Reconstruction of lossless molecular representations

Umit V. Ucak[1], Islambek Ashyrmamatov[1], and Juyong Lee[1,2]

[1]Department of Chemistry, Division of Chemistry and Biochemistry,Kangwon National University, Chuncheon, 24341, Republic of Korea
[2]Arontier co. Seoul, 06735, Republic of Korea
juyong.lee@kangwon.ac.kr

Abstract

SMILES is the most dominant molecular representation used in AI-based chemical applications, but also responsible for certain issues associated with its internal structure. Here, we exploit the idea that structural fingerprints may be used as efficient alternatives to unique molecular representations. For this purpose, we assessed the conversion efficiency of fingerprints back to the molecules. We successfully reconstructed molecules with the NMT approach, achieving a high level of accuracy. Our approach therefore brings structural fingerprints into play as strong representational tools in chemical NLP applications by restoring the connectivity information that is lost during the fingerprint transformation. This comprehensive study addresses the major limitation of structural fingerprints which precludes their implementations in NLP models. Our findings should enhance the efficiency of the models in generative and translational fields.

1 Introduction

SMILES [1] string is the most widely used linear representation describing chemical structure information. It uses several simple rules to convert a chemical structure into a character string. The notation allows multiple unique SMILES strings to be used to represent molecules. Since its inception, SMILES has undergone various extensions [2–5]. Among them, canonicalization algorithms, integration of isotopism and stereochemistry information (isomeric SMILES) can be highlighted [6–9].

Whereas this simplified line notation is superior to other one-dimensional representation schemes such as WLN [10], SLN [11], and IncHi [12], its internal structure leads to number of problems when using in NLP applications [13–16]. SMILES-based deep learning NLP models are prone to generate invalid SMILES strings [17–20] The invalidity errors can be attributed to the fragile structure of SMILES (strong dependence between tokens). Because language models formulates predictions one character at a time, a single character alteration often suffices to invalidate an entire SMILES string. In addition, predicted valid SMILES are not guaranteed to be chemically valid. Several attempts have been proposed to ensure syntactic and chemical validity of SMILES predictions [21–24]. In fact, the challenges that SMILES syntax poses prompted the development of alternative syntaxes such as DeepSMILES [25] and SELFIES [26].

Text generation and machine translation are among the mostly used NLP methods in chemistry. Their aim, by definition, is to generate meaningful sequences from meaningful tokens. Tokenization is therefore a pivotal preprocessing step in many NLP tasks. SMILES strings are merely meaningful as a whole so that any type of tokenization procedure must dissect them in an arbitrary fashion. From a chemist perspective, atom-wise or character-wise tokenization of SMILES do not produce fully interpretable tokens because many characters in SMILES strings are used to represent the topological characteristics, such as ring-closure or branches which do not correspond to physical atoms. Also, most SMILES tokens are indistinguishable due to the repetitiveness. Considering that the primary design purpose of SMILES was to serve as a universal exchange format, being unable to derive interpretable insights from tokenization is understandable. Nevertheless, despite the abovementioned challenges, SMILES has a prominent role to play in AI-based chemical models since these models require precise molecular descriptions in their output structures.

Model interpretability is crucial to both developers and users. Yet, it is often difficult to attribute meaning to the outcomes of deep learning methodologies due to their black box nature. The downside of

NLP methods in chemistry is their lack of interpretability [27]. An accurate interpretation requires careful consideration of both the molecular representation and the means of explaining the model. Although looking inside the model is easier with the state-of-the-art NLP models using attention mechanism via attention weight matrices, recent studies showed that the learned attention weights may not fully reflect the feature importance [28, 29]. In this regard, attribution maps such as integrated gradients [30] help us explain the important features contributing to the prediction.

Interpretability necessitates the existence of meaningful tokens since NLP models tend to learn the relationships between them. Interpretability of individual tokens is therefore highly desired. The chemical interpretability of conventional methods is hampered by the fact that SMILES tokens are not fully interpretable in a chemical sense. This fact contradicts the recent statement by Tu et al. [31], SMILES is indeed a highly efficient representation for capturing information about molecular structures, whereas issues only arise when SMILES are tokenized. Molecular fingerprints and substructural keys can be employed as an alternative to SMILES representation. They are designed to capture either chemical features, concepts or structural patterns which together yield interpretable set of tokens suitable for NLP applications.

Construction of molecular fingerprints is a lossy procedure. This is the main reason why fingerprints lead to stand-alone interpretable tokens. The fact that attention is a permutation-invariant operation [32], fingerprints fit well into attention mechanism. Regarding the representation itself, attention-based models such as Transformers can handle unconnected features of fingerprints [33, 34]. Thus, to address the major limitation of structural fingerprints which precludes their implementations in NLP models, we assessed the conversion efficiency of fingerprints back to the molecules. To accurately decode fingerprints back to lossless molecular representations we used a translation-based system, the Transformer architecture. In this study, we show that reconstruction of molecules is a practically plausible approach which delivers very high-accuracy needed for chemical applications. We illustrate our approach through 13 structural fingerprint examples, classified into five main categories, and show that certain fingerprints can be used directly in a NLP setting as alternatives to SMILES.

