SELFIES or SMILES? A case study in chemical reaction prediction.

Matthias Kellner, Tim Kircher
Machine Learning CS-433 : Project 2, EPFL

***Abstract*—Chemical reaction prediction is a key process of the drug development pipeline in medicinal chemistry. Accurate prediction of reaction products given a set of reactants is one important aspect of chemical reaction modelling. Recent work on reaction prediction is focused on neural machine translation with Simplified Molecular-Input Line-Entry System (SMILES) molecular representations. SMILES representations are not robust and a large fraction of randomly generated SMILES do not correspond to valid molecules. Due to this, a robust string-based molecular representation Self-referencing Embedded Strings (SELFIES) has been developed. This work explores the comparison of SMILES and SELFIES molecular representations in the context of chemical reaction prediction with transformer models.**

I. Introduction

The abstraction of complex physicochemical processes into rules and simple notations is a key instrument in the toolbox of any chemist. Common examples include skeletal visualisation of organic molecules and representation of reactions mechanisms by electron movement with arrows. Generally, these notations are useful for understanding chemical processes based on a set of heuristic rules. String-based representations of molecular graphs can describe these heuristics.An example for skeletal chemical formulae, corresponding string representations and the reaction prediction workflow are displayed in figure 1 Recently machine learning methods have been explored for chemical reaction prediction.[1], [2] Performing chemical reaction prediction in a supervised machine-learning setting with neural networks necessitates both a suitable model architecture and representation of chemical information.

A. Computational Chemical Reaction Prediction

Chemical reaction prediction describes mapping input molecules to output molecules[1]. Predicting the main product from a reaction mixture is the most simple task, where the input is the reactants and the output is the main product. Computational reaction prediction methods may be divided into template- and template-free methods[2], [3]. Template methods rely on handcrafted expert rules, while template-free methods do not require domain knowledge. In a big data setting, template-free methods are preferable because they are scalable. Template-free methods are further divided into graph-based methods and seq2seq methods. Graph-based methods utilise atom-mapping and map the location of an atom in the reactants, described by adjacency matrices, to the location of the atom in the products. seq2seq modelling utilises methods from neural machine translation to map the reactant string to the product string.

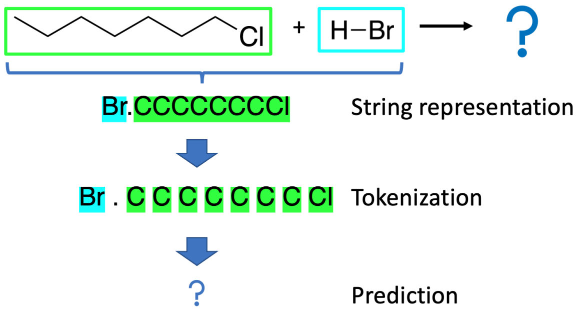


Figure 1. Reaction prediction workflow from skeletal formula of 1-chloroheptane with hydrobromic acid, corresponding string representation, tokenization and product prediction.

B. String-Based Molecular Representations

Application of seq2seq modelling to chemical problems requires a memory-efficient machine-readable representation of chemical information. SMILES is the most common string-based molecular representation and is widely used in computational chemistry.[2] SMILES is based on graph theory and molecular information is embedded into letters and symbols by a simple set of rules[4], [5]. Critically, the representation power of SMILES is not restricted to chemically feasible molecules.[3] It is possible that a SMILES, randomly generated from the SMILES grammar, to correspond to a non-valid molecule. This poses a significant design flaw for generative models. SELFIES are an alternative, robust string-based molecular representation.[3] Representation rules for SELFIES are derived from a systematically derived formal grammar. Each SELFIE string, even randomly generated SELFIE strings, corresponds to a valid molecule. Figure 2 shows the canonical SMILES and the corresponding SELFIE representation of 3,4-methylenedioxymethamphetamine.

