A COMPREHENSIVE REVIEW ON MOLECULAR STRUCTURE RECOGNITION USING DEEP LEARNING METHODS

Abstract

Information about chemical substances molecular structural formula is typically displayed as 2D visuals in scientific literature and journals. Unfortunately, these molecular images cannot be interpreted by machines. Ample need of molecular structure automation highly demands the conversion of graphic structure molecular representations into legible formats. This review provides a comprehensive overview of various methods and tools that have been published in the field of Chemical Structure Recognition using Deep Learning methods. Most of the published methods achieved good results for specific type of structures.

Keywords: Molecular Structure, Deep Learning

Authors

K. Ushaamurlidhar
Department of Computer Science
University College
Mangalore, Karnataka, India.
Ushaamurli6@gmail.com

Bharathi Pilar
Department of Computer Science
University College
Mangalore, Karnataka, India.
Bharathi.pilar@gmail.com

I. INTRODUCTION

The process of turning graphical representations of molecules into machine-readable chemical formula is known as molecular image recognition. Information extraction from the literature on chemistry comprises this essential process [1]. In scientific publications, text and graphics are used to convey chemical information. Manually obtaining the molecular structure name from the scientific documents will take more time and sometimes, it might be error-prone. The data format is in such a manner that they are not in a machine-readable form. Consequently, there is a requirement for automated chemical information extraction techniques due to the growth in the volume of chemical information being published [2]. As the number of research publications on chemical structure is growing dramatically, molecular image recognition is essential to many chemical sub-fields, such as synthetic science, study on natural products, medication development, etc. As a result, molecular image recognition is still rated highly demand [3]. In the molecular structure image, atom (character) locations are identified as nodes and all bonds- single, double and triple, dashed or wedged—are considered as edges. Character names will be extracted from the nodes (atom names) in molecular image recognition and counted for naming. The kind of molecule can be determined by the bond type. SMILES (Simplified Molecular Input Line Entry System) string and the InChI (International Chemical Identifier) string are two established two-line notations of chemical structures. However, in organic chemistry IUPAC (International Union of Pure and Applied Chemistry) nomenclature continues to be significant. The vast majority of chemical journals also demand IUPAC nomenclature for organic structures.

Significance of Automatic Recognition of Molecular Image: Naming an item, a thing, a person, or a place is crucial to daily life for simpler identification. Every single thing in the world is known by its name. We might understand the importance of a chemical substance when we think about its recognition.

- **Describe the Chemical Compound and Specify its Contents:** A name or label provides complete information regarding the chemical compound. It mainly includes the ingredients of the molecule, i.e., number and name of the atoms.

- **Identification of the Compound:** It is easier to identify a specific molecule or compound among many with the help of recognition and naming.

- **Grading (or Classification) of Compound:** When a single compound has different structures [Isomers], Molecule nomenclature process ensures that there is no ambiguity while concerning with chemical compound, helps to identify the type of compound and type of constituent atoms (IUPAC nomenclature). IUPAC nomenclature helps to differentiate them.

II. REVIEW OF THE RELATED LITERATURE

Molecular Image Recognition approaches can be categorized into two types: Rule based process and Data driven based process. The majority of molecular image recognition technologies in the past were rule- based and descended from optical character recognition. Although they are frequently used because the majority of them are public, their logical

principles frequently fall short of accounting for a variety of image styles [4]. Since the process of molecular image recognition is similar to that of image captioning, in this review we found that the majority of molecular image recognition methods have been adapted from captioning methods. Every day, we are confronted to a large number of images from various sources, including the internet, news bulletins, schematics in documents and advertisements. These contain visuals that visitors must interpret by themselves. Most of them do not contain an explanation, but humans can interpret them. If humans require automatic image captions, machines must read descriptions about the image. Due to its benefits, including strong feature extraction capabilities and excellent identification accuracy, deep learning is frequently used in image recognition [5]. One of the most important aspects of the image recognition system is feature extraction. There are two categories of image features; these are low-level and high-level image features [6]. Image processing extracts low-level image features like shape, colour, and texture, whereas high-level attributes represent the words or thoughts in an image. Furthermore, image characteristics used in existing image recognition algorithms can be divided into two categories: region-based features and global image features. Region-based features require image segmentation, whereas global image features are calculated from all images [6]. Visual features play an important part in the identification and recognition of objects, as well as the display of visual content. Texture, colour, SIFT, and other sorts of features have diverse qualities. Local features (SIFT [Scale Invariant Feature Transform], SURF [Speeded up robust features], shape, and so on) describe image regions, whereas global features represent the entire image. So, while local features are a subset of an image that may be used for object recognition, global features are a broadening of an image that can be utilised for object detection [7].