2 Method

2.1 Structural fingerprint representations

Structural fingerprints tested were taken from RDKit [35] implementations. They can be classified into five main groups which are reported in Table 1 along with corresponding sequence lengths and vocabulary size information. We generated a total of 13 different fingerprints to carry out our analyses. Binary variants of selected fingerprints are hashed to a fixed size of 2048, except Avalon. Fingerprints are optimized by its parameters to yield similar sequence lengths when necessary. We omitted sparse versions of atom-pair and ECFP4 from this calculation as the vocabulary space covered was huge, thus the token size.

1. *Predefined substructure* MACCS keys [36] converts a molecule into a bit vector with a fixed size of 166, in which each bit records the presence of a feature taken from a predefined dictionary of SMARTS patterns [37].

2. *Paths and feature classes* The Avalon enumerates paths and feature classes. We refer the reader to study of Gedeck et al. [38] for a thorough explanation of paths and feature classes covered.

3. *Path-based* The RDKit fingerprint is very similar to the Daylight fingerprint [37]. Hashed branched and linear subgraphs of size 4 is used. In both cases, minPath and maxPath parameters were set to 2 and 4 respectively. Hashed variant of atom pair fingerprint encodes all pairs of atoms with their environments as well as their bond distance [39]. Here it is used with the following parameters minLength=1, and maxLength=6.

4. *4-atom-paths* Topological torsion [40] encodes sequences of four bonded atoms such that the generated set of substructures have local character. It is used along with its hashed variant.

5. *Circular* The extended-connectivity fingerprint [41] ECFPx enumerates circular atom environments defined as topological neighborhood fragments up to a selected radius. The sparse and hashed variants are used separately. Feature-class fingerprints FCFPx include pharmacophoric features as invariants.

Table 1: Translation related statistics about domain-specific datasets generated by structural fingerprints used for the performance analysis are presented together with the targeted molecular representations, SMILES and SELFIES.

Abbreviations	Description	Dim	Sequence length		Token size
			Ave.	Max	
Predefined substructures					
MACCS		166	50	107	160
Paths and feature classes					
Avalon	Hashed	512	182	470	516
Path-based					
HashAP	Atom pair - hashed	2048	92	273	1998
RDK4	RDkit fingerprint - hashed	2048	83	288	2052
RDK4-L	RDK4 - with no branch	2048	58	209	2052
4-atom-paths					
TT	Topological torsion	sparse	32	124	54973
HashTT	TT - hashed	2048	31	118	2052
Circular					
AEs	Morgan radius 1	sparse	29	65	54076
ECFP0	Morgan radius 0 - hashed	2048	10	25	100
ECFP2	Morgan radius 1 - hashed	2048	28	64	2052
ECFP4	Morgan radius 2 - hashed	2048	47	103	2052
FCFP2	Feature-class of ECFP2	2048	20	51	1576
FCFP4	Feature-class of ECFP4	2048	36	86	2052
Unique Representation					
SMILES	Tokenized atom-wise		51	125	109
SELFIES	Generic tokenization		44	127	205

2.2 Model overview

In this study, we employed Transformer [42], a model architecture with a multi-head attention mechanism on each unit. Attention units allow the model to learn global dependencies between input and output. Transformer-based models can reach a high level of success in translation quality compared to generic seq-2-seq methods [13, 15, 17, 34]. In addition, the attention mechanism eliminates the dependence on the order of the input sequence. Therefore, models yield the same sequence of outputs regardless of the spatial connections between tokens. This property of attention mechanism makes Transformer-based models suitable for investigating fingerprint to molecule conversions.

Translation-based algorithms necessitate a huge corpus of diverse translation pairs for effective translation. For this purpose, we selected ChEMBL [43] (2.08M) dataset and added PubChem [44] compounds by maximizing the diversity of atom types based on atomic environments. This led to a total of 5 million training compounds. The dataset contains small and medium-sized molecules, 50 atom or less. Supporting Figure 1 illustrates the sequence length distribution of the dataset in terms of atom types. We tested our model against a diverse external test set consisting of 50K molecules. To have more realistic results, we preferred a more challenging dataset by not removing the stereochemical information though most fingerprints do not account for stereochemistry in RDKit.

2.3 Training and evaluation

We used Pytorch [45] Distributed Data-Parallel Training (DDP) module to train our models. Each model was trained with two GPUs up to 500K step which denotes the number of times the optimizer updates the parameters of the model. The hyperparameters of models were set the same as with the base-model of the original Transformer paper [42]. We employed Zero Redundancy Optimizer [46] (ZeRO) with Adam algorithm to optimize parameters of the models, ZeRO was developed to improve training speed by eliminating memory redundancies in data- and model-parallel training. The details of our key hyperparameters and hyperparameter space are described in Supplementary Table 1.

