iupacGPT: IUPAC-based large-scale molecular pre-trained model for property prediction and molecule generation

Jiashun Mao\textsuperscript{1,2}, Jianmin Wang\textsuperscript{1,2}, Kwang-Hwi Cho\textsuperscript{3,*}, Kyoung Tai No\textsuperscript{1,2,*}

1. The Interdisciplinary Graduate Program in Integrative Biotechnology and Translational Medicine, Yonsei University, Incheon 21983, Republic of Korea
2. Bioinformatics and Molecular Design Research Center (BMDRC), Incheon 21983, Korea
3. School of Systems Biomedical Science, Soongsil University, Seoul, Republic of Korea
* To whom correspondence should be addressed.

Abstract

The IUPAC (International Union of Pure and Applied Chemistry) nomenclature is a globally recognized unique naming system which assigns names to chemical compounds. As a form of molecular representation closest to natural language, it allows to estimate molecular data in a large-scale pre-trained paradigm by employing machine learning approaches for natural language processing (NLP). Although, SMILES is currently popular molecular representation used by most generative models, different molecular representation is suitable for different scenarios, and considering the advantages of IUPAC in terms of readability, it becomes meaningful to explore the difference of these two different molecular representations for molecular generation and regression/classification tasks. In this paper, we attempt to adapt the capabilities of transformer to a large IUPAC corpus by constructing a GPT-2-like language model named iupacGPT. For each task in addition to the molecular generation, we freeze model parameters and attach trainable lightweight networks to fine tune. The results show that pre-trained iupacGPT can capture general knowledge that can be successfully transferred to the downstream tasks such as molecule generation and binary classification and property regression prediction. What’s more, with a same setup, iupacGPT outperforms the model smilesGPT in term of the downstream tasks. Overall, transformer-like language models pretrained on IUPAC corpora are promising
alternatives that obtain more intuitive in terms of interpretability and semantic level than on SMILES corpora, and scale well with the pretraining data size.

Key word: IUPAC-based large-scale model; molecule generation; GPT-2; pre-training; molecular representation

1. Introduction

The IUPAC (International Union of Pure and Applied Chemistry) nomenclature is a globally recognized unique naming system which assigns names to chemical compounds. For instance, a compound, its IUPAC name is N-benzyl-5-(2-methylbutan-2-yl)-1,2-thiazol-3-amine. In contrast, SMILES (Simplified Molecular Input Line Entry System) is another naming system, which allocates symbolic representation to compounds, known as SMILES strings. For example, the SMILES string for the aforementioned compound is CCC(C)(C) c1cc (NCc2ccccc2) ns1. Its Canonical SMILES symbol is CCC(C)(C) C1=CC(=NS1) NCC2=CC=CC=C2. The SMILES strings of compounds are often intermingled and ambiguous. In other words, the same compound may obtain multiple SMILES strings before canonized. However, the IUPAC name for any compound is utterly unambiguous, that is, IUPAC harmonized chemical names globally, and established the nomenclature of organic chemistry that documented instruction on the unambiguous names for all compounds [1].

Currently, SMILES strings are the most popular alternative representations for organic compounds, and it is mainly emanated from chemistry literature to facilitate computer-assisted processing of chemical information. Wherein, it provides atom-based and structure-level information of compounds. But in terms of uniqueness and level of abstraction, the IUPAC name has significant advantages than SMILES strings. The IUPAC name not only includes atom-based, chemical group (fragment-level) and spatial-oriented information for an atom in the compound, but also looks like natural language that is readable and interpretable for chemists. Nevertheless, it is very hard to perform the naming process accurately to create IUPAC names manually, since
it involves a complex algorithm. Moreover, chemists are biased towards trivial names which is conceiving additional challenges for the appropriate conversion among several notations.

Nevertheless, there are some open-source tools for the structure-to-name translation. These tools can convert molecules from SMILES strings to IUPAC names and vice versa, such as OPSIN [2], Struct2IUPAC and IUPAC2Struct [1]. These tools can provide something like human chemical intuition and translate between SMILES strings and IUPAC chemical names. The latter is a transformer-based approach, and many more variants of the transformer model are applied successfully for direct translation from one sequence to other sequence system, even in different data pattern. Such data include molecular images, molecular sequences, and molecular crystal structure.

