Predicting Chemical Properties using Self-Attention Multi-task Learning based on SMILES Representation

Sangrak Lim, Yong Oh Lee

KIST Europe, Korea Institute of Science and Technology Europe Branch,
Campus E7 1, 66123 Saarbrücken Germany

Contents

I. Background

II. Method

III. Comparison with other models

IV. Results

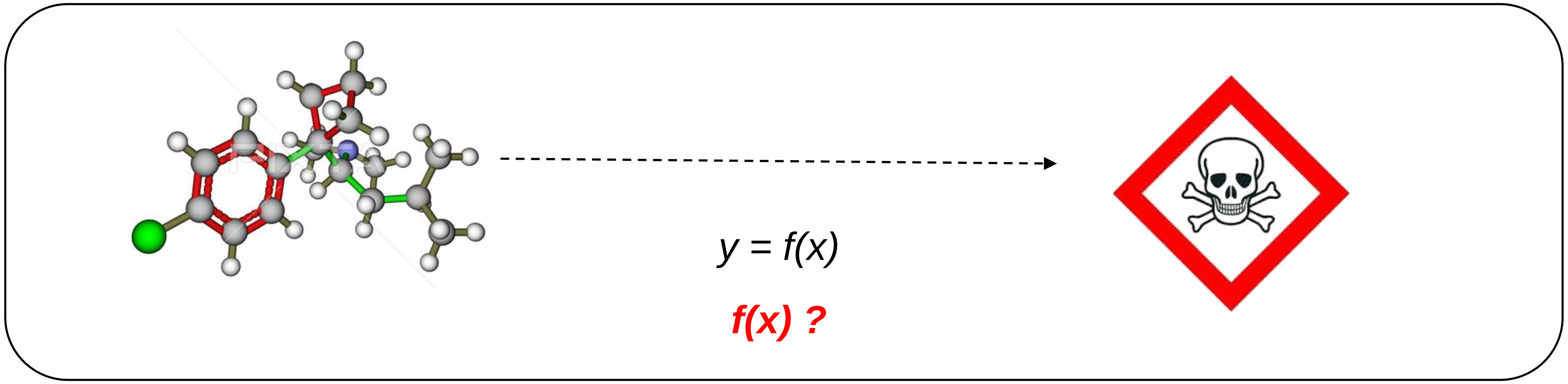
V. Conclusions

1. Background – QSAR model

- **Quantitative structure–activity relationship** (QSAR) models extract relationships from chemical structures and predict biological activities, such as toxicity, solubility, and so on.

- QSAR models are used in chemical and biological domain. The main applications include high throughput screening of chemicals for toxicity prediction and drug delivery.

- Previous QSAR models utilized molecular descriptors to represent chemical properties as vectors. The selection of proper molecular descriptors is challenging as the performance of QSAR model is highly dependent on descriptors.



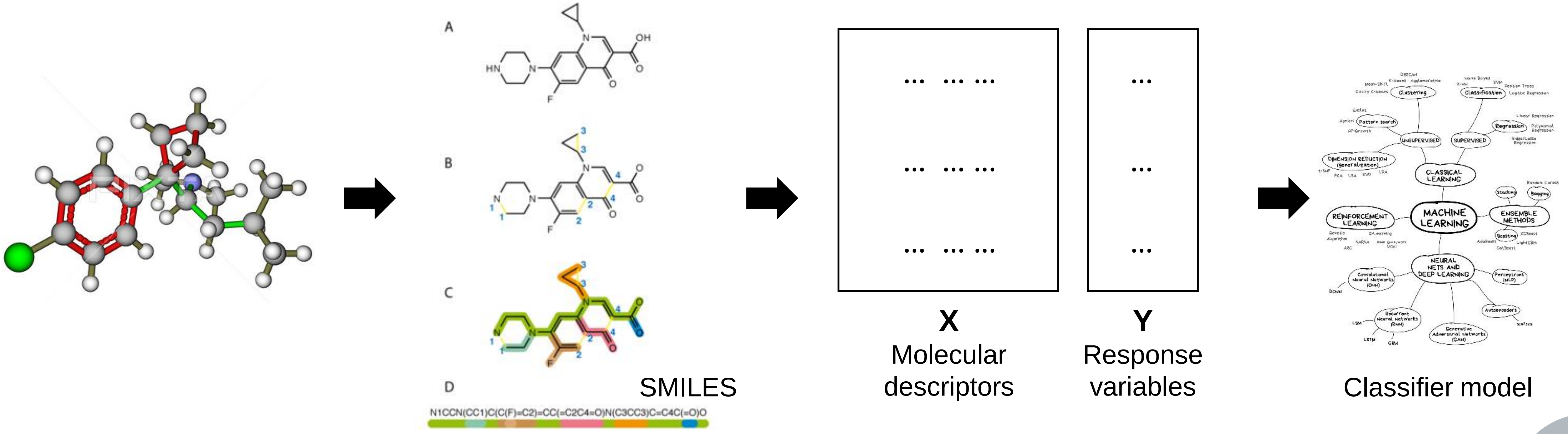
$y = f(x)$

f(x) ?

1. Background - Challenges

- For a deep learning models based on descriptors have two problems:

  1. Molecular descriptors require additional conversion processes from inputs, such as the Simplified Molecular Input Line Entry System (SMILES).
  2. The search space for certain substructures of chemicals converted into descriptors can be limited or ignored


A
B
C
D
N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O
… … …
…
MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
Clustering
Classification
Regression
Pattern search
DIMENSION REDUCTION (generalization)
REINFORCEMENT LEARNING
ENSEMBLE METHODS
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
Recurrent Neural Networks (RNN)
Generative Adversarial Networks (GAN)
Autoencoders
Perceptrons (MLP)