We set the number of tokens in one batch as 8000 per GPU. Due to our hardware limitations we could not set more than 8000. For fair comparison of fingerprints, the batch size was specified based on the average number of tokens in one batch, provided that the number of sentence pairs in one batch will vary in accordance with the sequence length of a fingerprint. We experimented several learning rate schedulers to extend the vanilla implementation [42]. We applied cyclic [47] learning rate, its decayed variant, and stochastic gradient descent with warm restarts [48] so as to see if scheduling algorithms affect the model performance (see Supplementary Figure 2). Among them, we selected the cyclic learning scheduler, provided a slightly better performance than the other techniques.

We evaluated the conversion efficiency with Tanimoto similarity matching. A further breakdown of the results were achieved by introducing simple string matching. The widely used Tanimoto coefficient operated on sparse Morgan fingerprint is selected as the similarity metric to represent the main results. Pairwise similarities between predictions and ground truths are computed at the end of each 25K step for every pair present in the test set. We used top-1 predictions to report conversion accuracy. We utilized the Python package named ccbmlib [49] to facilitate the generation of similarity value distributions of the all fingerprints.

3 Results and discussion

Conversion accuracy of each structural fingerprint to a unique molecular representation is illustrated in Figure 1 . As is apparent from Figure 1, SMILES conversion demonstrated favorable results in accuracy than SELFIES. In both translation attempts, top performing molecular representation became ECFP4, whereas the worst performance is observed in MACCS, omitting the ECFP0. ECFP0 tries to represent 5 million molecules by using only 100 tokens that are overgeneral and did not work in this translation context. Also, sparse versions perform better than hashed variants of the same fingerprints.

The performances of the structural fingerprints separately for SMILES and SELFIES prediction showed different dynamics during training. Convergence has achieved relatively at lower steps at SMILES translations than SELFIES. SMILES grammatical structure is easily learned, which allowed to compensate the fragility of the representation. On the other side, the drop in overall accuracy and the necessity of larger step size to reach convergence indicated that the correlations between fingerprint and SELFIES tokens are weaker than SMILES tokens. The Avalon's performance in SELFIES prediction broke

the general performance trend that might be attributed to the unusual cumulative distribution function (CDF) of Avalon.

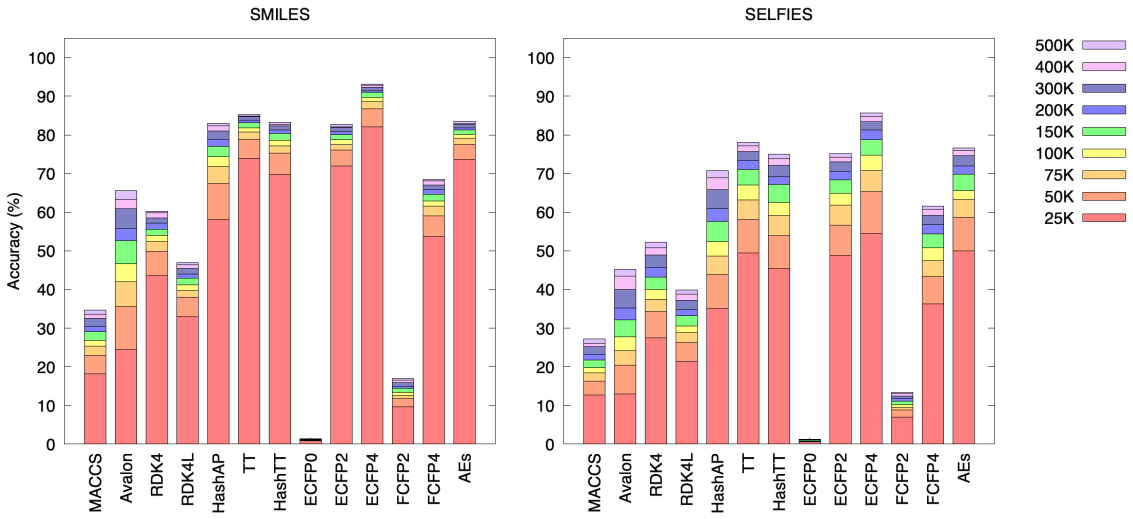

Figure 1: Conversion accuracy of each structural fingerprint to SMILES and SELFIES is demonstrated in cumulative column-stacked bar plot. The results are based on the Tanimoto exactness computed periodically during training with ECFP of radius 1 and dimension 2048.

