**Enhancing mRNA-lipid Nanoparticle Prediction via the Chemical Language Model and Multi-task Learning**

**Yiyang Wu1**, Defang Ouyang1.

University of Macau1, Macau SAR, China.

**Background and aims.** Traditional ionizable lipid screening for mRNA lipid nanoparticles (mRNA-LNPs) relies on resource-intensive trial-and-error experiments. While machine learning approaches have offered promise in accelerating the development of mRNA-LNPs, their applications are limited by small datasets.

**Methods.** To address the challenges posed by limited data, we developed FormulationLNP, a model integrating a chemical language model with multi-task learning to predict two key properties of mRNA-LNPs: the *in vivo* delivery efficiency and apparent pKa. For model development, we constructed the largest ionizable lipid structure dataset to date (~16,000 lipids) and compiled datasets for both target properties (Figure 1A). An ionizable lipid-tailored chemical language model was pre-trained with SMILES enumeration to learn comprehensive lipid representations (Figure 1B). Given the strong correlation between apparent pKa and *in vivo* delivery efficiency (1), a multi-task learning architecture was implemented to simultaneously fine-tune both prediction tasks (Figure 1C).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | ***In vivo* delivery efficiency** |  | **Apparent pKa** |
| **Model** |  | **ROC-AUC** |  | **ROC-AUC** |
| LightGBM-RDKit |  | 0.855±0.027 |  | 0.856±0.054 |
| DNN-ECFP6 |  | 0.852±0.033 |  | 0.855±0.038 |
| AttentiveFP |  | 0.794±0.023 |  | 0.680±0.129 |
| ChemBERTa |  | 0.837±0.038 |  | 0.819±0.038 |
| FormulationLNP (Ours) |  | **0.862±0.037** |  | **0.867±0.032** |

**Results.** FormulationLNP, trained with 5-fold data augmentation, showed excellent performance across 10 repeated experiments. It achieved ROC-AUC scores of 0.862±0.037 and 0.867±0.032 for *in vivo* delivery efficiency and apparent pKa, respectively, outperforming other baseline models (Table 1). Ablation studies revealed that pre-training contributed the most to model performance, while multi-task learning enhanced predictions for both tasks simultaneously. The model also exhibited strong generalization capability, achieving predictive accuracies of 0.882 and 0.758 on external test sets for *in vivo* delivery efficiency and apparent pKa, respectively. Furthermore, key ionizable lipid substructures associated with *in vivo* behavior were identified, offering valuable insights for rational ionizable lipid design.

**Figure 1.** The workflow of this study.

**Table 1.** Comparisons of models developed by different algorithms.

**Conclusion.** In conclusion, by integrating pre-training, data augmentation and multi-task learning, FormulationLNP effectively addressed the challenge of small mRNA-LNP datasets. This approach provides a powerful tool for predicting the *in vivo* behavior of mRNA-LNPs and will greatly accelerate the design and optimization of LNP delivery systems.

**References:** (1) Patel P., Ibrahim NM., Cheng K. Trends Pharmacol Sci. 2021:448–60.