Journal of Cheminformatics

RESEARCH

Open Access

Reconstruction of lossless molecular representations from fingerprints

Umit V. Ucak[1], Islambek Ashyrmamatov[2] and Juyong Lee[1,3*]

Abstract

The simplified molecular-input line-entry system (SMILES) is the most prevalent molecular representation used in AI-based chemical applications. However, there are innate limitations associated with the internal structure of SMILES representations. In this context, this study exploits the resolution and robustness of unique molecular representations, i.e., SMILES and SELFIES (SELF-referencIng Embedded strings), reconstructed from a set of structural fingerprints, which are proposed and used herein as vital representational tools for chemical and natural language processing (NLP) applications. This is achieved by restoring the connectivity information lost during fingerprint transformation with high accuracy. Notably, the results reveal that seemingly irreversible molecule-to-fingerprint conversion is feasible. More specifically, four structural fingerprints, extended connectivity, topological torsion, atom pairs, and atomic environments can be used as inputs and outputs of chemical NLP applications. Therefore, this comprehensive study addresses the major limitation of structural fingerprints that precludes their use in NLP models. Our findings will facilitate the development of text- or fingerprint-based chemoinformatic models for generative and translational tasks.

Keywords Fingerprints, SMILES, SELFIES, Neural Machine Translation

Introduction

The Simplified Molecular-Input Line-Entry System (SMILES) [1] is the most widely used linear representation for describing chemical structures. In SMILES, several simple rules are used to convert a chemical structure into a character string. This allows multiple unique SMILES strings to be used to represent molecules. Since its inception, SMILES has undergone various extensions [2–5], and among them, canonicalization algorithms, the integration of isotopism and the addition of stereochemical information (isomeric SMILES) are major milestones [6–9].

Although the simplified line notation of SMILES is superior to other one-dimensional representation schemes such as the Wiswesser Line Notation (WLN) [10], SYBYL line notation (SLN) [11], and International Chemical Identifier (InChI) [12], its internal structure leads to several problems when used in natural language processing (NLP) algorithms [13–15]. SMILES-based neural machine translation (NMT) models are prone to generate invalid SMILES strings [16, 17], which can be attributed to the fragile grammar (i.e., a strong dependence between tokens). Most notably, SMILES related issues seen in NMT models also occur in the most commonly used deep generative models such as variational autoencoders and generative adversarial networks that generate SMILES strings [18–21]. Because these models formulate predictions for one character at a time, a single-character alteration often suffices to invalidate an entire SMILES string. In addition, novel valid SMILES

*Correspondence:
Juyong Lee
nicole23@snu.ac.kr
[1] Research Institute of Pharmaceutical Science, College of Pharmacy, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
[2] Department of Chemistry, Kangwon National University, Chuncheon 24341, Republic of Korea
[3] Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

© The Author(s) 2023, corrected publication 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

strings generated by AI-models are not guaranteed to be chemically valid.

To address the aforementioned problem, several attempts have been made to ensure the syntactic and chemical validity of the SMILES predictions [22–27]. The challenges posed by the SMILES syntax have prompted the development of alternative syntaxes such as DeepSMILES [28] and SELF-referencIng Embedded strings (SELFIES) [29]. SELFIES is a new way of representing molecules that is receiving increasing attention from the scientific community and is being actively developed. Unlike SMILES, SELFIES units are enclosed by square brackets where no cuts is allowed within during tokenization, ensuring the generation of syntactically and semantically valid graphs. Multiple benchmarks have demonstrated that SELFIES outperforms alternative approaches in terms of validity and diversity of generated molecules.

The most commonly used NLP methods in chemistry are text generation and NMT. Particularly, these NLP methods aim to generate meaningful sequences from meaningful tokens. Therefore, tokenization is a pivotal preprocessing step in many NLP tasks. SMILES strings are meaningful as a whole, and any tokenization procedure must dissect these strings arbitrarily. From a chemist's perspective, the atom-wise or character-wise tokenization of SMILES strings does not produce fully interpretable tokens. This is because many characters in SMILES strings correspond to topological characteristics, such as the digits in ring opening and closures, or parenthesis enclosing branches, that do not correspond to physical entities. In addition, most SMILES tokens are indistinguishable owing to their repetitiveness and simplicity. Considering that the primary design purpose of SMILES is to serve as a universal exchange format, it is understandable that interpretable insights cannot be derived from tokenization.

Despite the challenges mentioned above, SMILES representation plays a prominent role in chemical language modeling because they are preferred over the generation of a set of fingerprint features (incomplete description of a molecule), as the latter would require extensive database searches to identify matches and is therefore not desired. There are currently few studies in the literature using fingerprints as model outputs. In the field of molecular generation, Kadurin et al. [30] first proposed the use of an adversarial autoencoder to generate novel compounds for cancer treatment. They used MACCS keys (166-bit long binary vectors) as the input–output data structure, together with the inhibition concentration of the molecules. The model was trained on cancer cell line assay data, and the generated fingerprints were used to screen compounds on PubChem to identify candidate molecules with anticancer properties. In the field of reaction route planning, our previous works have shown that fragmental and topological descriptors can be effectively used as the input–output data structure in end-to-end NMT pipelines [31, 32].

Furthermore, interpretability necessitates the existence of meaningful tokens because NLP models tend to learn the relationships between these tokens. Thus, the interpretability of an individual token is highly desirable. However, the chemical interpretability of conventional NLP methods is hampered as SMILES representations are not fully interpretable from a chemical perspective. Indeed, SMILES is a highly efficient system for capturing information about molecular structures, and issues arise only when SMILES are tokenized. This contradicts the recent statement by Tu et al. [33], who propounded that SMILES is inefficient in capturing structural information because SMILES augmentation can provide additional performance gains [34]. As an alternative to SMILES representations, molecular fingerprints and substructural keys can be employed. They are designed to capture chemical features, concepts, or structural patterns, yielding an interpretable set of tokens suitable for NLP applications.