Although several models are emerging to establish a mapping inner-relationship with SMILES string, such as smilesGPT[3], the representation using IUPAC name has not been explored yet. Hence, our method is the first exploration of the molecular representation for the chemical semantics language to generate novel molecular and the prediction in terms of molecular properties and activity classification.

In our work, considering the advantage of human-based IUPAC name, we integrate it to the transformer-based model named iupacGPT. For comparing to SMILES string representation, we first perform pre-training on the constructed IUPAC name dataset. Thereafter, the two tasks as the downstream task will be fine-tuning. Finally, 50k molecules will be generated for evaluation, 30k molecules will be used to the validation of molecular property and classification prediction.

2. Data and method

The IUPAC name of all molecules were downloaded from PubChem [4]. They were filtered using a series of criteria: i) filtering out the IUPAC strings containing disconnected ions or fragments; ii) compound standardization by RDKit [5] for the removal of salts and isotopes, and as well as charge neutralization. Upon filtration, a total of 97 million molecules were retained. Subsequently,
we calculated the essential properties of these molecules including molecular weight (MolWt), partition coefficient (LogP) [6], synthetic accessibility score (SAscore) [7], and quantitative estimate of drug-likeness (QED) [8] by RDKit tool. The filtration rules were parameterized as following:
(a) 50 <= MolWt <= 1000
(b) 3 <= sequence length < 1200
Right after the filtration, 90 million molecules were retained, wherein, we sorted out 89 million molecules as the training set.

In this study, we first applied Tokenization procedure on the sequence. Tokenization is a process of splitting the sequence into chunks and discretizing these chunks (also named tokens). It is a common pre-processing stage for any sequence model. Consequently, we constructed a rule-based IUPAC tokenizer with tokens in the IUPAC names to well-known functional groups and moieties. In order to understand the IUPAC tokenizers in depth, the readers are referred to these references [1, 9-11].

For each molecule, a particular property is usually a continuous value, and the bioactivity is a boolean value. Herein, we choose prediction of LogP value as regression task and use whether it is a natural product as a classification task.

After sequence tokenization, we encoded all the tokens as the incremental integer sequence from 3, specifically, 0 as padding mark, 2 as start mark, and 1 as end mark in the IUPAC sequence, but in the SMILES sequence, the start mark is 1, the end mark is 2 and the padding marks are the same both in IUPAC sequence and SMILES sequence.

In our proposed prediction model, a sequence prediction of probability distribution from the training set is obtained by the GPT-2, our iupacGPT model just obtain the decoder and cannot reserve IUPAC embedding representation by pre-training, it operates directly on IUPAC names, instinctively encodes rich chemical semantics information for organic chemists, and performs transfer learning by a configured text-to-text transformer model to capture inner-sequence
relationships. Our method does not require edited pairs of molecules for training, only single sequence.

Similar to the smilesGPT model, our model focuses on learning chemical semantic relationship in the IUPAC name space. In iupacGPT model, a variant of transformer model is included and derived from GPT-2, its total parameters are 1.5 billion parameters, the configure parameters can be reference here [12]. GPT-2 uses a high-performance variant of transformer, multi-layered Transformer Decoder as the base model for language modeling tasks. It also uses a self-attention mechanism which is capable of emphasizing critical details of the text. The framework of iupacGPT is shown in the Fig 1.
As a beginning, we tokenize iupac dataset to iupac batch, then fine tune on the pre-trained GPT-2 model and generate 50k iupac sequences, finally, filter them and convert them to smiles. For generation task, we calculated the metrics of validity, uniqueness(unique@1000, unique@10000), Frechet ChemNet Distance (FCD) [13], Internal diversity (IntDiv) [14], Novelty, the means of these metrics could be reference here [14]. For the regression task, we performed the means root-mean-square error (RMSE) metrics[15]. For the classification task, we computed the area under the receiver operating characteristic curve/precision recall curve (AUC/PRC) metrics[16].