Complexities and challenges of image captioning are handled by Deep-learning-based techniques [8]. Some of the molecular image recognition techniques have been discussed in this review. In 2019, Staker et al. [9] introduced end-to-end deep learning solutions for extracting molecular structure segments from documents as well as chemical structure predictions from these segmented images. This deep learning-based method is resistant to changes in image quality and style because it operates directly on raw pixels and doesn't require any manually created features. This architecture has used a convolutional neural network (CNN) architecture based on an open-source implementation of U-Net [10]. From the input documents, the segmentation model has recognised and extracted chemical structure images. For each extracted chemical structure image, the structure prediction model has produced a computer-readable SMILES string. The results showed that it was possible to anticipate low resolution image of chemical structures using an automated method because all of the images used in the presented results were heavily down sampled. Where, the prediction network uses an encoder-decoder architecture, in which a CNN encodes images including molecular graphs of a chemical compound to a fixed-length latent space, and then a recurrent neural network (RNN) decodes them back to a sequence of SMILES letters [11]. In 2020 olden of et al. [12] provided a deep learning approach that comprises of three classification models that predict atom positions, bonds, and charges after a segmentation model. They used both segmentation and classification models. These networks were built on the foundation of CNN. Yu and Koltun's concept of dilated convolution [13] is used to create the segmentation networks. A chemical compound's molecular structure representations served as the segmentation network's training data, created with RDKit [14] and CHEMBL [15] data sources. The segmentation network's output is fed into classification networks, which locate atoms, bonds, and charges. The accuracy of each network was tested independently, and the

classification networks were shown to perform significantly better than the segmentation networks. In 2021, Weir et al. presented DL based algorithms for identifying hydrocarbon structures automatically [16]. Here authors adopt neural image captioning, in which an image is sent into a neural network (NN) and a caption is generated. An image of hand-drawn hydrocarbon molecule is fed into this programme, and the predicted SMILES string is returned. In the proposed method, convolutional neural network encoder and a long short term memory decoder having beam search and attention make up the NN architecture. The trained data-driven models were combined with ensemble learning to improve the accuracy of the constituent models and acquire information on, when the model likely fails. A data-source of 50 tagged RDKit images achieved an out-of-sample (test set) accuracy of more than 90%, and a collection of 500 images got a maximum accuracy of 98 percent. This indicates that the chosen NN architecture is capable of learning SMILES sequence from hydrocarbon images rendered by a machine. In 2021, Clevert et al. [17] introduced a molecular op tical recognition system based on machine learning. It uses an encoder model using convolution neural network. They employed the CDDD (Continuous Data Driven Descriptor) decoder to convert a graphical structure of a chemical compound into a SMILES sequence. The tests indicate strong rebuilding efficiencies for chemical compounds with up to 35 atoms and robustness across a variety of datasets. The study used the auto-encoder architecture created by Winter et al. [18]. Clevert et al. presented a model which is accurate and fast combining deep CNN learning from molecule depictions and a pre-trained decoder using which the latent representation is translated in to SMILES representation of the molecules. They used ChEMBL and PubChem data-sources, for which several structural representations at various resolutions and with various conventions were taken. In order to successfully determine the CDDD embedding of the chemicals represented, the presented technique was trained on a variety of graphical representations of compounds. DECIMER (Deep lEarning for Chemical Image Recognition) is a deep learning system that converts a raster image of a chemical compound discovered in scientific documents into SMILES [19]. They employed an autoencoder based network with TensorFlow 2.0 at the backend. It is built on existing showand-tell [20] deep neural networks. The Inception V3 CNN encoder network, which has one fully connected layer and a RELU- Rectified Linear Unit activation function, is used in this scenario. Their decoding system is an RNN with two fully connected layers and a gated recurrent unit (GRU). Rajan et al. [2] presented DECIMER Segmentation; it is the first deep learning-based open-source tool for automatically recognising and segmenting structural image of chemical compound from the academic document. It accepts PDF files as input and produces segmented chemical structure diagrams in grayscale. The various sections of the input PDF document are transformed to individual PNG images. The working process of this technique consists of two main components. Mask R-CNN architecture recognises chemical structure renderings, during the detection stage generates masks that identify those renderings on the input page. Then, in a post-processing stage, any possibly flawed masks are stretched. In 2021 Krasnov et al. developed a Transformer based artificial neural architecture to translate between SMILES and IUPAC chemical notations: Struct2IUPAC and IUPAC2Struct [21]. They adopted the transformer architecture, which included six encoder and decoder layers and eight attention heads. The attention dimension was 512, and the feed forward layer’s dimension was 2048. They trained the developed model on PUBCHEM [22] dataset. Transformer can generate several versions of a sequence using beam search. The model-SMILES to IUPAC names converter, running in production mode with beam size = 5, achieved 98.9% accuracy on a subset of 1,00,000 random molecules from the test set. Encoder runs only once to read SMILES input, whereas decoder processes each output token.