Several studies have recently explored the conversion of the extended-connectivity fingerprint ECFP [35] to SMILES representation. Within the context of data sharing and confidentiality, Le et al. [36] suggested the NeuralDecipher model. The model deduces the molecular structure of compounds using a two-step process involving a feedforward neural network model that predicts a compact vector representation of the compounds given their ECFP, and a pre-trained model that converts this representation into SMILES. NeuralDecipher showed a success rate of 69%. Kwon et al. [37] proposed a data-driven evolutionary molecular design methodology using a genetic algorithm, a recurrent neural network (RNN), and a deep neural network to evolve ECFP vectors of seed molecules and reconstruct chemically valid molecular structures in SMILES format. The model showed a success rate of 62.4%. Cofala and Kramer also used a genetic algorithm to demonstrate the ability to reconstruct molecules similar to the specified target or even the original molecule from ECFP representations [38]. Their method also showed a reconstruction rate of 58% ∼ 68%. Overall, these studies show the potential for using ECFPs as a starting point for generating molecular structures in SMILES representation, either through direct prediction or through genetic algorithms and evolutionary design techniques.

Because the construction of molecular fingerprints is a lossy procedure, the use of fingerprints leads to the generation of stand-alone interpretable tokens. Moreover,

fingerprints are well suited to the attention mechanism because attention is a permutation-invariant operation [39]. Furthermore, attention-based models, such as transformers, can handle the unconnected features of fingerprints [31, 32]. Thus, we assessed the efficiency of the back-conversion of fingerprints to molecules to overcome the significant limitations of structural fingerprints that preclude their implementation in NLP models. For this purpose, we employed a translation-based system, namely the transformer architecture, to decode fingerprints accurately into lossless molecular representations. We aim to demonstrate that the reconstruction of molecules from molecular fingerprints is a practical and highly accurate approach for various chemical applications. Finally, we illustrate our approach using thirteen structural fingerprint examples, classified into five main categories. We show that certain fingerprints can be used directly in an NLP setting as alternatives to SMILES and SELFIES representations.

Results and discussion

Structural fingerprint representations

Structural fingerprints were obtained from RDKit [40] implementations. They can be classified into five main groups, as reported in Table 1 along with the corresponding sequence lengths and vocabulary size information. We generated thirteen different fingerprints for our analysis. Binary variants of the selected fingerprints were hashed to a fixed size of 2048, except for Avalon. Fingerprints were optimized based on their parameters to yield similar sequence lengths when necessary. We omitted sparse versions of atom pairs and ECFP4 from this calculation because the vocabulary space covered, and thus the token size, was considerably large.

1. *Predefined substructure* MACCS keys [41] converts a molecule into a bit vector with a fixed size of 166, in which each bit records the presence of a feature obtained from a predefined dictionary of SMARTS patterns [42].
2. *Paths and feature classes* The Avalon enumerates paths and feature classes. We refer the reader to Gedeck et al. [43] for a thorough explanation of paths and feature classes covered.
3. *Path-based* The RDKit fingerprint is very similar to the Daylight fingerprint [42]. Hashed branched and linear subgraphs of size 4 were used. In both cases, the minPath and maxPath parameters were set to two and four, respectively. The hashed variant of the

Table 1 Translation-related statistics regarding the domain-specific datasets generated by the structural fingerprints used for the performance analysis, 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

atom pair fingerprint encodes all pairs of atoms with their environments and their bond distances [44]. Here, it was used with the following parameters: minLength=1, and maxLength=6.

4. *4-atom-paths* Topological torsion [45] encodes sequences of four bonded atoms, so that the generated set of substructures has a local character. It was used along with its hashed variants.
5. *Circular* ECFPx [35] enumerates circular atom environments, defined as topological neighborhood fragments, up to a selected radius (x). The set of all circular fragments, that is atom environments, is denoted as AEs. Feature-class fingerprints FCFPx include pharmacophoric features as invariants.

Model overview

In this study, we employed Transformer [46], a model architecture with a multi-head attention mechanism for each unit. Transformer-based models can achieve highly successful translation quality compared to generic seq-2-seq methods [13, 16, 17, 32], thanks to attention units allowing the model to learn global dependencies between inputs and outputs. In addition, the attention mechanism eliminates the dependence on the order of the input sequence. Therefore, the models yield the same sequence of outputs regardless of the spatial connections between the tokens. This property of the attention mechanism renders Transformer-based models suitable for investigating fingerprint-to-molecule conversion.

Translation-based algorithms require a large corpus of diverse translation pairs for an effective translation. For this purpose, we selected the ChEMBL [47] (2.08 M) dataset and extended it to include PubChem [48] compounds by maximizing the variety of atom-types based on the atomic environments. Atom-types refer to the features obtained by sparse ECFP of radius zero. This resulted in 5,050,000 small- and medium-sized molecules (those with 50 heavy atoms or less) that maximally represent available drug-like chemical space, considering that most current drugs are small organic molecules of natural or synthetic origin [49]. Figure 1 illustrates the normalized molecular weight distribution of our training dataset, along with several drug and natural product libraries. From this large pool, we randomly selected and separated 50,000 molecules for testing purposes. To obtain more realistic results, we used a challenging dataset, which retains the stereochemical information. However, we note that most of fingerprints in RDKit do not account for stereochemistry.

