A Transformer-Convolutional Neural Network Based Framework for Predicting Ionic Liquid Properties

Guzhong Chen[1], Zhen Song[1], and Zhiwen Qi[1]

[1]East China University of Science and Technology

September 13, 2022

Abstract

Of central importance to evaluate the suitability of ionic liquids (ILs) for a process is the accurate estimation of IL properties related to target performances. In this work, a versatile deep learning method for predicting IL properties is developed. Molecular fingerprints are derived from the encoder state of a Transformer model pre-trained on the PubChem database, which allows transfer learning from large-scale unlabeled data and significantly improves generalization performance for developing models with small datasets. Employing the pre-trained molecular fingerprints, convolutional neural network (CNN) models for IL properties prediction are trained and tested on 11 databases. The obtained Transformer-CNN models present superior performance to state-of-the-art models in all cases and enable property prediction of millions of ILs shortly. The application of the proposed models is exemplified by searching CO2 absorbent from a huge database of 8,333,096 synthetically feasible ILs, which is by far the most high-throughput IL screening in literature.

Hosted file

`Graphical Abstract.emf` available at https://authorea.com/users/344984/articles/585940-a-transformer-convolutional-neural-network-based-framework-for-predicting-ionic-liquid-properties

A Transformer-Convolutional Neural Network Based Framework for Predicting Ionic Liquid Properties

Guzhong Chen,[a,b] Zhen Song,[a,*] Zhiwen Qi[a,*]

[a] *State Key laboratory of Chemical Engineering, School of Chemical Engineering, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China*

[b] *Process Systems Engineering, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, D-39106 Magdeburg, Germany*

Corresponding authors: songz@ecust.edu.cn (Z. S.); zwqi@ecust.edu.cn (Z. Q.)

Abstract:

Of central importance to evaluate the suitability of ionic liquids (ILs) for a specific process is the accurate estimation of IL properties related to the target performance. In the present work, a versatile deep learning method for predicting various properties of ILs is developed. Molecular fingerprints are derived from the encoder state of a Transformer model pre-trained on the PubChem database, which allows transfer learning from large-scale unlabeled data and significantly improves generalization performance for developing models with small datasets. Employing the pre-trained molecular fingerprints, convolutional neural network (CNN) models for IL properties prediction are trained and tested on 11 different databases. The obtained Transformer-CNN models present superior performance to state-of-the-art models in all these cases and enable the property prediction of millions of ILs in a short time. The application of the proposed models is exemplified by extensively searching CO_2 absorbent from a huge database of 8,333,096 synthetically feasible ILs, which is by far the most high-throughput IL screening in literature.

Keywords: IL properties prediction, deep learning, Transformer, convolutional neural network, high-throughput IL screening, CO_2 capture

1 INTRODUCTION

Ionic liquids (ILs) are molten salts comprised of cations and anions which remain in liquid state around ambient temperature. In recent years, ILs have attracted considerable attention in various applications due to their unique physicochemical properties such as negligible volatility, high thermal and electrochemical stability, and wide liquidus range.[1,2] More importantly, ILs also offer the potential to tune their physical and chemical properties by judicious selection of their cations and anions. Due to this tunable character, ILs can be designed to offer favorable properties to meet specific requirements for a given application. The challenge, however, is to evaluate various IL properties related to the target performance and identify optimal ILs from the near-infinite combinations of the possible ions and functional groups.[3,4]

So far, IL selection toward a specific process mainly relies on laborious trial-and-error experiments and experiences. However, such approaches are not only very time-consuming but also limited to a small IL chemical space, leaving many potentially promising structures unexplored. Alternatively, thermodynamic, transport, and EHS (environment, health, and safety) related properties of ILs can be estimated by computational methods,[5] following which targeted IL design/screening can be performed. Traditional thermodynamic models, such as Equations of State (EoSs)[6] and group contribution models (GCMs)[7,8] are commonly used for estimating IL properties. Nevertheless, both the two schemes are prone to inherent weakness of limited predictive power and/or insufficient accuracy.[7] Another computational method for IL property prediction is the quantitative structure-property relationship (QSPR) approach, wherein the property of interest is quantitatively correlated with certain structural descriptors of molecules.[8,9] However, the databases used to correlate the properties of ILs are often small and the molecular descriptors used to represent ILs are often diverse, making

many developed QSPR models only cover a limited applicability domain and not easily integrable with each other in IL design/screening framework. As a quantum mechanics based predictive approach, the conductor-like screening model for real solvents (COSMO-RS)[10] has been shown to be a relatively robust predictive method for IL-involved mixture properties such as activity coefficients of molecular solutes,[11,12] gas absorption capacity,[13,14] and lignin and cellulose solubilities.[15] However, the COSMO-RS approach requires computationally expensive (especially for complex IL structures) density functional theory calculations and has been proven to be qualitative rather than quantitative for some IL-involved systems.[16]

In addition to the above methods, machine learning (ML) techniques for molecular property prediction have recently gained in popularity in cheminformatics and promoted broad applications of data-driven models in chemical engineering studies.[17–21] With the availability of IL property databases such as the ILThermo[22], there has been a sharp rise in the use of data-driven methods for modelling IL properties.[23–30] Among these ML models, different types of molecular descriptors have been used for IL representation. For example, Paduszyński has developed ML based models for estimating density (ρ)[31] and dynamic viscosity (η)[32] of pure ILs by group contribution (GC) descriptors; Song et al. have modeled CO_2 solubility in ILs by artificial neural network (ANN) and support vector machine (SVM) algorithm, both adopting GC descriptors of ILs.[26] Despite the notably improved accuracy, these models still suffer from the inherent weaknesses of GC approach while covering only a small number of functional groups due to the limitation of the IL properties database for model development. Zhu et al.[33] and Peng et al.[27] utilized the screening charge density distribution area (S_σ) of the COSMO-RS approach as an a priori quantum chemistry descriptor for modeling cytotoxicity of ILs towards the leukemia rat cell line IPC-81.

Although the COSMO-based descriptor can be theoretically used for any IL, the computational cost of such descriptor for complex ILs could be very high, which is contrary to the goal of quickly predicting IL properties by ML to a certain extent. Two structural descriptors, the ECFP4 circular fingerprints and the Coulomb matrix have been chosen by Low et al.[24] for melting point prediction of ILs. Although such structural descriptors are easy to calculate and can theoretically be used for any IL, the prediction accuracy of the kernel ridge regression (KRR) model is not satisfactory. Beyond these structural descriptors, one eventually needs to add descriptors obtained by quantum mechanical calculations (e.g., molecular orbital energies and/or σ-profiles of ILs) to obtain higher model prediction accuracy.[24] To sum up, almost all currently reported ML models for IL property prediction employ manually designed IL descriptors, which usually require expert knowledge for the types of ILs and the properties to be modeled. This kind of IL descriptors and the ML models developed thereon could work well for specific tasks, but may not generalize well for others.[34]

In very recent years, ML methods, especially deep neural networks (DNN), have evolved rapidly. DNN-based ML systems have aroused great interest by overcoming obstacles of conventional models and obtaining high prediction quality for complex tasks.[35–38] The growth of deep learning (DL) has provided excellent flexibility and performance to learn molecular fingerprints from data, without explicit guides from experts.[39–41] In our previous work, a DNN based recommender system (RS) for predicting the solute-in-IL infinite dilution activity coefficient (γ^{∞}) was developed without any manually designed fingerprint. Instead, neural network embeddings were employed for mapping each IL and solute, which can be learned automatically as part of the neural network for γ^{∞} prediction.[42] The γ^{∞} prediction accuracy of this model exceeds all ML models that use manually designed fingerprints. However, due to the

matrix completion essence of the selected method,[43] the developed RS model mainly applies to already covered ILs and solutes in the database.

To develop DL approaches with strong extrapolation performance, a sufficiently large labeled training database is desirable. In many areas (e.g., image classification), the labeled sample number could easily reach several millions or even more. However, it is not the case for IL properties prediction, for which the available databases (i.e., experimental or theoretically calculated IL properties) are far smaller than such a scale and do not always cover a wide range of functionalities. This challenge of developing strong extrapolative model based on a small dataset is also originally encountered in natural language processing (NLP), which has almost unlimited unlabeled data while only a tiny portion of labeled data.[44] To address this challenge in NLP, extensive efforts have been devoted by researchers, among which the pre-training and fine-tuning approach[45] is very encouraging. In this approach, word representations are derived from statistics gathered from large unlabeled corpus of text data by pre-training; these pre-trained representations provide distributional information about words that typically improve the generalization of models learned on a limited amount of data by fine-tuning.