SMILES

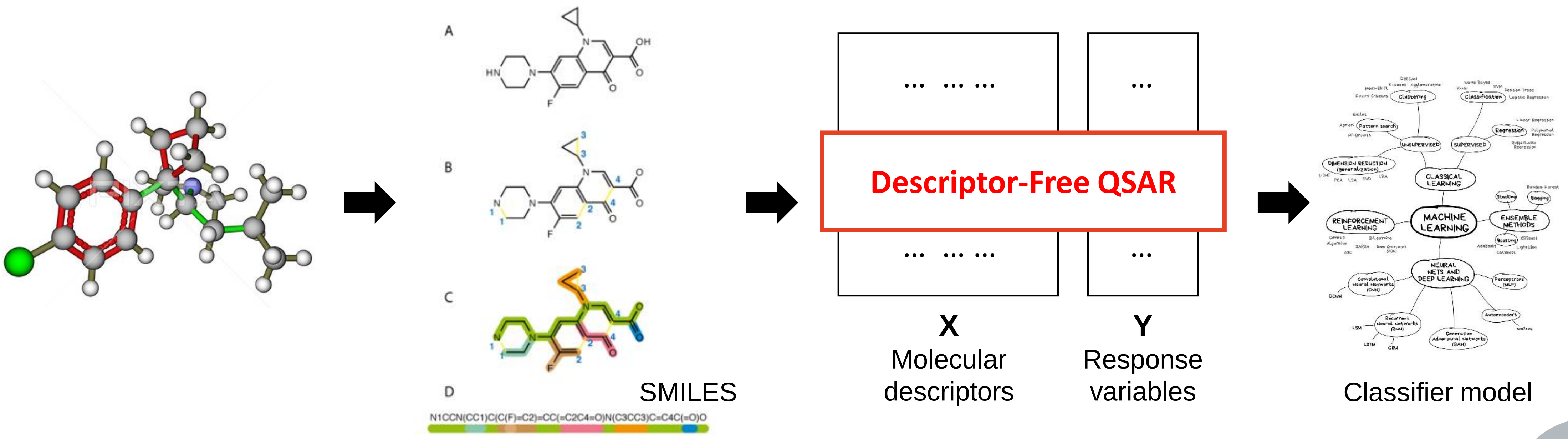
X
Molecular descriptors

Y
Response variables

Classifier model

1. Background - Proposal

- We present a Natural Language Processing (NLP) based QSAR model that utilizes SMILES as direct input.
- We explored the structural differences of existing transformer-variant models and proposed a new self-attention based model.
- The representation learning performance of our self-attention module was evaluated in a multi-task learning environment using several chemical datasets.