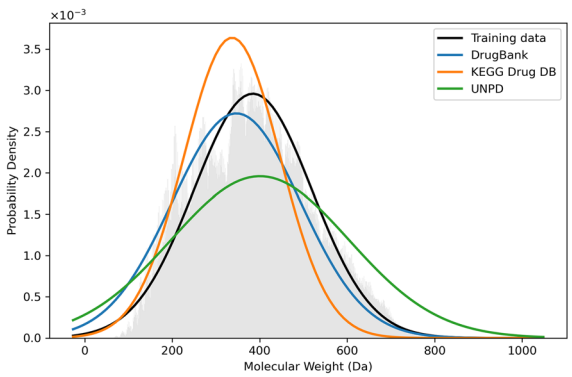

Fig. 1 The normalized molecular weight distribution of our training dataset along with several drug and natural product libraries such as KEGG DRUG Database, DRUGBANK and Universal Natural Product Database (UNPD). The training dataset consisted of five million small- and medium-sized molecules of approximately 50 heavy atoms or less that maximally represent available drug-like chemical space

Model performance

The conversion accuracy of each structural fingerprint into unique molecular representations, namely SMILES and SELFIES strings, is illustrated in Fig. 2. The SMILES conversion demonstrated more favorable results in terms of accuracy compared to the SELFIES conversion. In both translation attempts, the top-performing molecular representation was ECFP4. The highest accuracy reached 93.1%, indicating that the model reflects an optimal level of fragment specificity within a fixed-length vector. Alongside the performance of ECFP4, TT, HashAP, and AEs yielded competitive accuracy, whereas the worst performance was observed in MACCS, omitting ECFP0. It should be noted that ECFP0 attempts to represent five million molecules using only 100 tokens so that the produced fragments are overly-general. ECFP0 did not function well in this translation task. Additionally, sparse versions perform better than hashed variants of the same fingerprint, as in the cases of the TT-HashTT and AEs-ECFP2 pairs.

The performances of the structural fingerprints for the SMILES and SELFIES reconstruction showed different dynamics during training. Near-convergence was achieved at a lower number of steps for SMILES compared to SELFIES (learning was quicker, as evident from the relative bar heights after 100 K steps; see Fig. 2). Accordingly, the SMILES grammatical structure can be easily learned, compensating for the fragility of the representation. On the other hand, the decrease in the overall accuracy and the necessity for a more significant step size to reach convergence of SELFIES indicated that the correlations between the fingerprints and SELFIES tokens were weaker than those between the fingerprints and SMILES tokens. The performance of Avalon in the

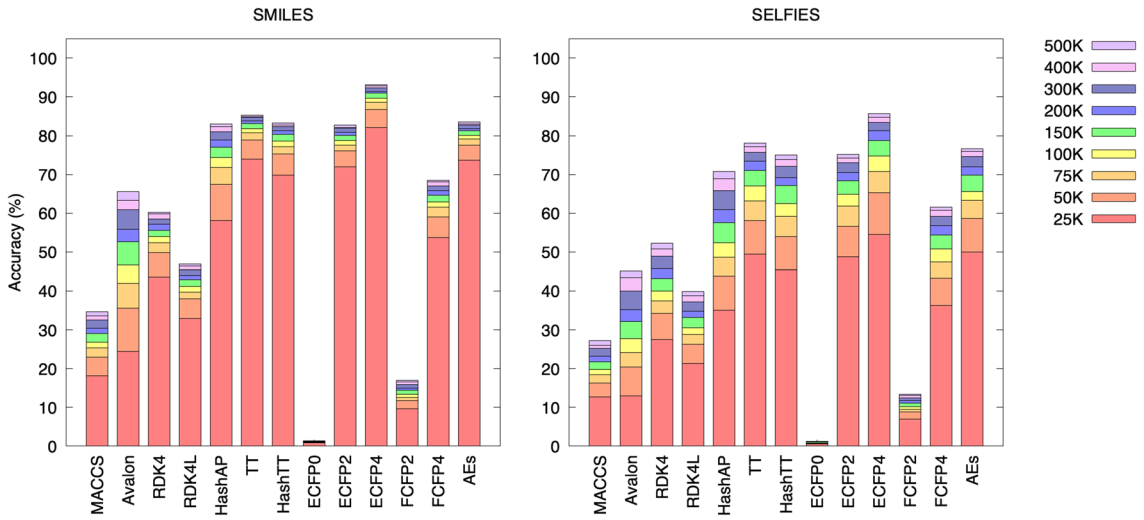

Fig. 2 Conversion accuracy of each structural fingerprint to SMILES (left) and SELFIES (right) demonstrated using cumulative column-stacked bar plots along with the number of training steps, from 25K to 500K steps (right color map). The results are based on the Tanimoto exactness, the percentage of $T_c = 1.0$ reconstructions, computed periodically during training with a sparse form of an extended connectivity fingerprint (ECFP) of radius 1. Each bar represents the progress over the iterations for the given step intervals

SELFIES prediction differed from the general performance trend, which may be due to its unusual cumulative distribution function (CDF).

The mean Tanimoto score (T_c) is important as it reflects the overall conversion quality. However, similarity metrics generally have different scales for different types of fingerprints. Therefore, it is not ideal 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 possible only when the similarity value corresponding to a reference significance score is presented. Considering this, we generated the CDFs of all fingerprints and obtained T_c values with a significance of 0.99. Figure 3 illustrates the mean T_c scores (vertical lines) within the training step interval [25K-500K] coupled with a fixed p-value of 0.01 (horizontal lines). The small horizontal lines in the Figure 2 were determined for each fingerprint using the method proposed by Vogt and Bajorath [50] to model the distribution of similarity values for various fingerprints in RDKit. These lines represent Tanimoto coefficients for a p-value of 0.01, which allowed us to assess the level of learning.

Lower T_c values for the reference significance score, and higher mean T_c values at convergence were observed as characteristics of high-performing fingerprints (ECFP4, ECFP2, FCFP4, AEs, HashAP, TT and HashTT). As shown in Fig. 3, the ECFP4-SMILES conversion yielded the best overall result, with a mean T_c of 0.98. AEs was the next in terms of performance, having a mean T_c of 0.97. The performances of HashAP, TT, and HashTT were comparable to that of AEs, with mean T_c scores of 0.96, 0.96, and 0.95, respectively. In contrast, the RDKit variants-SELFIES conversion performed poorly relative to the other path-based fingerprints.