The structure of molecule sequences is shown to be very similar to the structure of natural language sentences when molecules are represented by the simplified molecular-input line-entry system (SMILES)[46,47] sequence. There are already several millions of molecules (including ILs) that have been synthesized in laboratory, which can be readily retrieved from online databases such as the PubChem and ChEMBL. Pre-training can also exploit such large-scale unlabeled databases to learn the representations of molecules or molecular fragments, and then the pre-trained model can be fine-tuned to downstream molecular property prediction tasks using a relatively smaller set of labeled data. Winter et al.[48] developed a pre-trained sequence-to-sequence

(seq2seq) model for predicting molecular properties based on recurrent neural networks (RNNs) by translating equivalent chemical representations. Gómez-Bombarelli et al.[39] used variational autoencoders (VAE) to get continuous representation of molecules in a latent space, and molecular properties were then predicted by decoding SMILES from the learned representations. The Transformer model[49] that comprises an encoder-decoder architecture (more parallelizable and superior as opposed to seq2seq) have also been used in the cheminformatics field for molecular properties prediction[50,51] and reaction prediction[52,53], achieving better model performance based on small databases when comparing with other pre-training approaches.[34] Our recent work on quickly predicting surface charge density profiles (σ-profile) and cavity volumes (V_{COSMO}) of molecules also utilized a pre-trained SMILES Transformer as molecular fingerprints.[54] These studies have demonstrated the success of pre-training methods for predicting various molecular properties including physical properties (melting point, aqueous solubility), QM calculated properties (σ-profile, V_{COSMO}), molecular orbital properties (HOMO, LUMO), and EHS related properties (mutagenicity, toxicity). Therefore, it could be highly expected that the predictive modeling of IL properties based on small databases can also be achieved in a similar manner.

With these observations, we propose a pre-training and fine-tuning two-stage framework for IL properties prediction as outlined in Figure 1. Importantly, our model does not make use of any manually designed or selected molecular fingerprint. Instead, a self-attention mechanism is deployed to learn the high-dimensional structure of an IL from a given raw sequence, that is, the SMILES. The large unlabeled IL SMILES database is first taken for the unsupervised pre-training of the self-attention mechanism (Molecular Transformer model), to obtain the encoder-decoder architecture that can well capture the complex structure of a molecule from its canonical SMILES.

Following that, the encoder of the Transformer is integrated with a convolutional neural network (CNN) architecture for supervised training of predictive models of IL properties. By simply switching the labeled IL property dataset (and adding other inputs such as temperature and pressure if needed), predictive models for various IL properties can be developed based on the proposed framework.

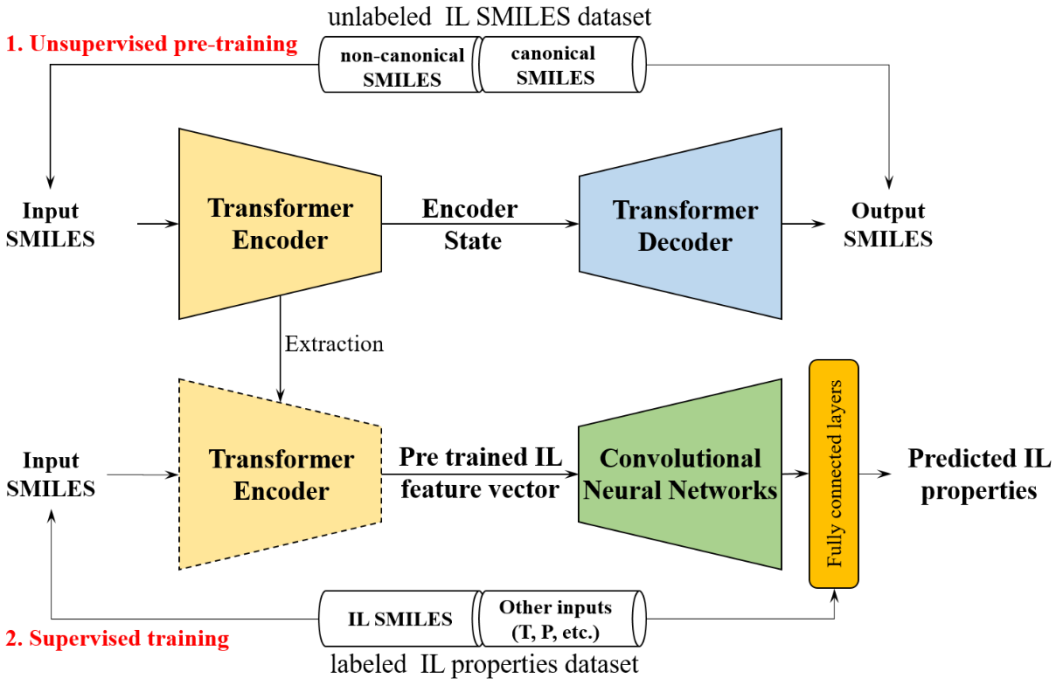


Figure 1. Schematic outline of the Transformer-CNN framework for IL properties prediction.

The rest of the paper is organized as follows: Section 2 provides an overview of the pre-training and IL properties datasets, the prediction problem, and the two-stage modelling framework. Section 3 briefly introduces the Transformer architecture used for molecular representation learning and the CNN architecture used for supervised learning of IL properties. In Section 4, we present the modeling results on both IL representation learning and different IL properties datasets, and comparisons are made with reference works. In Section 5, the developed IL properties prediction models are applied in a high-throughput IL screening task for CO_2 capture. Finally, a summary of this work is given in Section 6.

2 OVERVIEW OF DATASETS AND MODELLING PROBLEM

2.1 Pre-training dataset

To get the SMILES Transformer model for IL encoding, the PubChem[55] compound database is used. The original dataset (*ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound/*) contains a total of 108,923,995 molecules as well as their canonical SMILES representations. Due to the limited memory size of the computer used, we cannot train the Transformer model on the entire PubChem compound database. Therefore, considering the aim of predicting IL properties, the molecules containing '+' and/or '-' symbols in the SMILES are first screened to form a subset of IL-like molecules, retaining 10,243,410 molecules. As summarized in Figure 2, the length of 92.10% of the SMILES strings in the subset of IL-like molecules is below 100 characters. To make the input length of the model not too large, molecules with SMILES string length less than or equal to 100 characters (9,434,070 molecules) are kept to form the pre-training dataset. This pre-training dataset is augmented 10 times (an optimal value reported by Tetko et al.[56]) up to 94,340,700 non-canonical SMILES strings using the SMILES enumerator to increase the performance of DNN models that can be developed.[57]

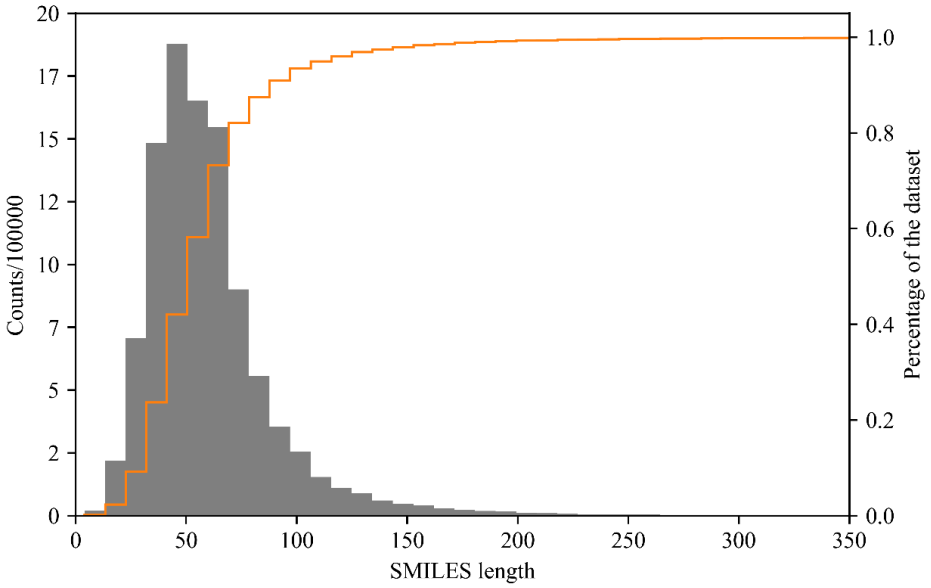


Figure 2. Length distribution of canonical SMILES representations of molecules in IL-liked subset.

2.2 IL properties datasets

The IL properties datasets used in this work are derived from several recently published scientific works as listed in Table 1. From these datasets, only ILs with SMILES string length less than or equal to 100 characters (consistent with the pre-training dataset) are kept. The involved properties of ILs can be divided into two types. One type is the properties related only to IL molecular structure, such as melting point, glass transition temperature, thermal decomposition temperature, and cytotoxicity. The other type relates to not only IL molecular structure but also conditions such as temperature and pressure, including heat capacity, refractive index, density, viscosity, surface tension, CO_2 solubility, and thermal conductivity. When dealing with the latter type of IL properties, as the random splitting of the entire dataset may cause overestimation of models by separating data points of the same ILs (with only difference in temperature and/or pressure) into both the training and test sets, a more rigorous IL-based dataset splitting strategy as used in our previous study is adopted[42]. By using this method, data points of the same IL at different temperatures and pressures only enter the same subset during the splitting of training and test sets, which can avoid data leakage and give a more reliable test score.

Table 1. IL properties involved in this work.

