https://doi.org/10.1038/s42004-025-01423-3

fragSMILES as a chemical string notation for advanced fragment and chirality representation

Check for updates

Fabrizio Mastrolorito[1], Fulvio Ciriaco[2], Maria Vittoria Togo[1], Nicola Gambacorta[1], Daniela Trisciuzzi[1], Cosimo Damiano Altomare[1], Nicola Amoroso[1], Francesca Grisoni[3,4] & Orazio Nicolotti[1]

Generative models have revolutionized de novo drug design, allowing to produce molecules on-demand with desired physicochemical and pharmacological properties. String based molecular representations, such as SMILES (Simplified Molecular Input Line Entry System) and SELFIES (Self-Referencing Embedded Strings), have played a pivotal role in the success of generative approaches, thanks to their capacity to encode atom- and bond- information and ease-of-generation. However, such 'atom-level' string representations could have certain limitations, in terms of capturing information on chirality, and synthetic accessibility of the corresponding designs.

In this paper, we present fragSMILES, a novel fragment-based molecular representation in the form of string. fragSMILES encode fragments in a 'chemically-meaningful' way via a novel graph-reduction approach, allowing to obtain an efficient, interpretable, and expressive molecular representation, which also avoids fragment redundancy. fragSMILES contributes to the field of fragment-based representation, by reporting fragments and their 'breaking' bonds independently. Moreover, fragSMILES also embeds information of molecular chirality, thereby overcoming known limitations of existing string notations. When compared with SMILES, SELFIES and t-SMILES for de novo design, the fragSMILES notation showed its promise in generating molecules with desirable biochemical and scaffolds properties.

Molecular representations based on strings are getting increasing attention in the molecular machine learning community, especially in combination with the so-called 'chemical language models' (CLMs), e.g., for de novo molecule design[1–3] and synthesis planning[4,5]. The Simplified Molecular Input Line Entry System (SMILES)[6] notation is the most well-established of such notations. By traversing the two-dimensional graph of a molecule, a SMILES string encodes atoms and bond information using predefined characters (Fig. 1a)[7]. Thanks to the increasing success of chemical language modeling[8,9], several alternatives have been proposed to overcome some of the limitations of SMILES[10], e.g., the Self-referencing Embedded Strings[11] (SELFIES), which enforce the generation of molecular strings corresponding to valid molecules. While these string notations have successfully led to experimentally-validated de novo designs[12–14], they are not devoid of limitations. In particular, due to the 'linearization' of molecular graph information, fragments are not univocally represented in 'atom-level' strings like SMILES, and information on atomic neighborhoods can be distributed in different parts of the string (Fig. 1b). Moreover, de novo designs based on 'atom-level' string representations like SMILES might be affected by a limited synthetic accessibility[15,16].

A complementary solution for storing and processing molecular information is constituted by representing molecules as a collection of fragments[17–21], via fragmentation algorithms like BRICS[20,22]. Several fragment-based de novo design approaches have been proposed, which can constrain the generation of new molecules that are easier to synthesize[16]. Several works have focused on how to develop molecular strings at the level of molecular fragments, e.g., t-SMILES[23], aimed to reduce the number of invalid sequences, although different sequences can be used to describe a given fragment.

Noteworthy, the strategy to encode fragments as stand-alone words has been documented in the literature, e.g., SMILES Pair Encoding (SPE)[24] or

[1]Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari Aldo Moro, Bari, Italy. [2]Dipartimento di Chimica, Università degli Studi di Bari Aldo Moro, Bari, Italy. [3]Institute for Complex Molecular Systems and Eindhoven Artificial Intelligence Systems Institute, Department of Biomedical Engineering, Eindhoven University of Technology, AZ Eindhoven, Netherlands. [4]Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, Utrecht, The Netherlands.
e-mail: f.grisoni@tue.nl; orazio.nicolotti@uniba.it

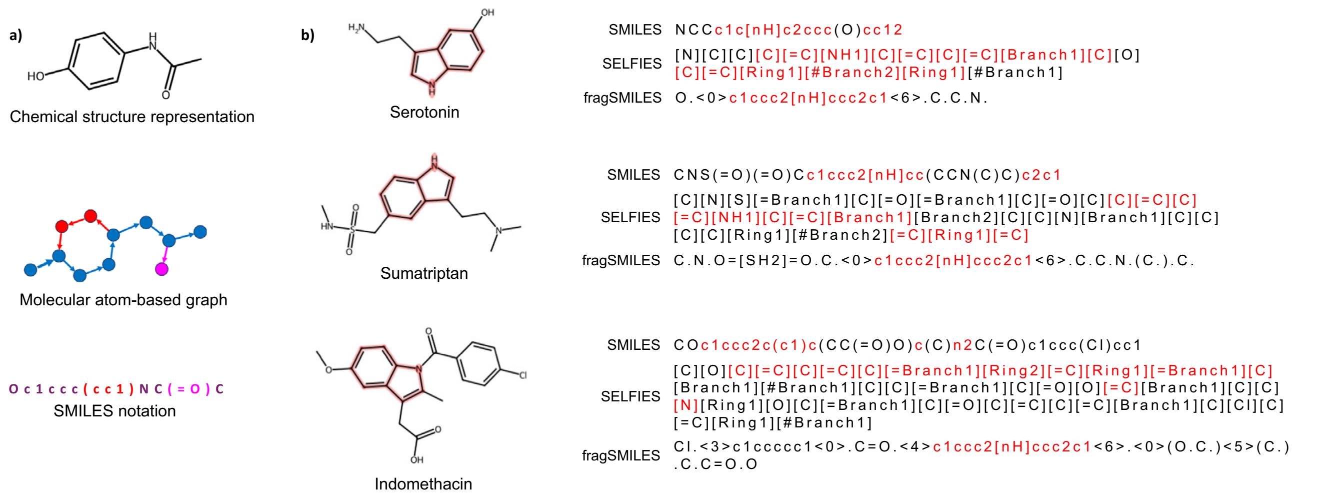

a)
HO
H
N
O
Chemical structure representation
Molecular atom-based graph
O c 1 c c c (c c 1) N C (= O) C
SMILES notation
b)
H_2N
OH
N
H
Serotonin
SMILES NCCc1c[nH]c2ccc(O)cc12
SELFIES [N][C][C][C][=C][NH1][C][=C][C][=C][Branch1][C][O][C][=C][Ring1][#Branch2][Ring1][#Branch1]
fragSMILES O.<0>c1ccc2[nH]ccc2c1<6>.C.C.N.
H
N
HN
S
O
O
N
Sumatriptan
SMILES CNS(=O)(=O)Cc1ccc2[nH]cc(CCN(C)C)c2c1
SELFIES [C][N][S][=Branch1][C][=O][=Branch1][C][=O][C][C][=C][C][=C][NH1][C][=C][Branch1][Branch2][C][C][N][Branch1][C][C][C][C][Ring1][#Branch2][=C][Ring1][=C]
fragSMILES C.N.O=[SH2]=O.C.<0>c1ccc2[nH]ccc2c1<6>.C.C.N.(C.).C.
O
N
O
Cl
O
OH
Indomethacin
SMILES COc1ccc2c(c1)c(CC(=O)O)c(C)n2C(=O)c1ccc(Cl)cc1
SELFIES [C][O][C][=C][C][=C][C][=Branch1][Ring2][=C][Ring1][=Branch1][C][Branch1][#Branch1][C][C][=Branch1][C][=O][O][=C][Branch1][C][C][N][Ring1][O][C][=Branch1][C][=O][C][=C][C][=C][Branch1][C][Cl][C][=C][Ring1][#Branch1]
fragSMILES Cl.<3>c1ccccc1<0>.C=O.<4>c1ccc2[nH]ccc2c1<6>.<0>(O.C.)<5>(C.).C.C=O.O

Fig. 1 | Molecular string notations based on graph traversal. a An example of graph traversal to obtain a SMILES string for the molecule paracetamol. **b** Examples of SMILES, SELFIES and fragSMILES of three known drugs sharing the indole fragment highlighted in red.