Predictive performance is often susceptible to bias if the fingerprints representing the input sequences are used to compute the similarity score. To minimize the selection bias, multiple fingerprints were used, as listed in Table 1. The Tanimoto exactness of each model, the percentage of predictions under the condition that T_c equals unity, was computed across 15 different fingerprints (by including explicit bit vector type of the ECFP2 and ECFP4), and is presented as a matrix in Fig. 4. This approach was essential 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 confirmed the fingerprint dependency of the results. Figure 4 highlights the high performance and robustness of the 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 its hashed variant, HashAP, ECFP2 and AEs, performed similarly regardless of the choice of similarity metric. An analysis of the fingerprint dependency of SELFIES is shown in the Additional File 1: Figure S2.

Breakdown of the top-1 accuracy

A complete breakdown of the top-1 accuracy results over the 50 K test set for the top-performing structural fingerprints is presented in Table 3, wherein the total accuracy

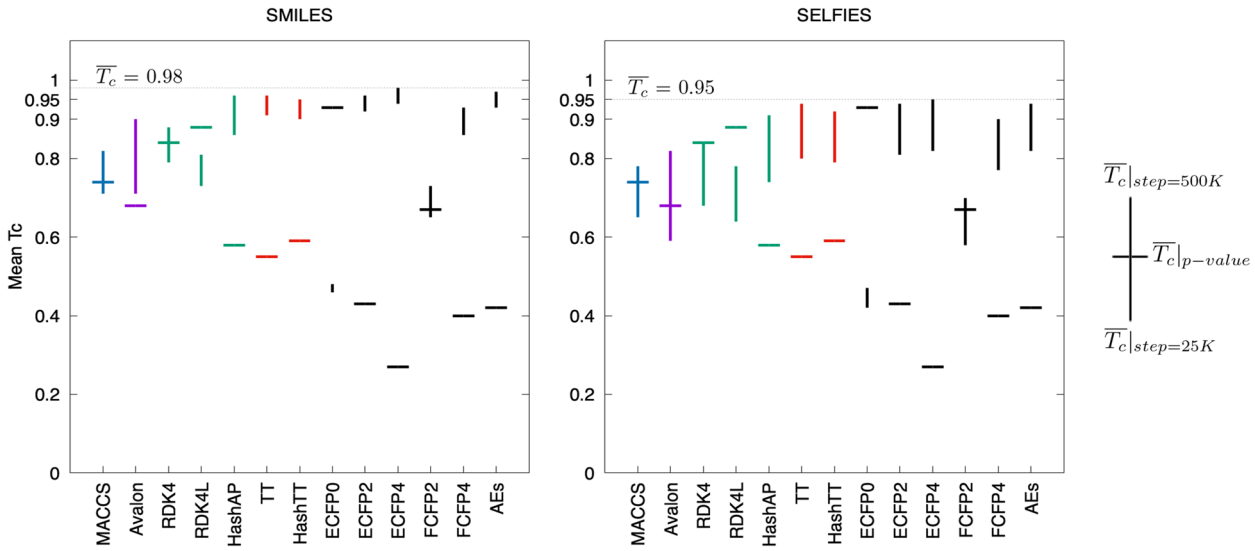

Fig. 3 Mean Tanimoto coefficients for each type of conversion along with the reference significance score to assess the actual performance of structural fingerprints. Horizontal lines represent the similarity values of each fingerprint corresponding to a p-value of 0.01. Vertical lines show the continuum, which starts at 25K step and ends with convergence

is given based on Tanimoto exactness. We further separated the total accuracy into major components, using a simple string comparison. Here, we note that identical structures based on the Tanimoto metric can be categorized depending on whether they are sourced from identical strings, stereochemistry, canonicalization, or other characteristics, including chain length and symmetry properties. The invalidity rates and mean Tanimoto scores are listed in Table 3.

A large fraction of our test set (i.e., ~30%) incorporates stereochemistry, and the obtained results indicate that the models account for stereochemical information. However, they struggle to achieve an accurate picture of relative atom orientations. Indeed, for the best-performing fingerprint, ECFP4, the stereochemical errors equaled ~20%. Therefore, we examined the stereochemically-inconsistent predictions by removing the stereochemical information to determine whether these predictions were string-exact relative to the ground truths. In most cases, the models treat reverse (or opposite) stereochemistry as cis/trans or clockwise/anti-clockwise. Moreover, predictions featuring stereochemistry also existed even when the ground truths possessed no stereocenters, or vice versa.

Our dataset was not subjected to canonicalization before training to investigate the full capacity of the SMILES representation. Our models could produce noncanonical instances of ground-truth SMILES representations, and the rates of predicting chemically equivalent SMILES representations varied from 1.6 to 4.8%, depending on the fingerprint type. In addition, it should