Property	Number of data points	Number of ILs	Data Source
melting point $T_m(K)$	2,212	2,212	Low et al.[24]
glass transition temperature T_g(℃)	609	609	Venkatraman et al.[29]
thermal decomposition temperature T_d(℃)	1,223	1,223	Venkatraman et al.[29]
heat capacity $\ln(C_p)$	9,083	236	Venkatraman et al.[29]
refractive index n_D	3,009	464	Venkatraman et al.[29]
density ρ	31,167	2,257	Paduszyński[25]
viscosity $\ln(\eta)$	15,368	1,964	Paduszyński[28]

surface tension γ	2,972	331	Venkatraman et al.[29]
CO_2 solubility x_{CO_2}	10,116	124	Song et al.[26]
cytotoxicity towards the leukemia rat cell line IPC-81 $log_{10}(EC_{50})$	326	326	Wang et al.[58]
thermal conductivity λ	454	73	Venkatraman et al.[30]

2.3 Prediction problem formulation

The objective of this work is to predict the properties of ILs only from its structure represented by SMILES (and the temperature and/or pressure if needed). However, it can be seen from Table 1 that although there is a considerable amount of data points (hundreds to tens of thousands) on the properties of ILs, the number of ILs involved (dozens to thousands) is not enough for many properties. With such small datasets, the application of traditional ML methods and conventional molecular descriptors for property prediction may be largely limited. To introduce additional molecular representation information to solve the problem of small datasets, this work formulates the IL properties prediction problem as a pre-training and fine-tuning two-stage framework. The motivations of this method are two-folded: (i) to build a powerful semi-supervised framework utilizing the essential information in unlimited unlabeled data to improve the prediction performance with limited labeled data; (ii) to enable the proposed model framework to predict various properties of ILs only by replacing the database used for supervised training.

To accurately predict IL properties from relatively small labeled datasets, we first perform an auxiliary text translation task based on sequence to sequence learning[59] with SMILES as text representations of millions of molecular structures to get molecular fingerprints of ILs. Employing the pre-trained molecular fingerprints (as well as temperature and/or pressure if needed) as inputs, models for predicting IL properties can be trained on small labeled datasets as listed in Table 1.

3 METHODS

3.1 Transformers

The SMILES Transformer model used in this work is based on the Transformer architecture originally constructed by Vaswani et al.[49] for neural machine translation (NMT) tasks. Similar to the seq2seq model used for molecular property prediction[48,60] and reaction prediction[61,62], Transformer is also based on the encoder-decoder architecture[63]. The main architectural difference from seq2seq models is that the RNN component is completely removed, and it is fully based on the attention mechanism combined with positional embedding for encoding sequential information. In the following, a brief description of the encoder-decoder architecture, attention mechanism, and positional encoding that comprise the building blocks of a Transformer is given.

3.1.1 Encoder–decoder architecture

As an instance of the encoder-decoder architecture, the overall architecture of the transformer is presented in Figure. 3. The encoder maps an input sequence of symbol representations $(x_1, \dots, x_n)$ to a sequence of continuous representations $\mathbf{z} = (z_1, \dots, z_n)$. Given $\mathbf{z}$, the decoder then generates an output sequence $(y_1, \dots, y_m)$ of symbols, one element at a time. At each step, the model is auto-regressive, consuming the previously generated symbols as additional input when generating the next. In the encoder, the multi-head attention layers attend the input sequence and encode it into a hidden representation carrying the essential information, namely encoder state. The decoder consists of two types of multi-head attention layers: the first type is masked and attends only the preceding outputs of the decoder, while the second type multi-head attention layer attends encoder states as well as the output of the first decoder attention layer. It basically combines the information of the source sequence with the target sequence that has been produced so far. The SMILES Transformer model in this work

utilized 3 Transformer blocks for both encoder and decoder, that is, N is 3 in Figure 3.

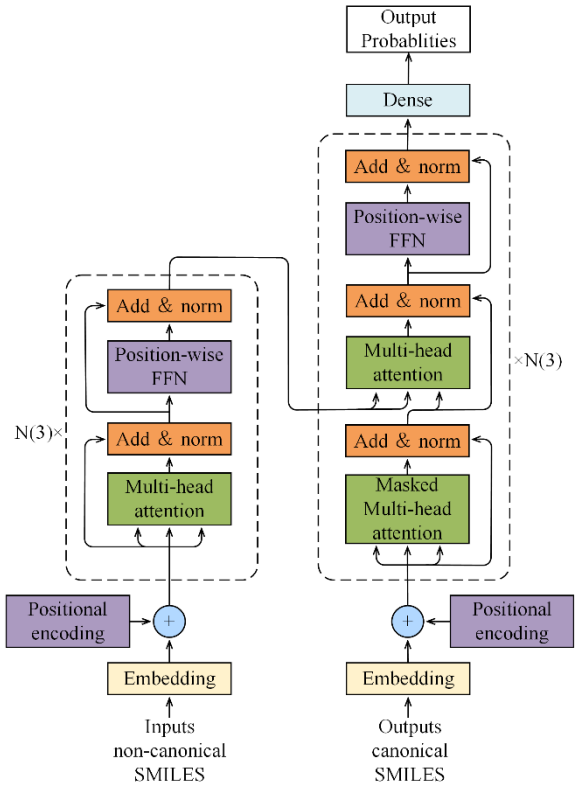


Figure 3. Architecture of the SMILES Transformer model used in this work. The left-half corresponds to the encoder while the right-half corresponds to the decoder.

3.1.2 Multi-head Attention

As the most important part of the Transformer architecture, the attention mechanism allows the model to focus on different tokens in the sequence at different stages of the network, enabling it to discover multiple relationships between groups of tokens. The attention function used here is called Scaled-Dot Product Attention[49] and can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. The input consists of queries and keys of dimension d_k, and values of dimension d_v. The dot products of the query with all keys are computed, and then divided by $\sqrt{d_k}$. A Softmax function is then

applied to obtain the weights on the values. In practice, the attention function is computed on a set of queries simultaneously, packed together into a matrix Q. The keys and values are also packed together into matrices K and V. The matrix of outputs is:

$$\text{Attention}\ (Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The attention score as computed above determines the importance that should be given to different parts of an input sequence in the current context. In order to allow the model to jointly factor in information from different representation subspaces at different positions, multi-headed attention is used. Multiple attention scores are first calculated in parallel and then concatenated and projected using a linear transformation as:

$$\text{MultiHead}\ (Q, K, V) = \text{Concat}\ (\text{head}_1, \ldots, \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

where the projections are parameter matrix $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

3.1.3 Positional encoding

Unlike RNNs that recurrently process tokens of a sequence one by one, self-attention ditches sequential operations in favor of parallel computation. To use the sequence order information, absolute or relative positional information is injected by adding positional encoding to the input representations. Positional encodings can be either learned or fixed. In the following, a fixed positional encoding based on sine and cosine functions is described.

Suppose that the input representation $X \in \mathbb{R}^{n \times d}$ contains the d-dimensional embeddings for n tokens of a sequence. The positional encoding outputs $X+P$ using a positional embedding matrix $P \in \mathbb{R}^{n \times d}$ of the same shape, whose element on the i^{th} row and the $(2j)^{\text{th}}$ and the $(2j+1)^{\text{th}}$ column is:

$$p_{i,2j} = \sin\left(\frac{i}{10000^{2j/d}}\right)$$
$$p_{i,2j+1} = \cos\left(\frac{i}{10000^{2j/d}}\right)$$

In the positional embedding matrix *P*, rows correspond to positions within a sequence and columns represent different positional encoding dimensions.

3.2 CNN for IL properties prediction

The core of CNN is to capture local features. For ILs (represented by SMILES), local features are sliding windows composed of several symbols (representing atoms, bonds, charges, etc.), similar to N-gram[64]. The advantage of CNN is that it can automatically combine and filter N-gram features to obtain molecular information at different levels of abstraction. In this section, a CNN based neural network is built to predict IL properties with input embedding from the pre-trained IL SMILES Transformer. To be specific, the encoder part of the pre-trained IL SMILES Transformer model is utilized for generating latent representations of input ILs (as molecule fingerprints). For example, for an IL with *n* symbols in its SMILES string, the encoder produces the latent representation matrix with dimensions *n* (output size). Since different ILs have different length of SMILES, the input size of the downstream model can vary from case to case. Therefore, the Text-CNN structure[64] originally developed for sentence classification is used for the downstream IL properties prediction model as such structure can deal with distinct input lengths.

As shown in Figure. 4, after taking the encoder state of the pre-trained IL SMILES Transformer, the CNN mainly uses a one-dimensional convolutional layer and a max-over-time pooling layer[64]. The input of the CNN model is a matrix of $n \times k$, where *n* is the number of symbols in an IL SMILES and k is the dimension of the vector corresponding to each symbol. $x_i \in \mathbb{R}^k$ is used here to represent the k dimension embedding of the *i*th symbol in the IL SMILES string. On the input matrix $n \times k$, a

kernel $w \in \mathbb{R}^{hk}$ and a window $x_{i:i+h-1}$ are used to perform convolution operation to generate a feature c_i, that is to say, $c_i = f(w \cdot x_{i:i+h-1} + b)$. Here, $x_{i:i+h-1}$ represents a window of $h \times k$ formed by row *i* to row *i*+*h*-1 of the input matrix, which is formed by splicing $x_i, x_{i+1}, \ldots, x_{i+h-1}$; h denotes the number of symbols in the window; *w* is a $h \times k$-dimensional weight matrix (so the number of parameters that a filter needs to learn is *hk*); *b* is the offset parameter and *f* is a non-linear function; $w \cdot x_{i:i+h-1}$ is the dot product operation. The filter is applied to the SMILES string, moving from top to bottom one step at a time ($i = 1 \ldots n - h + 1$). For example, c_1 is obtained by convolution operation on $x_{1:h}$, c_2 is obtained by convolution operation on $x_{2:h}$, etc., and $c = [c_1, c_2, \ldots, c_{n-h+1}]$ obtained by splicing them together is the feature map of the CNN model. Each convolution operation is equivalent to a feature vector extraction. By defining different windows, different feature vectors can be extracted to form the output of the convolutional layer.