SELFIES were developed for inverse molecular design valid chemical structures are important[3], [6]. Additionally, SELFIES have shown promising results in neural machine translation of IUPAC structures to molecular representations[7] and image recognition tasks of skeletal molecular formulae[8]. Finally, seq2seq reaction prediction is an important application of molecular representations to

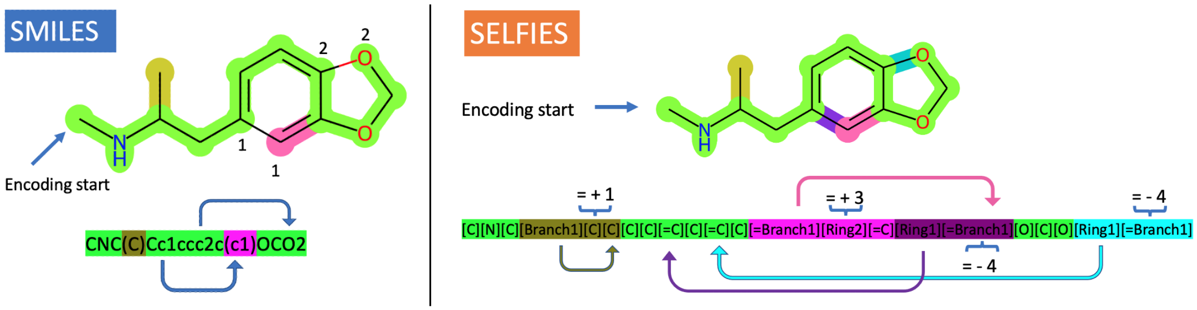

SMILES
2
2
O
O
N
H
1
1
Encoding start
CNC(C)Cc1ccc2c(c1)OCO2
SELFIES
Encoding start
O
O
N
H
= + 1
= + 3
= - 4
[C][N][C][Branch1][C][C][C][C][=C][C][=C][C][=Branch1][Ring2][=C][Ring1][=Branch1][O][C][O][Ring1][=Branch1]
= - 4

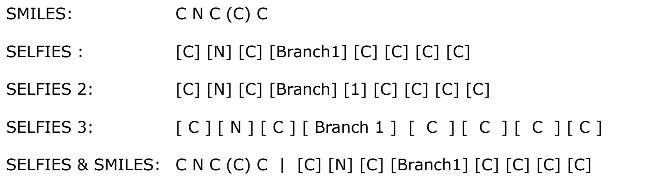
Figure 2. Canonical SMILES and corresponding SELFIES encoding of 3,4-methylenedioxymethamphetamine.

chemical modelling, as well. To the best of our knowledge, there exists no work on the comparison of molecular representation for seq2seq reaction prediction. This work compares molecular representations for the seq2seq modelling based on neural machine translation with transformer models.

II. Data

We used the USPTO-480k chemical reactions dataset in this work. It is derived from patent mining by Lowe et. al [9][10] and used as a benchmark set in the research community. The reactions are described by SMILES and are separated into products and reactants. The data set contains a total of 480,000 reactions. The data set is split into a train, validation and test set containing 410000, 30000 and 40000 reactions. The data is publicly available at https://github.com/coleygroup/Graph2SMILES. The data is tokenized according to the tokenizer available at https://github.com/pschwllr/MolecularTransformer. SELFIES are generated from SMILES and tokenized by the SELFIES package https://github.com/aspuru-guzik-group/selfies. SMILES Molecules are canonicalized and checked for validity by RDKit. The conversion of canonicalized from SMILES to SELFIES and back to SMILES yielded different strings before and after the conversion for molecules containing hypervalent atoms such as phosphorous or transition metals. To circumvent this problem and ensure a closed-loop for conversion between SELFIES and SMILES, hyper-valencies are allowed for SELFIES strings.

Two SELFIES tokenization and different data augmentation approaches are explored. For the baseline model SELFIES are tokenized according to Krenn et. al[3]. Separating connected tokens, such as [Branch1] or [Ring1] is explored in SELFIES 2. Further tokenization of all SELFIES tokens is performed in SELIFES 3. Both additional tokenization approaches try to decrease the information content per SELFIES token. For data augmentation, a training set containing SMILES reactants and SELFIES products and vice-versa are explored, where SMILES and SELFIES molecular representations are separated by ”|” and the representations do not share any tokens. Finally, a training set containing both SMILES and SELFIES representations of each reactant and product is investigated. Figure 3 shows an example for the tokenization of N-methylbutan-2-amine.



SMILES:	C N C (C) C
SELFIES :	[C] [N] [C] [Branch1] [C] [C] [C] [C]
SELFIES 2:	[C] [N] [C] [Branch] [1] [C] [C] [C] [C]
SELFIES 3:	[C] [N] [C] [Branch 1] [C] [C] [C] [C]
SELFIES & SMILES:	C N C (C) C | [C] [N] [C] [Branch1] [C] [C] [C] [C]

Figure 3. Tokenization approaches exemplified for N-methylbutan-2-amine.