SMILES

Descriptor-Free QSAR

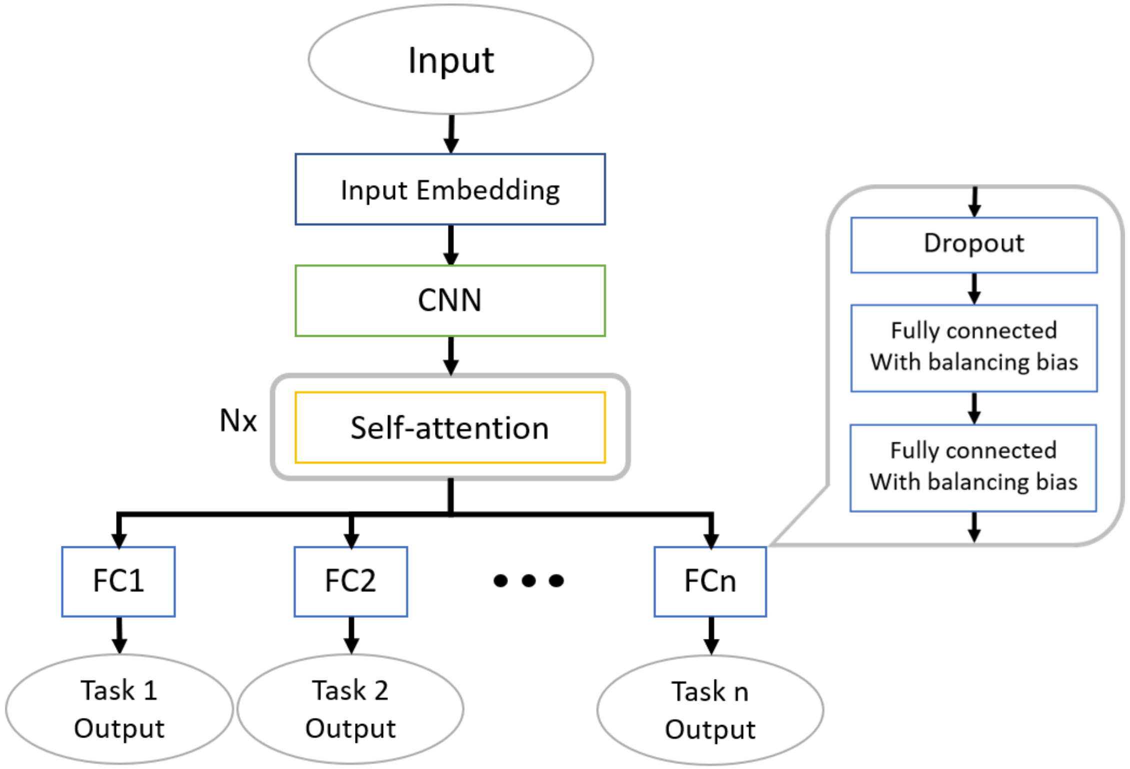
X
Molecular descriptors

Y
Response variables

Classifier model

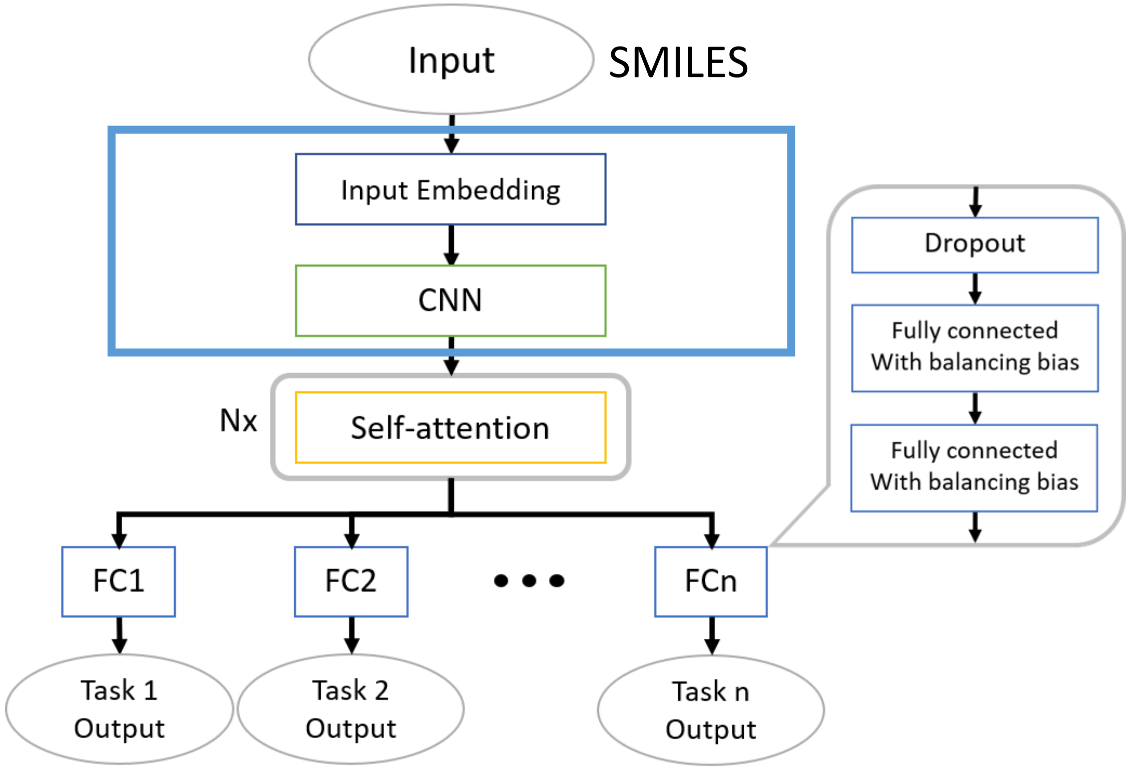
2. Method

- Component 1: SMILES embedding and feature extraction with CNN layers
- Component 2: Self-attention
- Component 3: Multi-task learning



2. Method – SMILES embedding and feature extraction with CNN layers

- A CNN layer serves as a shared hidden layer for multi-task learning.
- Input is a SMILES format. ➔ No chemical descriptors required.
- A shared hidden layer is useful when target of the multi-task learning is closely related tasks.



The CNN layer stores shared hidden features even though the self-attention layer is located ahead of the discrete output layer.

Ablation Study

- Excluding the CNN layer leads to lower performance.
- Compared to RNN layer, CNN layer is faster at converging.

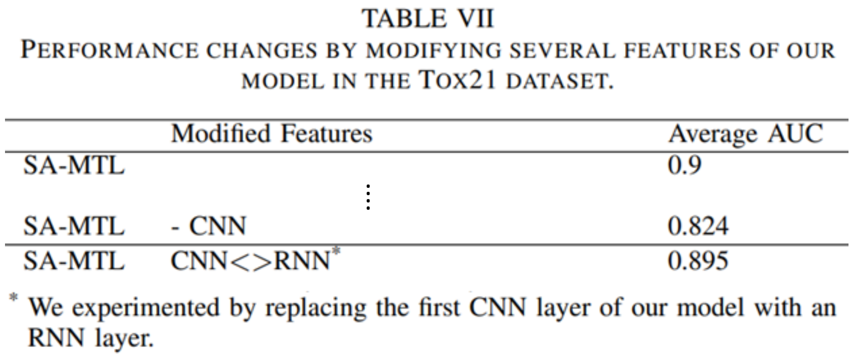


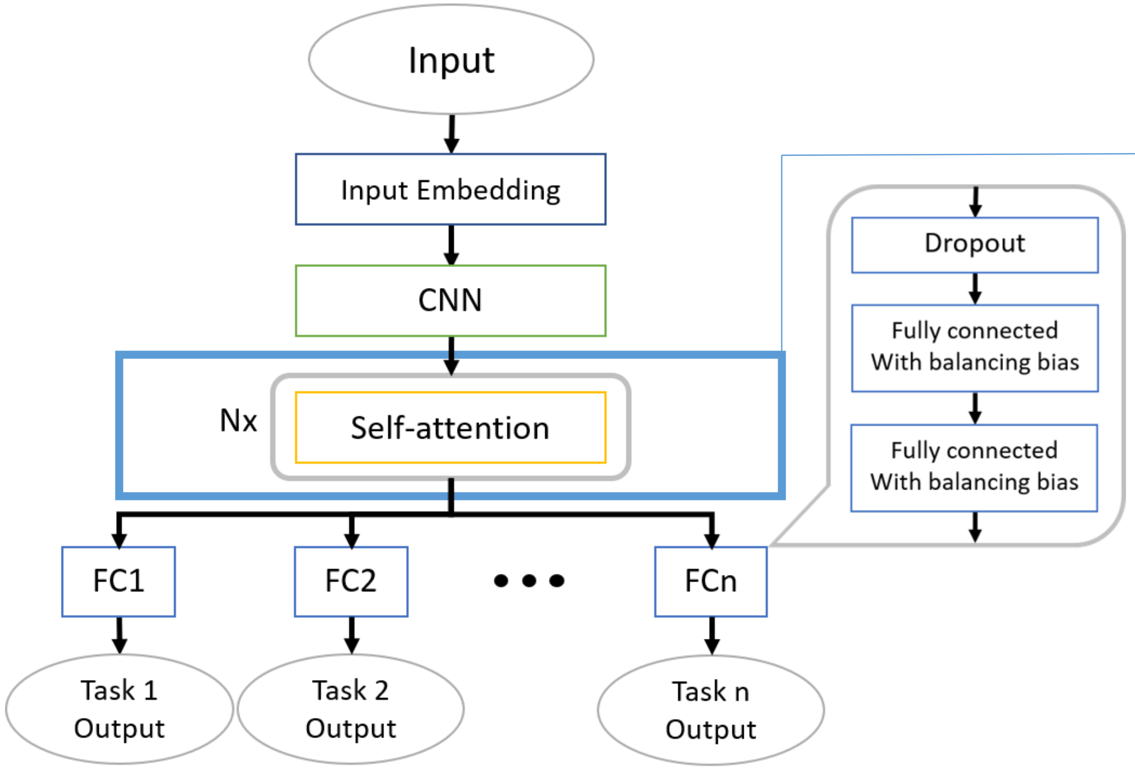
TABLE VII
PERFORMANCE CHANGES BY MODIFYING SEVERAL FEATURES OF OUR MODEL IN THE TOX21 DATASET.