Mean Tanimoto score is quite critical since it reveals the overall conversion quality. However, similarity metrics in general indicate different meanings in different fingerprints. It is therefore unlikely to rationalize a specific similarity value as a performance evaluation indicator for various fingerprints. A global comparison of all fingerprints within a fair framework is only possible when the similarity value corresponding to a reference significance score is also presented. For this purpose, we generated the CDFs of each fingerprint and found Tc values having a significance of 0.99. The Figure 2 illustrates mean Tc scores (vertical lines) within the training step interval [25K-500K] coupled with a fixed significance score (horizontal lines).

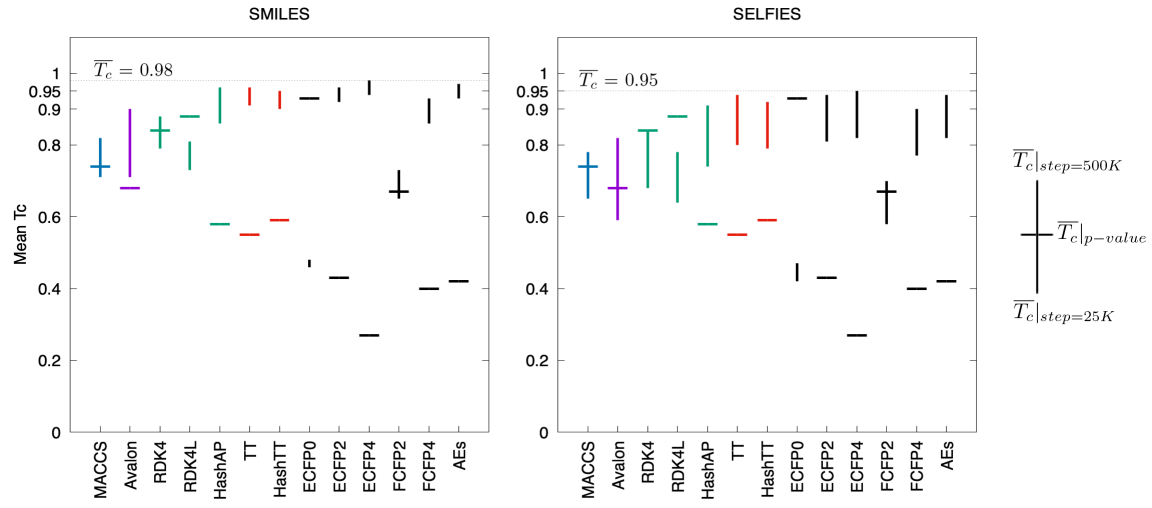

Figure 2: Mean Tanimoto coefficients are given for each type of conversion along with the reference significance score to assess the real performance of structural fingerprints. Horizontal lines represent similarity values of each fingerprint corresponding to a p-value of 0.01. Vertical lines show the continuum, which starts at 25K step and ends with convergence.

Lower Tc values relative to reference significance score, and higher mean Tc values at convergence were

the characteristics of high-performing fingerprints. Based on the Figure 2, ECFP4-SMILES conversion yielded the best overall result. Atomic environments was the runner-up high-performing fingerprint, having a mean Tc of 0.97. The performance of HashAP, TT, and HashTT were comparable to AEs, with mean Tc scores 0.96, 0.96, and 0.95, respectively. RDKit variants-SELFIES conversion performed poorly relative to the other path-based fingerprints.

Each prediction might have introduced a bias to the results if Tanimoto score is computed with the same fingerprint used in the original model. For this reason, multiple fingerprints, as described in Table 1, are utilized, to minimize the selection bias fingerprints provide. Tanimoto exactness of each model is computed across 15 different fingerprints, and presented as a matrix in Figure 3. This approach was important to our assessment, as it decoupled the robustness of the models from the effectiveness and bias of the fingerprints. The enhanced prediction accuracies of MACCS, RDK4, RDK4-L, and ECFP2 fingerprints clearly confirm the existence of fingerprint dependency of results. The Figure 3 highlighted the high performance and the robustness of ECFP4-SMILES model. The true performance of each model, averaged over 15 fingerprints, is presented in Table 2. Ultimately, our top performing models such as ECFP4, TT and hashed variant, HashAP, ECFP2 and AEs remain neutral regardless of the choice of similarity metric.

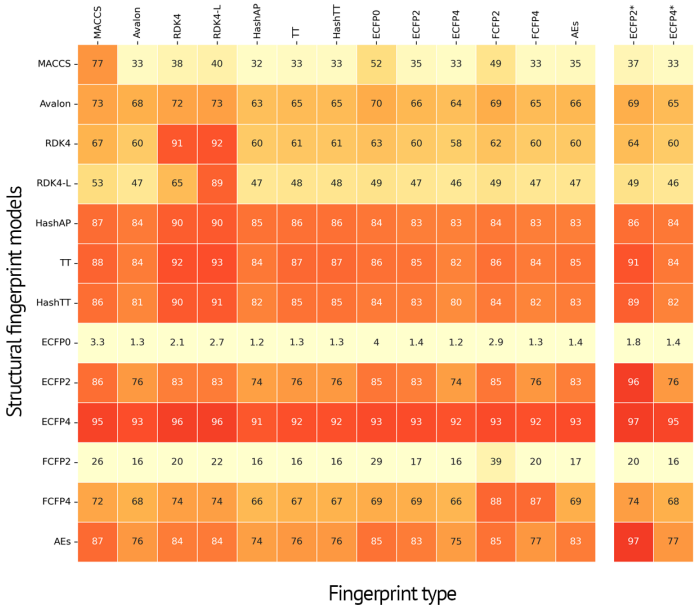

Figure 3: Tanimoto exactness (%) of each fingerprint transformation to SMILES computed across 15 different fingerprint encodings. The results show the robustness and the effect of selection bias.