III. Computational Methods

The data and reaction predictions were handled and analysed with the RDkit [11] cheminformatics toolkit and the RDkit compatible SELFIES code (available on github).[12]

The model used in this work is based on the carbohydrate transformer by Schwaller et. al [13]. The transformer architecture is implemented with the OpenNMT-py framework [14]. Hyperparameters are taken from https://github.com/rxn4chemistry/OpenNMT-py/tree/carbohydrate_transformer. The transformer utilises an autoregressive encoder-decoder architecture with multiheaded attention and positional feed-forward layers. The model has 4 hidden layers with 2048 neurons and 8 attention heads. For training, the ADAM optimizer is utilised with 8000 warm-up steps. The models are trained for 250000 steps. A dropout of 0.1 and gradient normalisation over 4 batches are utilised for regularisation. In contrast to the original NMT model, label smoothening is set to 0. Models are evaluated by predicting reaction products by beam search with a beam width of 5. Due to long model training time (>24 h per model), limited computational resources and the limited time frame of the semester project hyperparameter optimisation is not explored and the identical set of hyperparameters is utilised for all models.

IV. Results

We trained the transformer model on the USPTO-480K reaction data set. The resulting Top-1 and Top-5 validation set accuracies on SMILES and SELFIES using different tokenization and data augmentation approaches are evaluated. Top 1 accuracy denotes the model accuracy when predicting a single product and Top 5 accuracy denotes the number of correct predictions when 5 possible reaction products are predicted by beam search. For augmented SELFIES & SMILES the SMILES and SELFIES the output is a string containing a predicted SELFIES and a predicted SMILES string separated by ”|”. The SELFIES and SMILES strings are evaluated separately (augmented SMILES and augmented SELFIES). The molecular transformer[2], trained and validated on the identical data set, is included as a benchmark.

Molecular Representation	Top 1 accuracy [%]	Top 5 accuracy [%]
SMILES	86.0	92.3
SELFIES	75.8	86.1
SELFIES 2	73.3	83.9
SELFIES 3	68.8	80.2
Augmented SMILES	83.8	89.4
Augmented SELFIES	74.9	86.2
Molecular Transformer [2]	88.8	94.4

Table I
RESULTS FOR TOKENIZATION AND DATA AUGMENTATION APPROACHES ON THE VALIDATION SET.

Table I shows that SMILES molecular representation has the highest validation accuracy of the analysed molecular representations. SELFIES top-1 accuracy is significantly lower by 10 %. Further SELFIES tokenization approaches did not yield an increased accuracy. Furthermore, data set augmentation did not increase accuracy. The accuracy for both SELFIES and SMILES with the augmented data set is slightly lower than without augmentation. The model is not able to transfer the context of SELFIES to SMILES or vice-versa. All models are inferior to the benchmark transformer model. This is attributed to the lack of hyperparameter optimisation and possibly the increased number of parameters of our model that was adopted from the newer Carbohydrate Transformer and not the reference Molecular Transformer. Our best performing SMILES model has a test accuracy of 86.4 %.

SELFIES and SMILES models are further analysed to investigate the difference in model performance. Figure 4 shows the Top-1 model accuracy and model steps. It also shows the progression towards the final accuracy over the first 125,000 steps. Figure 4 shows that the SMILES model converges significantly faster than the SELFIES model. Also, it shows that both models are converged after 250,000 steps, as the model accuracy does not increase after 125,000 steps. Both models reach their final accuracy after 125,000 steps. Next, the prediction for SMILES and SELFIES model for each reaction in the validation set is compared. Figure 5 shows a Venn diagram for shared true products, products only identified in the SMILES model and products only identified in the SELFIES model.

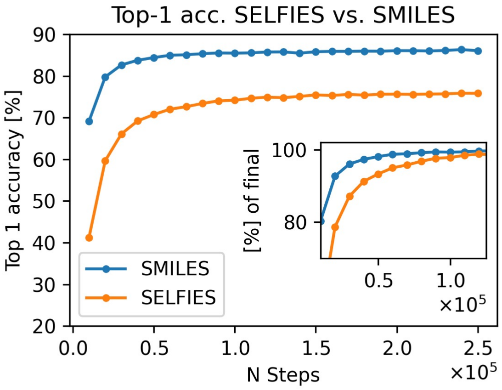


Figure 4. Top-1 Validation Accuracy of the SELFIES and SMILES Model.