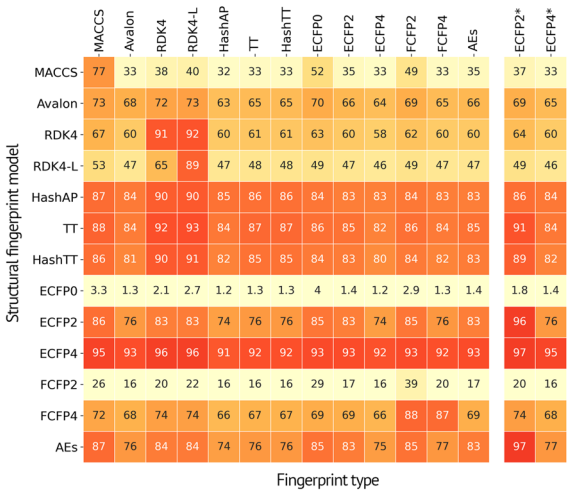

Fig. 4 Percentages of reconstructed SMILES strings from a source fingerprint (y-axis) with $T_c = 1.0$, the Tanimoto exactness, computed with the respective fingerprints (x-axis). The consistent values across a row reflect the robustness and high quality of reconstructed SMILES strings, while significant variations of values represent the fingerprint bias in T_c calculation ECFP2* and ECFP4* represent explicit bit versions

be noted that the Kekule forms play an important role in non-canonical predictions because switches in the Kekule representations can alter SMILES enumerations. SELFIES provided robust conversions regarding invalidity rates, with no invalid cases, as expected. Furthermore, SMILES performed comparably well, with only 0.2-—

Table 2 Overall performance (%) of fingerprint decoders, computed as the average Tanimoto exactness score across 15 fingerprintsOverall 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	39.6	67.3	65.2	51.6	85.1	86.6	84.6	1.9	80.8	93.6	20.3	71.7	81.3
SELFIES	31.2	46.6	56.7	44.1	72.6	79.5	76.4	1.6	73.6	86.2	16.3	64.7	75.0

0.3% invalidity rates. Representative predictions displaying the changes in stereochemistry, kekule forms, and enumerations are provided in Additional File 1: Table S1.

Interpretability

Translation-based models require a detailed quantitative study of the relationships between the translated pairs. To establish a thorough explanation of the model, we evaluated the correlated features obtained using the integrated gradients and attention weights, commonly used to explain the relationship between tokens (Fig. 5). As a form of gradient-based feature importance measure, integrated gradients reveal relevant features more reliably than attention weights. Recent findings showed that attention weights are often uncorrelated with gradient-based methods [51, 52]. Therefore, we recognized attention weights as a valuable supplementary tool to address the interpretability problem. Although the interpretation of attribution matrices for each combination is highly intricate, an explainable path exists between the AEs and the reconstruction of the SMILES string.

The matrices shown in Fig. 5 can be interpreted in two ways: First, the column-wise approach reflects the effect of an input feature on the prediction. Based on this approach, our results indicated that the high-attribution AEs at positions 9 and 11 were the most salient fragments for predicting the SMILES substring of the nitro groups (Fig. 5b). In particular, the AE at position 11, with a radius of 0, made a decisive contribution specifically to the oxygen atoms of the nitro group because the negatively charged oxygen is in resonance with the geminal oxygen. Second, the row-wise approach reflects salient input features attributed to a specific part of the prediction. For example, the higher attention values in the row of chlorine atoms (Fig. 5c) highlight three atomic environments, all containing chlorine, including the central atoms at radii 0 and 1.

Conclusion

Structural fingerprints were exploited as alternatives to unique molecular representations. We successfully rebuilt the molecules with a high level of precision, that is, >90% for the top-performing fingerprints. Consequently, structural fingerprints can be used as strong representational tools in chemistry-related NLP applications after restoring the connectivity information lost during fingerprint transformation. Therefore, our diverse selection of fingerprints provided an unbiased examination

Table 3 Detailed breakdown (%) of top-1 accuracy on 50 K test set for the top-performing structural fingerprints belonging to five sub-categories

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

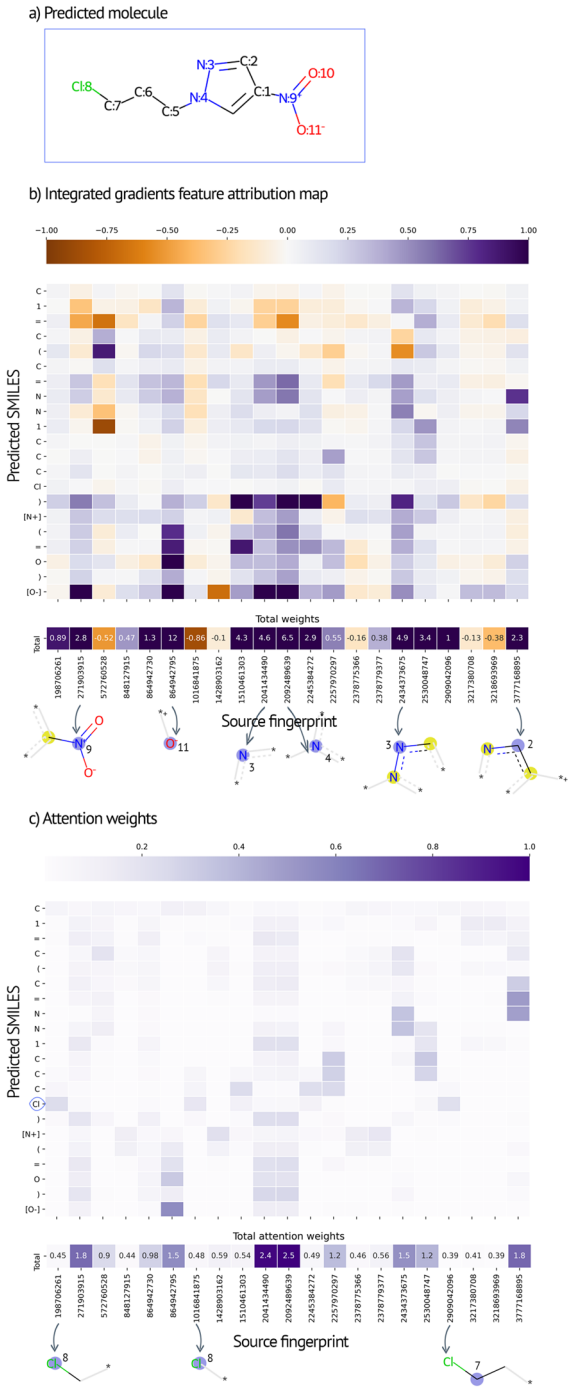

Fig. 5 Correlated features of the **a** predicted SMILES given with atomic indices obtained by **b** integrated gradients and **c** attention weight matrices. The most salient fragments (atom indices attached to the central atoms for easy recognition) are interpreted column-wise and row-wise

of the overall conversion performance. Our results indicated that AEs, ECFP4, topological torsion, and atom-pair fingerprints are ideal candidates for developing NLP tools with molecules.

