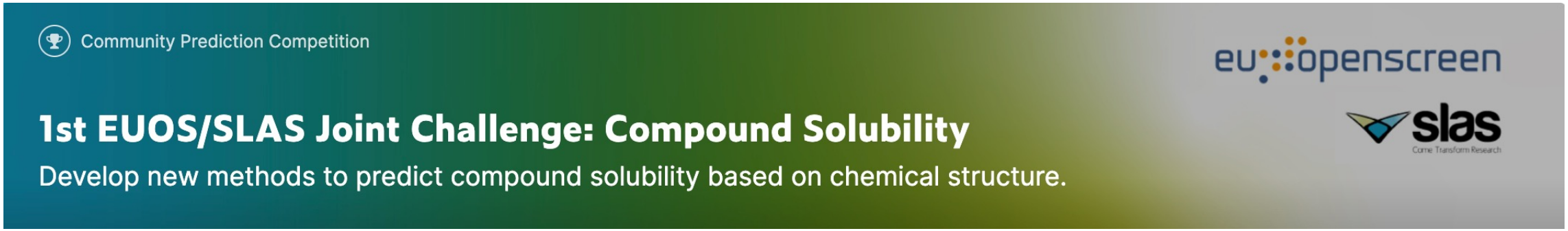
Tokyo Tech



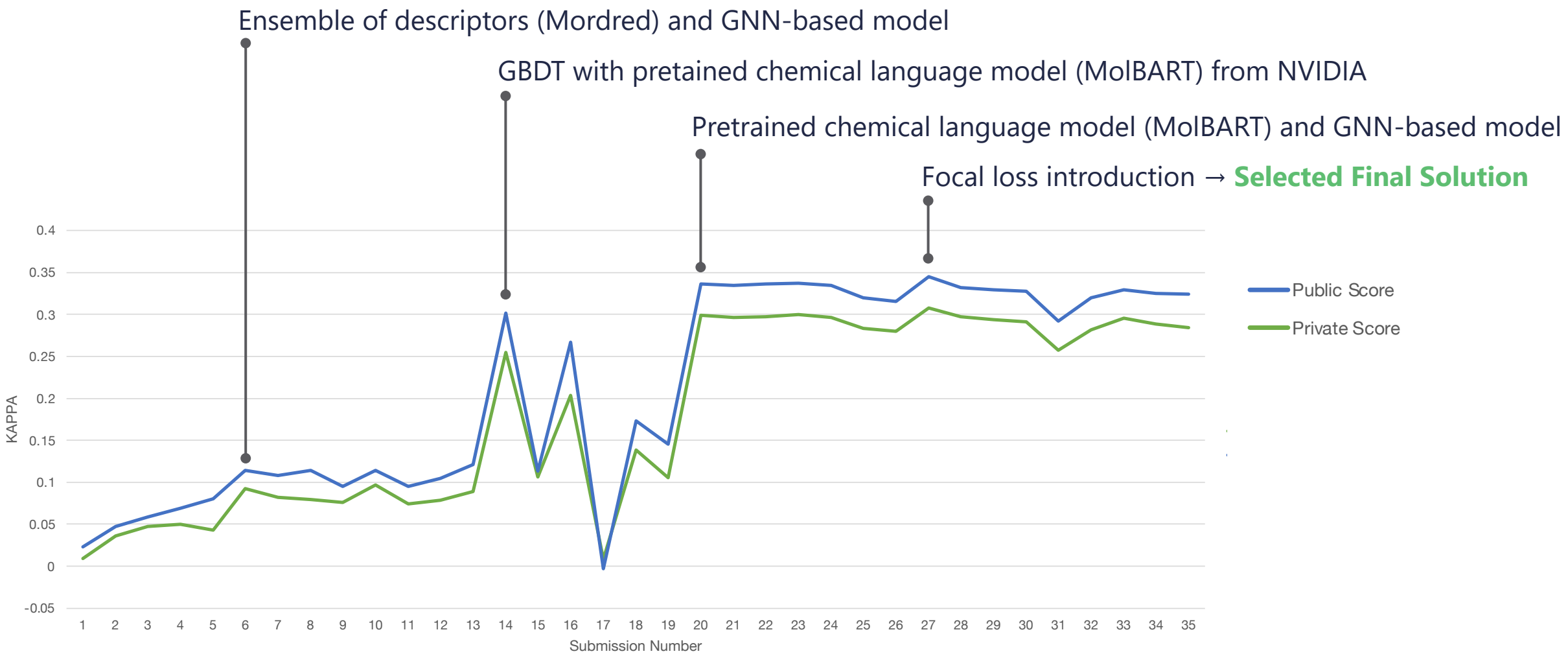
Solution from olab

Apakorn | Furui | Ohue

Ohue laboratory https://www.li.c.titech.ac.jp/en/

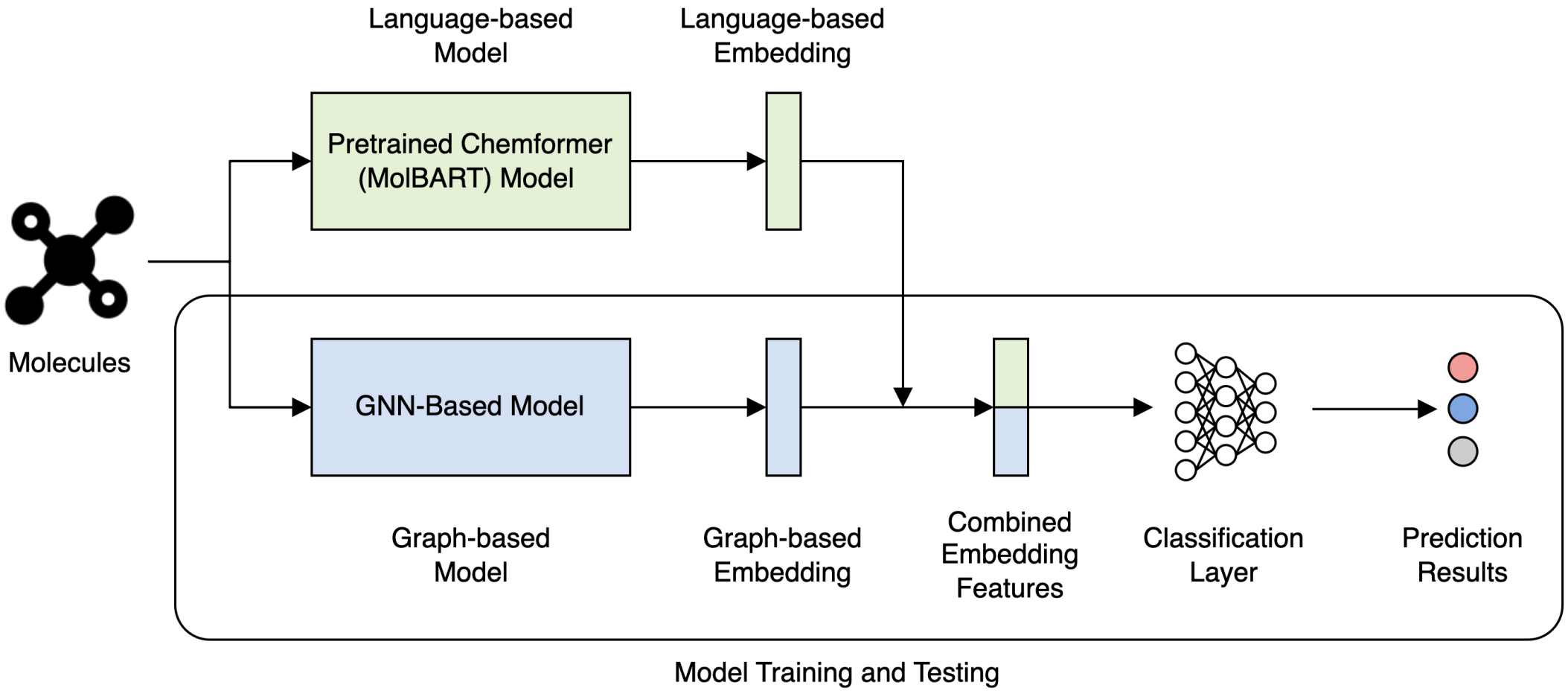
Story of our solution development

Our solution development from submission history on Kaggle



Overview workflow of our solution

Combination of pretrained chemical language-based model with graph-based model



Overview workflow of our solution

[1] Irwin R., et al. Mach. Learn.: Sci. Technol., 2022
[2] https://github.com/NVIDIA/MegaMolBART