	Modified Features	Average AUC
SA-MTL		0.9
	⋮	
SA-MTL	- CNN	0.824
SA-MTL	CNN<>RNN*	0.895

* We experimented by replacing the first CNN layer of our model with an RNN layer.

2. Method – Self-attention

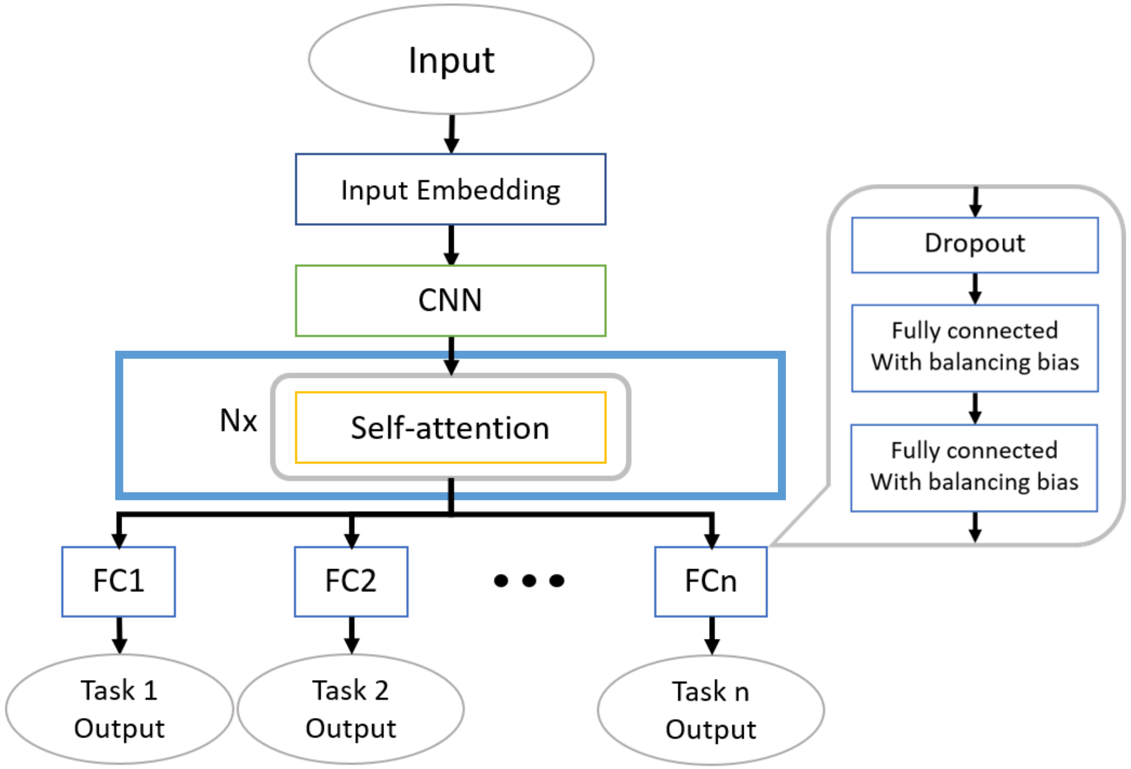
- A Self-attention module focuses on long-range dependencies of a given input.
- Pseudo-code of the self-attention is shown below.
- No pre-training (see the comparison with other study part)



$$Q, K, V = linear(x), linear(x), linear(x)$$
$$\mathrm{attn}_{\mathrm{output}} = softmax(\frac{QK^T}{\sqrt{n}})V$$
$$ln1 = layernorm(x + \mathrm{attn}_{\mathrm{output}})$$
$$\mathrm{FFN}_{\mathrm{output}} = linear\left(relu(linear(ln1))\right)$$
$$ln2 = layernorm(ln1 + \mathrm{FFN}_{\mathrm{output}})$$

2. Method – Self-attention

- A Self-attention module focuses on long-range dependencies of a given input.
- Pseudo-code of the self-attention is shown below.
- No pre-training (see the comparison with other study)



The Self-attention module, as well as multi-task learning, is an essential component.

Ablation Study

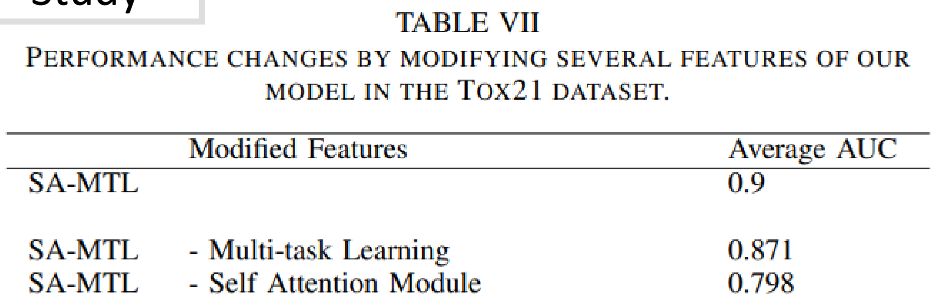
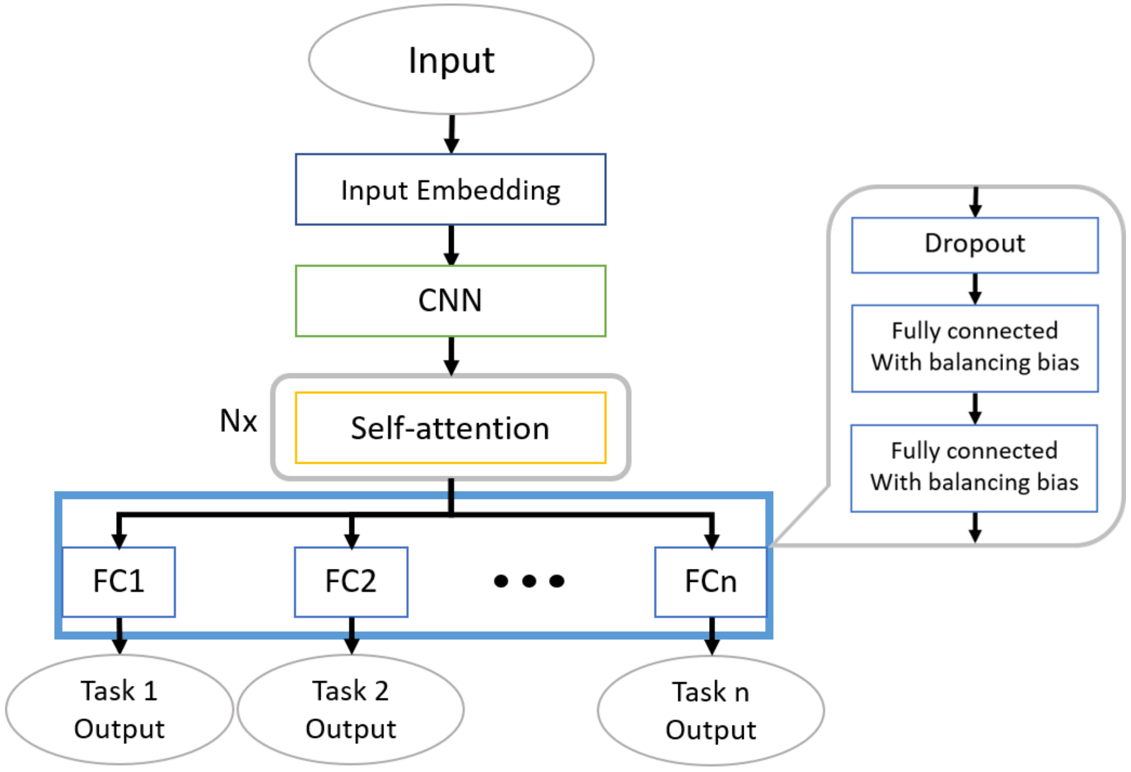