Handsel et al. proposed a Sequence-to-Sequence ML based technique for anticipating the IUPAC name of a chemical compound from its standard InChI format [23]. They used Transformer as encoder and LSTM as decoder, the model has used two stacks of transformers, comparable to the NNs used in the state-of the-art machine conversion. This model has analysed the InChI and predicted the IUPAC name letter by letter, while in neural machine translation, input and output contents were tokenized into words or sub-words. The proposed method was trained on a freely available dataset of 10 million InChI/IUPAC name pairs obtained from the National Library of Medicine's online PubChem service. The model attained test-set accuracies of 95 percent (character-level) and 91 percent (object-level) after five days of training on a Tesla K80 GPU (whole name). With the exception of macrocyclics, the Model did quite well on organics. The encoder sends the decoder a numerical representation of the InChI. A start token is used to seed the decoder, and its output is iteratively re-input until an end token is predicted. Stereoisomers are specified by an extra layer in the InChI representation. The inchi2iupac model can successfully label enantiomers and diastereomers, even when their InChI differs by a single character. STOUT (SMILES-TO-IUPAC-name translator) is a DL neural machine translation method that predicts a SMILES sequence from the IUPAC name and generates the IUPAC name for a given chemical compound from its SMILES sequence, was developed by Rajan et al. [24] based on language translation and language understanding. They implemented an autoencoder–decoder architecture with Python 3 and they used TensorFlow 2.3.0 as the backend. RNNs with Gated Recurrent Units (GRU) are used in both the encoder and decoder networks. The encoder receives the input strings, while the decoder receives the output strings. The encoder output and hidden state are generated by the encoder network. The attention weight is determined by the network's deployed attention mechanism. The context vector is created by combining the encoder output with attention weights. Meanwhile, an embedding layer processes the decoder output. The embedding layer's output and the context vector are combined and sent to the decoder's GRUs. They used Adam optimizer with a learning rate of 0.0005 as optimization method and sparse categorical cross-entropy as the loss function. Yanchi Li et al. [25] build a unique generating model called ICMDT (Image Captioning Model based on Deep TNT) by stacking Deep TNT blocks and decoding with the native transformer-decoder. The proposed method translates molecule structure to InChI text. In 2015, Heller et al. described the procedure for translating chemical image into InChI text. There are three major steps in the general workflow of derivation of InChI from structural data: (a) normalization of input structure, that is, converting the supplied structural data into internal data structures conforming to the InChI chemical model; (b) canonicalization of atomic numbering, which accounts for atomic equivalence/inequivalence relations appearing under this model; and (c) serialization, that is, generating the final sequence of symbols, an InChI string. There is an optional fourth step (d); hashing of the InChI string and producing a compact InChIKey [26]. The Levenshtein distance, a metric for measuring the similarity of two sequences, is used to evaluate the final outcome. Local and global information are modelled independently in TNT block [27]. The inner-transformer block processes the local information (pixel embedding) first, after which the local information is attached to the respective patches and the outside transformer block processes the global information (patch embedding). Simple position encoding is used to combine patch-level data into global data, to get accurate data they re-model the patch-level information to better reflect the global and local information of chemical pictures. They combine numerous patch-level messages into a single large patch before processing them with a transformer-block. This greatly simplifies the integration of location coding and global data known as deep TNT block. In 2022 Yoo et al. proposed a