For the pooling layer, this work uses max-over-time pooling, which is to filter out the largest feature from the feature vector generated by each sliding window, and then these features are spliced together to form a fixed-length vector representation. Therefore, the input of the max-over-time pooling layer can have different time steps on each dimension, namely deal with variable input lengths from the IL SMILES Transformer encoder. After a dropout layer to deal with overfitting, the pooling result is then concatenated with other inputs (e.g., temperature and pressure if necessary) for IL properties prediction. Finally, the data go through fully connected layers and convert to the output layer that contains one neuron for prediction of IL properties. It is worth noting that as the prediction of the 11 different IL properties involved in this work are all regression problems, only one neuron is needed in the output layer here; if there are IL related classification or multiple regression problems, one can also easily set the

output layer neurons to the required number.

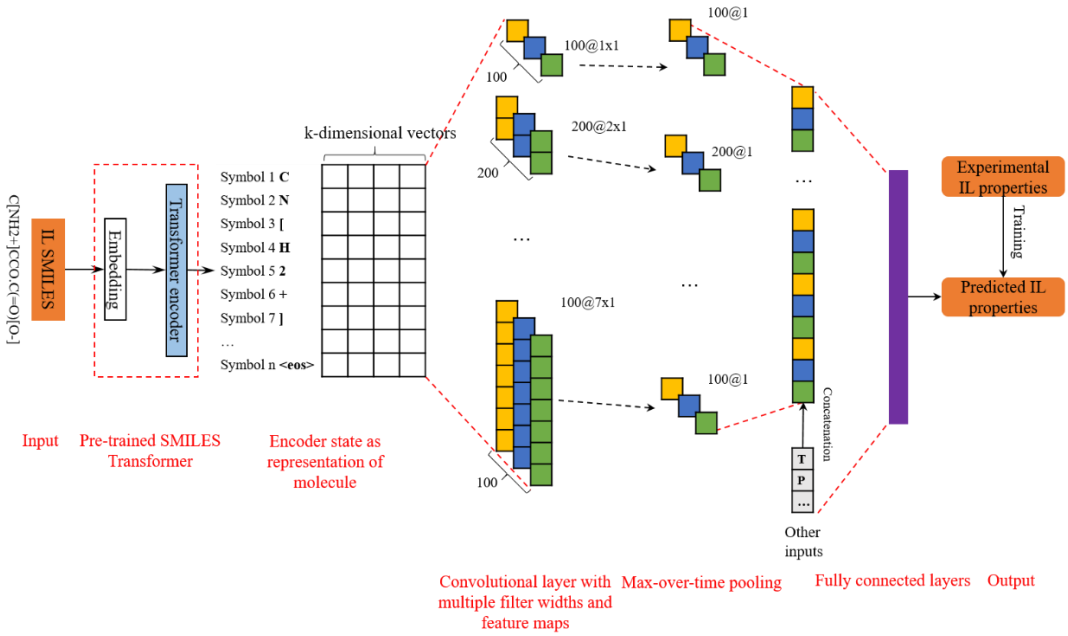


Figure 4. Architecture of the Transformer-CNN model for IL properties prediction. The model is mainly composed of two parts. The left side is the pre-trained IL SMILES Transformer, which takes the input IL SMILES and output the encoder state as representation of IL. Then, this IL representation enters the CNN model as input for IL properties prediction.

3.3 Implementation details

This work takes RDKit (*http://www.rdkit.org*) for processing IL SMILES and for the generation of canonical SMILES used in pre-training. For the implementation and training of the proposed IL SMILES Transformer-CNN model, the MXNet library[65] with GPU acceleration (on a single RTX2080Ti and CUDA 10.1) and GluonNLP toolkit[66] is employed.

3.3.1 Pre-training IL SMILES Transformer

To use the SMILES representations as the input and output of the Transformer model, the SMILES strings are tokenized into characters and then encoded in a one-hot vector representation (values are zero everywhere except the position of the current token that is set to one). In this work, the character-level tokenization[40,42] where every single character appears in SMILES is tokenized separately is used. The vocabulary is

built by using the MXNet library[55] and the GluonNLP toolkit[56], which contain all 71 possible characters in the SMILES of 9,434,070 molecules in the pre-training dataset. The characters as well as their indexes in the vocabulary are: [('<unk>', 0), ('<pad>', 1), ('<bos>', 2), ('<eos>', 3), ('c', 4), ('C', 5), ('(', 6), (')', 7), ('O', 8), ('1', 9), ('=', 10), ('N', 11), ('[', 12), (']', 13), ('2', 14), ('-', 15), ('+', 16), ('n', 17), ('3', 18), ('H', 19), ('@', 20), ('F', 21), ('S', 22), ('.', 23), ('l', 24), ('/', 25), ('4', 26), ('s', 27), ('B', 28), ('#', 29), ('r', 30), ('o', 31), ('\', 32), ('P', 33), ('I', 34), ('5', 35), ('i', 36), ('a', 37), ('K', 38), ('e', 39), ('Z', 40), ('L', 41), ('U', 42), ('Y', 43), ('6', 44), ('u', 45), ('R', 46), ('T', 47), ('M', 48), ('A', 49), ('g', 50), ('t', 51), ('b', 52), ('W', 53), ('d', 54), ('f', 55), ('V', 56), ('h', 57), ('7', 58), ('G', 59), ('p', 60), ('8', 61), ('m', 62), ('9', 63), ('E', 64), ('D', 65), ('%', 66), ('y', 67), ('0', 68), ('*', 69), ('X', 70), ('k', 71)]. The meaning of each character in SMILES can be found in the original literature[35]. The non-canonical SMILES and canonical SMILES share the same vocabulary in this work.

Similar to NMT tasks, the SMILES Transformer model is trained on a translation task of non-canonical SMILES to canonical SMILES. Considering the much smaller size of the vocabulary used here (71 different characters for all involved SMILES strings) than common NMT tasks (several thousands of different words) and the less complicacy of SMILES canonicalization task, the numbers of Transformer block, heads in multi-head attention and units for the output is decreased from 6, 8 and 512 to 3, 4 and 128, respectively, with reference to the original paper[39]. A dropout rate of 0.1 (the same as the original paper) is used for model regularization. The Transformer model is trained for 10 epochs by Adam optimizer[57] with a base learning rate of 0.001. The learning rate is multiplied by a factor of 0.5 for each epoch after four epochs of training.

The Masked Softmax Cross Entropy Loss[49] is used as the loss function for pre-training and is implemented by the *gluonnlp.loss.MaskedSoftmaxCELoss*() function. To

validate the pre-training model, 100,000 and 100,000 SMILES are randomly split from the pre-training dataset after augmentation (contains 94,340,700 SMILES) to form the validation set and test set, respectively, while the rest of the pre-training dataset is kept as model training set.

3.3.2 Fine-tuning using CNN for IL properties prediction

To train the Transformer-CNN model for IL properties prediction, the 11 IL property datasets are utilized to train 11 sets of weights of the same CNN structure, while the weights of the IL SMILES Transformer encoder are pre-trained on the PubChem dataset and frozen in all the 11 models. For the model development, 10-fold cross-validations (CVs) on each of the 11 IL property datasets are first performed to determine the model hyper parameters (dropout rate and the size of fully connected layers). The mean squared error (MSE) function (L2 loss) is used as the loss function for all the 11 IL properties. Optimal values of the hyper parameters are obtained by performing an extensive grid search (output size of fully connected layer: 128, 256, 512, 1024; dropout rate: 0.05, 0.1, 0.3, 0.5, 0.7).

4 RESULTS AND DISCUSSION

4.1 Performance of the IL SMILES Transformer model

To assess the performance of the IL SMILES Transformer model, two measures that capture different aspects of the model performance are considered. One is the BLEU (Bilingual Evaluation Understudy) score[67], which is a standard metric used for the evaluation of a given translation (the output canonical SMILES of the SMILES Transformer) against the reference translation (the original canonical SMILES). As shown in Figure 5, both the validation and test BLEU scores evolve with the training epochs and become stable at nearly 99.5 after the eighth epoch. The second measure is the translation accuracy computed by the perfect matches between the predicted and

the actual canonical SMILES. From Table 2, over 94% correctly canonicalized SMILES are achieved by the model for both the validation and test set. Even for molecules with stereo- or cis/trans conformers, the translation accuracy for both sets is still higher than 84% and 90%, respectively. Moreover, although 6% of the predicted SMILES is not a perfect match with the actual canonical SMILES, the average similarity (calculated using the SequenceMatcher routine in python that matches the longest continuous matching sub-sequence) between these 6% predicted SMILES and their canonical SMILES is 89.75% on the test set and 89.94% on the validation set. This similarity value shows that even 6% of the predicted SMILES do not match perfectly, they are still very close to the actual canonical SMILES. These results well demonstrate the high performance of the pre-trained SMILES Transformer model in capturing molecular features from IL SMILES.