TABLE VII
PERFORMANCE CHANGES BY MODIFYING SEVERAL FEATURES OF OUR MODEL IN THE TOX21 DATASET.

	Modified Features	Average AUC
SA-MTL		0.9
SA-MTL	- Multi-task Learning	0.871
SA-MTL	- Self Attention Module	0.798

2. Method – Multi-tasking learning

- Discrete output layers produce outputs for multiple tasks
- Shape : [batch size, sequence size, hidden size] => [batch size]
- A balancing bias is applied to rectify the class-imbalance in the data



Replacing the discrete output layer with max pooling leads to lower performance.

Ablation Study

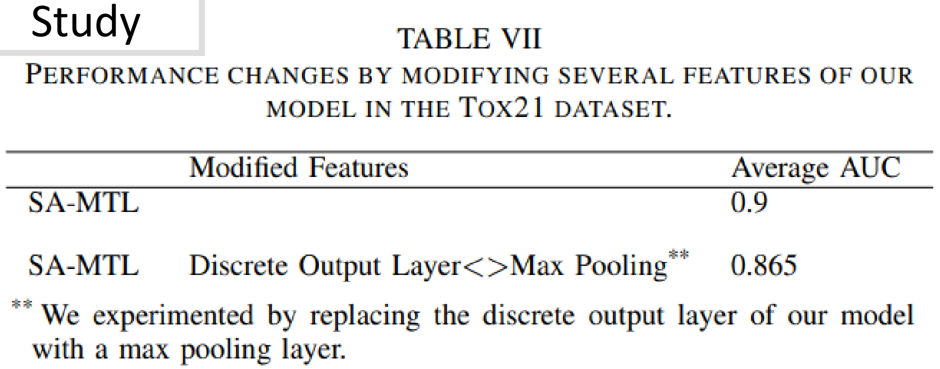


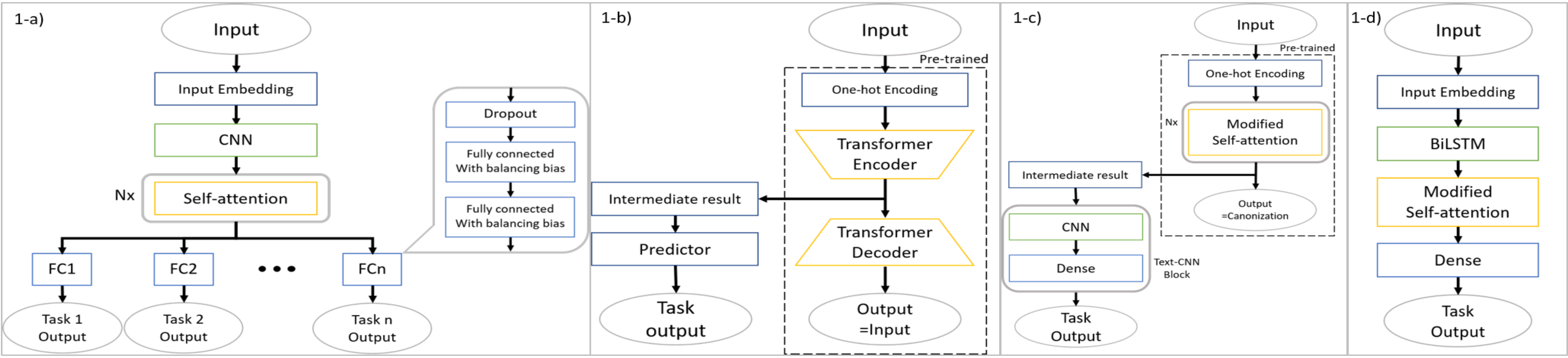
TABLE VII
PERFORMANCE CHANGES BY MODIFYING SEVERAL FEATURES OF OUR MODEL IN THE TOX21 DATASET.

	Modified Features	Average AUC
SA-MTL		0.9
SA-MTL	Discrete Output Layer<>Max Pooling**	0.865