Group SELFIES[25]. Such a shift from atom- to fragment-based notation in chemical language modeling resembles the character- to word-level[26] shift in natural language processing (NLP), and it has the potential to yield more efficient and expressive representations[27,28]. However, current fragment-based approaches have several limitations, e.g. they can generate redundant structural motifs by annotating parts of identical substructures as they were different fragments. This might yield inefficient representation learning, since the same chemical information is annotated inconsistently[29–31]. Other works have used sets of fragments for de novo design. A notable example is SAFE[32], which obtains an unordered sequence of interconnected fragment blocks.

To overcome these gaps in molecular representation field, here we propose fragSMILES, a novel 'chemical-word'-level molecular string notation. Our representation is based on graph reduction, i.e., the simplification of molecular graphs by collapsing selected atoms and bonds into single fragments, thereby reducing the complexity of the graph while retaining key structural and functional information[19,33]. The reduced graph is then converted into a string-based notation, which we named fragSMILES.

Based on an interpretable tokenization technique, our 'chemical word'-level representation allows to easily identify building blocks. This allows (a) each fragment to be univocally encoded, irrespective of its molecular neighborhoods, (b) achieving a richer 'fragment semantics', while simplifying how fragment linkers are encoded, and (c) obtaining shorter sequences for chemical language modeling, that still preserve key chemical information. In what follows, after introducing the theory of fragSMILES, we show that this new notation advances the state of the art thanks to a better specification of molecular chirality, along with a good capacity to explore the chemical space when combined with de novo design algorithms, e.g., to achieve desirable physico/biochemical properties, synthesizability and scaffold novelty. Additionally, fragSMILES can improve many drug discovery related tasks[34] in a broader sense, such as database storing[35], chemical reaction predictions[36] and bioactivity prediction[37].

Results and discussions

Representing molecules as fragSMILES

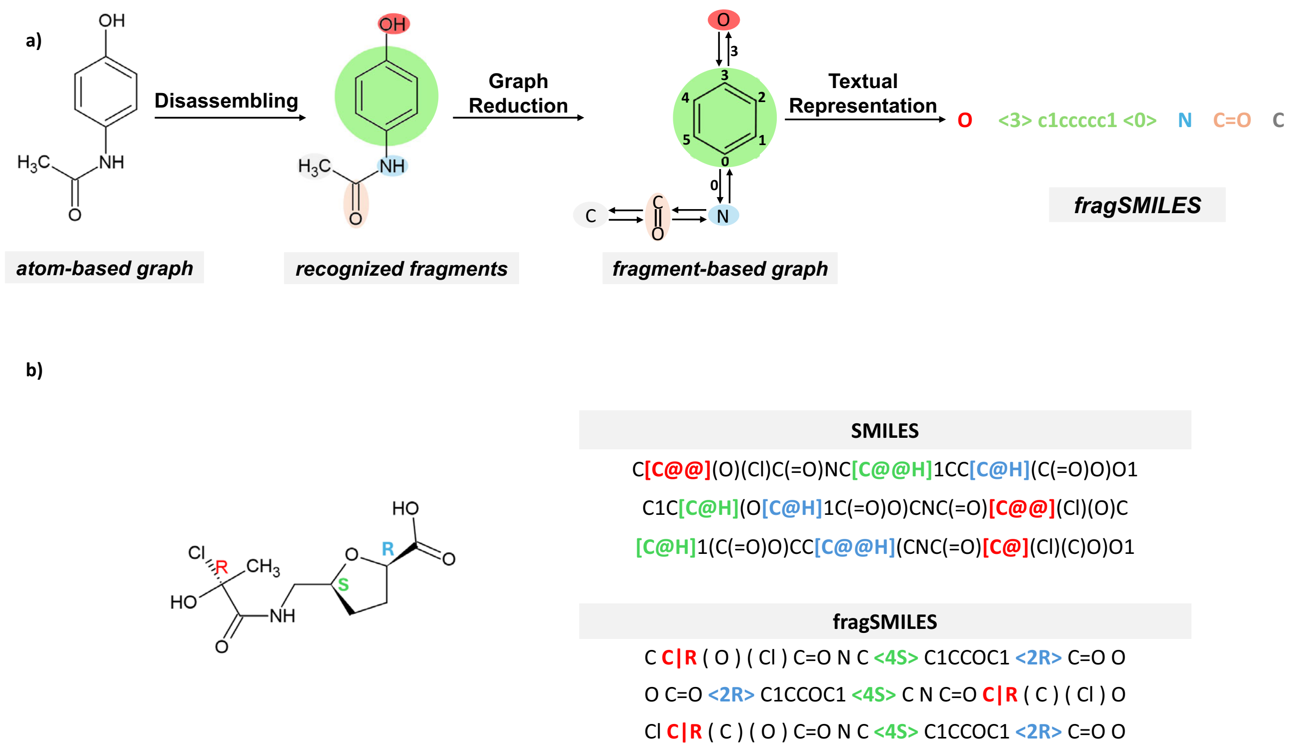
The fragSMILES notation is generated via molecular graph reduction (Fig. 2a). In particular, the conversion of molecular graphs to fragSMILES consists of three phases (a):

1. *Disassembling*. This phase 'breaks' a given molecule according to customizable set of cleavage rules. Breaking bonds are constituted by either (a) exocyclic single bonds, except for bonds between oppositely charged atoms (e.g., nitro groups), or (b) user-customizable rules, including for example rotatable bonds. In our approach, molecular fragments were obtained by cleaving all the exocyclic single bonds but other user-defined fragmentation rules could also be set (e.g., BRICS)[20,22].
2. This phase leads to a set of unique fragments. Each unique fragment is also annotated with indexes to track the 'breaking' bonds.
3. *Graph reduction*. This phase condenses molecular fragments and the 'breaking' bonds into the respective nodes and edges of a reduced molecular graph. Bidirected edges specify the correct fragment combination (i.e., by identifying which atoms belong to the 'breaking' bonds).
4. *Conversion into fragSMILES*. The reduced molecular graph is then converted into string notation, where, for interpretability, the SMILES alphabet is used to identify fragment-level nodes.