Table 2. Number of perfect matches between the predicted and the actual canonical SMILES on the test and validation set.

Strings	All	Correctly canonicalized
All on test set	100,000	94,329 (94.33%)
All on validation set	100,000	94,243 (94.24%)
Stereo (with @) on test set	14,777	12,668 (85.73%)
Stereo (with @) on validation set	14,945	12,679 (84.84%)
Cis/trans (with / or \) on test set	9,226	8,426 (91.33%)
Cis/trans (with / or \) on validation set	4,123	3,750 (90.95%)

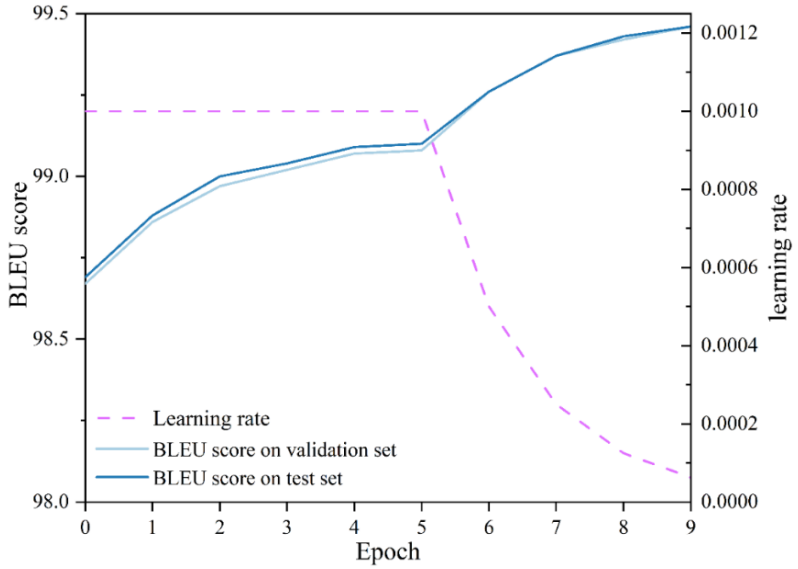


Figure. 5. Learning curves: learning rate (axes bottom and right) and BLEU score (axes bottom and left) on the validation and test set.

4.2 Performance on IL properties modeling

To prove the performance of the proposed IL SMILES Transformer-CNN model for predicting IL properties, state-of-the-art models in recent literature[24–26,28–30,58] are chosen for comparison. For the sake of a fair comparison, this work trains all the IL SMILES Transformer-CNN models on the same IL property databases as in the corresponding references. Moreover, two different test set split strategies are also compared for IL properties related to temperature and/or pressure. One is the direct split of all data points as used in the references and the other is the more rigorous split by different ILs.

The comparative results for the 11 IL properties are summarized in Table 3. As can be seen, for the properties only related to the molecular structure of ILs (namely melting point, glass transition temperature, thermal decomposition temperature, and cytotoxicity towards the leukemia rat cell line IPC-81), the prediction error (mean absolute error, MAE) of the models proposed in this paper is overall lower than that of the reference models reported in literature. These results indicate that the pretrained

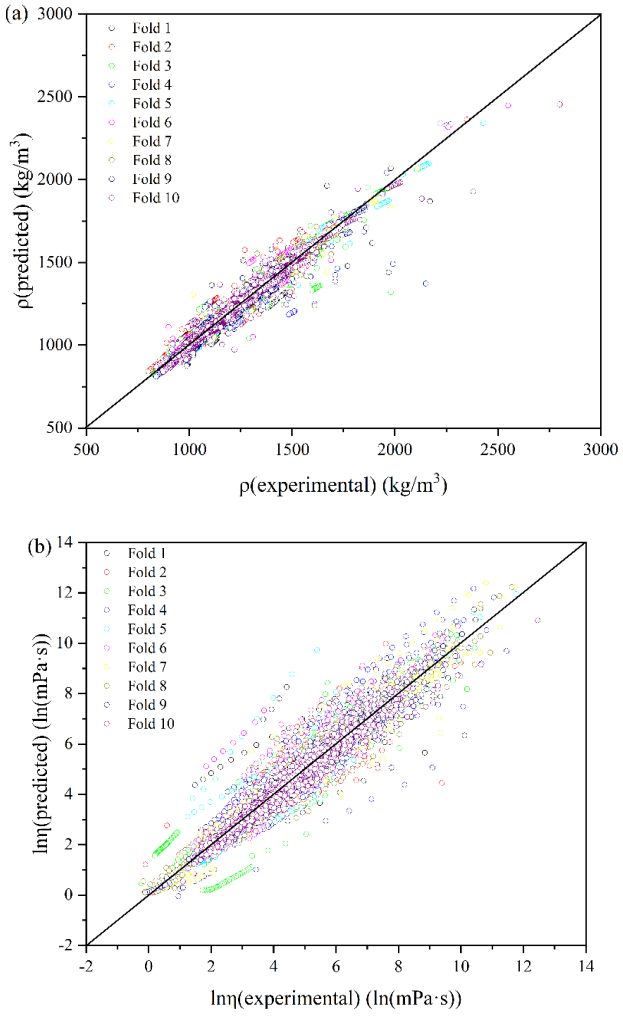
SMILES Transformer can extract the structural features of ILs better than the various descriptors used in literature, especially when the database of IL properties is small (e.g., toxicity and thermal conductivity). In this sense, the adopted pre-training and fine-tuning method in this paper is a good way to solve the problem of insufficient experimental data in predictive property modelling. For properties also related to temperature and/or pressure (namely heat capacity, refractive index, density, viscosity, surface tension, CO_2 solubility, and thermal conductivity), the Transformer-CNN models outperform all the reference models on the test set split by data points. Notably, the Transformer-CNN models still have comparable and even lower MAE (for the properties of density, viscosity, and thermal conductivity) on the test set rigorously split by ILs than the reference models do on the test set split non-rigorously by data points. This comparison proves that it can achieve higher prediction accuracy by using CNN structure to handle different types of input than the non-neural network reference methods.

It should be mentioned that some of the above references have also tried to use neural network based methods in their model development; however, the neural network methods constructed in these references cannot achieve better prediction accuracy compared with the models listed in Table 3. The reason is that the database size of most of such IL properties database is not large enough to train a neural network model with a high enough prediction accuracy, leading to the selection of other statistical ML methods instead. The results here prove that pre-training on unlabeled datasets can solve the problem of insufficient labeled data, and the predictive model constructed thereon by neural network based methods can obtain higher accuracy than statistical ML methods and can better handle multiple types of input features.

Table 3. Comparison of the models reported in literature with the Transformer-CNN method proposed in this work for estimating the 11 IL properties.

Property	Number of data points	Number of ILs	Descriptor	Method	Test MAE (split by data points)	Test MAE (split by ILs)	Source
$T_m(K)$	2212	2212	ECFP4 and CM	KRR	\	29.78	Low et.al.[24]
				T-CNN	\	**11.15**	This work
$T_g(℃)$	609	609	charge distributions and geometrical indices	Cubist	\	12	Venkatraman et.al.[29]
				T-CNN	\	**6.77**	This work
$T_d(℃)$	1223	1223	charge distributions and geometrical indices	RF	\	25	Venkatraman et.al.[29]
				T-CNN	\	**19.19**	This work
$\ln(\eta)$	15368	1964	group contributions	LSSVM	0.42	\	Paduszyński[28]
				T-CNN	**0.17**	**0.35**	This work
ρ	31167	2257	group contributions	LSSVM	29.76	\	Paduszyński[25]
				T-CNN	**12.31**	**16.46**	This work
$\ln(C_p)$	9083	236	charge distributions and geometrical indices	GBM	0.19	\	Venkatraman et.al.[29]
				T-CNN	**0.18**	0.28	This work
γ	2972	331	charge distributions and geometrical indices	GBM	0.0027	\	Venkatraman et.al.[29]
				T-CNN	**0.0014**	0.0030	This work
n_D	3009	464	charge distributions and geometrical indices	GBM	0.011	\	Venkatraman et.al.[29]
				T-CNN	**0.0047**	0.015	This work
x_{CO_2}	10116	124	group contributions	SVM	0.024	\	Song et.al.[26]
				T-CNN	**0.022**	0.057	This work
$log_{10}(EC_{50})$	326	326	structural descriptors	SVM	\	0.1935	Wang et.al.[58]
				T-CNN	\	**0.1126**	This work
λ	454	73	charge distributions and geometrical indices	GBM	0.009	\	Venkatraman et.al.[30]
				T-CNN	**0.0034**	**0.0061**	This work

To more vividly show the predictive performance of the Transformer-CNN models proposed in this paper, the density, viscosity, and cytotoxicity of ILs are taken as examples to inspect the test results in more detail. As seen in Figure 6, the test set points of each fold in the 10-fold cross-validation for the density, viscosity, and cytotoxicity are distributed almost evenly in a close region around the diagonal. These examples prove that the CNN model can well predict different types of IL properties by fine-tuning on the corresponding IL properties dataset based on the IL features obtained by the pre-trained Transformer encoder. The detailed 10-fold-cross-validation results for all 11 IL properties can be seen in Table S1-S18 (Supporting Information), which also agree with this conclusion.



