Review ci-2022-00779z

Few-Shot Learning for Low-Data Drug Discovery

The authors propose a new architecture, prototypical networks (PN), for few-shot learning in drug discovery. The new architecture is compared to literature networks that have been reimplemented for consistency. PN is shown to be the best performing on the Tox21 dataset. For MUV standard ML models perform better, whilst for DUD-E data biases impact the results.

The work is of relevance to the readership and would be suitable for publication after the following points have been addressed.

1) Presentation.
    a. The article is generally well written and clear. However, there is significant duplication, particularly in the method descriptions which are repeated in the Related Work and Methods sections. So much so that some sentences/paragraphs are identical. Perhaps much of the detail could be moved to methods to make the Related Work section more precise?
    b. As another example, Figure 11 is redundant given Figure 10.
2) Train/test/prediction
    a. The language is sometimes difficult to follow. E.g. p. 25 line 44/45. This talks about training on a query prediction. P. 23 line 17 talks about training on samples from test with remaining data used for testing. It would be better to used prediction in the latter case.
    b. P. 21 talks about “learning” the embeddings. What is being learned?
    c. Are all “test” molecules included in the initial GCN embedding or just the support set for “training” the few-shot model? i.e. is this step imparting “test” information into the model?
3) Results
    a. On Tox21, the size of the support set has very little impact on the results. The ROC_AUC varies in the third decimal place and PR-AUC by about 0.01. This doesn’t seem logical. Even if one accepts the premise that a network can learn from a single example in each class, increasing the data ten-fold must have a more significant impact. Do the authors have an explanation for this?
    b. What is the correlation between the training and test endpoints in Tox21, particularly the SR assays used in Test and Train?
        i. Compound overlap/similarity
        ii. Activity overlap
    c. The authors explain away the MUV results by the fact that the compounds are deliberately chosen to be dissimilar. But isn’t this the general use case where Hit ID has discovered new lead molecules that are distinct from anything else. What are the limits of applicability of these methods in a real-world setting?
    d. Would an alternative explanation be that the methods do not work and that there is some inherent bias in the Tox21 data as for DUD-E?