Table 2: Overall performance (%) of fingerprint decoders computed as the average Tanimoto exactness score across 15 fingerprints.

	MACCS	Avalon	RDK4	RDK4L	HashAP	TT	HashTT	ECFP0	ECFP2	ECFP4	FCFP2	FCFP4	AEs
SMILES	0.40	0.67	0.65	0.52	0.85	0.87	0.85	0.02	0.81	0.94	0.20	0.72	0.81
SELFIES	0.31	0.47	0.57	0.44	0.73	0.79	0.76	0.02	0.74	0.86	0.16	0.65	0.75

A complete breakdown of top-1 accuracy results over 50K test set for the top performer structural fingerprints are presented in Table 3. Total accuracy is given based on Tanimoto exactness. We further breakdown the accuracy by simple string comparison. Identical structures based on Tanimoto metric can be categorized, depending on whether they are sourced from identical strings, stereochemistry, canonicalization, or others related to chain length and symmetry properties. Invalidity rates and mean Tanimoto scores are also reported in Table 3.

A large fraction of our test set, approx. %30, incorporates stereochemistry. The results indicate that the models account for stereochemical information, yet struggle to achieve accurate picture of relative atom orientations. Stereochemistry errors on the test set was about 20 percent for the best performing fingerprints. We examined stereochemically inconsistent predictions to see if they are string exact to the ground truths by removing stereochemical information. We found that in most cases models treat stereochemistry in reverse forms as cis/trans or clockwise/anti-clockwise. There were also predictions featuring stereochemistry even though ground truths had no stereo-center, or vice versa.

Our dataset was not subjected to canonicalization prior to training to ensure full capacity of the SMILES representation. Our models were able to produce non-canonical instances of the ground truth SMILES. The rates of predicting chemically-equivalent SMILES vary from 1.6 to 4.8 percent depending on the fingerprint type. Kekule forms play an important role in non-canonical predictions since switches in kekule representations can alter the SMILES enumerations. With regard to invalidity rates, SELFIES, as expected, provided a totally robust conversions with no invalid cases. SMILES performed comparably well with invalidity rates of about %0.2-0.3. Representative predictions displaying changes in stereochemistry, kekule forms, and enumerations are given in Supplementary table 2.

Table 3: A detailed breakdown (%) of top-1 accuracy on 50K test set for the top performer structural fingerprints belong to five sub-categories. All components are given..

Representation	Components	MACCS	Avalon	HashAP	TT	AEs	ECFP4
SMILES	$T_c = 1.0$	34.7	65.6	83.1	85.2	83.5	93.1
	String exact	22.3	44.7	58.7	57.8	52.1	64.6
	Stereo	8.2	14.9	19.2	19.2	18.0	21.2
	Non-canonical	1.6	3.5	4.3	4.2	3.7	4.8
	Others	2.6	2.6	0.8	4.0	9.6	2.5
	Invalid	0.2	0.4	0.3	0.3	0.3	0.2
	$\overline{T_c}$	81.9	90.5	95.5	96.3	96.7	98.1
SELFIES	$T_c = 1.0$	27.2	45.2	70.7	78.0	76.6	85.6
	String exact	17.7	31.3	50.9	54.0	49.1	60.5
	Stereo	5.9	9.3	15.2	16.7	19.9	18.5
	Non-canonical	1.5	2.8	4.0	4.1	3.6	4.7
	Others	2.2	1.7	0.6	3.3	8.0	1.9
	Invalid	no invalid predictions					
	$\overline{T_c}$	77.8	81.5	90.7	93.9	94.4	95.1

Translation-based models require the studying of relationships between translated pairs in more precise quantitative ways. In order to establish a thorough explanation of the model, we evaluated the correlated features obtained by both integrated gradients and attention weights as illustrated in Figure 4. As a form of gradient-based feature importance measure, integrated gradients reveal relevant features more reliably than the attention weights. Recent findings showed that attention weights are often uncorrelated with gradient-based methods [50, 51]. Thus, we recognized attention weights as valuable stand-alone supplementary tool for addressing the interpretability problem. Interpreting attribution matrices for each combination is highly intricate, however, there is an explainable path between the AEs and the reconstruction of SMILES string.