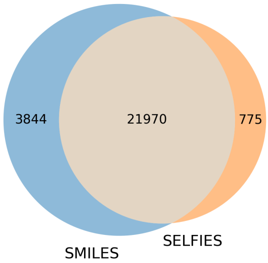


Figure 5. Correctly identified products for both models, only SMILES and only SELFIES model.

Figure 5 shows that the models for both molecular representations share a large number of correct predictions. Additionally, the SMILES model is able to predict a significant number of reactions correctly for which the SELFIES prediction is false. The number of reactions which the SELFIES model exclusively predicts correctly is far lower. This leads to the question if the difference in prediction quality is related to the chemical reaction space or the string representation encoding.

To address the first question of the relation between model proficiency and the chemical reaction space, the validation set is mapped in a tree plot[15]. Figure 7 (Appendix) shows a tree plot of the validation set. Similar reactions are are clustered and the color of a reaction data point denotes if the reaction is predicted correctly in no model, in both models, only in the SMILES model and only in the SELFIES model. Generally, 73 % of all reactions are predicted correctly by both SELFIES and SMILES model and thus the majority of the plot is green. For the case of reactions predicted only by SELFIES or only by SMILES, no clusters can be identified. Reactions that are exclusively predicted by one of the two models share no common characteristics in the chemical space.

Next, the number of chemically valid products are investigated. SELFIES are a robust molecular representation and excel in generative tasks because all SELFIES strings correspond to valid molecules. Table II shows number of chemically valid molecules predicted by the SELFIES and SMILES model. Invalid molecules are detected by RDKit utilising the DetectChemistryProblems function. SELFIES are converted to SMILES to check for validity.

Molecular Representation	Valid Molecules [%]
SMILES	99.3
SELFIES	98.3

Table II
PERCENTAGE OF CHEMICALLY VALID MOLECULES PREDICTED BY THE MODEL ON THE VALIDATION SET ACCORDING TO VALENCY RULES.

Surprisingly, not all SELFIES correspond to valid molecules. The proportion of chemically invalid molecules is higher for SELFIES than for SMILES. We attribute this to the hyper-valency settings. When hypervalencies are not allowed during the translation of predicted SELFIES to SMILES, 99.99% of predicted molecules follow standard valency rules. For the case of the SMILES model, 0.7% of the predicted products are invalid. Chemical validity is a main advantage of the SELFIE representation. If hypervalent atoms are present in the data set, interchangeability between SELFIES and SMILES is not possible without allowing for hyper-valency. Unfortunately, this impairs the advantage of the robust SELFIES representation.

Exploration of the chemical space shows that reaction chemistry and model proficiency are not strongly related. We believe the prediction quality for SELFIES is impaired by token ambiguity. SELFIES tokens are ambiguous and either encode an atom or a position, depending on the context. Tokens following "[Branch]" or "[Ring]" indicate the size of the respective element. "[Branch]" and "[Ring]" tokens include a number, for example "[Branch1]", which denotes the number of following tokens used for indexing. An example of ambiguity is the "[C]" token, the most frequent token in the training set, which indicates the index 0 or a carbon atom. The string "[Branch1] [C] [C]" references a branch of size 1 with 1 carbon atom. Figure 6 shows the identical molecule as figure 2 on the left. The string representations of 2 and 6 differ slightly, as a different encoding start position is chosen to exemplify SELFIES ambiguity.[1]On the right, a carbon atom is exchanged for a phosphine group.

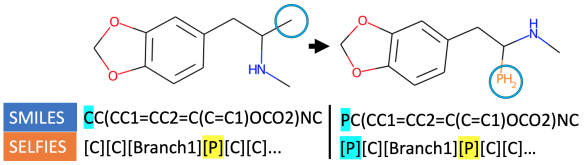

O
O
HN
H
N
O
O
PH_2
SMILES
SELFIES
CC(CC1=CC2=C(C=C1)OCO2)NC
[C][C][Branch1][P][C][C]...
PC(CC1=CC2=C(C=C1)OCO2)NC
[P][C][Branch1][P][C][C]...

Figure 6. This figure shows the ambiguity of the SELFIES tokens. After a point mutation (a CH3 to PH2 group), the P or phosphorous token only appears once in the SMILES string, but twice in the SELFIES string, encoding the branch length, as well.