Pretrained Chemformer (MolBART) Model

Pretrain Chemformer (MolBART) Model [1][2]

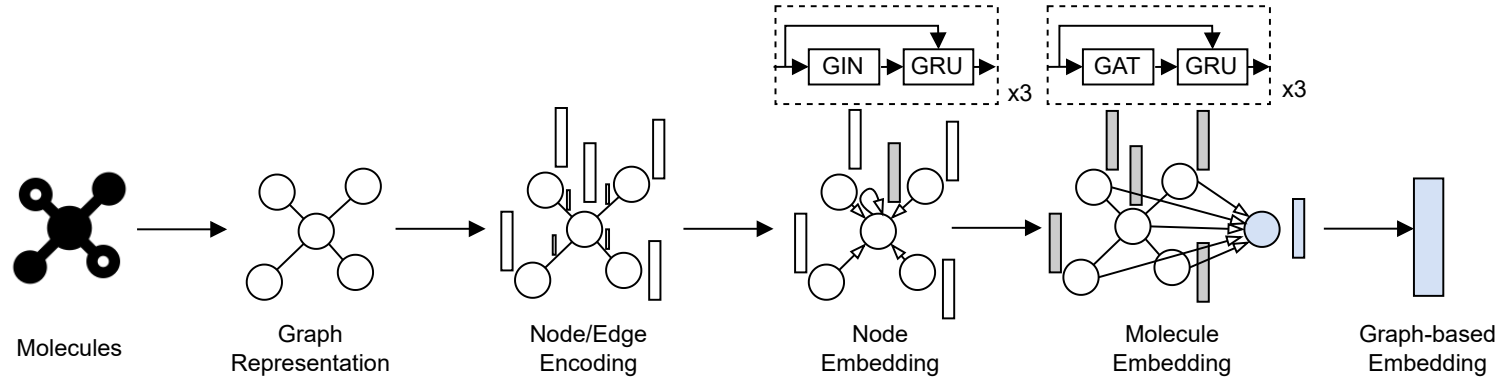
- Transfromer model from NVIDIA's NeMo-Megatron framework
- Pre-training on approximately 1.45 billion molecules from ZINC-15 database using SMILES language encoding

GNN-Based Model

GNN-Based Model

- Learn node embedding with GIN and molecule embedding and GAT
- All learning layers apply GRU to enhance embedding information

GIN: Graph Isomorphism Network
GAT: Graph Attention Network
GRU: Gated Recurrent Unit

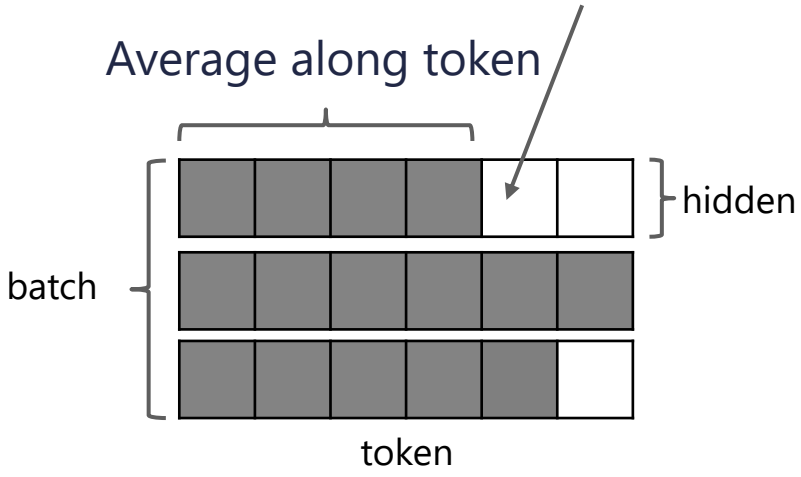


Issues in chemical language model

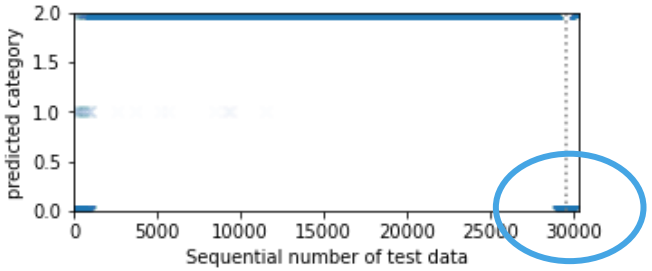
Problems in latent space generation step