The matrices in Figure 4 can be interpreted in two ways. Firstly, the column-wise approach reflects the effect of an input feature over the whole prediction. The high-attribution AEs at positions 9 and 11 were the most salient fragments for predicting the SMILES sub-string of nitro group. Especially, the AE at position 11 with a radius 0 made decisive contribution specifically to the oxygens of nitro group because, here, the negatively charged oxygen is in resonance with the geminal oxygen. Secondly, the row-wise approach reflects the salient input features attributing to a specific part of the prediction. For example, the higher attention values at the row of chlorine atom (Figure 4c) highlighted three atomic

environments, all containing chlorine, including as central atoms at radius 0 and 1.

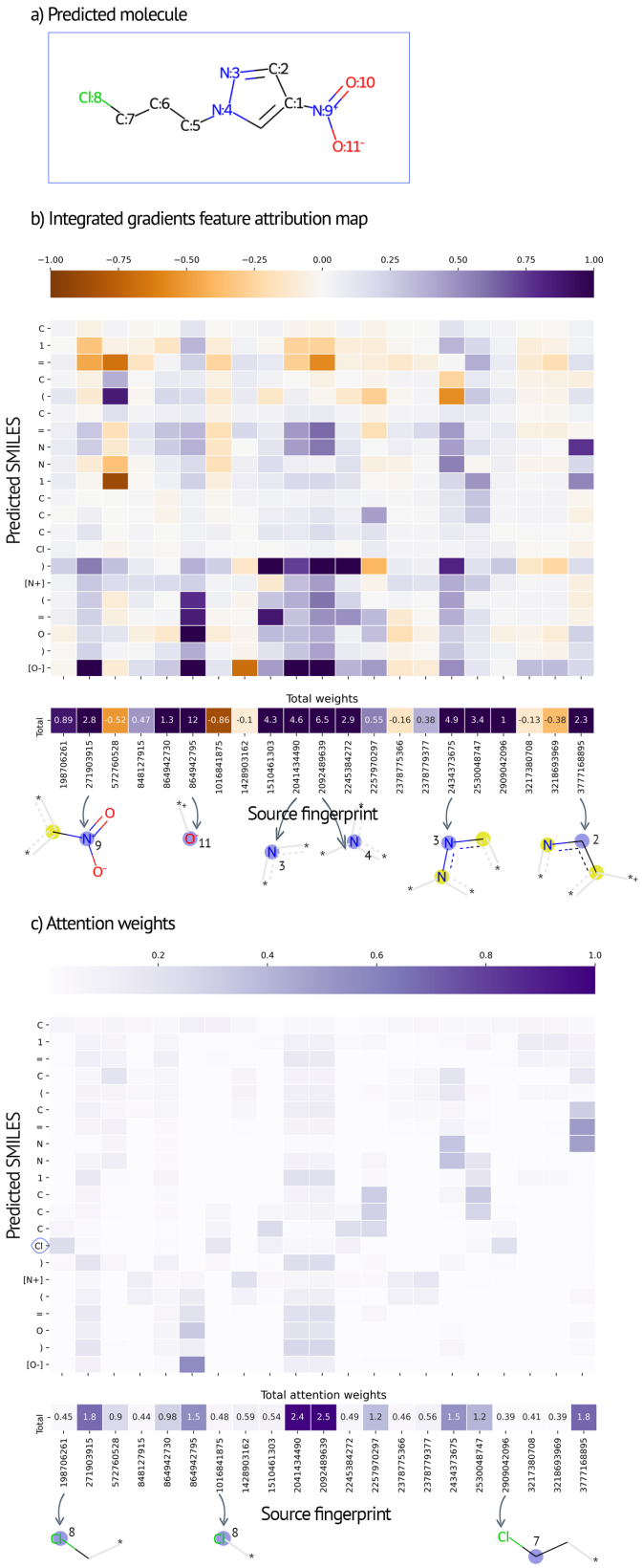

Figure 4: Correlated features of the a) predicted SMILES obtained by b) integrated gradients and c) attention weight matrices are illustrated.

We have exploited the idea of structural fingerprints as alternatives to unique molecular representations. We have successfully rebuild molecules with a high level of precision, precisely higher than 90% for top performing fingerprints. As a result, structural fingerprints come into play as strong representational tools in chemistry related NLP applications after restoring the connectivity information which is lost during the fingerprint transformation. Our diverse selection of fingerprints have provided an unbiased examination of the overall conversion performance. Atom environments, ECFP4, topological torsion and atom-pairs fingerprints have been presented as ideal candidates for developing NLP tools with molecules.

A complete breakdown of accuracies per fingerprint class has been presented in detail. Such an analysis provided invaluable insights into critical factors effecting the conversion process such as stereochemistry as a noticeable limitation. Since the model have struggled to treat stereochemistry, more research is required to fully address this issue. We have assessed the interpretability of our conversion approach by evaluating methods that compute and extract the most salient features for a prediction. Attribution maps have revealed that the model focuses on the right fragments to reconstruct the molecule. We believe our findings could help improve the quality of the outcomes by offering ways to develop more efficient chemical models in generative and translational fields.