** We experimented by replacing the discrete output layer of our model with a max pooling layer.

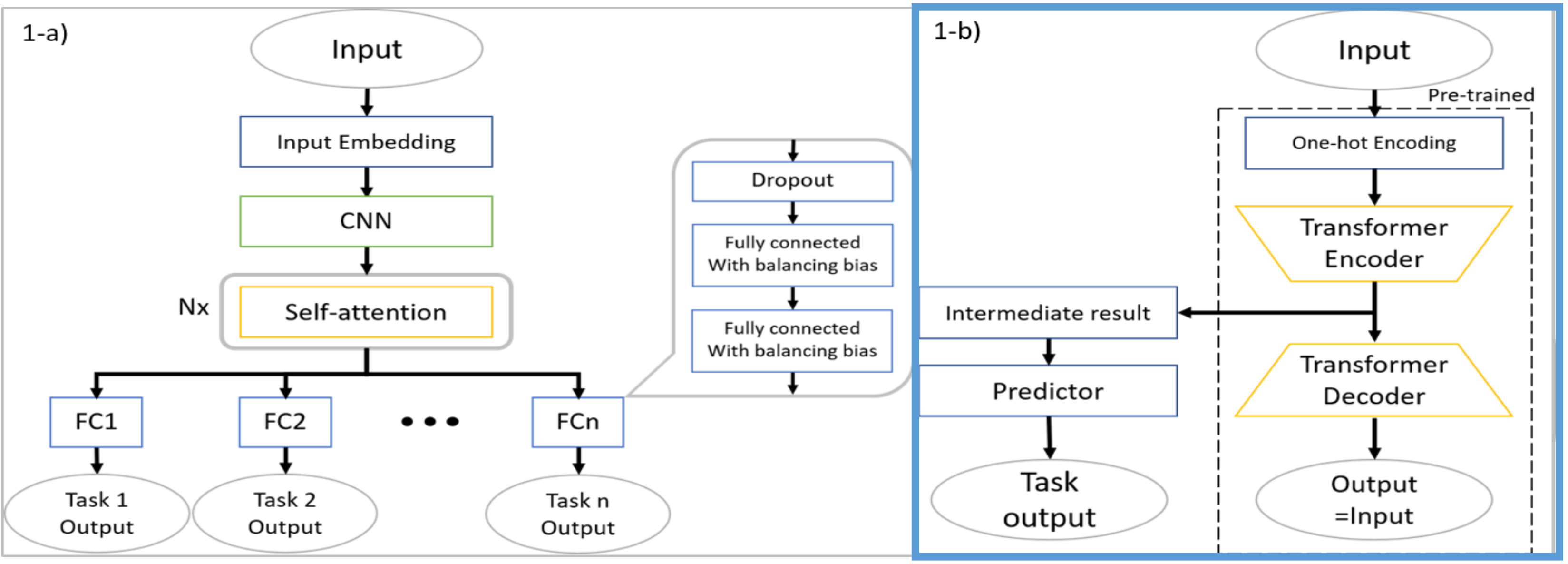
3. Comparison with other studies

- 1-b) Smiles Transformer Model: The Smiles Transformer model uses the intermediate result obtained from the pre-training step.
- 1-c) Transformer-CNN Model: The Transformer-CNN model also implemented the pre-training approach. The model contains text-CNN block for several CNN layers.
- 1-d) BiLSTM-SA Model: The concept of the BiLSTM-SA model implemented a self-attention module without the multi-task learning scheme.



3. Comparison with other studies

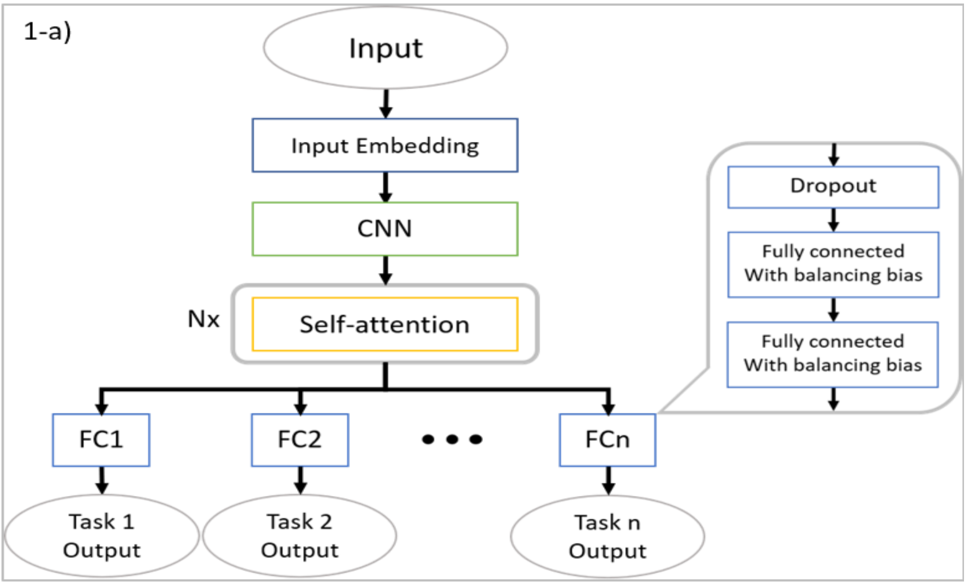
- 1-b) Smiles Transformer Model: The Smiles Transformer model uses the intermediate result obtained from the pre-training step.

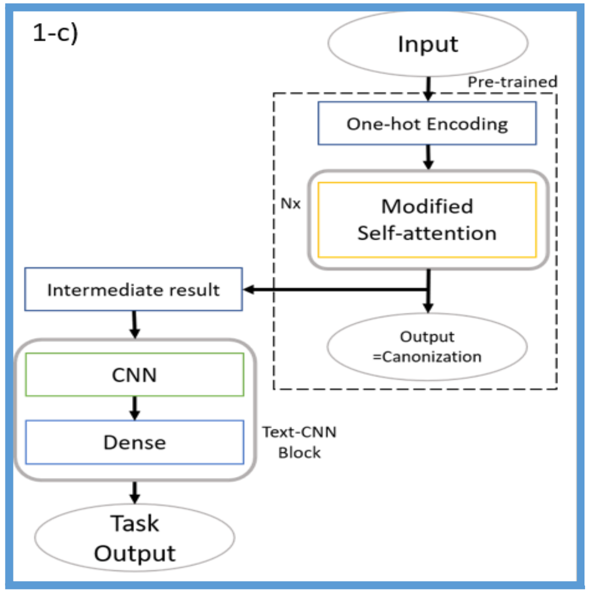


Smiles Transformer

3. Comparison with other studies

- 1-c) Transformer-CNN Model: The Transformer-CNN model also implemented the pre-training approach. The model contains text-CNN block for several CNN layers.