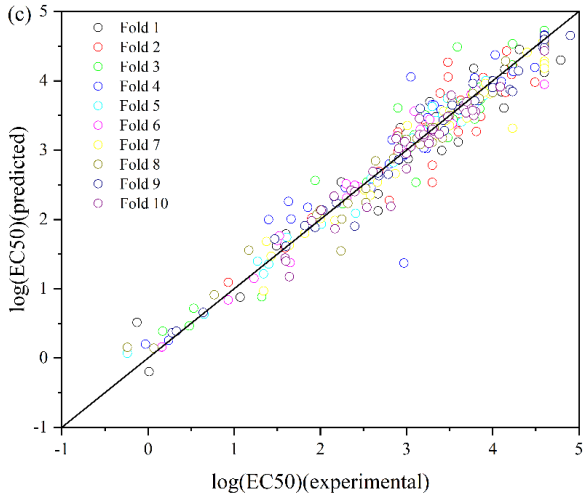


Figure. 6. Parity plots of predicted versus experimental data with the proposed Transformer-CNN method. Each color representants the test set results of one fold in 10-fold-cross validation. (a) Density, (b) Viscosity, (c) Cytotoxicity.

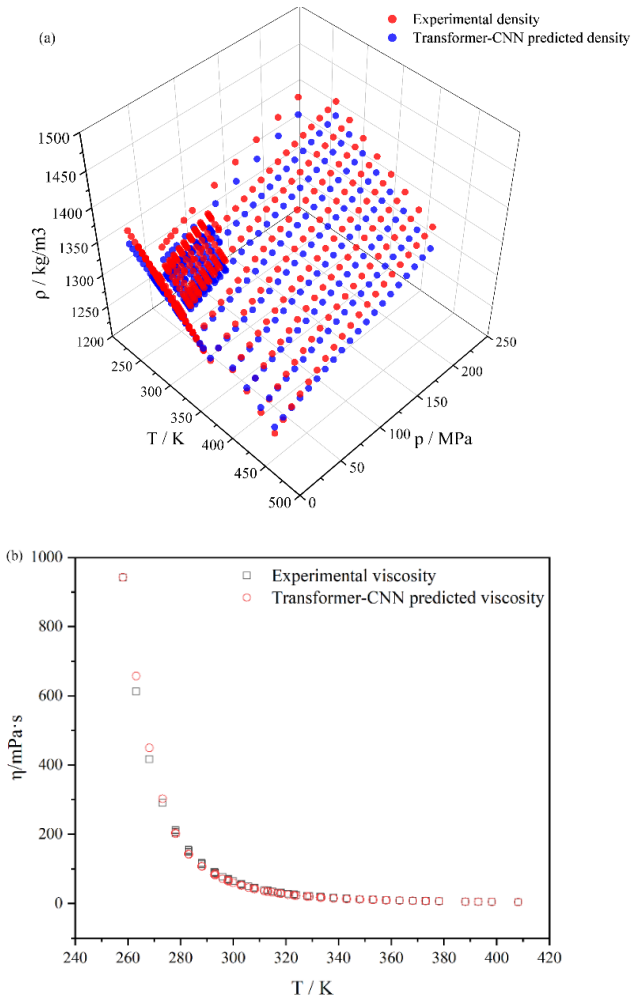


Figure. 7. Predicted versus experimental density (a) and viscosity (b) of 1-hexyl-3-methylimidazolium bistriflamide as a function of temperature and/or pressure.

To further demonstrate that the Transformer-CNN model can well handle different inputs for property prediction, 1-hexyl-3-methylimidazolium bistriflamide ([C6C1Im][NTf2]) is selected as a representative to examine its predicted η-T and ρ-T-P relationship. As seen in Figure 7a, the predicted viscosity by the Transformer-CNN model well resembles the experimental data over a wide range of temperature. As for the density of [C6C1Im][NTf2], the Transformer-CNN model also provides very satisfactory prediction as compared to the experimental data over a wide range of temperature and pressure (up to T = 450 K and P = 200 MPa, respectively). These two examples clearly show that different types of inputs namely the IL molecular structure and temperature and/or pressure are properly handled by the Transformer-CNN model. It is worth mentioning that very few previously reported ML models have scrutinized whether the temperature and/or pressure dependence of such IL properties could be correctly captured.

To provide a more detailed insight into the performance of the Transformer-CNN model, IL density is again selected as a representative property to analyze the model predictions for each possible combination of cationic and anionic families. The corresponding AARE values are obtained by averaging the test set results in 10-fold cross-validation. As shown in Figure 8, the AAREs for most of the involved anionic and cationic combinations are below 5%, which again prove that the model has a high prediction accuracy for IL density. Moreover, such a prediction accuracy model is found to be dependent on the moieties forming IL. For instance, the AAREs for the imidazolium-based ILs are all lower than 5%, with 13 of the 15 anionic families below 3%; low AAREs are also observed for carboxylates ILs, except when the paired cationic moiety is guanidinium. The highest AARE of 15.4% is obtained for cyclic sulfonium cations combined with common inorganics, as such a combination only appears once

in the entire dataset (the density prediction in this case is fully extrapolated in cross-validation). To wrap up, the detailed analyses of the density prediction well demonstrate that the Transformer-CNN model could reasonably predict IL properties for different IL families.

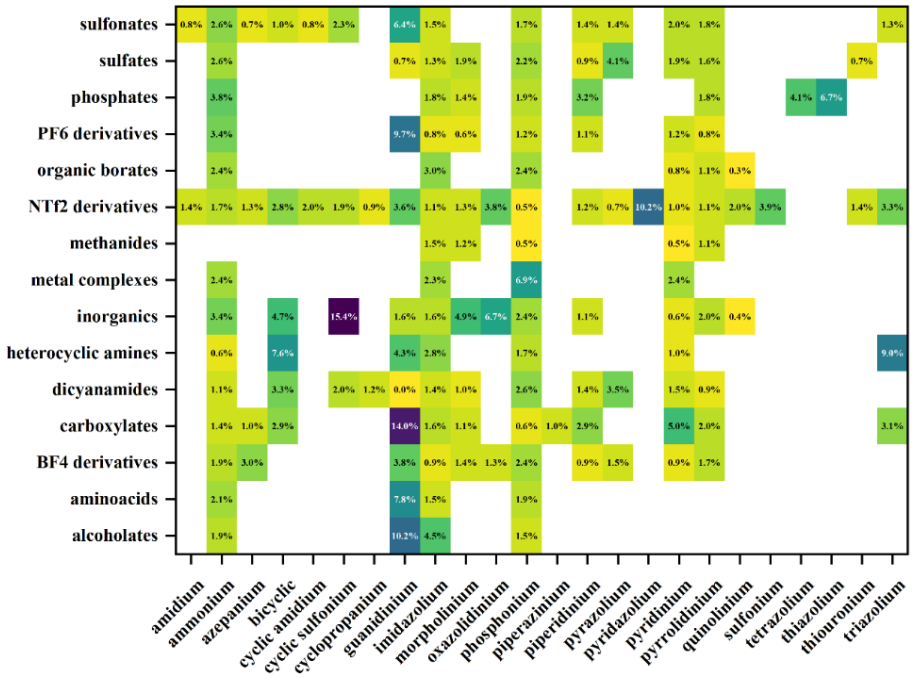


Figure. 8. Average absolute relative errors (AAREs) between Transformer-CNN predicted and experimental density for different combinations of cationic and anionic families of ILs. Empty cell means that the experimental data have been not available yet.

5. Model application case study: CO_2 absorbent screening

From the IL SMILES Transformer-CNN models obtained above, the 11 IL properties can be reliably and quickly predicted, allowing for many applications such as the high-throughput IL screening toward different processes. Herein, the screening of ILs as CO_2 absorbent is presented as an illustrative case study.

When screening ILs for CO_2 capture processes, the following IL properties are of great importance: (1) The capacity of IL to absorb CO_2 can be evaluated by the gas solubility in ILs at the desired absorption temperature, while the desorption performance of IL can be estimated by the difference in the CO_2 solubility at the desired

absorption temperature and desorption temperature, respectively. (2) The melting point, viscosity, thermal decomposition temperature, toxicity, and heat capacity of ILs should be considered as constraints because all these properties determine the feasibility and suitability of ILs as absorbent.[68–70] To be specific, the melting point limits the lowest absorption temperature of ILs as liquid CO_2 absorbents; the thermal decomposition temperature limits the highest temperature for CO_2 desorption; the energy consumption of solvent regeneration can be assessed from the heat capacity of IL; the toxicity is a key factor related to the potential EHS impacts of ILs. All the above properties can be calculated by the Transformer-CNN models developed in this work.