The 'syntactic' rules of the obtained fragSMILES are the following:

- The nodes representing the fragments are expressed as canonical SMILES, to preserve interpretability of the corresponding chemical information.
- The edges are provided with numerical indexes, with the notation "< *index* >" to indicate the connecting atoms between neighboring fragments. In the case of fragments including only a single linking atom (i.e., a single atom having replaceable hydrogen atoms), the numerical index is not shown.
- Branches are described in parentheses. Before and after the opening parenthesis, numerical indexes indicate how the branches are connected to the fragments.
- The configuration of chiral centers (Fig. 2b) is indicated as suffix tag to the indices in the case of connector atoms (e.g., <2 R > C1CCOC1 < 4S> to represent the tetrahydrofuran with two given connector carbon atoms in *R* and S absolute configurations) or as special suffix in the case of non-connector atoms (e.g., C1CC2CNC2CN1 | 2R5S to represent the 3,8-diazabicyclo[4.2.0]octane with the two bridgehead carbon atoms in *R* and *S* absolute configurations). The stereochemical configuration of fragments having unspecified connector atoms is also reported as suffix (e.g., C | R to represent the carbon in R absolute configuration).

Like SMILES strings, any molecule can be represented as multiple fragSMILES (depending on the starting point to traverse the reduced graph). This aspect can be leveraged for data augmentation. Alternatively, fragSMILES can be also canonicalized, via graph traversal and fragment prioritization rules (see Materials and Methods). Moreover, Fig. 2b shows that, irrespective of the graph traversal order, in fragSMILES the chiral centers are consistently annotated, unlike in the case of SMILES.



SMILES

C**[C@@]**(O)(Cl)C(=O)NC**[C@@H]**1CC**[C@H]**(C(=O)O)O1

C1C**[C@H]**(O**[C@H]**1C(=O)O)CNC(=O)**[C@@]**(Cl)(O)C

[C@H]1(C(=O)O)CC**[C@@H]**(CNC(=O)**[C@]**(Cl)(C)O)O1

fragSMILES

C **C|R** (O) (Cl) C=O N C **<4S>** C1CCOC1 **<2R>** C=O O

O C=O **<2R>** C1CCOC1 **<4S>** C N C=O **C|R** (C) (Cl) O

Cl **C|R** (C) (O) C=O N C **<4S>** C1CCOC1 **<2R>** C=O O

Fig. 2 | Graph reduction framework and fragSMILES representation. **a** Generation of fragSMILES based on disassembling, graph reduction and textual representation; **b** chiral center configurations for a generic chemical structure specified as SMILES, and fragSMILES notations (traditional stereochemical labels are highlighted in red, green and blue).

Effect of fragmentation and tokenization rules

Several tokenization techniques can be employed to split molecular portions, textually represented, into computer-manageable 'units' (tokens)[24,38]. fragSMILES notation separates molecular fragments, their connector indices and branching brackets, and considers them as tokens.

We compared the effect of two fragmentation rules for fragSMILES, based on the cleavage of: (a) all the exocyclic single bonds[17], and of (b) all the rotatable bonds[39]. To this end, we used ZINC-250K[40] database, which contains 249,414 molecules. Cleaving all exocyclic single bonds generated a more compact vocabulary of token types (5869 *vs* 13,035), thus mitigating the risk of redundancy[29–31] (Table 1). Moreover, such fragmentation rule awards the occurrence of 'generic' (e.g., amino group, carbon atoms, and phenyls) instead of specific tokens (e.g., aniline and the toluene). An example of different fragmentation rules for a generic molecular structure is depicted in Supplementary Fig. 1.

As a result, the fragmentation based on exocyclic single bonds constitutes a good trade-off between (a) the number of tokens necessary to represent a molecule, and (b) the number of token types. Hence, fragSMILES works as syllable-level language, which is capable of better performances[41] and lower complexity[42].

For completeness, we implemented in the t-SMILES[23] a word-level tokenization, where tokens encode fragments. By comparing the vocabulary size and the number of tokens for t-SMILES word-level, fragSMILES rotatable bonds and fragSMILES exocyclic single bonds, we noticed that t-SMILES, due to its sequence lengths, was unsuitable for fragment tokenization although its vocabulary was larger than ~12 K (Supplementary Fig. 2).

Based on these observations, the default fragmentation strategy for fragSMILES uses exocyclic single bonds.

Compact encoding of molecular information via fragSMILES

The fragSMILES approach allows to encode molecules with a smaller number of elements ('tokens'[43]) than SMILES and SELFIES strings given that its tokenization is based on fragments descending from reduced molecular graph. Notably, other tokenization techniques could return shorter sequences[24] and whose structure is not always related to any particular molecular fragment (e.g., Byte Pair Encoding)[38].

When analyzed on ZINC-250K[40] database, fragSMILES returned an average length of 17 tokens (Fig. 3), remarkably smaller than the length of Group-SELFIES[25], SELFIES[11] and SMILES (which have an average of 30, 37 and 44 tokens, respectively). Representations with fewer tokens in combination with deep learning have the potential advantage to reduce computational complexity and memory usage[44].

Noteworthy, the top occurring tokens are those made by the single carbon atom fragment, irrespective if it is a terminal methyl or a polysubstituted carbon, and connectivity tags as *<0>* or *<3 >*. Rare tokens are instead made by cyclic fragments provided with unambiguous chirality and occurring in very few molecular structures. The high frequency of fragments such as single carbon or nitrogen atoms demonstrates that fragSMILES captures mainly the occurrence of non-redundant tokens. For instance, aniline and toluene are represented by two different structural tokens. They are the amino group and the aromatic ring for the aniline and the carbon atom and the aromatic ring for the toluene (Supplementary Fig. 3).

De novo molecule design with fragSMILES

We benchmarked fragSMILES for de novo molecule design in comparison with SMILES, SELFIES and t-SMILES. To this end, we employed Recurring Neural Networks (RNNs)[45] with Long Short Term Memory (LSTM) cells[46], as implemented in the MOSES benchmarking platform[47]. While many architectures for de novo design exist, LSTMs have been widely adopted in this field and extensively validated in the wet-lab[3,12,48–50]. The MOSES platform was chosen as it provides a curated dataset, predefined metrics, and a ready-to-use LSTM model for benchmarking on de novo drug design purposes.

Table 1 | fragSMILES chemical words obtained through different molecular fragmentation rules

Fragm. Rule	Vocab. size	N (amino group)	C (carbon atom)	c1ccccc1 (phenyl)	Nc1ccccc1 (aniline)	Cc1ccccc1 (toluene)
Exocyclic single bonds	5869	247,050	815,799	203,681	0	0
Rotatable bonds	13,035	191,522	258,876	102,798	843	18,981

Vocabulary size (number of total tokens) for fragSMILES, according to the chosen fragmentation rule, applied to ZINC-250k molecules. The occurrence of some representative token types is also reported.