In this study, a complete breakdown of the accuracy per fingerprint class is presented in detail. Such an analysis provides invaluable insights into the critical factors affecting the conversion process, such as stereochemistry, which was a noticeable limitation of the model proposed herein. As this model has struggled to treat stereochemistry, additional research is required to fully address this issue. Moreover, we assessed the interpretability of our conversion approach by evaluating the methods that compute and extract the most salient features for prediction. The attribution maps revealed that the model focused on the correct fragments for reconstructing the molecule. Finally, our findings could help improve the quality of outcomes by offering ways to develop more efficient chemical models in the fields of deep generative modelling and neural machine translation.

Method

Training

The Pytorch [53] Distributed Data-Parallel Training (DDP) module was employed to train our models. Each model was trained with two GPUs up to a 500 K step, which denotes the number of times the optimizer updates the parameters of the model. The hyperparameters of the models were set similar to ones used in the original Transformer publication [46]. The encoder and decoder used a stack of six identical layers consisting of eight heads with a 512 dimensional multi-head attention mechanism, followed by a 2048 dimensional fully connected feed-forward layer. In contrast, we used a normalization layer before each sub-layer, and with the outputs of the encoder and decoder by following The Annotated Transformer [54]. A dropout layer, with a dropout rate of 0.1, was applied to the output of each sub-layer to avoid overfitting.

Even though the attention mechanism is a permutation-invariant operation and fingerprint features are unconnected, we did not remove the positional encoding in our final models because a previous study [39] stated that it was preferable over non-positional encoding. We had tested this claim by training our models without positional encoding. The results of our models with positional encoding showed slightly better performance (in a range of 0.2–0.9 percent) compared to those without positional encoding, consistent with the previous

findings [39, 55] Though not significant, what considered at positional encoding was the default order of fingerprints features, e.g., index-order for hashed fingerprints. In addition, a zero-redundancy optimizer (ZeRO) [56] with the Adam algorithm was employed to optimize the parameters of the models. This was done to improve the training speed by eliminating memory redundancies during data- and model-parallel training. A negative log-likelihood function was used as the loss function.

We set the number of tokens in one batch to 8000 per GPU. Owing to hardware limitations, this number could not be exceeded. For a fair comparison of the 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 varied according to the fingerprint sequence length. To extend the performance of the standard transformer implementation [46], we experimented with several learning-rate schedulers. In addition to testing stochastic gradient descent with warm restarts [57], we designed a decayed variant of the cyclic learning rate because the importance of scheduling is well emphasized by Karpov [58]. The behaviors of the schedulers are shown in Additional File 1: Figure S1. The cyclic learning scheduler was ultimately selected as the most appropriate scheduler because it provided a slightly superior performance compared to the other techniques. For the cyclic rate scheduler, the constant factor parameter and the warm-up step size were set to 5 and 5000, respectively. The learning rate decreased from 0.001 to 3.9e−12 at each 25 K step and increased to its maximum again.

Evaluation

Conversion efficiency was evaluated using Tanimoto similarity matching. Further breakdown of the results was achieved by introducing simple string matching. The widely used Tanimoto coefficient, which operated on the sparse Morgan fingerprint, was selected as the similarity metric to represent the main results. Pairwise similarities between the predictions and ground truths were computed at the end of each 25K step for each pair present in the test set. Top-1 predictions were used to report the conversion accuracy, and the Python package ccbmlib [50] was employed to facilitate the generation of similarity value distributions for all fingerprints.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-023-00693-0.

Additional file 1: Figure S1. We tried four different learning rate schedulers. CylicLR in reference to Karpov et al., its decay variant that is designed in this study, the scheduler used in standard Transformer paper, and stochastic gradient descentwith warm restarts (SGDR). The cyclic learning scheduler was selected due to its slightly superior performance compared to the other techniques. The constant factor parameter and the warm-up step size were set to 5 and 5000, respectively. The learning rate decreased from 0.001 to 3.9e-12 at each 25K steps and jumped to its maximum again. **Figure S2.** Each cell shows the Tanimoto exactness (%) of selected fingerprint transformation to SELFIES (y-axis) computed at the respective fingerprint encodings. The consistency in color code reflects the robustness, while the jumps represent the effect of selection bias. ECFP2* and ECFP4* represent explicit bit versions. **Table S1.** Case 1: Ground Truth has stereo information but prediction has in reverse form. Case 2 : Ground Truth has stereo information but prediction does not. Case 3 : Ground Truth has no stereo information but prediction does. Case 4 : Enumerations are different. Case 5 : Ground Truth is not in kekulized form but prediction is.

Author contributions

UVU and JL conceived and designed the study. UVU and IA processed data, trained the models and analyzed the results. UVU, IA., and JL discussed and interpreted the results, wrote and reviewed the manuscript.

Funding

This work was supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (Nos. NRF-2022M3E5F3081268, NRF-2022R1C1C1005080 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.).