The SELFIES representation on the left includes a "[Branch1]" token followed by a "[P]" token representing a branch of 4 tokens. After introducing a phosphine group, the SELFIES string on the right contains two "[P]" with vastly different meanings. The SMILES string contains no ambiguity regarding atoms and positions. The inferior performance of SELFIES when compared to SMILES for transformer applications is mirrored by similar work, where SMILES and SELFIES are compared for chemical image recognition.[16].

V. Summary & Outlook

In this work, we find that the SELFIES molecular graph representation is outperformed by the more established SMILES representation in reaction prediction using transformer models.
Further analysis of the prediction outcomes shows no correlation between prediction accuracy and selected reaction types for both representations. Conclusively reaction prediction models using SMILES and SELFIES predictions share a majority of correctly predicted reactions. We attribute the slower convergence and lower prediction accuracy of the SELFIES model to the ambiguous encoding of ring- and branch closing connections in SELFIES strings. Albeit necessary for generative tasks, such as the discovery of novel molecular scaffolds, this ambiguity most likely can not be captured by the transformer model. So far we have not made use of the opportunities that SELFIES offer through the mutation of chemical structures exploited in generative applications. In further work, ambiguous tokens could be replaced by tokens that are not part of the SELFIES grammar, as the validity of the SELFIES after point mutation does not have to be ensured in prediction applications.

[1] When using string-based molecular graph descriptors, the encoding start can be freely chosen, resulting in varying strings depending on the encoding start. Cheminformatics toolkits such as the RDkit have algorithms implemented that can transform all of the possible representation strings of one molecular graph into one unique, so-called canonical string. The different SMILES encoding position has been chosen to illustrate the ambiguity of the [P] SELFIES token, after transforming the aforementioned non-canonical SMILES.

References

[1] T. I. Madzhidov, A. Rakhimbekova, V. A. Afonina, T. R. Gimadiev, R. N. Mukhametgaleev, R. I. Nugmanov, I. I. Baskin, and A. Varnek, "Machine learning modelling of chemical reaction characteristics: yesterday, today, tomorrow," *Mendeleev Communications*, vol. 31, no. 6, pp. 769–780, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959943621002959

[2] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction," *ACS Central Science*, vol. 5, no. 9, pp. 1572–1583, 2019, pMID: 31572784. [Online]. Available: https://doi.org/10.1021/acscentsci.9b00576

[3] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation," *Machine Learning: Science and Technology*, vol. 1, no. 4, p. 045024, oct 2020. [Online]. Available: https://doi.org/10.1088/2632-2153/aba947

[4] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *J. Chem. Inf. Comput. Sci.*, vol. 28, pp. 31–36, 1988.

[5] D. Weininger, A. Weininger, and J. L. Weininger, "Smiles. 2. algorithm for generation of unique smiles notation," *J. Chem. Inf. Comput. Sci.*, vol. 29, pp. 97–101, 1989.

[6] Alan Aspuru-Guzik, "Molecular graph representations and selfies: A 100% robust molecular string representation," 2021. [Online]. Available: https://aspuru.substack.com/p/molecular-graph-representations-and

[7] "Stout: Smiles to iupac names using neural machine translation," *Journal of Cheminformatics*, vol. 13, no. 1, p. 34, 2021.

[8] K. Rajan, A. Zielesny, and C. Steinbeck, "Decimer: towards deep learning for chemical image recognition," *Journal of Cheminformatics*, vol. 12, no. 1, p. 65, 2020.

[9] D. M. Lowe, "Extraction of chemical structures and reactions from the literature," Thesis, University of Cambridge, Oct. 2012, accepted: 2013-07-23T08:23:10Z. [Online]. Available: https://www.repository.cam.ac.uk/handle/1810/244727

[10] W. Jin, C. W. Coley, R. Barzilay, and T. Jaakkola, "Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network," *arXiv:1709.04555 [cs, stat]*, Dec. 2017, arXiv: 1709.04555. [Online]. Available: http://arxiv.org/abs/1709.04555

[11] "RDKit." [Online]. Available: http://www.rdkit.org/

[12] "SELFIES," Dec. 2021, original-date: 2019-05-31T16:33:37Z. [Online]. Available: https://github.com/aspuru-guzik-group/selfies

[13] G. Pesciullesi, P. Schwaller, T. Laino, and J.-L. Reymond, "Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates," *Nature Communications*, vol. 11, no. 1, p. 4874, 2020.