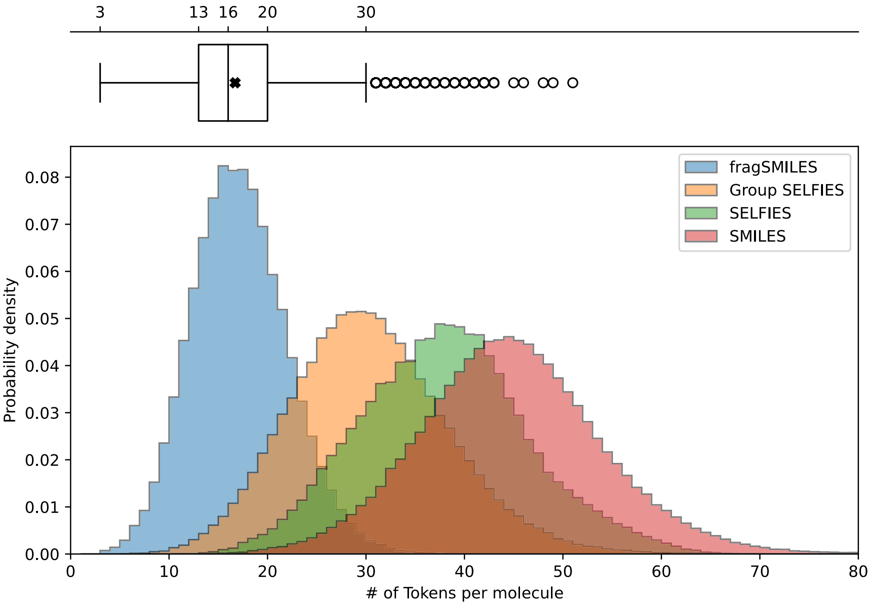


Fig. 3 | Encoding length benchmarking. Token number distribution of fragSMILES, Group SELFIES, SELFIES and SMILES, computed on ZINC-250K molecules. The boxplot represents the length distribution of fragSMILES (whiskers indicate 1st and 3rd quartiles, median, the central line and the cross indicate the median and mean values, respectively, and circles indicate the outliers).

Models were trained on the MOSES dataset (approximately 1.9 M molecules) using a five-fold cross validation and on each of the representations separately.

The trained models were used to sample 6000 molecules for each representation (i.e., fragSMILES, SELFIES, SMILES, and t-SMILES), for each cross-validation fold, obtaining 30,000 in total for each representation. For each set of generated strings, we computed (a) validity, capturing the number of 'chemically valid' molecules generated from each representation (b) uniqueness, capturing the number of non-duplicated molecules, and (c) novelty, quantifying the number of de novo designs that were not present in the training set (Table 2). Detailed explanations about the metrics are reported in the Materials and Methods. In this context, fragSMILES strings showed an intermediate behavior between SELFIES and t-SMILES (best) *vs* SMILES (worst). SELFIES showed consistently better values of validity and uniqueness as t-SMILES, while the latter showed best novelty value, and fragSMILES reached statistically significant better than SMILES strings for validity and uniqueness (Table 2). Validity was computed on the initial 6000 generated samples, while uniqueness and novelty were computed on the number of valid, and valid and unique molecules, respectively.

To further investigate the quality of the generated de novo designs, we sampled an additional set of 6000 (*per* 5-fold of the cross-validation) novel molecules for each representation. We computed an array of metrics to quantify the similarity of the designs to the known molecules. In particular, we computed (Table 2):

- Fréchet ChemNet Distance (FCD)[51], which captures the differences of biological and chemical properties between de novo designs and a reference set of molecules (test set molecules here, Table 2). The lower the FCD, the more similar two sets are.
- *Physicochemical properties*, i.e., lipophilicity (logP)[52], Synthetic Accessibility (SA)[53], Quantitative Estimation of Drug-likeness (QED)[54,55] and molecular weight (MW). The differences of these properties between de novo designs and training set molecules were reported as Wasserstein-1 distance[56] (the lower, the more similar the two sets).

In this context, fragSMILES showed consistently superior performance across all the analyzed metrics, with statistically significant improvements in two out of five cases with SMILES, four out of five cases with SELFIES and four out of five cases with t-SMILES (Table 2). These results suggest that fragSMILES are an ideal trade-off between chemical space exploration (validity, uniqueness, novelty) and the identification of designs with desirable properties (i.e., similar to the training set molecules).

Considering that the training set contains drug-like molecules, fragSMILES generated new molecules with a desirable property profile in terms of logP, QED, and MW. A similar consideration can be extended to the SA values. For the sake of comparison, results are shown in Supplementary Figs. 4 and 5.

Chemical space exploration with fragSMILES

In what follows, we focus our de novo design efforts using a training set of 270,408 bioactive molecules[57] (for $K_d/K_I/IC_{50}/EC_{50} < 1\ \mu M$) from ChEMBL[58] v22. Unlike the previously used MOSES molecules, this set contains stereochemistry information, and is aimed at a specific task, namely, generating drug-like molecules. This test was used to elucidate various properties of fragSMILES: (a) the effect of string augmentation on the quality of de novo designs, (b) the representation capability to capture chirality, and (c) the potential to explore novel molecular scaffolds. These aspects are discussed below. Supplementary Fig. 6 shows how alternative representations of fragSMILES are obtained for the same molecule.

fragSMILES augmentation

'Atom-level' representations like SMILES and SELFIES are non-univocal, since they can be obtained starting from any non-hydrogen atom, and by reading the graph in different directions. As a result, a molecule can be represented by multiple valid strings for training purposes: such 'data augmentation' can improve the quality of molecules produces via chemical language modeling[59–61]. The new fragSMILES notation can also be augmented: since any node of the reduced graph can be the starting point for

Table 2 | Quality of de novo designs generated using SMILES, SELFIES, t-SMILES and fragSMILES, employing an RNN trained on MOSES

	6000 (x5 fold) sampled strings			6000 (x5 fold) sampled novel molecules				
Notation	**Validity (↑)**	**Uniqueness (↑)**	**Novelty (↑)**	**FCD•10^1 (↓)**	**ΔlogP•10^1 (↓)**	**ΔSA•10^2 (↓)**	**ΔQED•10^2 (↓)**	**ΔMW (↓)**
SMILES	5790 ± 20 (97%)	5790 ± 20 (100%)*	5270 ± 40* (91%)*	3.9 ± 0.3*	1.2 ± 0.3*	7 ± 2	1.0 ± 0.4	6 ± 2
SELFIES	**6000** ± **0*** **(100%)***	**5999** ± **1*** **(100%)***	5550 ± 50* (93%)*	10.0 ± 0.6*	1.7 ± 0.7*	21.9 ± 0.9*	2.5 ± 0.4*	4.2 ± 0.9
t-SMILES	**6000** ± **0*** **(100%)***	5966 ± 6* (99%)*	**5740** ± **20*** **(96%)***	6.7 ± 0.3*	0.9 ± 0.2	15 ± 2*	1.8 ± 0.5*	6 ± 1*
fragSMILES	5810 ± 10 (97%)	5800 ± 10 (100%)	5160 ± 20 (89%)	**3.2** ± **0.2**	**0.7** ± **0.2**	**4** ± **2**	**0.8** ± **0.4**	**3.8** ± **0.3**

The metrics are reported as average ± standard deviation. Validity, uniqueness and novelty were computed on a set of 6000 strings, with 5-fold cross-validation. The other properties were computed on 6000 novel designs, with 5-fold cross-validation: Fréchet ChemNet Distance (FCD); octanol-water partitioning coefficient (logP), Synthetic Accessibility (SA), Quantitative Estimation of Drug-likeness (QED) and molecular Weight (MW). logP, SA, QED and MW were reported as the Wasserstein-1 distance to the properties of the training set molecules (the lower, the better). Arrows indicate the optimal directionality of each metric (↑: the higher, the better; ↓: the lower, the better), and * indicates statistically significant differences (t-test, α = 0.05) with relative values of fragSMILES notation. The best value of each metric is indicated in boldface, while underlining indicates the second-best performance.

traversal, multiple fragSMILES can be obtained from the same molecule. This property allows to perform data augmentation, as for SMILES and SELFIES. Furthermore, the reduced graph allows to explicitly account for the chiral centers of stereoisomers. In this respect, the corresponding fragSMILES reports the absolute configurations, tokenized as *R* and *S*, respectively. Thus, the chiral centers in fragSMILES are univocally assigned, irrespective of the order of their substituents in the string notation. Unlike SMILES[62], this makes fragSMILES tokens univocal and invariant to graph traversal order once the chiral centers have been defined.