Availability of data and materials

The data that support the findings of this study, the source code, and the associated trained models are all available at the MolForge GitHub repo: https://github.com/knu-lcbc/MolForge.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 3 December 2022 Accepted: 4 February 2023
Published: 23 February 2023

References

1. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comp Sci 28(1):31–36. https://doi.org/10.1021/ci00057a005
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 Feb 2022
3. OpenSMILES. Home Page https://opensmiles.org. Accessed 10 Dec 2021
4. Lin T-S, Coley CW, Mochigase H, Beech HK, Wang W, Wang Z, Woods E, Craig SL, Johnson JA, Kalow JA, Jensen KF, Olsen BD (2019) Bigsmiles: A structurally-based line notation for describing macromolecules. ACS Cent Sci 5(9):1523–1531. https://doi.org/10.1021/acscentsci.9b00476. (**PMID: 31572779**)
5. Drefahl A (2011) CurlySMILES: a chemical language to customize and annotate encodings of molecular and nanodevice structures. J Cheminformatics 3(1):1–7. https://doi.org/10.1186/1758-2946-3-1
6. Morgan HL (1965) The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. J Chem Doc 5(2):107–113. https://doi.org/10.1021/c160017a018

7. Weininger D, Weininger A, Weininger JL (1989) Smiles. 2. Algorithm for generation of unique smiles notation. J Chem Inf Comp Sci 29(2):97–101. https://doi.org/10.1021/ci00062a008
8. O'Boyle NM (2012) Towards a Universal SMILES representation—a standard method to generate canonical SMILES based on the InChI. J Cheminformatics 4(9):1–14. https://doi.org/10.1186/1758-2946-4-22
9. Schneider N, Sayle RA, Landrum GA (2015) Get your atoms in order-an open-source implementation of a novel and robust molecular canonicalization algorithm. J Chem Inf Model 55(10):2111–2120. https://doi.org/10.1021/acs.jcim.5b00543. (**PMID: 26441310**)
10. Wiswesser WJ (1982) How the WLN Began in 1949 and How It Might Be in 1999. J Chem Inf Model 22(2):88–93. https://doi.org/10.1021/ci00034a005
11. Homer RW, Swanson J, Jilek RJ, Hurst T, Clark RD (2008) SYBYL line notation (SLN): a single notation to represent chemical structures, queries, reactions, and virtual libraries. J Chem Inf Model 48(12):2294–2307. https://doi.org/10.1021/ci7004687
12. Heller S (2014) InChI—the worldwide chemical structure standard. J Cheminformatics 6(S1):1–9. https://doi.org/10.1186/1758-2946-6-s1-p4
13. Lin K, Xu Y, Pei J, Lai L (2020) Automatic retrosynthetic route planning using template-free models. Chem Sci 11(12):3355–3364. https://doi.org/10.1039/c9sc03666k
14. Skalic M, Jiménez J, Sabbadin D, De Fabritiis G (2019) Shape-Based Generative Modeling for de Novo Drug Design. J Chem Inf Model 59(3):1205–1214. https://doi.org/10.1021/acs.jcim.8b00706
15. Kwon Y, Lee J (2021) MolFinder: an evolutionary algorithm for the global optimization of molecular properties and the extensive exploration of chemical space using SMILES. J Cheminformatics 13(1):1–14. https://doi.org/10.1186/s13321-021-00501-7
16. Liu B, Ramsundar B, Kawthekar P, Shi J, Gomes J, Luu Nguyen Q, Ho S, Sloane J, Wender P, Pande V (2017) Retrosynthetic reaction prediction using neural sequence-to-sequence models. ACS Cent Sci 3(10):1103–1113. https://doi.org/10.1021/acscentsci.7b00303
17. Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, Lee AA (2019) Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci 5(9):1572–1583. https://doi.org/10.1021/acscentsci.9b00576
18. Brown N, Fiscato M, Segler MHS, Vaucher AC (2019) Guacamol: benchmarking models for de novo molecular design. J Chem Inf Model 59(3):1096–1108. https://doi.org/10.1021/acs.jcim.8b00839. (**PMID: 30887799**)
19. Lim J, Ryu S, Kim JW, Kim WY (2018) Molecular generative model based on conditional variational autoencoder for de novo molecular design. J Cheminformatics 10(1):31. https://doi.org/10.1186/s13321-018-0286-7
20. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A (2018) Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 4(2):268–276. https://doi.org/10.1021/acscentsci.7b00572. arXiv:1610.02415
21. Alperstein Z, Cherkasov A, Rolfe JT (2019) All SMILES variational autoencoder. arXiv. doi:1048550/arxiv.1905.13343
22. Zheng S, Rao J, Zhang Z, Xu J, Yang Y (2020) Predicting retrosynthetic reactions using self-corrected transformer neural networks. J Chem Inf Model 60(1):47–55. https://doi.org/10.1021/acs.jcim.9b00949
23. Duan H, Wang L, Zhang C, Guo L, Li J (2020) Retrosynthesis with attention-based NMT model and chemical analysis of "wrong" predictions. RSC Adv 10(3):1371–1378. https://doi.org/10.1039/c9ra08535a
24. Kim E, Lee D, Kwon Y, Park MS, Choi YS (2021) Valid, plausible, and diverse retrosynthesis using tied two-way transformers with latent variables. J Chem Inf Model 61(1):123–133. https://doi.org/10.1021/acs.jcim.0c01074
25. Bilsland AE, McAulay K, West R, Pugliese A, Bower J (2021) Automated generation of novel fragments using screening data, a dual SMILES autoencoder, transfer learning and syntax correction. J Chem Inf Model 61(6):2547–2559. https://doi.org/10.1021/acs.jcim.0c01226
26. Dai H, Tian Y, Dai B, Skiena S, Song L (2018) Syntax-directed variational autoencoder for structured data. arXiv. doi:1048550/arxiv.1802.08786 . arXiv:1802.08786
27. Kusner MJ, Paige B, Hernández-Lobato JM (2017) Grammar variational autoencoder. In: Precup D, Teh YW, eds. Proceedings of the 34th international conference on machine learning. Proceedings of machine learning research, vol 70, pp 1945–1954. https://proceedings.mlr.press/v69/kusner17a.html
28. O'Boyle NM, Dalke A (2018) DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures. ChemRxiv, 1–9. https://doi.org/10.26434/chemrxiv.7097960
29. Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A (2020) Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. Mach Learn Sci Technol 1(4):045024. https://doi.org/10.1088/2632-2153/aba947
30. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A (2017) druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. Mol Pharm 14(9):3098–3104. https://doi.org/10.1021/acs.molpharmaceut.7b00346
31. Ucak UV, Ashyrmamatov I, Ko J, Lee J (2022) Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. Nat Commun 13(1):1186. https://doi.org/10.1038/s41467-022-28857-w
32. Ucak UV, Kang T, Ko J, Lee J (2021) Substructure-based neural machine translation for retrosynthetic prediction. J Cheminformatics 13(1):1–15. https://doi.org/10.1186/s13321-020-00482-z
33. Tu Z, Coley CW (2021) Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. arXiv:2110.09681 [cs]. Accessed 2022-02-10
34. Tetko IV, Karpov P, Van Deursen R, Godin G (2020) State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. Nat Commun 11(1):1–11. https://doi.org/10.1038/s41467-020-19266-y
35. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50(5):742–754. https://doi.org/10.1021/ci100050t
36. Le T, Winter R, Noé F, Clevert D-A (2020) Neuraldecipher—reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures. Chem Sci 11(38):10378–10389. https://doi.org/10.1039/d0sc03115a
37. Kwon Y, Kang S, Choi Y-S, Kim I (2021) Evolutionary design of molecules based on deep learning and a genetic algorithm. Sci Rep 11(1):17304. https://doi.org/10.1038/s41598-021-96812-8
38. Cofala T, Kramer O (2022) An evolutionary fragment-based approach to molecular fingerprint reconstruction. In: Proceedings of the genetic and evolutionary computation conference, pp 1156–1163. https://doi.org/10.1145/3512290.3528824
39. Jaegle A, Gimeno F, Brock A, Zisserman A, Vinyals O, Carreira J (2021) Perceiver: general perception with iterative attention. Preprint at arXiv:2103.03206
40. Landrum G (2016) RDKit: open-source cheminformatics software. https://github.com/rdkit/rdkit/releases/tag/Release_2020_03_1
41. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comp Sci 42(6):1273–1280. https://doi.org/10.1021/ci010132r
42. James CA, Weininger D, Delany JD (2002) Daylight Theory Manual. Daylight Chemical Information Systems Inc. https://daylight.com/dayhtml/doc/theory/index.html
43. Gedeck P, Rohde B, Bartels C (2006) 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(5):1924–1936. https://doi.org/10.1021/ci050413p
44. Smith DH, Carhart RE, Venkataraghavan R (1985) Atom Pairs as molecular features in structure-activity studies: definition and applications. J Chem Inf Comp Sci 25(2):64–73. https://doi.org/10.1021/ci00046a002
45. Nilakantan R, Bauman N, Venkataraghavan R, Dixon JS (1987) Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. J Chem Inf Comp Sci 27(2):82–85. https://doi.org/10.1021/ci00054a008
46. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 2017-Decem(Nips): 5999–6009
47. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR (2016) The ChEMBL database in 2017. Nucleic Acids Res 45(D1):945–954. https://doi.org/10.1093/nar/gkw1074
48. Bolton EE, Wang Y, Thiessen PA, Bryant SH (2008) Chapter 12 - pubchem: Integrated platform of small molecules and biological activities. In:

Annual reports in computational chemistry, vol 4, pp 217–241. https://doi.org/10.1016/S1574-1400(08)00012-1
49. Decherchi S, Cavalli A (2020) Thermodynamics and kinetics of drug-target binding by molecular simulation. Chem Rev 120(23):12788–12833. https://doi.org/10.1021/acs.chemrev.0c00534
50. Vogt M, Bajorath J (2020) Ccbmlib—a python package for modeling tanimoto similarity value distributions. F100Research.https://doi.org/10.12688/f1000research.22292.1
51. Grimsley C, Mayfield E, RS Bursten J (2020) Why attention is not explanation: surgical intervention and causal reasoning about neural models. In: Proceedings of the 12th language resources and evaluation conference, pp 1780–1790. European Language Resources Association, Marseille, France. https://aclanthology.org/2020.lrec-1.220
52. Jain S, Wallace BC (2019) Attention is not explanation. In: Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, vol 1 (Long and Short Papers), pp 3543–3556. Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1357.https://aclanthology.org/N19-1357
53. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) Pytorch: An imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R, eds. Advances in neural information processing systems, pp 8024–8035
54. Rush A (2018) The annotated transformer. In: Proceedings of workshop for NLP open source software (NLP-OSS), pp 52–60. Association for Computational Linguistics, Melbourne, Australia. https://doi.org/10.18653/v1/W18-2509
55. Xiong R, Yang Y, He D, Zheng K, Zheng S, Xing C, Zhang H, Lan Y, Wang L, Liu T (2020) On layer normalization in the transformer architecture. arxiv:2002.04745
56. Rajbhandari S, Rasley J, Ruwase O, He Y (2020) Zero: Memory optimizations toward training trillion parameter models. In: International conference for high performance computing, networking, storage and analysis, SC 2020-November, 1–24. https://doi.org/10.1109/SC41405.2020.00024.arXiv:1910.02054
57. Loshchilov I, Hutter F (2017) SGDR: stochastic gradient descent with warm restarts. 5th international conference on learning representations, ICLR 2017—conference track proceedings, pp 1–16
58. Karpov P, Godin G, Tetko IV (2019) A transformer model for retrosynthesis. In: Artificial neural networks and machine learning—ICANN 2019: workshop and special sessions, pp 817–830. Springer, Cham

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions