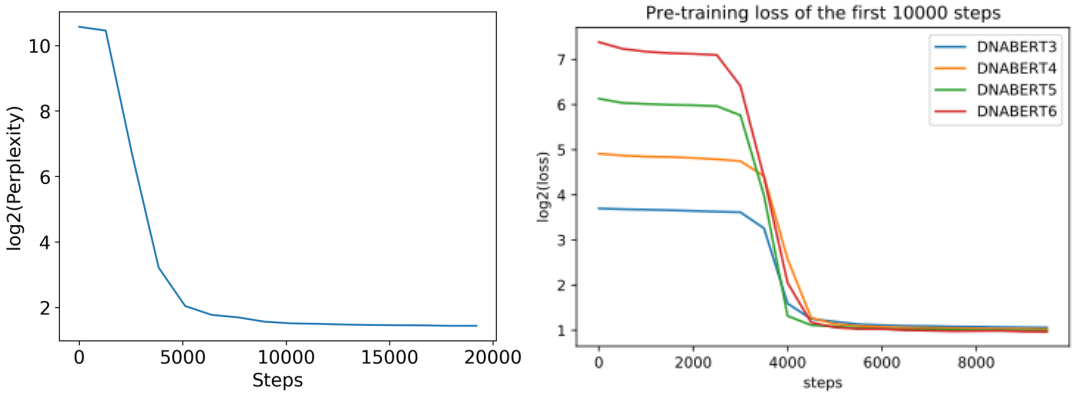
**Questions by reviewer aUkC with additional figure/table:**

- (1) How well is the pre-training conducted? Can you provide the training curve and evaluation (e.g. perplexity) of the pre-trained model? How much does it differ from the DNABERT model?

  Reply:

  Starting from the last question - DNABERT was pre trained on the entire human genome. Since the vast majority of the human genome does not code for genes, such pre-training is not ideal for our purposes as most of its effort would be spent on irrelevant genomic regions which are quite different in their composition. In addition, when we initially tried to use DNABERT we found the model to be extremely finicky to hyper parameter settings. These two observations led us to pre-train a separate, lighter, model, only on splice sites region sequences. In terms of the training curve, our pretraining had much higher perplexity at the beginning and took longer to converge to a similar loss. As shown in Figure1(see uploaded pdf), on the right hand side is the plot from the DNABERT paper while on the left is the loss during pre-training our model. We will include this plot now as a supplementary figure. We note though that the DNABERT "steps" and our iterations are not exactly comparable. The DNABERT batch size is 2K while ours is only 80. We therefore scaled the plot on the left accordingly. Still, the difference in batch sizes may still play a role in the observed results.

  **Figure 1: Pretraining curves. They show the pretraining curves for TrASPr(left) and for different k-mers of DNABERT(right)**




- (2) How do you include conservation values for each k-mer for the Transformer model? Since the conservation values are not considered in the pre-training, the inputs for the Tranformer encoder will be different from the pre-training and possibly incur out-of-distribution problems.

  Reply:

  The conservation values reflect the conservation level of each base. Therefore, we treat it as a "position related" feature. In the model, we first feed it to an embedding layer and then add it to the other embeddings as the input to the model.

- (3) Do the Pangolion and TrASPr share the same training data? If not, is it possible to compare the TrASPr with the Pangolion trained with the same training data? It would more clearly show the effect of the proposed methods.

  Reply:

Yes. To have a fair comparison, we split the training and test in the same way.

- (4) How important is the pre-training? Is it possible to use the pre-trained DNABERT instead? How does the model perform with a randomly-initialized Transformer model? How important is the 6-mer tokenization compared to 1-mer tokenization?

  Reply:

  We used the DNABERT pretrained model and fine tuned it on the MGP dataset. Compared to the AE+MLP model, the performance was generally worse in most tissue pairs and the model was very finicky to parameter settings. Therefore, we decided to pretrain a lighter Transformer model with splice site data and then the performance improved as shown in the paper.

  The importance of the choice of k-mer was tested in a preliminary experiment. We tested 4, 5 and 6-mer on the simpler task of just splice site prediction (same task as in the SpliceAI algorithm). We saw a monotonic increase in accuracy from the 4-mer (0.949) to the 6-mer (0.976) and therefore used 6-mer. Of note, the sequences typically associated with an RNA recognition motif (RRM) are ~5b long which is why we did not test very short k-mers. Inline with that, we saw 5 and 6-mer performed similarly with only a slight advantage for the 6-mer.

- (5) How important is the Transformer architecture? How does the model perform with other model types? How important are the event features?

  Reply:

  We added ablation studies and tested different architecture as shown in table 1 below:

  **Table 1: Ablation study.** Full_TrASPr is the model with full features and loaded pretrained transform- ers. TrASPr_noPretrain has the same structure and input as the full TrASPr model but training from scratch. TrASPr_noFeature is the pretrained transformer but with no extra features and conservation values. TrASPr_LSTM indicates the model which replaced transformers with Bidirectional LSTM with no features as well.

| | Full TrASPr | TrASPr_noPretrain | TrASPr_noFeature | TrASPr_LSTM |
|---|---|---|---|---|
| AUPRC | 0.2845 0.2867 | 0.1889 0.1946 | 0.1762 0.1577 | 0.0674 0.0698 |
| Spearman | 0.3373 | 0.2040 | 0.2964 | 0.2222 |
| AUROC | 0.8774 0.8836 | 0.8516 0.8528 | 0.7896 0.7617 | 0.7070 0.7189 |

  As can be seen in the table, when the model is trained from scratch, it has a hard time to extract more useful information from inputs and converges slower than the full TrASPr. Therefore, even though it is similar to the full model on AUROC, it performs much worse in the other metrics. After removing the features, its performance significantly drops on prediction values but the ranking evaluation metrics doesn't decrease too much. We want to note this difference in the TrASPr_noFeature performance to also push back on the suggestion by reviewer aUkC that we simply applied an existing BERT model. Finally, if we use Bidirectional LSTM to replace transformers, the performance gets dramatically worse even compared to the no feature model. RNA splicing depends on complex regulatory elements around the splice sites. These results

indicate the advantage of transformers on information extraction from context at least on the splicing prediction task.

- (6) It seems TrASPr is used as an Oracle for both training and evaluation of the sequence design. Wouldn't it produce over-optimistic results?

  Reply:

  This is a point worth clarifying. The evaluation of BOS derived mutations' overlap with known sequence motifs (Fig5, right panel) is not based on TrASPr but rather known motifs of the splice factors which were shown experimentally to both bind the region and affect splicing when knocked down. In the second BOS evaluation, which assesses how many sequences perturbed PSI compared to a naive search there are of course no guarantees beyond the general accuracy we demonstrated for TrASPr. In that case it's indeed quite possible that the assessment is over-optimistic. We will make sure to convey this point in the revised manuscript. Nonetheless, we note that in the above both the naive and the BOS searches being compared are subject to the same evaluation.

- (7) How important is the LSBO? Can you provide comparisons with other baseline methods?

  Reply:

  Please see response to reviewer XeDB regarding other baselines to compare against.

- (8) The explanations of the biological problem and interpretations of the experiment results should be easier and more intuitive to understand. In addition, more detailed and friendly backgrounds should be included as supplementary.

  Reply:

  We very much agree and the reviews generally reflect that we did not do a good job at this. In the final version w\e will work to improve the explanation of the biological problem, the challenges involved, what are the main contributions/significance, and the experimental analysis/results.