In this case study, a virtual library of 8,333,096 (219,216 cations combined with 38 anions) synthetically feasible ILs as suggested by Venkatraman et al.[30] is used as the initial candidate database. By using the Transformer-CNN models, the CO_2 solubility of ILs at 298 K and 328 K (P = 1 bar) are calculated for evaluating the absorption and desorption performance of ILs; the heat capacity (C_p (J/K)), and viscosity (η (mPa·s)), under 1 bar and 298 K, as well as the melting point ($T_m\,(K)$), cytotoxicity ($log_{10}(EC_{50})$), and thermal decomposition temperature ($T_d\,(K)$) are also predicted. As the calculation speed of the Transformer-CNN model for IL properties is very fast, a database of the seven properties for all the 8,333,096 candidate ILs is obtained in around 14 hours (2 hours per property) on a laptop equipped with a RTX3070 GPU. Apply the following constraints namely T_m< 298 K, T_d> 150 °C, $log_{10}(EC_{50})$> 3, and η< 100 mPa·s, a high-throughput screening over the entire IL database is performed, which retains 18 ILs meeting all the four constraints (as illustrated in Figure 9, see detailed information of these ILs in the Supporting information Table S19). Among them, 8 ILs are basically located on the pseudo pareto front of all the candidate ILs in terms of the potential absorption and desorption performance. It should be noted that

the four ILs in the lower right corner of Figure 9 are excluded due to too small solubility of CO_2 at the absorption temperature. Of course, the selected constraints could be properly relaxed if one would like to keep a larger set of ILs for further study.

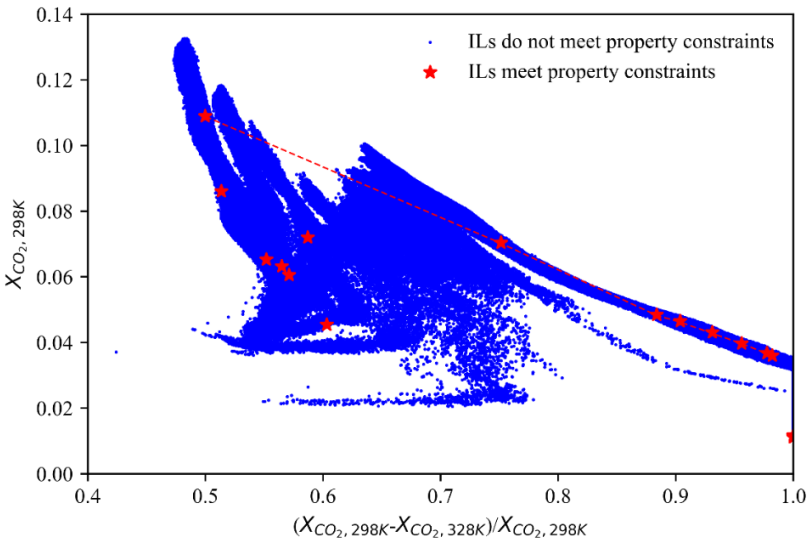


Figure 9. Illustration of the high-throughput IL screening in terms of the potential absorption and desorption performance.

The predicted properties of the 8 retained ILs are listed in Table 4 and their molecular structures are shown in Figure 10. These 8 ILs are highly worth investigating by experiment in future studies as they are survivals from 8,333,096 candidates. It is worth mentioning that this case study is for the first time that such a huge database of ILs is considered for a high-throughput solvent screening toward a specific process, which benefits from both the high prediction accuracy and fast calculation speed of the proposed Transformer-CNN models.

Table 4. Predicted properties of the 8 retained ILs from high-throughput IL screening.

IL ID	C_p(J/K)	$T_m(K)$	$log_{10}(EC_{50})$	$X_{CO_2,298K}$	$X_{CO_2,328K}$	η(mPa·s)	$T_d(K)$
3219685	539.04	296.92	3.48	0.0704	0.0175	76.85	212.54
3252213	630.81	292.92	3.41	0.0360	0.0006	94.40	263.32
3257267	487.29	285.92	3.14	0.0485	0.0056	48.70	169.71
3257305	541.49	262.31	3.11	0.0465	0.0045	59.35	185.38
3257343	577.75	262.85	3.21	0.0367	0.0008	74.32	183.15
3257533	623.95	294.67	3.40	0.0431	0.0030	74.57	171.23
3258445	521.51	284.80	3.40	0.0397	0.0017	52.46	201.99

3259535	803.36	275.87	3.21	0.1089	0.0545	86.40	191.64

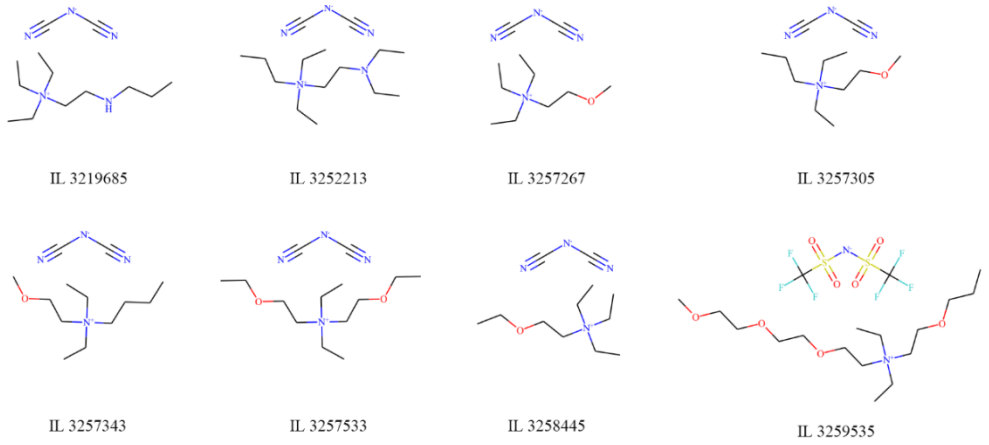


Figure 10. Molecular structures of the 8 retained ILs from high-throughput IL screening.

6 Conclusion

In this work, a novel pre-training and fine-tuning two-stage framework is proposed to better exploit numerous unlabeled molecular data and overcome problems caused by limited training data in IL properties modelling. The pre-trained SMILES Transformer utilizes the power of unlabeled molecular data through a large-scale (9,434,070 molecules) pre-training through a translation task of non-canonical SMILES to canonical SMILES. The labeled IL properties datasets could be easily fine-tuned using CNN architecture with the pre-trained SMILES Transformer as IL feature extractor. It is found that the proposed neural networks can also handle multiple types of input features very well. In experiments on 11 diverse benchmark datasets, the proposed Transformer-CNN method surpasses various state-of-the-art methods reported in literature. Moreover, the prediction of IL properties by the proposed Transformer-CNN model is very computationally efficient, which enables high-throughput IL screening toward a specific task. As a case study, a large virtual library of 8,333,096 synthetically feasible ILs is used for CO_2 absorbent screening. Seven IL properties closely related to CO_2 capture process performance are calculated using the proposed Transformer-CNN

model for all the 8,333,096 ILs, finally retaining 8 ILs that meet all desired constraints.

The model proposed in this work provides a one-stop solution for IL researchers, that is to use the same Transformer-CNN model structure to predict all IL properties with good prediction accuracy. With the support of the reported models, high-throughput IL screening could be applied to other chemical engineering processes in future work. Besides, the Transformer-CNN approach could also be extended to develop other important molecular property models, which are currently limited by small available databases.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (NSFC) under the grants of 22278134, 21CAA01709, and 22208098. Guzhong Chen acknowledges the financial support of the China Scholarship Council (CSC) for his joint Ph.D. program (No. 202106740022) with the Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany.

REFERENCES

1. Rogers RD, Seddon KR. Ionic Liquids--Solvents of the Future? *Science*. 2003;302(5646):792-793.
2. Bezold F, Roehrer S, Minceva M. Ionic Liquids as Modifying Agents for Protein Separation in Centrifugal Partition Chromatography. *Chem Eng Technol*. 2019;42(2):474-482.
3. Niedermeyer H, P. Hallett J, J. Villar-Garcia I, A. Hunt P, Welton T. Mixtures of ionic liquids. *Chem Soc Rev*. 2012;41(23):7780-7802.
4. Plechkova NV, Seddon KR. Applications of ionic liquids in the chemical industry. *Chem Soc Rev*. 2008;37(1):123-150.
5. Izgorodina EI. Towards large-scale, fully ab initio calculations of ionic liquids. *Phys Chem Chem Phys*. 2011;13(10):4189.
6. Maia FM, Tsivintzelis I, Rodriguez O, Macedo EA, Kontogeorgis GM. Equation of state modelling of systems with ionic liquids: Literature review and application with the Cubic Plus Association (CPA) model. *Fluid Phase Equilibria*. 2012;332:128-143.

7. Hosseini SM, Mulero A, Alavianmehr MM. Predictive methods and semi-classical Equations of State for pure ionic liquids: A review. *J Chem Thermodyn*. 2019;130:47-94.
8. Coutinho JAP, Carvalho PJ, Oliveira NMC. Predictive methods for the estimation of thermophysical properties of ionic liquids. *RSC Adv*. 2012;2(19):7322.
9. Das RN, Roy K. Advances in QSPR/QSTR models of ionic liquids for the design of greener solvents of the future. *Mol Divers*. 2013;17(1):151-196.
10. Klamt A, Eckert F. COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids. *Fluid Phase Equilibria*. 2000;172(1):43-72.
11. Paduszyński K. An overview of the performance of the COSMO-RS approach in predicting the activity coefficients of molecular solutes in ionic liquids and derived properties at infinite dilution. *Phys Chem Chem Phys*. 2017;19(19):11835-11850.
12. Gerlach T, Müller S, Smirnova I. Development of a COSMO-RS based model for the calculation of phase equilibria in electrolyte systems. *AIChE J*. 2018;64(1):272-285.
13. Liu X, Zhou T, Zhang X, et al. Application of COSMO-RS and UNIFAC for ionic liquids based gas separation. *Chem Eng Sci*. 2018;192:816-828.
14. Zeng S, Zhang X, Bai L, et al. Ionic-Liquid-Based CO_2 Capture Systems: Structure, Interaction and Process. *Chem Rev*. 2017;117(14):9625-9673.
15. Casas A, Palomar J, Alonso MV, Oliet M, Omar S, Rodriguez F. Comparison of lignin and cellulose solubilities in ionic liquids by COSMO-RS analysis and experimental validation. *Ind Crops Prod*. 2012;37(1):155-163.
16. Izgorodina EI, Seeger ZL, Scarborough DLA, Tan SYS. Quantum Chemical Methods for the Prediction of Energetic, Physical, and Spectroscopic Properties of Ionic Liquids. *Chem Rev*. 2017;117(10):6696-6754.
17. Lee JH, Shin J, Realff MJ. Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Comput Chem Eng*. 2018;114:111-121.
18. Su Y, Wang Z, Jin S, et al. An architecture of deep learning in QSPR modeling for the prediction of critical properties using molecular signatures. *AIChE J*. 2019;65(9):e16678.
19. Venkatasubramanian V. The promise of artificial intelligence in chemical engineering: Is it here, finally? *AIChE J*. 2019;65(2):466-478.

20. Sivaram A, Venkatasubramanian V. XAI-MEG: Combining symbolic AI and machine learning to generate first-principles models and causal explanations. *AIChE J*. 2022;68(6):e17687.

21. Chiang LH, Braun B, Wang Z, Castillo I. Towards artificial intelligence at scale in the chemical industry. *AIChE J*. 2022;68(6):e17644.

22. Dong Q, Muzny CD, Kazakov A, et al. ILThermo: A Free-Access Web Database for Thermodynamic Properties of Ionic Liquids. *J Chem Eng Data*. 2007;52(4):1151-1159.

23. Ding Y, Chen M, Guo C, Zhang P, Wang J. Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties. *J Mol Liq*. 2021;326:115212.

24. Low K, Kobayashi R, Izgorodina EI. The effect of descriptor choice in machine learning models for ionic liquid melting point prediction. *J Chem Phys*. 2020;153(10):104101.

25. Paduszyński K. Extensive Databases and Group Contribution QSPRs of Ionic Liquids Properties. 1. Density. *Ind Eng Chem Res*. 2019;58(13):5322-5338.

26. Song Z, Shi H, Zhang X, Zhou T. Prediction of CO2 solubility in ionic liquids using machine learning methods. *Chem Eng Sci*. 2020;223:115752.

27. Peng D, Picchioni F. Prediction of toxicity of Ionic Liquids based on GC-COSMO method. *J Hazard Mater*. 2020;398:122964.

28. Paduszyński K. Extensive Databases and Group Contribution QSPRs of Ionic Liquids Properties. 2. Viscosity. *Ind Eng Chem Res*. 2019;58(36):17049-17066.

29. Venkatraman V, Evjen S, Lethesh KC, Raj JJ, Knuutila HK, Fiksdahl A. Rapid, comprehensive screening of ionic liquids towards sustainable applications. *Sustain Energy Fuels*. 2019;3(10):2798-2808.

30. Venkatraman V, Evjen S, Chellappan Lethesh K. The Ionic Liquid Property Explorer: An Extensive Library of Task-Specific Solvents. *Data*. 2019;4(2):88.

31. Paduszyński K. Extensive Databases and Group Contribution QSPRs of Ionic Liquids Properties. 1. Density. *Ind Eng Chem Res*. 2019;58(13):5322-5338.

32. Paduszyński K. Extensive Databases and Group Contribution QSPRs of Ionic Liquids Properties. 2. Viscosity. *Ind Eng Chem Res*. 2019;58(36):17049-17066.

33. Zhu P, Kang X, Zhao Y, Latif U, Zhang H. Predicting the Toxicity of Ionic Liquids toward Acetylcholinesterase Enzymes Using Novel QSAR Models. *Int J Mol Sci*. 2019;20(9):2186.

34. Honda S, Shi S, Ueda HR. SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery; 2019; *ArXiv:191104738*

35. Sivaram A, Das L, Venkatasubramanian V. Hidden representations in deep neural networks: Part 1. Classification problems. *Comput Chem Eng*. 2020;134:106669.

36. Das L, Sivaram A, Venkatasubramanian V. Hidden representations in deep neural networks: Part 2. Regression problems. *Comput Chem Eng*. 2020;139:106895.

37. Wen H, Su Y, Wang Z, et al. A systematic modeling methodology of deep neural network-based structure-property relationship for rapid and reliable prediction on flashpoints. *AIChE J*. 2022;68(1):e17402.

38. Xing Y, Dong Y, Goergakis C, et al. Automatic data-driven stoichiometry identification and kinetic modeling framework for homogeneous organic reactions. *AIChE J*. 2022;68(7):e17713.

39. Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci*. 2018;4(2):268-276.

40. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des*. 2016;30(8):595-608.

41. Xu Z, Wang S, Zhu F, Huang J. Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*. ACM; 2017:285-294.

42. Chen G, Song Z, Qi Z, Sundmacher K. Neural recommender system for the activity coefficient prediction and UNIFAC model extension of ionic liquid-solute systems. *AIChE J*. 2021;67(4):e17171.

43. Hayer N, Jirasek F, Hasse H. Prediction of Henry's law constants by matrix completion. *AIChE J*. 2022;68(9):e17753.

44. Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A. Advances in Pre-Training Distributed Word Representations. Published online December 26, 2017.

45. Dai AM, Le QV. Semi-supervised Sequence Learning. In: *Advances in Neural Information*

Processing Systems. Vol 28. Curran Associates, Inc.; 2015.

46. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Model*. 1988;28(1):31-36.
47. Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. *J Chem Inf Comput Sci*. 1989;29(2):97-101.
48. Winter R, Montanari F, Noé F, et al. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem Sci*. 2019;10(6):1692-1701.
49. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. In: *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates, Inc.; 2017.
50. Karpov P, Godin G, Tetko IV. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J Cheminformatics*. 2020;12(1):17.
51. Wang S, Guo Y, Wang Y, Sun H, Huang J. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM; 2019:429-436.
52. Mann V, Venkatasubramanian V. Predicting chemical reaction outcomes: A grammar ontology-based transformer framework. *AIChE J*. 2021;67(3):e17190.
53. Schwaller P, Laino T, Gaudin T, et al. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent Sci*. 2019;5(9):1572-1583.
54. Chen G, Song Z, Qi Z. Transformer-convolutional neural network for surface charge density profile prediction: Enabling high-throughput solvent screening with COSMO-SAC. *Chem Eng Sci*. 2021;246:117002.
55. Kim S, Thiessen PA, Bolton EE, et al. PubChem Substance and Compound databases. *Nucleic Acids Res*. 2016;44(D1):D1202-D1213.
56. Tetko IV, Karpov P, Bruno E, Kimber TB, Godin G. Augmentation Is What You Need! In: Tetko IV, Kůrková V, Karpov P, Theis F, eds. *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*. Lecture Notes in Computer Science. Springer International Publishing; 2019:831-835.

57. Bjerrum EJ. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules; 2017; *ArXiv:170307076*

58. Wang Z, Song Z, Zhou T. Machine Learning for Ionic Liquid Toxicity Prediction. *Processes*. 2021;9(1):65.

59. Sutskever I, Vinyals O, Le QV. Sequence to Sequence Learning with Neural Networks. In: *Advances in Neural Information Processing Systems*. Vol 27. Curran Associates, Inc.; 2014.

60. Sattarov B, Baskin II, Horvath D, Marcou G, Bjerrum EJ, Varnek A. De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping. *J Chem Inf Model*. 2019;59(3):1182-1196.

61. Liu B, Ramsundar B, Kawthekar P, et al. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent Sci*. 2017;3(10):1103-1113.

62. Nam J, Kim J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions; 2016; *ArXiv:161209529*

63. Cho K, van Merrienboer B, Bahdanau D, Bengio Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches; 2014; *ArXiv:14091259*

64. Kim Y. Convolutional Neural Networks for Sentence Classification; 2014; *ArXiv:1408.5882*

65. Chen T, Li M, Li Y, et al. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems; 2015; *ArXiv:151201274*

66. Guo J, He H, He T, et al. GluonCV and GluonNLP: Deep Learning in Computer Vision and Natural Language Processing; 2020; *ArXiv:1907.04433*

67. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2002:311-318.

68. Zhao Y, Gani R, Afzal RM, et al. Ionic liquids for absorption and separation of gases: An extensive database and a systematic screening method. *AIChE J*. 2017;63(4):1353-1367.

69. Zheng S, Zeng S, Li Y, et al. State of the art of ionic liquid-modified adsorbents for CO2 capture and separation. *AIChE J*. 2022;68(2):e17500.

70. Taheri M, Zhu R, Yu G, Lei Z. Ionic liquid screening for CO2 capture and H2S removal from gases: The syngas purification case. *Chem Eng Sci*. 2021;230:116199.