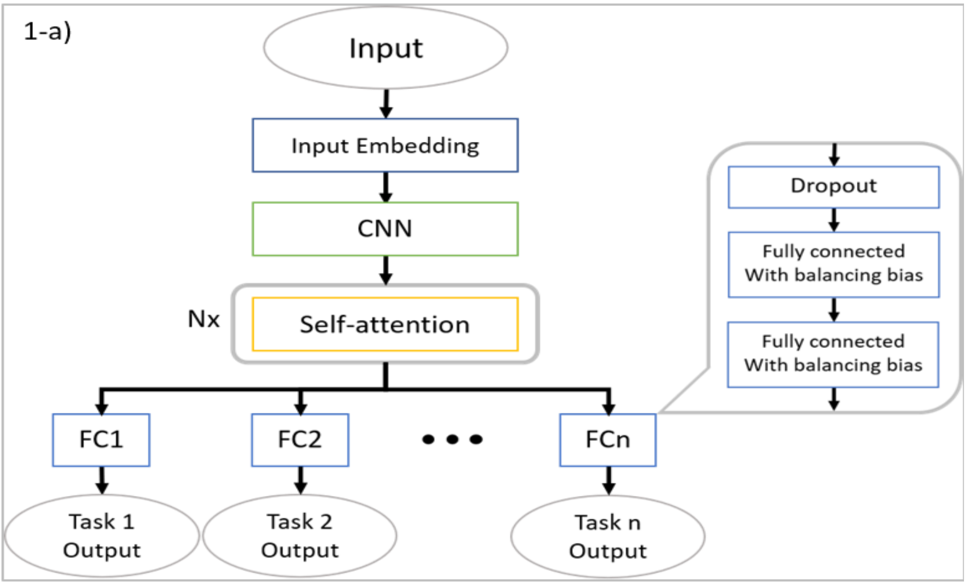


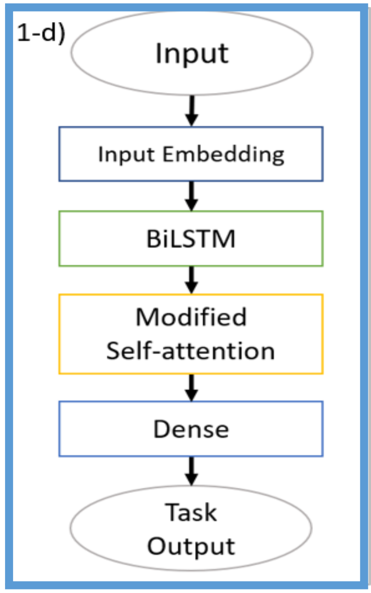


Transformer-CNN

3. Comparison with other studies

- 1-d) BiLSTM-SA Model: The concept of the BiLSTM-SA model implemented a self-attention module without the multi-task learning scheme.





BiLSTM-SA Model

4. Results – Tox21

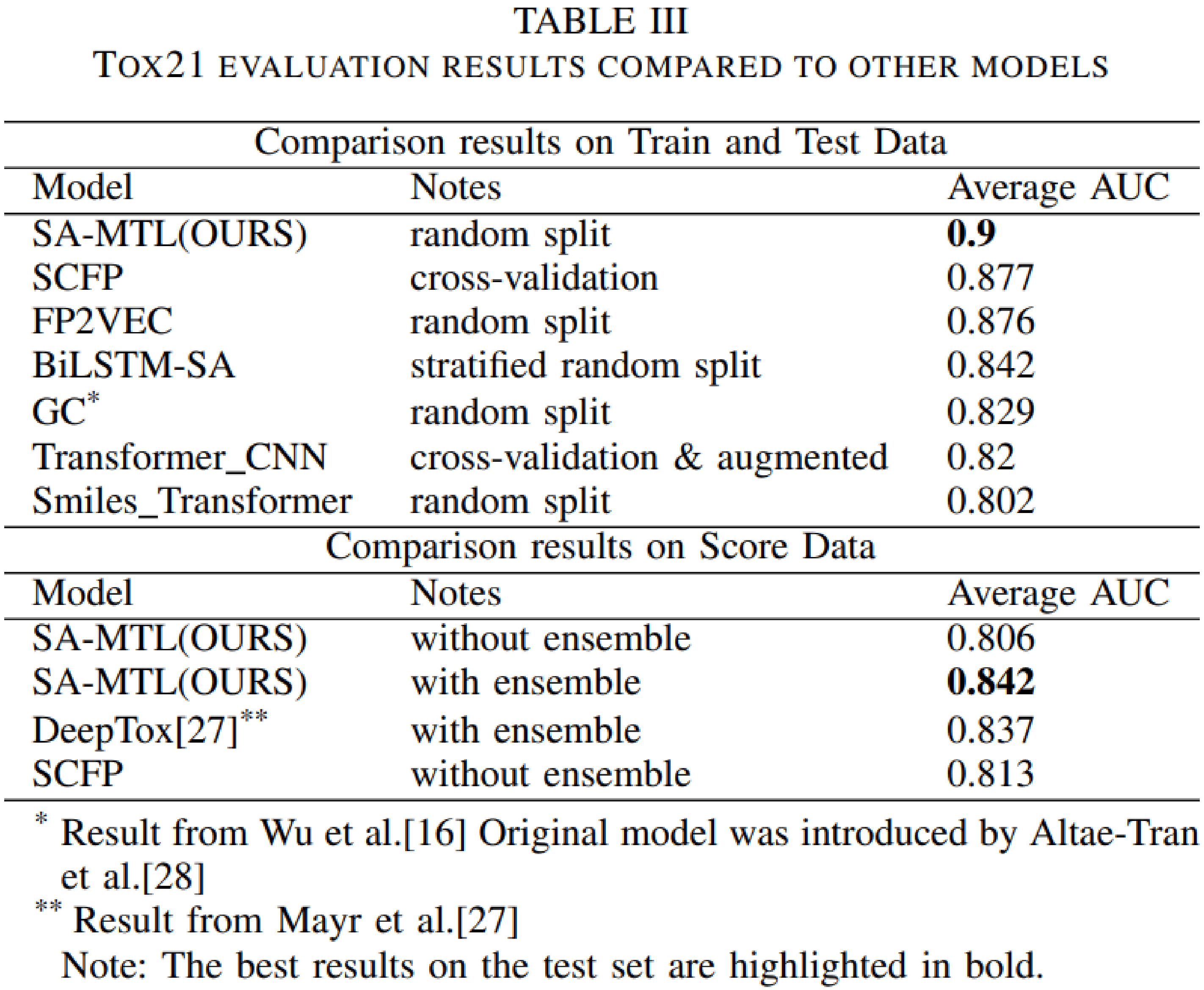


TABLE III
TOX21 EVALUATION RESULTS COMPARED TO OTHER MODELS

Comparison results on Train and Test Data		
Model	Notes	Average AUC
SA-MTL(OURS)	random split	**0.9**
SCFP	cross-validation	0.877
FP2VEC	random split	0.876
BiLSTM-SA	stratified random split	0.842
GC*	random split	0.829
Transformer_CNN	cross-validation & augmented	0.82
Smiles_Transformer	random split	0.802
Comparison results on Score Data		
Model	Notes	Average AUC
SA-MTL(OURS)	without ensemble	0.806
SA-MTL(OURS)	with ensemble	**0.842**
DeepTox[27]**	with ensemble	0.837
SCFP	without ensemble	0.813

* Result from Wu et al.[16] Original model was introduced by Altae-Tran et al.[28]

** Result from Mayr et al.[27]

Note: The best results on the test set are highlighted in bold.

