

Citations

Aaronson DS, Horvath CM (2002). "A road map for those who don't know JAK-STAT". Science. **296**

Devlin, J, Chang, M-W, Lee, K, Toutanova, K (2018). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” https://arxiv.org/abs/1810.04805

ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction Seyone Chithrananda, Gabriel Grand, Bharath Ramsundar