deep learning model to extract molecular structures from images. The encoder extracts data from images in order to create graphs. This model's input image has a fixed resolution of 800 by 800 pixels. They virtually divided the image into 25 pieces horizontally and vertically, resulting in 625 fragments of 32 × 32 pixels each. After that, they assign each piece's 2D position information, which is then concatenated with the raw image features. The ResNet-34's output features are moulded into a 1D sequence. Then they're put into the Transformer encoder [28], which captures the relationship between two atoms in the image that are far apart. The encoder's final output is utilised as the graph decoder's input. The Transformer based decoder generates the final graph in an auto regressive approach; at step t, it encodes the sub-graph formed at step t-1 before generating a new node and its related edges to the current ones [4]. In 2022 Khokhlov et al. proposed a Transformer based ANN to convert organic structure to SMILEs form [29]. The input for the model is 384x384 in size. ResNet-50 [30] used as a CNN block. The output of the ResNet module was 2048x12x12 in size. The final two residual blocks of ResNet-50 were eliminated, and the resulting shape, which is quite similar of the traditional transformer dimension, is 512x48x48 obtained. Other transformer decoder settings were derived from classical architecture. In 2022 Zhanpeng Xu et al. [3] proposed an end-toend model called SwinOCSR, which works on the basis of a Swin Transformer architecture. On this technique Transformer model has been utilized to transform chemical data information from publications and documents into Deep SMILES. It has used the Swin Transformer architecture as the foundation for retrieving image characteristics. Transformer encoders, decoders and a backbone make up the framework of the SwinOCSR model. The Swin Transformer serves as the basis for the backbone. The backbone first obtains a high dimensional patch sequence by extracting image attributes from an input molecular image. The Transformer encoder is then fed the patch sequence and positional embedding to produce a representation sequence. The appropriate Deep SMILES are then decoded by the Transformer decoder.

III. DATASET

The existing datasets of 2D chemical structure which seem useful are listed below.

1. **PUBCHEM:** PubChem [18] is a chemical information resource at the U.S. National Centre for Biotechnology Information (NCBI). It was released in 2004. It contains multiple substance descriptions and small molecules with fewer than 100 atoms and 1,000 bonds.

2. **USPTO:** A collection of 4852 images and molecule descriptions based on US Patent Office (USPTO) data, obtained from Rajan et al. [9] the average resolution of the images is 649 × 417 pixels. The dataset contains structures of chemical compounds with an average size of 28 atoms, ranging between 10 and 96 atoms [26].

3. **UoB:** Rajan et al. [9] provided 5716 images and molecular descriptions of chemical structures developed by the University of Birmingham. The images have an average resolution of 762 x 412 pixels. The molecules in this data set are quite tiny, with an average of only 13 atoms and a range of 4 to 34 [26].

4. **CLEF:** Based on the Conference and Labs of the Evaluation Forum (CLEF) test set, Rajan et al. [9] provided a collection of 711 image data and chemical descriptions. The images have a resolution of 1243 x 392 pixels on average. The molecules in the dataset range in size from 4 to 42 atoms on average [26].

5. **CHEMBL:** ChEMBL [13] is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs. It is maintained by the European Bioinformatics Institute (EBI). It was released in 2009. It contains 2.2M compounds.

6. **JPO:** Rajan et al. [9] provided a collection of 365 images and molecular descriptions based on Japanese Patent Office (JPO) data. This data collection includes a lot of textual labels, including Japanese characters, as well as irregular elements like line thickness variations. Furthermore, several of the images are of low quality. The images have an average resolution of 607X373 pixels. Molecules have an average of 20 atoms, ranging from 5 at the smallest to 43 at the largest [26].

IV. STEPS IN METHODOLOGY

The general process for molecular image recognition is as shown in Figure 1.