4 Data Availability

The data that support the findings of this study are generated by using RDKit software tools and are available in the MolForge GitHub repo: https://github.com/knu-lcbc/MolForge. Source data are provided with this paper.

5 Code Availability

The source code of this work and associated trained models are available at the MolForge GitHub repo: https://github.com/knu-lcbc/MolForge

6 Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (Nos. NRF-2019M3E5D4066898, NRF-2018R1C1B600543513 and NRF-2020M3A9G7103933 to I.A. and J.L.). This work was also supported by the Korea Environment Industry & Technology Institute (KEITI) through the Technology Development Project for Safety Management of Household Chemical Products, funded by the Korea Ministry of Environment (MOE) (KEITI:2020002960002 and NTIS:1485017120 to U.V.U. and J.L.).

References

[1] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comp. Sci.* **28**, 31–36 (1988).

[2] ChemAxon Extended SMILES and SMARTS CXSMILES and CXSMARTS Documentation. https://docs.chemaxon.com/display/docs/chemaxon-extended-smiles-and-smarts-cxsmiles-and-cxsmarts.md#src-1806633_ChemAxonExtendedSMILESandSMARTS-CXSMILESandCXSMARTS-Fragmentgrouping. Accessed: 10 Feburary 2022.

[3] OpenSMILES. Home Page https://opensmiles.org. Accessed: 10 December 2021.

[4] Lin, T.-S. *et al.* Bigsmiles: A structurally-based line notation for describing macromolecules. *ACS Cent. Sci.* **5**, 1523–1531 (2019). URL https://doi.org/10.1021/acscentsci.9b00476. PMID: 31572779.

[5] Drefahl, A. CurlySMILES: A chemical language to customize and annotate encodings of molecular and nanodevice structures. *J. Cheminformatics* **3**, 1–7 (2011).

[6] Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. chem. doc.* **5**, 107–113 (1965). URL https://doi.org/10.1021/c160017a018. https://doi.org/10.1021/c160017a018.

[7] Weininger, D., Weininger, A. & Weininger, J. L. Smiles. 2. algorithm for generation of unique smiles notation. *J. Chem. Inf. Comp. Sci.* **29**, 97–101 (1989). URL https://doi.org/10.1021/ci00062a008. https://doi.org/10.1021/ci00062a008.

[8] O'Boyle, N. M. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *J. Cheminformatics* **4**, 1–14 (2012).

[9] Schneider, N., Sayle, R. A. & Landrum, G. A. Get your atoms in order—an open-source implementation of a novel and robust molecular canonicalization algorithm. *J. Chem. Inf. Model.* **55**, 2111–2120 (2015). URL https://doi.org/10.1021/acs.jcim.5b00543. PMID: 26441310, https://doi.org/10.1021/acs.jcim.5b00543.

[10] Wiswesser, W. J. How the WLN Began in 1949 and How It Might Be in 1999. *J. Chem. Inf. Model.* **22**, 88–93 (1982).

[11] Homer, R. W., Swanson, J., Jilek, R. J., Hurst, T. & Clark, R. D. SYBYL line notation (SLN): A single notation to represent chemical structures, queries, reactions, and virtual libraries. *J. Chem. Inf. Model.* **48**, 2294–2307 (2008).

[12] Heller, S. InChI – the worldwide chemical structure standard. *J. Cheminformatics* **6**, 1–9 (2014).

[13] Liu, B. *et al.* Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Central Science* **3**, 1103–1113 (2017).

[14] Skalic, M., Jiménez, J., Sabbadin, D. & De Fabritiis, G. Shape-Based Generative Modeling for de Novo Drug Design. *J. Chem. Inf. Model.* **59**, 1205–1214 (2019).

[15] Lin, K., Xu, Y., Pei, J. & Lai, L. Automatic retrosynthetic route planning using template-free models. *Chem. Sci.* **11**, 3355–3364 (2020).

[16] Kwon, Y. & Lee, J. MolFinder: an evolutionary algorithm for the global optimization of molecular properties and the extensive exploration of chemical space using SMILES. *J. Cheminformatics* **13**, 1–14 (2021). URL https://doi.org/10.1186/s13321-021-00501-7.

[17] Schwaller, P. *et al.* Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).

[18] Cao, N. D. & Kipf, T. Molgan: An implicit generative model for small molecular graphs (2018). 1805.11973.

[19] Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. Guacamol: Benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019). URL https://doi.org/10.1021/acs.jcim.8b00839. PMID: 30887799.

[20] Lim, J., Ryu, S., Kim, J. W. & Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J. Cheminformatics* **10**, 31 (2018). URL https://doi.org/10.1186/s13321-018-0286-7.