On the selected molecular set, RNN models were trained for each string notation (SMILES, SELFIES, t-SMILES and fragSMILES notations), and on their single- (canonical representation) and five-fold augmented versions. 61,870 fragSMILES representations were not augmentable until to five-folds, because they were composed of fully linear graphs. As far as fragSMILES is concerned, we calculated an average of 10 alternative representations as maximum number *per* molecule of the Zinc250K dataset. Our results align with what observed for SMILES strings, where augmentations larger than 10 do not lead to remarkable improvements in the model quality[8,60,63].

Five-fold cross validation was carried out, and each fold was used to generate 6000 strings to compute validity, uniqueness and novelty (Table 3). The (bio)chemical properties previously discussed (Table 2) were also evaluated, using a pool of 6000 novel designs per each fold (Table 3).

In agreement with existing literature[59,63,64], the augmentation improved the quality of de novo designs for SMILES and SELFIES in terms of novelty, uniqueness and validity. This is also visible on fragSMILES (Table 3). The same trend in performance was observed (i.e., SELFIES and t-SMILES yielding the best values of validity, uniqueness, and novelty), with fragSMILES de novo designs being the most similar to the reference molecules in terms of chemical and biological properties (FCD in Table 3). In most cases, string augmentation led to a small decrease in (bio)chemical similarity to the reference set (Supplementary Figs. 7-9).

Capturing chirality with fragSMILES

The previously generated 6000 strings for each fold were used to analyze the capacity of each representation to capture chirality. In particular, for each of the representations, the 6000 strings were filtered to a subset containing tokens referring to chirality information (i.e., 1770 ± 70 for SMILES × 1 and 1800 ± 200 for SMILES × 5, 1820 ± 40 for SELFIES × 1 and 1900 ± 100 for SELFIES × 5, 1840 ± 90 for t-SMILES × 1 and 1900 ± 100 for t-SMILES × 5, 1770 ± 90 for fragSMILES × 1 and 2000 ± 100 for fragSMILES × 5). Each of these strings was then converted into a molecule (if possible), and the number of (in)valid, unique and novel molecules was quantified.

It is important to note that the conversion of fragSMILES into a valid molecule can only happen when chiral fragments (tokens) preserve their stereocenters (i.e., they have four different substituents). Several SMILES, SELFIES and t-SMILES strings can contain errors as they report achiral carbon atoms as chiral (Table 3, Supplementary Fig. 10), which become syntactically valid only after sanitization and canonization with RDKit[65].

In this context, fragSMILES produce a significantly higher number of molecules with unambiguously annotated and correct chirality (Table 3), thereby advancing upon known limitations of existing molecular strings (Supplementary Fig. 11)[62,66].

Exploration of novel scaffolds with fragSMILES

Finally, we performed a scaffold analysis on the 6000 novel molecules previously generated per fold. In particular, Bemis-Murcko scaffolds[67] were computed to identify the number of unique and novel scaffolds (Table 4).

In this context, SELFIES outperformed both SMILES, t-SMILES and fragSMILES. In their non-augmented version, SMILES and fragSMILES show comparable performance of scaffold uniqueness (no statistical difference achieved) unlike t-SMILES that performed worse. When five-fold augmentation is performed, SMILES outperforms fragSMILES, with a 3% higher scaffold novelty. t-SMILES achieved the same trend as the non-augmented version.

To further elucidate the characteristics of the novel scaffolds, we computed the total number of novel scaffolds containing new fragments compared to the training set molecules (reported as "No. novel fragments" and computed based on fragSMILES fragments, in Table 4).

SMILES, SELFIES and t-SMILES can generate a higher number of new scaffolds compared to fragSMILES. However, many of these scaffolds are new primarily because they include cyclic fragments that were not present in the training set. This behavior is likely due to their character-level tokenization[68], which allows a very high number of possible combinations, independently on the chemical relevance. Noteworthy, fragSMILES can also generate new cyclic systems and promote scaffold novelty (Supplementary Table 1) if an atom-level tokenization is set. On the other hand, fragSMILES on word-level tokenization is more effective at capturing the scaffolds present in the training set and generating genuinely novel cores by recombining different cyclic elements. This ability to create novel cores through recombination bears potential for de novo molecule design, as it ensures that the generated molecules are more likely to possess desired chemical properties, to be chemically stable and synthetically feasible.

Conclusions

This work introduced fragSMILES, a novel 'chemical-word'-level notation for molecules. Unlike previous fragment-based representations, fragSMILES possesses desirable qualities, i.e., it (a) reports fragments and their 'breaking' bonds independently, (b) allows canonical encoding of fragments without redundancy, and (c) strikes an ideal balance between sequence

Table 3 | Comparison of SELFIES, SMILES, t-SMILES and fragSMILES, across different augmentation levels and based on various properties, for a set of generated strings (using a ChEMBL subset, across five cross-validation folds)