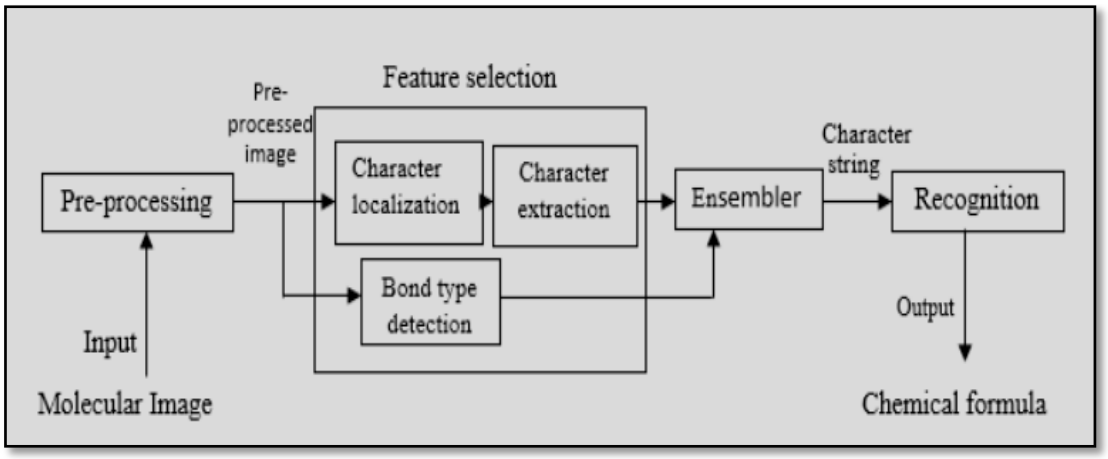

Figure 1: Block Diagram of the Process.

Where, the molecular image is fed into the pre-processing phase. The features (character name and bond type) are selected from the pre-processed image. Ensembler receives the feature and generates the character string. Recognition phase recognizes the chemical formula. The pre-processing steps involves gray-scale conversion, denoising and normalization of the input image. After preprocessing the images are passed to the feature selection process. The feature selection process extracts the information required for the recognition of structural image. Once the features are selected, those features are passed to the ensembler. Then, ensembler performs classification of the selected features and generates character string. Finally, it generates chemical formula for the input image.

V. CONCLUSIONS

Superior quality scientific literature data promotes the need for automated perception process from printed scientific literature in order to advance research efforts. In chemistry, this entails converting pictures of chemical structures into a syntax that computers can understand. This study aims to provide an overview of the literature on deep learning-based molecular structure recognition. We discussed about the creation and development of techniques that enable the automatic extraction of chemical data from the scientific articles. However, there is a need for further progress.

Acknowledgements: The authors with high appreciation acknowledge all authors of Molecular Structure Recognition using Deep Learning method for sharing their valuable visions to the progress of science and technology.

Funding: First author acknowledge NFPWD UGC-INDIA for providing National Fellowship to carry out Research work.

REFERENCES

[1] Y. Qian, Z. Tu, J. Guo, C. W. Coley, and R. Barzilay, “Robust molecular image recognition: A graph generation approach,” arXiv preprint arXiv:2205.14311, 2022.
[2] K. Rajan, H. O. Brinkhaus, M. Sorokina, A. Zielesny, and C. Steinbeck, “Decimer-segmentation: Automated extraction of chemical structure depictions from scientific literature,” Journal of Cheminformatics, vol. 13, no. 1, pp. 1–9, 2021.
[3] Z. Xu, J. Li, Z. Yang, S. Li, and H. Li, “Swinocsr: end-to-end optical chemical structure recognition using a swin transformer,” Journal of Cheminformatics, vol. 14, no. 1, pp. 1–13, 2022.
[4] S. Yoo, O. Kwon, and H. Lee, “Image-to-graph transformers for chemical structure recognition,” arXiv preprint arXiv:2202.09580, 2022.
[5] L. Li, “Application of deep learning in image recognition,” in Journal of Physics: Conference Series, vol. 1693, no. 1. IOP Publishing, 2020, p. 012128.
[6] M. M. Adnan, M. S. M. Rahim, A. Rehman, Z. Mehmood, T. Saba, and R. A. Naqvi, “Automatic image annotation based on deep learning models: a systematic review and future challenges,” IEEE Access, 2021.
[7] P. Bhagat and P. Choudhary, “Image annotation: Then and now,” Image and Vision Computing, vol. 80, pp. 1–23, 2018.
[8] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, “A comprehensive survey of deep learning for image captioning,” ACM Computing Surveys (CsUR), vol. 51, no. 6, pp. 1–36, 2019.
[9] J. Staker, K. Marshall, R. Abel, and C. M. McQuaw, “Molecular structure extraction from documents using deep learning,” Journal of chemical information and modeling, vol. 59, no. 3, pp. 1017–1029, 2019.
[10] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.
[11] K. Rajan, H. O. Brinkhaus, A. Zielesny, and C. Steinbeck, “A review of optical chemical structure recognition tools,” Journal of Cheminformatics, vol. 12, no. 1, pp. 1–13, 2020.
[12] M. Oldenhof, A. Arany, Y. Moreau, and J. Simm, “Chemgrapher: optical graph recognition of chemical compounds by deep learning,” Journal of chemical information and modeling, vol. 60, no. 10, pp. 4506–4517, 2020.
[13] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” arXiv preprint arXiv:1511.07122, 2015.
[14] Rdkit: open-source cheminformatics. http://http://www.rdkit.org. [Accessed: 2022-03- 29].
[15] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrian-Uhalte ´ et al., “The chembl database in 2017,” Nucleic acids research, vol. 45, no. D1, pp. D945–D954, 2017.