[14] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-source toolkit for neural machine translation," in *Proceedings of ACL 2017, System Demonstrations*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 67–72. [Online]. Available: https://www.aclweb.org/anthology/P17-4012

[15] P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino, and J.-L. Reymond, "Mapping the space of chemical reactions using attention-based neural networks," 2020.

[16] K. Rajan, C. Steinbeck, and A. Zielesny, "Performance of chemical structure string representations for chemical image recognition using transformers," Oct. 2021. [Online]. Available: https://doi.org/10.33774/chemrxiv-2021-7c9wf-v2

VI. Appendix

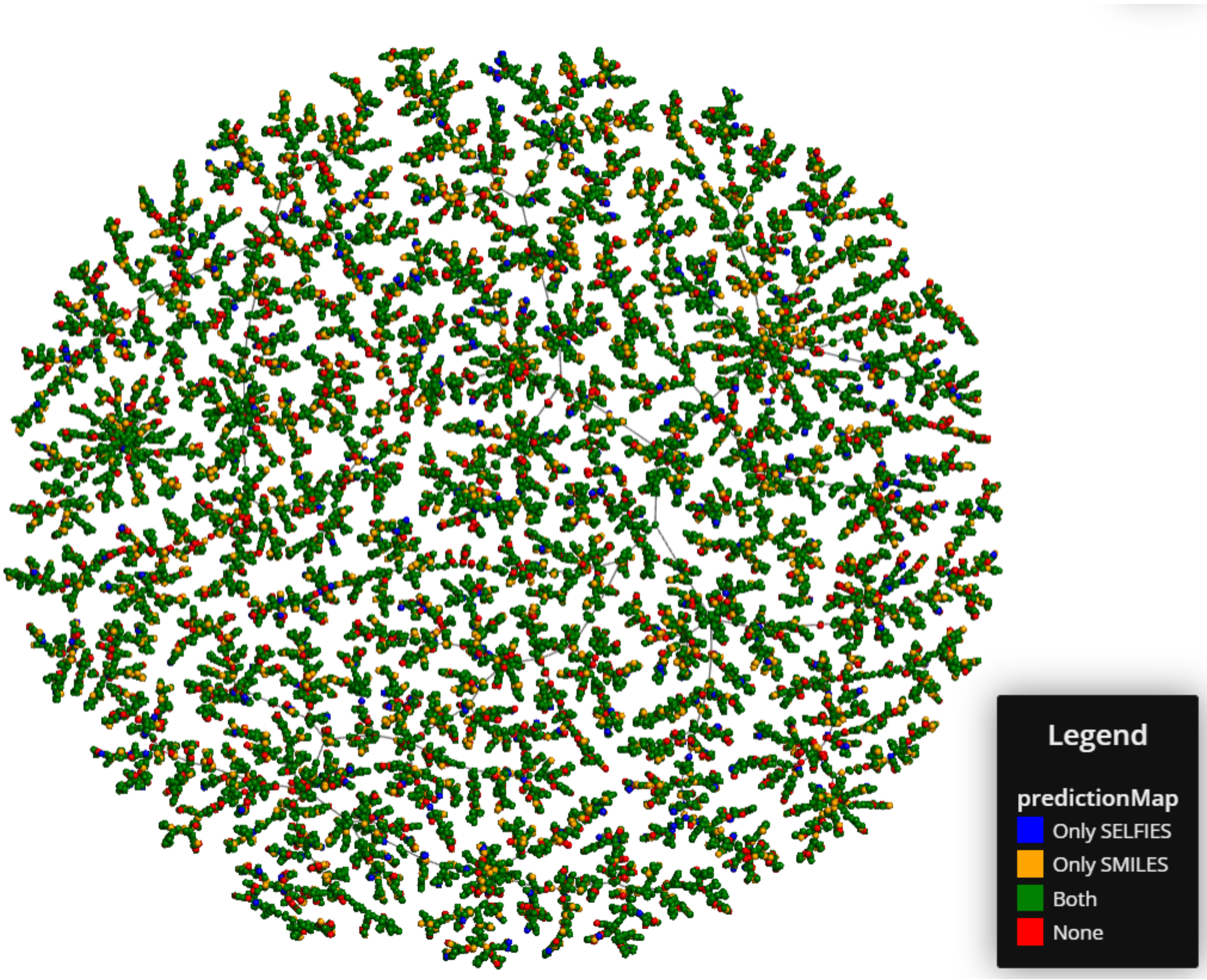


Figure 7. Correctly identified reactants for both models, only SMILES and only SELFIES model and none grouped by product similarity.