**Experimental Setup**

The implemented experimental code is based on the open-source machine learning framework “Pytorch” (https://pytorch.org). The employed variant transformer models are based on the open-source deep learning platform “Hugging Face library” (https://huggingface.co). All experiments were performed on Windows 10 operating system, an Ubuntu 20.04.4 LTS operating system of an Intel(R) Xeon(R) Gold 6226R CPU 64 cores, 2.9GHz CPU and 256G memory, and single A100 GPU, 48G memory.

3. Result

**Analysis of model performance**

To evaluate the efficiency of iupacGPT over existing models, we calculated the same metrics from the generated molecules by iupacGPT and the retrained smilesGPT by MOSES platform [14]. The experimentally compared results have been summarized in Table 1.
Herein, each metric depends on generated set and reference (from MOSE platform) set. We have performed retaining four decimals for each value, where the Valid and Unique@1000 and Unique@10000 indicate that iupacGPT is overall better than smilesGPT in Table 1. Relative to smilesGPT, our model performs the better novelty, FCD and internal diversity by introducing the IUPAC name representation. It indicates that iupacGPT could be improved by learning the hierarchical and inner-sequences relationship between atoms and chemical groups. Nevertheless, the smiles-based representation model can only learn the relationship between atoms and bonds.

**Comparison of chemical properties distribution**

Chemical properties distribution is an efficient presentation for visually evaluating the generative model. To gain further insight into the performance of iupacGPT, we conducted extensive experiments to systematically compare the properties distribution of the molecules generated by iupacGPT and smilesGPT.
After generating 50k molecules, we calculated these properties including molecular weight, natural products likeness (NP-likeness), LogP, SAscore, QED, and quantitative estimate of protein-protein interaction targeting drug-likeness (QEPPI) [17-19].

In Fig 2, the pale-yellow curve represents the property distribution of molecules generated by the smilesGPT, the other curve colored by purple represent the property distribution of molecules generated iupacGPT. For MolWt distribution, all the compounds are less than 800 and the peak of all curves is about 300. Compared to iupacGPT, the molecules generated by smilesGPT have overall larger MolWt, in addition, the LogP and QED distribution curves show similar characteristics. Furthermore, Figure 2 shows a clear difference on the property’s distribution curves of QEPPI and NP-likeness and SAscore between two model. The entire curve for iupacGPT show right shift than smilesGPT in term of NP-likeness and SAscore, in turn, left shift for QEPPI.
curve. However, we cannot suppose that the distribution curve can indicate which model is better, since each model has own target and merit.

**Classification prediction analysis**

In this section, we conduct the binary classification prediction of whether it is a nature product (is_np). To evaluate the transfer performance of the model on the classification task after pre-training, we conducted experiments on a small dataset of 30k with is_np values, splitting the data in the ratio of 0.8:0.1:0.1 for the training set, validation set, and test set, respectively, and ran three separate experiments, the results of which are shown in Table 2.

<table>
<thead>
<tr>
<th>model</th>
<th>Train auc/prc</th>
<th>Valid auc/prc</th>
<th>Test auc/prc</th>
<th>Mean auc/prc</th>
</tr>
</thead>
<tbody>
<tr>
<td>iupacGPT</td>
<td>0.963/0.964</td>
<td>0.977/0.978</td>
<td>0.982/0.983</td>
<td>0.982(+/-0.001)/0.983 (+/-0.001)</td>
</tr>
<tr>
<td>smilesGPT</td>
<td>0.838/0.789</td>
<td>0.831/0.736</td>
<td>0.897/0.845</td>
<td>0.907(+/-0.021)/0.854(+/- 0.028)</td>
</tr>
</tbody>
</table>

Where *Mean* represents the average of three experiments run after a 10-fold cross-test, the other metric is the average of the best results of three experiments. *auc* means area under the receiver operating characteristic curve from prediction scores. *prc* means precision recall curve. Higher *auc/prc* is better for a model.

In table 2, the test auc/prc value of iupacGPT is close to 0.982, the mean auc/prc value by 10-fold cross-test is about 0.982/0.983, on the contrary, it is close to 0.907/0.854 for smilesGPT. It is obvious that iupacGPT is better than smilesGPT in this classification task. To this end, we argued that our model is reliable enough to predict nearly the experiential is_np value.