[16] H. Weir, K. Thompson, A. Woodward, B. Choi, A. Braun, and T. J. Martınez, "Chempix: automated recognition of hand-drawn hydrocarbon structures using deep learning," Chemical science, vol. 12, no. 31, pp. 10 622–10 633, 2021.
[17] D.-A. Clevert, T. Le, R. Winter, and F. Montanari, "Img2mol–accurate smiles recognition from molecular graphical depictions," Chemical science, vol. 12, no. 42, pp. 14 174–14 181, 2021.
[18] R. Winter, F. Montanari, F. Noe, and D.-A. Clevert, ´ "Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations," Chemical science, vol. 10, no. 6, pp. 1692– 1701, 2019.
[19] K. Rajan, A. Zielesny, and C. Steinbeck, "Decimer: towards deep learning for chemical image recognition," Journal of Cheminformatics, vol. 12, no. 1, pp. 1–9, 2020.
[20] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in international conference on machine learning. PMLR, 2015, pp. 2048–2057.
[21] L. Krasnov, I. Khokhlov, M. Fedorov, and S. Sosnin, "Struct2iupac–transformer-based artificial neural network for the conversion between chemical notations," 2021.
[22] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu et al., "Pubchem 2019 update: improved access to chemical data," Nucleic acids research, vol. 47, no. D1, pp. D1102–D1109, 2019.
[23] J. Handsel, B. Matthews, N. Knight, and S. Coles, "Translating the molecules: adapting neural machine translation to predict iupac names from a chemical identifier," 2021.
[24] K. Rajan, A. Zielesny, and C. Steinbeck, "Stout: Smiles to iupac names using neural machine translation," Journal of Cheminformatics, vol. 13, no. 1, pp. 1–14, 2021.
[25] Y. Li, G. Chen, and X. Li, "Automated recognition of chemical molecule images based on an improved tnt model," Applied Sciences, vol. 12, no. 2, p. 680, 2022.
[26] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi, "Inchi, the iupac international chemical identifier," Journal of cheminformatics, vol. 7, no. 1, pp. 1–34, 2015.
[27] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," Advances in Neural Information Processing Systems, vol. 34, 2021.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
[29] I. Khokhlov, L. Krasnov, M. V. Fedorov, and S. Sosnin, "Image2smiles: Transformer-based molecular optical recognition engine," Chemistry-Methods, vol. 2, no. 1, p. e202100069, 2022.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

ABBREVIATIONS AND ACRONYMS

[1] IUPAC : International Union of Pure and Applied Chemistry
[2] SMILES : Simplified Molecular Input Line Entry System
[3] InChI : International Chemical Identifier
[4] CNN : Convolutional Neural Network
[5] RNN : Recurrent Neural Network
[6] ResNet : Residual Networks
[7] RDKit : An open-source toolkit for cheminformatics
[8] DECIMER : Deep lEarning for Chemical Image Recognition
[9] STOUT : SMILES-TO-IUPAC-name Translator
[10] RELU : Rectified linear activation unit
[11] Deep TNT : Deep Transformer-iN-Transformer
[12] SwinOCSR : Shifted window Optical Chemical Structure Recognition
[13] ICMDT : Image Captioning Model based on Deep TNT
[14] GRU : Gated Recurrent Unit
[15] LSTM : Long short-term memory
[16] CLEF : Conference and Labs of the Evaluation Forum
[17] JPO : Japanese Patent Office
[18] USPTO : US Patent Office