	6000 (x5 fold) sampled strings			6000 (x5 fold) sampled novel molecules					6000 (x5 fold) sampled strings (chiral set)			
Notation	Validity (↑)	Uniqueness (↑)	Novelty (↑)	FCD•10^1 (↓)	ΔlogP•10^1 (↓)	ΔSA•10^2 (↓)	ΔQED•10^2 (↓)	ΔMW (↓)	Invalidity (↓)	Validity (↑)	Uniqueness (↑)	Novelty (↑)
SMILES 1×	4930 ± 70* (82%)*	4920 ± 70* (100%)	4770 ± 60* (97%)	8 ± 1*	**0.8 ± 0.3**	5 ± 3	2 ± 1	14 ± 4	400 ± 40* (22%)*	1370 ± 40 (78%)*	1370 ± 40 (100%)	1320 ± 40* (96%)
SELFIES 1×	**6000 ± 0*** **(100%)***	**5999 ± 2*** **(100%)***	**5971 ± 2*** **(100%)***	55 ± 2*	2.0 ± 0.9	74 ± 4*	1.9 ± 0.3	**5 ± 3**	670 ± 40* (37%)*	1150 ± 20* (63%)*	1150 ± 20* (100%)*	1140 ± 20* (99%)*
t-SMILES 1×	**6000 ± 0*** **(100%)***	5880 ± 10* (98%)*	5860 ± 10* (100%)*	15.6 ± 0.8*	2 ± 1	**5 ± 1**	3.8 ± 0.5*	38 ± 3*	1010 ± 50* (55%)*	830 ± 50* (45%)*	830 ± 50* (100%)*	830 ± 50* (100%)*
fragSMILES 1×	5280 ± 20 (88%)	5270 ± 30 (100%)	5110 ± 40 (97%)	**6.9 ± 0.5**	1.1 ± 0.6	5 ± 3	**1 ± 1**	9 ± 5	**330 ± 30 (19%)**	**1440 ± 70** **(81%)**	**1440 ± 60** **(100%)**	**1400 ± 60** **(97%)**
SMILES 5×	5300 ± 40* (88%)*	5300 ± 40* (100%)*	5280 ± 40 (100%)*	9.9 ± 0.7*	**1.1 ± 0.4**	6 ± 2	2 ± 2	15 ± 9	320 ± 50 (17%)*	1500 ± 100 (83%)*	1500 ± 100 (100%)*	1500 ± 100 (100%)*
SELFIES 5×	**6000 ± 0*** **(100%)***	**6000 ± 0*** **(100%)***	**5997 ± 1*** **(100%)***	34 ± 1*	1.2 ± 0.5	53 ± 2*	1.7 ± 0.5	**5 ± 2**	520 ± 40* (27%)*	1380 ± 80* (73%)*	1380 ± 80* (100%)*	1370 ± 80* (100%)*
t-SMILES 5×	**6000 ± 0*** **(100%)***	5930 ± 10* (99%)*	5880 ± 10* (99%)*	13.7 ± 0.6*	1.4 ± 0.6	**5 ± 2**	3 ± 1*	36 ± 4*	1000 ± 100* (53%)*	890 ± 60* (47%)*	890 ± 60* (100%)*	880 ± 60* (99%)*
fragSMILES 5×	5420 ± 60 (90%)	5410 ± 60 (100%)	5300 ± 60 (98%)	**7.2 ± 0.6**	1.5 ± 0.7	**5 ± 2**	**1.5 ± 0.7**	7 ± 4	**290 ± 30 (15%)**	**1700 ± 100** **(85%)**	**1700 ± 100** **(100%)**	**1600 ± 100** **(98%)**

For each metric, the string sampling strategy is reported. (FCD = Fréchet ChemNet Distance; logP = octanol-water partitioning coefficient, SA = Synthetic Accessibility; QED = Quantitative Estimation of Drug-likeness; MW = molecular weight; Δ = Wasserstein-1 distance to the training set). * Indicates statistically significant differences (t-test, α = 0.05) with relative values of fragSMILES notation. The best value of each metric is indicated in boldface.

length and vocabulary size. Our systematic analysis shows that fragSMILES possess desirable properties for de novo design and a good capacity to explore the chemical space while preserving desirable physico- and biochemical properties. Importantly, the fragSMILES notation excels in capturing molecular chirality, a critical aspect often overlooked by traditional string-based representations[62].

Thanks to its 'chemical-word'-level character and expressive representation of chemical information, we expect the fragSMILES notation to advance current capabilities of chemical language modeling, not only for de novo molecule design. Specifically, it could improve but reaction properties or molecule properties[69,70], synthesis planning[15], prediction of challenging bioactivity prediction tasks, e.g., fields involving activity cliffs[37] and chiral activity cliffs[71], due to fragSMILES improved detection of chirality. Moreover, its textual representation could help database storing[35] and fragment-based molecule searching, avoiding substructure searching by employing the use of slower graph-based algorithms.

By setting word-level or atom-level tokenization rules, fragSMILES can be employed to effectively explore uncharted regions of the chemical space. The applicability of fragSMILES can be further extended to new tasks in the molecular sciences, e.g., by incorporating additional fragmentation rules or incorporating additional chemical information relative to the fragments. Finally, we expect the development of neural network architectures tailored to word-level processing to further propel the potential of fragment-based notations for generative artificial intelligence in chemistry.

Methods

Graph reduction procedure for fragSMILES

After interpreting atom-based molecular graphs as RDKit (v. 2023.9.5)[65] '`Mol`' objects, fragSMILES are obtained with the following procedure:

1. *Cleavage bond definition.* Cleavage bonds pattern can be defined and customized via SMARTS[72] notation.
2. *Molecule fragmentation.* Based on the defined cleavage bonds, molecules are divided into fragments. In this work, this was performed via the '`Chem.FragmentOnBonds`' function of RDKit.
3. *Conversion into a reduced graph.* All information on obtained fragments (nodes) and cleavage bonds (edges) are converted into a bidirectional graph. In this work, this was handled via NetworkX[73] package (v. 3.2.1) and interpreted as a bidirectional graph carrying all attributes for nodes and edges.

fragSMILES canonicalization

The canonicalization of the reduced graph is achieved via the following steps:

1. longest paths are recognized;
2. paths that branch out earlier along the way are retained;
3. paths with more numerous branching are retained;
4. Equal paths are compared by the sequence of their component nodes. Each node, depending on the fragment it represents, is identified by a unique numeric ID that places it in a ranking. The path with the most importance is considered.

All codes for tool employing are written as Python language and available on GitHub link https://github.com/f48r1/chemicalgof or Zenodo https://doi.org/10.5281/zenodo.12700298.

Data sources and preprocessing

The following datasets and preprocessing steps were used:

- *ZINC-250K database*[40] was obtained from repository of a recent work[25]. It contained 249,414 molecules but some of them were discarded because they contained only one fragment according to the fragmentation framework. Specifically, 86 and 864 molecules were discarded when the exocyclic single bond rule and rotatable bond rule were applied, respectively.
- MOSES dataset[47] consisted of 1,936,962 molecules whose SMILES data were converted to SELFIES and fragSMILES representations. 269

Table 4 | Scaffold evaluation analysis

	6000 (× 5 fold) sampled novel molecules		
Notation	**Scaffold Uniqueness (↑)**	**Scaffold Novelty (↑)**	**No. novel fragments**
SMILES 1×	5490 ± 80 (92%)	4600 ± 100* (84%)*	820 ± 20* (18%)*
SELFIES 1×	**5570 ± 40*** **(93%)***	**5160 ± 70*** **(93%)***	2820 ± 100* (55%)*
t-SMILES 1×	5330 ± 50* (90%)*	4730 ± 40* (89%)*	670 ± 40* (14%)*
fragSMILES 1×	5500 ± 50 (92%)	4440 ± 70 (81%)	0 ± 0 (0%)
SMILES 5×	5610 ± 70 (94%)	5000 ± 100* (90%)*	1060 ± 40* (21%)*
SELFIES 5×	**5620 ± 40*** **(94%)***	**5240 ± 60*** **(93%)***	2300 ± 60* (44%)*
t-SMILES 5×	5380 ± 40* (91%)	4730 ± 40* (88%)*	600 ± 50* (13%)*
fragSMILES 5×	5510 ± 80 (92%)	4400 ± 100 (81%)	0 ± 0 (0%)

Comparison of SELFIES, SMILES, t-SMILES and fragSMILES, across different augmentation levels and based on scaffold evaluation metrics, for a set of novel generated molecules (using a ChEMBL subset, across five cross-validation folds). * Indicates statistically significant differences (t-test, α = 0.05) with relative values of fragSMILES notation. The best value of each metric is indicated in boldface.

molecules were discarded because they were composed of a single cyclic fragment.

- *The bioactive ChEMBL subset* was obtained by a recent work[57]. It contained around 650 K structures from ChEMBL v. 22 having $K_d/K_I/IC_{50}/EC_{50} < 1\ \mu M$.