[21] Zheng, S., Rao, J., Zhang, Z., Xu, J. & Yang, Y. Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *J. Chem. Inf. Model.* **60**, 47–55 (2020).

[22] Duan, H., Wang, L., Zhang, C., Guo, L. & Li, J. Retrosynthesis with attention-based NMT model and chemical analysis of "wrong" predictions. *RSC Advances* **10**, 1371–1378 (2020).

[23] Kim, E., Lee, D., Kwon, Y., Park, M. S. & Choi, Y. S. Valid, Plausible, and Diverse Retrosynthesis Using Tied Two-Way Transformers with Latent Variables. *J. Chem. Inf. Model.* **61**, 123–133 (2021).

[24] Bilsland, A. E., McAulay, K., West, R., Pugliese, A. & Bower, J. Automated Generation of Novel Fragments Using Screening Data, a Dual SMILES Autoencoder, Transfer Learning and Syntax Correction. *J. Chem. Inf. Model.* **61**, 2547–2559 (2021).

[25] O'Boyle, N. M. & Dalke, A. DeepSMILES: An adaptation of SMILES for use in machine-learning of chemical structures. *ChemRxiv* 1–9 (2018).

[26] Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **1**, 045024 (2020).

[27] Kovács, D. P., McCorkindale, W. & Lee, A. A. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nat. Commun.* **12**, 1–9 (2021). URL http://dx.doi.org/10.1038/s41467-021-21895-w.

[28] Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. *34th International Conference on Machine Learning, ICML 2017* **7**, 5109–5118 (2017). 1703.01365.

[29] Wiegreffe, S. & Pinter, Y. Attention is not not explanation. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* 11–20 (2020). 1908.04626.

[30] Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In Precup, D. & Teh, Y. W. (eds.) *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, 3319–3328 (PMLR, 2017). URL https://proceedings.mlr.press/v70/sundararajan17a.html.

[31] Tu, Z. & Coley, C. W. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *arXiv:2110.09681 [cs]* (2021). URL http://arxiv.org/abs/2110.09681.

[32] Jaegle, A. *et al.* Perceiver: General Perception with Iterative Attention. Preprint at http://arxiv.org/abs/2103.03206 (2021).

[33] Ucak, U. V., Ashyrmamatov, I., Ko, J. & Lee, J. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nat. Commun.* **13**, 1186 (2022). URL https://doi.org/10.1038/s41467-022-28857-w.

[34] Ucak, U. V., Kang, T., Ko, J. & Lee, J. Substructure-based neural machine translation for retrosynthetic prediction. *J. Cheminformatics* **13**, 1–15 (2021).

[35] Landrum, G. Rdkit: Open-source cheminformatics software (2016).

[36] Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comp. Sci.* **42**, 1273–1280 (2002).

[37] James, C. A., Weininger, D. & Delany, J. D. Daylight theory manual. daylight chemical information systems inc. (2002).

[38] Gedeck, P., Rohde, B. & Bartels, C. QSAR - How good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J. Chem. Inf. Model.* **46**, 1924–1936 (2006).

[39] Smith, D. H., Carhart, R. E. & Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comp. Sci.* **25**, 64–73 (1985).

[40] Nilakantan, R., Bauman, N., Venkataraghavan, R. & Dixon, J. S. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comp. Sci.* **27**, 82–85 (1987).

[41] Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

[42] Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017-Decem**, 5999–6009 (2017).

[43] Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2016).

[44] Bolton, E. E., Wang, Y., Thiessen, P. A. & Bryant, S. H. Chapter 12 - pubchem: Integrated platform of small molecules and biological activities. In *Ann. Rep. Comp. Chem.*, vol. 4 of *Annual Reports in Computational Chemistry*, 217–241 (Elsevier, 2008).

[45] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. *et al.* (eds.) *Adv. Neural Inf. Process. Syst.*, 8024–8035 (Curran Associates, Inc., 2019).

[46] Rajbhandari, S., Rasley, J., Ruwase, O. & He, Y. Zero: Memory optimizations toward training trillion parameter models. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC* **2020-Novem**, 1–24 (2020). 1910.02054.

[47] Karpov, P., Godin, G. & Tetko, I. V. A transformer model for retrosynthesis. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*, 817–830 (Springer International Publishing, Cham, 2019).

[48] Loshchilov, I. & Hutter, F. SGDR: Stochastic gradient descent with warm restarts. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings* 1–16 (2017).

[49] Vogt, M. & Bajorath, J. Ccbmlib - A python package for modeling tanimoto similarity value distributions. *F1000Research* **9** (2020).

[50] Grimsley, C., Mayfield, E. & R.S. Bursten, J. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 1780–1790 (European Language Resources Association, Marseille, France, 2020). URL https://aclanthology.org/2020.lrec-1.220.

[51] Jain, S. & Wallace, B. C. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3543–3556 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019). URL https://aclanthology.org/N19-1357.