4. Results – ablation study

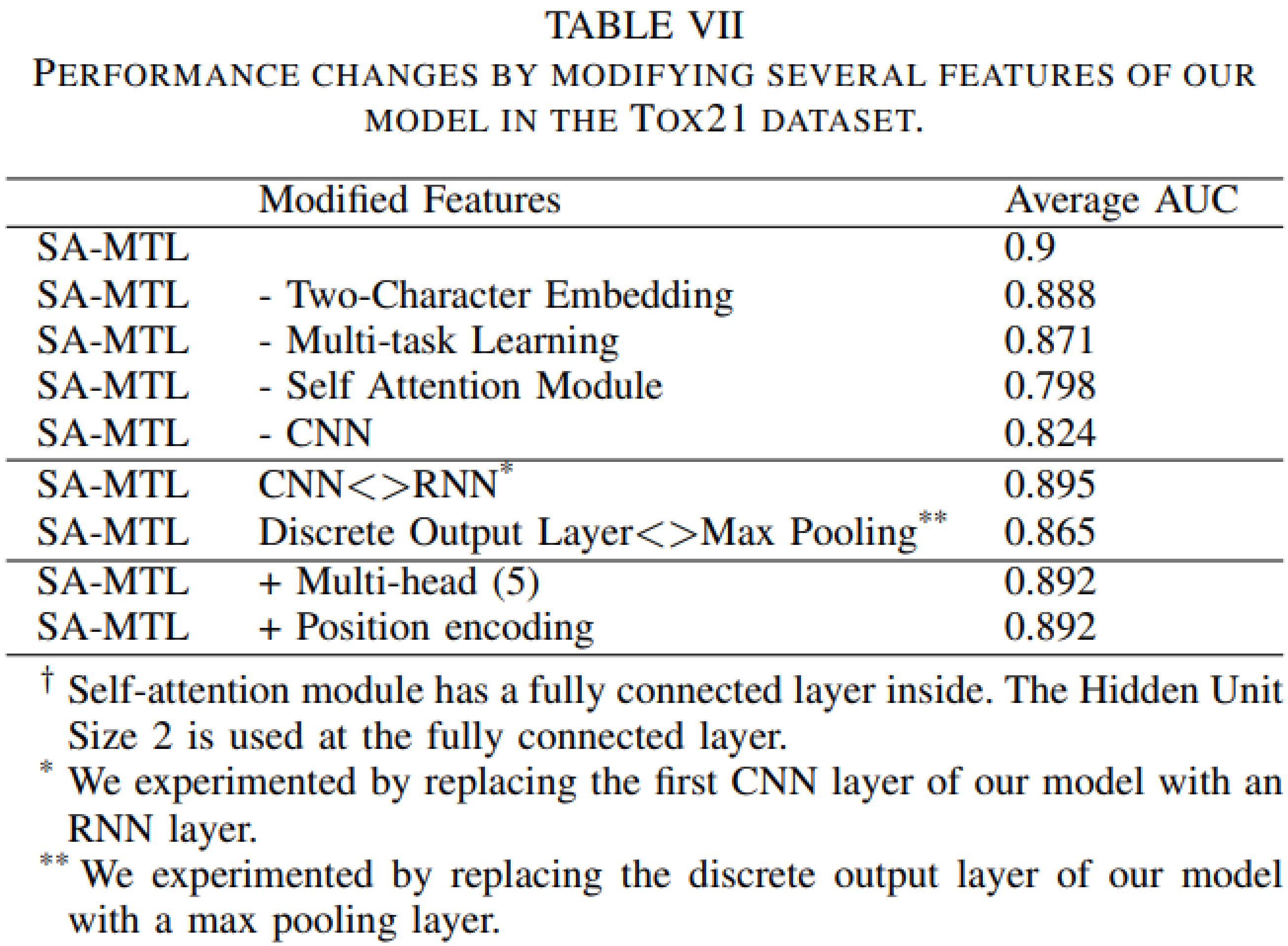


TABLE VII

Performance changes by modifying several features of our model in the Tox21 dataset.

	Modified Features	Average AUC
SA-MTL		0.9
SA-MTL	- Two-Character Embedding	0.888
SA-MTL	- Multi-task Learning	0.871
SA-MTL	- Self Attention Module	0.798
SA-MTL	- CNN	0.824
SA-MTL	CNN<>RNN[*]	0.895
SA-MTL	Discrete Output Layer<>Max Pooling[**]	0.865
SA-MTL	+ Multi-head (5)	0.892
SA-MTL	+ Position encoding	0.892

[†] Self-attention module has a fully connected layer inside. The Hidden Unit Size 2 is used at the fully connected layer.

[*] We experimented by replacing the first CNN layer of our model with an RNN layer.

[**] We experimented by replacing the discrete output layer of our model with a max pooling layer.

5. Conclusions

- We proposed Self-attention Multi-Task learning (SA-MTL) QSAR model which is a descriptor-free as SMILES is the direct input of the model.
- We described structural differences of our model and other transformer-variant models and showed the influence of such a structural change on learning.
- Our SA-MTL model exhibited the state-of-the-art performance in the Tox21 and several other datasets.

Acknowledgement

This study is supported by National Research Council of Science & Technology (NST) grant by the Korea government (MSIP) (No. CAP-17- 01-KIST Europe) and the Basic research grant (12001).

THANK YOU

KIST Europe Forschungsgesellschaft mbH
Campus E7.1
66123, Saarbrücken, Germany