For ZINC and the bioactive ChEMBL molecules, isotope information was removed, and salts charges were neutralized keeping the heavier organic part. Geometric stereochemical information at the double bond ("/" and "\" characters) were removed but retaining optical stereochemical ("@" and "@@" characters) one. Finally, SMILES were canonized, and duplicates were removed. For ChEMBL, only molecules containing 10 to 32 fragSMILES tokens were retained, obtaining 270,408 molecules. MOSES dataset was not preprocessed further. All datasets used in this work listed molecules as SMILES strings, which were used to obtain canonical SELFIES and fragSMILES notations.

Model training and hyperparameter optimization

The RNN architectures used in this work were taken from the MOSES[47] benchmarking platform. Default settings for parameterization do not provide a customizable embedding size. Therefore, we extended a customizable setting of embedding size for RNN model trained on fragSMILES representations, useful for word-level NLP[26]. Sampling was performed via *softmax* function with a temperature of T = 1.

For each representation, each cross-validation fold and each level of augmentation, the following hyperparameters were optimized: number of hidden layers (2, 3), number of hidden units per layer (256, 512), batch size (256, 512). Adam optimizer and a learning rate of 0.001 were used. Cross-validation loss was used for model optimization, in combination with early stopping at loss convergency. The details of the optimized hyperparameters can be found in Supplementary Tables 2 and 3.

The values of the evaluation metrics shown in the main text refer to models adopting hyperparameters that maximized novelty metric for SMILES notation at the sampling phase. Notably, the trend of performance as the hyperparameters changed was also observed by the other notations.

Evaluation metrics

For each model, strings were sampled at the early-stopping epoch. All generated strings were converted into canonical SMILES to assess their validity, uniqueness and novelty. In particular, validity was calculated on the total number of sampled SMILES strings that could be converted to chemically-valid molecules (by RDKit '`Chem.MolFromSmiles`'). Uniqueness was calculated on the valid (canonical) SMILES that were not duplicated. Novelty was calculated on the unique canonical SMILES that were not already included in the training set.

All other metrics were calculated by using software available in MOSES. Molecular scaffolds were computed via the RDKit (v. 2023.9.5) '`Chem.Scaffolds.MurckoScaffold`' module of RDKit package.

For the sake of completeness, all the results of the metrics provided by MOSES are reported as Supplementary Data 1.

Data availability

The data employed to conduct our analysis are available on GitHub, at the following URL https://github.com/f48r1/fragsmiles.

Code availability

The code for graph reduction and obtaining fragSMILES is available in Python language on GitHub, at the following URL: https://github.com/f48r1/chemicalgof, and on Zenodo at the following https://doi.org/10.5281/zenodo.12700298. All code to reproduce our analysis and processing steps can be found on GitHub, at the following https://github.com/f48r1/fragsmiles.

Received: 23 July 2024; Accepted: 21 January 2025;
Published online: 29 January 2025

References

1. Grisoni, F. Chemical language models for de novo drug design: Challenges and opportunities. *Curr. Opin. Struct. Biol.* **79**, 102527 (2023).
2. Alberga, D. et al. De novo drug design of targeted chemical libraries based on artificial intelligence and pair-based multiobjective optimization. *J. Chem. Inf. Model.* **60**, 4582–4593 (2020).
3. Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131 (2018).
4. Schwaller, P. et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
5. Öztürk, H., Özgür, A., Schwaller, P., Laino, T. & Ozkirimli, E. Exploring chemical space using natural language processing methodologies for drug discovery. *Drug Discov. Today* **25**, 689–705 (2020).
6. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
7. Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **29**, 97–101 (1989).
8. Skinnider, M. A., Stacey, R. G., Wishart, D. S. & Foster, L. J. Chemical language models enable navigation in sparsely populated chemical space. *Nat. Mach. Intell.* **3**, 759–770 (2021).