- Latent space is obtained by averaging the encoder output of the language model along the tokens, since the length of the tokens depends on the length of the SMILES
  - Padding features that should not be included in the averaging process were also included.
  - Information on compounds in the same batch is shared in the padding portion.
  - If latent spaces are generated for the same batch with the same solubility category, they can be easily identified by the similarity of the latent spaces.
- In this competition, compounds with high solubility were lined up at the end of the test data.
  - Easily identified by similarities in latent space
  - Performance was unintentionally supported by order of test data

Padding features contain information shared within batch



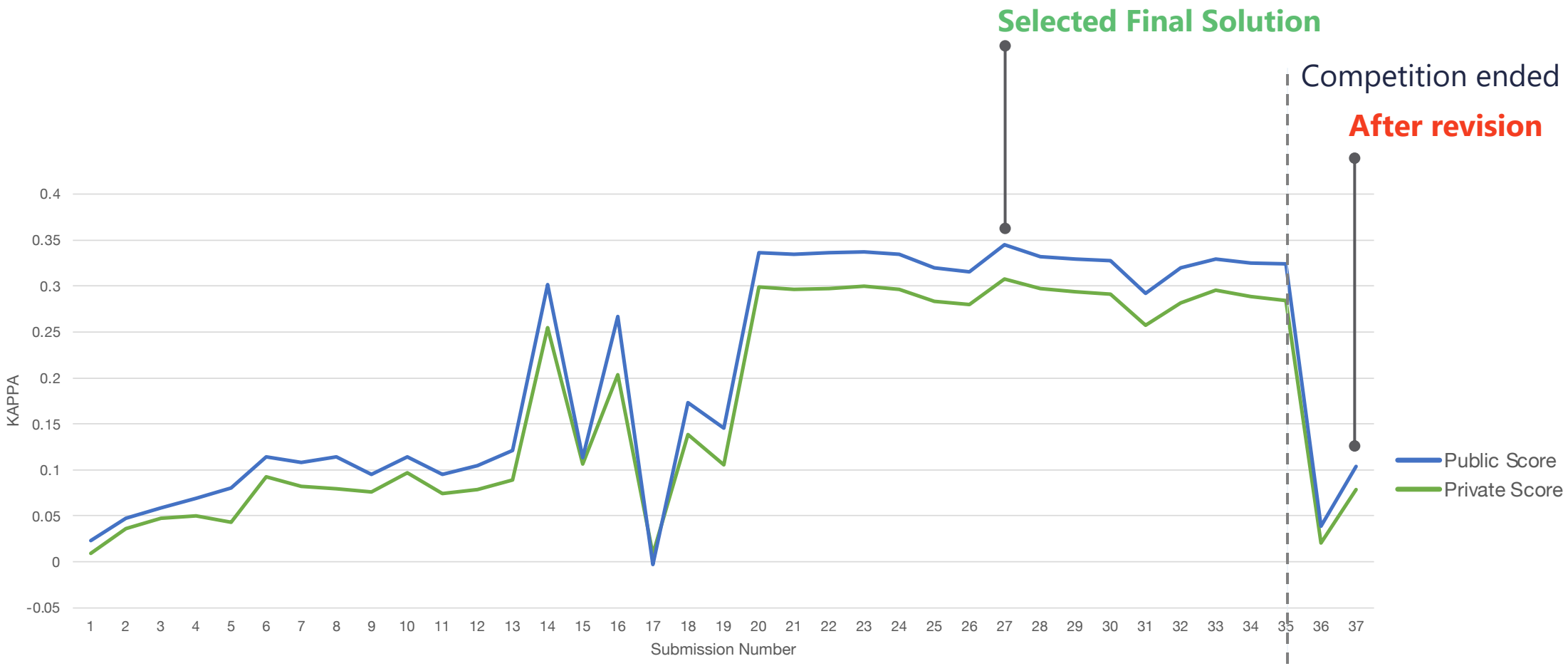
[BATCH_SIZE, TOKEN_NUM, HIDDEN_DIM]
↓
[BATCH_SIZE, HIDDEN_DIM]



We examined why it is possible to predict test data in the wrong way.