Maybe, you could argue that we should compare on more classification tasks and compare with more models. However, in this paper our object is simply to compare the differences between iupac and smiles representations at two different semantic levels, which requires that the
conditions such as models and datasets must be uniform. If asked to compare the performance of different classification tasks under the same conditions, then another research topic is involved, on which classification tasks these two representations perform well or badly, which is beyond the scope of this paper.

**Property prediction analysis**

In this section, we conduct the property prediction of LogP. To prepare the input data before fine-tuning, at first, we converted the SMILES strings to IUPAC names and calculated the LogP value for each molecule. Thereafter, the selected molecules were fed to the model and consequently output the predicted LogP property.

To evaluate the transfer performance of the model on the regression task type after pre-training, we conducted experiments on a small dataset of 30k with LogP values, splitting the data in the ratio of 0.8:0.1:0.1 for the training set, validation set, and test set, respectively, and ran three separate experiments. After calculating root mean square error (RMSE) of LogP value, the results of which are shown in Table 3.

<table>
<thead>
<tr>
<th>model</th>
<th>Train RMSE</th>
<th>Valid RMSE</th>
<th>Test RMSE</th>
<th>Mean RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>iupacGPT</td>
<td>0.724</td>
<td>0.524</td>
<td>0.524</td>
<td>0.524 (+/-0.008)</td>
</tr>
<tr>
<td>smilesGPT</td>
<td>1.040</td>
<td>0.859</td>
<td>0.871</td>
<td>0.880 (+/-0.029)</td>
</tr>
</tbody>
</table>

Where **Mean** represents the average of three experiments run after a 10-fold cross-test, the other metric is the average of the best results of three experiments. **RMSE** means root-mean-square error, less RMSE is better for a model.

In table 3, the test rmse value of iupacGPT is close to 0.524, the mean rmse value by 10-fold cross-test is about 0.524, on the contrary, it is close to 0.88 for smilesGPT. Conclusively, it is obvious that iupacGPT is better than smilesGPT in this regression task.
In general, we compared generated molecular distribution of six properties. Although, the molecules generated by iupacGPT are structurally similar with the fine-tuning sets. Yet, they showed diversity in terms of the specific properties such as is_np and LogP. For the prediction of property regression and classification, iupacGPT has shown better performance than smilesGPT, especially on the LogP and is_np.

4. Conclusion

IUPAC is a more promising representation in sequence-based molecular generative models than the current smiles representation to small molecules, especially in tasks that integrate with natural language.

In this study, first, we trained our model and other deep generative molecular models on a large dataset of diverse molecules, which were contained about 90 million compounds. The goal of this process was to train the models to learn how to generate chemically valid molecules and compare performance with each other. For the prediction of regression properties and classification task, our model iupacGPT show better performance than smilesGPT. Especially, for classification task, the chemical functional group at the semantic level is relevant to determine whether it is a natural product, thus, iupac name can directly contribute to the judgment.

It is important to highlight some limitations of our study. Firstly, the metrics for the assessment of generative designing models continue to evolve and it is not feasible to explore all metrics that have been reported in literature. It is critical that performance assessments of generative models are always taken in the context of the metrics that are applied in the goals of a given project. Second, gold-standard validation sets for generative molecular designing problems do not exist. Thus, assessing difference between the generated molecules and the reference sets, we recognize that some otherwise biologically significant molecules may be missed.

There are several directions for future studies, which we can preclude from our study. As our proposed iupacGPT is a general approach for single modal molecular generation and single property prediction, it would be interesting to apply it to other domains and problems, for instance,
learning of inner relationship of SMARTS. Moreover, multi-objective and multi-task prediction of molecular properties is a very challenging and meaningful problem.

Author contributions
J.M., J.W. conceived the idea, J.M. wrote the article, developed the codes, and K.H.C. draw the figures. K.T.N. provided the financial support.

Declaration of interests
The authors declare they have no conflict of interest.

Acknowledgements
This work was supported by Yonsei University graduate school "Integrative Biotechnology & Translational Medicine".

Data and code availability
The implemented code and experimental dataset are available online at https://github.com/AspirinCode/iupacGPT.

Reference