9. Özçelik, R. et al. Chemical language modeling with structured state space sequence models. *Nat. Commun*. **15**, 6176 (2024).
10. O'Boyle, N. & Dalke, A. DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures. (2018).
11. Krenn, M. et al. SELFIES and the future of molecular string representations. *Patterns* **3**, 100588 (2022).
12. Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inform.* **37**, 1700153 (2018).
13. Li, X., Xu, Y., Yao, H. & Lin, K. Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors. *J. Cheminformatics* **12**, 1–13 (2020).
14. Moret, M. et al. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nat. Commun.* **14**, 114 (2023).
15. Gao, W. & Coley, C. W. The synthesizability of molecules proposed by generative models. *J. Chem. Inf. Model.* **60**, 5714–5723 (2020).
16. Seo, S., Lim, J. & Kim, W. Y. Molecular generative model via retrosynthetically prepared chemical building block assembly. *Adv. Sci.* **10**, 2206674 (2023).
17. Jinsong, S., Qifeng, J., Xing, C., Hao, Y. & Wang, L. Molecular fragmentation as a crucial step in the AI-based drug development pathway. *Commun. Chem.* **7**, 20 (2024).
18. Ivanov, N. N., Shulga, D. A. & Palyulin, V. A. Decomposition of small molecules for fragment-based drug design. *Biophysica* **3**, 362–372 (2023).
19. Pogány, P., Arad, N., Genway, S. & Pickett, S. D. De novo molecule design by translating from reduced graphs to SMILES. *J. Chem. Inf. Model.* **59**, 1136–1146 (2019).
20. Podda, M., Bacciu, D. & Micheli, A. A deep generative model for fragment-based molecule generation. *International conference on artificial intelligence and statistics* (2020).
21. Kong, Y. et al. Integrating concept of pharmacophore with graph neural networks for chemical property prediction and interpretation. *J. Cheminformatics* **14**, 52 (2022).
22. Taleongpong, P. & Brooks P. Improving Fragment-Based Deep Molecular Generative Models. in *ICML'24 Workshop ML for Life and Material Science: From Theory to Industry Applications* (2024).
23. Wu, J.-N. et al. t-SMILES: a fragment-based molecular representation framework for de novo ligand design. *Nat. Commun.* **15**, 4993 (2024).
24. Li, X. & Fourches, D. SMILES pair encoding: a data-driven substructure tokenization algorithm for deep learning. *J. Chem. Inf. Model.* **61**, 1560–1569 (2021).
25. Cheng, A. H. et al. Group SELFIES: a robust fragment-based molecular string representation. *Digit. Discov.* **2**, 748–758 (2023).
26. Asudani, D. S., Nagwani, N. K. & Singh, P. Impact of word embedding models on text analytics in deep learning environment: a review. *Artif. Intell. Rev.* **56**, 10345–10425 (2023).
27. Khurana, D., Koli, A., Khatter, K. & Singh, S. Natural language processing: state of the art, current trends and challenges. *Multimed. Tools Appl.* **82**, 3713–3744 (2023).
28. Chuang, K. V., Gunsalus, L. M. & Keiser, M. J. Learning molecular representations for medicinal chemistry. *J. Med. Chem.* **63**, 8705–8722 (2020).
29. Zhang, Q. et al. Scientific Large Language Models: A Survey on Biological & Chemical Domains. Preprint at http://arxiv.org/abs/2401.14656 (2024).
30. Chen, J.-A., Niu, W., Ren, B., Wang, Y. & Shen, X. Survey: exploiting data redundancy for optimization of deep learning. *ACM Comput. Surv.* **55**, 1–38 (2023).
31. Yang, H., Lou, C., Li, W., Liu, G. & Tang, Y. Computational approaches to identify structural alerts and their applications in environmental toxicology and drug discovery. *Chem. Res. Toxicol.* **33**, 1312–1322 (2020).
32. Noutahi, E., Gabellini, C., Craig, M., Lim, J. S. C. & Tossou, P. Gotta be SAFE: a new framework for molecular design. *Digit. Discov.* **3**, 796–804 (2024).
33. Harper, G., Bravi, G. S., Pickett, S. D., Hussain, J. & Green, D. V. S. The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *J. Chem. Inf. Comput. Sci.* **44**, 2145–2156 (2004).
34. Liao, C., Yu, Y., Mei, Y. & Wei, Y. From Words to Molecules: A Survey of Large Language Models in Chemistry. Preprint at https://arxiv.org/abs/2402.01439 (2024).
35. Feller, D. The role of databases in support of computational chemistry calculations. *J. Comput. Chem.* **17**, 1571–1586 (1996).
36. Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Mach. Learn. Sci. Technol.* **2**, 015016 (2021).
37. Van Tilborg, D., Alenicheva, A. & Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. *J. Chem. Inf. Model.* **62**, 5938–5951 (2022).
38. Leon, M., Perezhohin, Y., Peres, F., Popovič, A. & Castelli, M. Comparing SMILES and SELFIES tokenization for enhanced chemical language modeling. *Sci. Rep.* **14**, 25016 (2024).
39. Scharfer, C. et al. Torsion angle preferences in druglike chemical space: a comprehensive guide. *J. Med. Chem.* **56**, 2016–2028 (2013).
40. Irwin, J. J. & Shoichet, B. K. ZINC: a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182 (2005).
41. Park, K., Lee, J., Jang, S. & Jung, D. An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks. Preprint at http://arxiv.org/abs/2010.02534 (2020).
42. Chen, W. et al. How Large a Vocabulary Does Text Classification Need? A Variational Approach to Vocabulary Selection. Preprint at http://arxiv.org/abs/1902.10339 (2019).
43. Mahmoud, H. H., Hafez, A. M. & Alabdulkreem, E. Language-independent text tokenization using unsupervised deep learning. *Intell. Autom. Soft Comput.* **35**, 321 (2023).
44. Hu, X., Chu, L., Pei, J., Liu, W. & Bian, J. Model complexity of deep learning: a survey. *Knowl. Inf. Syst.* **63**, 2585–2619 (2021).
45. Gupta, A. et al. Generative recurrent networks for de novo drug design. *Mol. Inform.* **37**, 1700111 (2018).
46. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
47. Polykovskiy, D. et al. Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Front. Pharmacol.* **11**, 565644 (2020).
48. Yang, Y. et al. Discovery of highly potent, selective, and orally efficacious p300/CBP histone acetyltransferases inhibitors. *J. Med. Chem.* **63**, 1337–1360 (2020).
49. Grisoni, F. et al. Combining generative artificial intelligence and on-chip synthesis for de novo drug design. *Sci. Adv.* **7**, eabg3338 (2021).
50. Skinnider, M. A. Invalid SMILES are beneficial rather than detrimental to chemical language models. *Nat. Mach. Intell.* **6**, 437–448 (2024).
51. Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S. & Klambauer, G. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *J. Chem. Inf. Model.* **58**, 1736–1741 (2018).
52. Wildman, S. A. & Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873 (1999).
53. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminformatics* **1**, 1–11 (2009).
54. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98 (2012).

55. Shultz, M. D. Two decades under the influence of the rule of five and the changing properties of approved oral drugs: miniperspective. *J. Med. Chem.* **62**, 1701–1714 (2018).
56. Panaretos, V. M. & Zemel, Y. Statistical aspects of Wasserstein distances. *Annu. Rev. Stat. Appl.* **6**, 405–431 (2019).
57. Grisoni, F., Moret, M., Lingwood, R. & Schneider, G. Bidirectional Molecule Generation with Recurrent Neural Networks. *J. Chem. Inf. Model.* **60**, 1175–1183 (2020).
58. Gaulton, A. et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).
59. Arús-Pous, J. et al. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminformatics* **11**, 71 (2019).
60. Bjerrum, E. J. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. *ArXiv abs/1703.07076*, (2017).
61. Nigam, A., Friederich, P., Krenn, M. & Aspuru-Guzik, A. Augmenting genetic algorithms with deep neural networks for exploring the chemical space. *ArXiv Prepr. ArXiv190911655* (2019).
62. Yoshikai, Y. Difficulty in chirality recognition for Transformer architectures learning chemical structures from string representations. *Nat. Commun*. **15**, 1197 (2024).
63. Moret, M., Friedrich, L., Grisoni, F., Merk, D. & Schneider, G. Generative molecular design in low data regimes. *Nat. Mach. Intell.* **2**, 171–180 (2020).
64. Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Data augmentation strategies to improve reaction yield predictions and estimate uncertainty. *chemrxiv* (2020).
65. Landrum, G. RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org/ (2016).
66. Tom, G., Yu, E., Yoshikawa, N., Jorner, K. & Aspuru-Guzik, A. Stereochemistry-aware string-based molecular generation. Preprint at https://doi.org/10.26434/chemrxiv-2024-tkjr1 (2024).
67. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
68. Langevin, M., Minoux, H., Levesque, M. & Bianciotto, M. Scaffold-constrained molecular generation. *J. Chem. Inf. Model.* **60**, 5637–5646 (2020).
69. Heid, E. & Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J. Chem. Inf. Model.* **62**, 2101–2110 (2022).
70. Born, J. & Manica, M. Regression Transformer enables concurrent sequence regression and generation for molecular language modelling. *Nat. Mach. Intell.* **5**, 432–444 (2023).
71. Schneider, N., Lewis, R. A., Fechner, N. & Ertl, P. Chiral cliffs: investigating the influence of chirality on binding affinity. *ChemMedChem* **13**, 1315–1324 (2018).
72. Schmidt, R. et al. Comparing molecular patterns using the example of SMARTS: theory and algorithms. *J. Chem. Inf. Model.* **59**, 2560–2571 (2019).
73. Hagberg, A. & Conway, D. Networkx: Network analysis with python. *URL Httpsnetworkx Github Io* (2020).

Acknowledgements

FG acknowledges the support from the Centre for Living Technologies. The authors acknowledge Derek van Tilborg for his comments on the manuscript draft.

Author contributions

Conceptualization: F.M., M.T., O.N., F.C., N.A., N.G., D.T., F.G. Data curation: F.M. Formal analysis: F.M. Investigation: F.M., M.T. Methodology: F.M., F.C., N.A. Software: F.M. Visualization: F.M. Writing, review and editing: all the authors.

Competing interests

The authors declare no competing interests. Francesca Grisoni is an Editorial Board Member for Communications Chemistry, but was not involved in the editorial review of, or the decision to publish this article.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42004-025-01423-3.

Correspondence and requests for materials should be addressed to Francesca Grisoni or Orazio Nicolotti.

Peer review information *Communications Chemistry* thanks Kenneth Atz, Emmanuel Noutahi, and Giustino Sulpizio for their contribution to the peer review of this work.