After revision of chemical language model

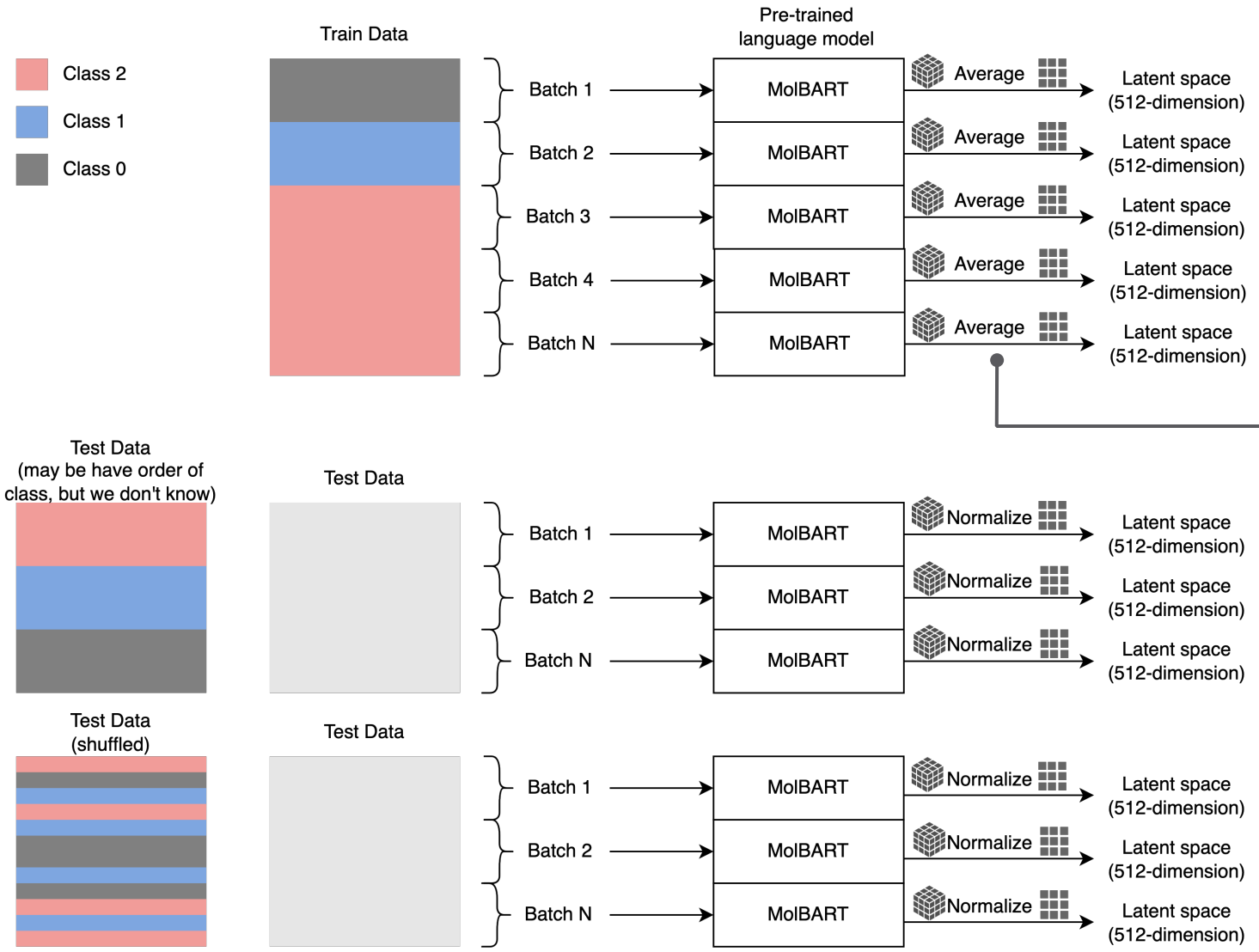
We submitted late submission with our revised version of solution again



Tokyo Tech

Thank you very much

Embedding leakage



Because of our mistaken configuration, this step calculates specifically for each batch, so when batch contains same class of compounds, the latent space is generated with bias. (Latent space leakage, latent space generation unintentionally gains some advantages when the batch containing same class compounds)

Good Performance

Cannot reproduce same good performance

Generating process of embbedings