# Bio-inspired Chemical Space Exploration of Terpenoids

Tao Zeng[1], B. Andes Hess, Jr. [2], Fan Zhang[1], Ruibo Wu[1,*]

[1]*School of Pharmaceutical Sciences, Sun Yat-sen University, Guangzhou 510006, P.R. China*

[2]*Department of Chemistry, Vanderbilt University, Nashville, TN 37235 (USA)*

* **E-mail**: wurb3@mail.sysu.edu.cn

**Abstract**

Many computational methods are used to expand the open-ended border of chemical spaces. Natural products and their derivatives are an important source for drug discovery, and some algorithms are devoted to rapidly generating pseudo-natural products, while their accessibility and chemical interpretation were often ignored or underestimated, thus hampering experimental synthesis in practice. Herein, a bio-inspired strategy (named TeroGen) is proposed, in which the cyclization and decoration stage of terpenoid biosynthesis were mimicked by meta-dynamics simulations and deep learning models respectively, to explore their chemical space. In the protocol of TeroGen, the synthetic accessibility is validated by reaction energetics (reaction barrier and reaction heat) based on the GFN2-xTB methods. Chemical interpretation is an intrinsic feature as the reaction pathway is bioinspired and triggered by the RMSD-PP method in conjunction with an encoder-decoder architecture. This is quite distinct from conventional library/fragment-based or rule-based strategies, by using TeroGen, new reaction routes are feasibly explored to increase the structural diversity. For example, only a rather limited number of sesterterpenoids in our training set is included in this work, but our TeroGen would predict more than 30000 sesterterpenoids and map out the reaction network with super efficiency, ten times as many as the known sesterterpenoids (less than 2500). In sum, TeroGen not only greatly expands the chemical space of terpenoids but also provides various plausible biosynthetic pathways, which are crucial clues for heterologous biosynthesis, bio-mimic and chemical synthesis of complicated terpenoids.

## Introduction

A compound library is essential for drug discovery, and the structural diversity is key to the original innovation of drug design. It has been estimated that $10^{60}$ small organic molecules could possibly populate the chemical space[1,2], which is much larger than the $10^8$ actual compounds found so far[3]. Natural products make up only a tiny fragment ($<10^6$) of the chemical space[4], whose importance for medicinal chemistry[5-7] is widely admired, mostly owing to their diverse structures[8]. In order to expand the chemical space of natural products, many experimental approaches have been reported to accelerate the discovery of natural products[9-11], or synthesize pseudo-natural products inspired by natural products[12-14]. The virtual libraries of natural products or pseudo-natural products were also generated rapidly by computational methods such as recursive atom-based enumeration[15] and reaction rule-based exploration[16].

Molecule generation is widely used to explore the chemical space. Recent developments in deep learning have resulted in various generative models for *de novo* structure generation such as recurrent neural networks (RNN), variational autoencoder (VAE) and generative adversarial networks (GAN), in which the molecules were basically represented by a simplified molecular input line entry specification (SMILES) or molecular graphs.[17,18] Segler et al.[19] proposed an RNN model to generate focused molecule libraries correlated with the training data. MolGAN[20] is an implicit generative model for higher validity and novelty molecular graphs using reinforcement learning. In addition, the scaffold-based molecular design has also evoked the interest of researchers in this filed, by which derivative compounds retaining a particular scaffold can be generated as needed. Lim et al.[21] developed a VAE model that accepts a molecular scaffold as input and extends it to generate derivative molecules. There is also a SMILES-based generative architecture that can generate molecules for a scaffold by specifying its attachment points.[22] For natural products, Zheng et al.[23] reported a quasi-biogenic molecule generator with RNN that is able to generate focused libraries biased on a specified scaffold. These deep generative models can generate a large set of novel and even customized structures, but the synthesizability is not considered in most cases, which hampers their utility.[24] To consider the accessibility, the reaction-based generative models have been reported recently[25,26], while the reliability of the results relies on the generality/specificity of the reaction rules, which is very difficult to make a trade-off[27].

In contrast to the above data-driven methods, chemical reaction mechanism exploration based on chemical theory, e.g. quantum chemistry, will guarantee the depth, reliability and accuracy in the process.[28] Mechanisms have always been essential for the elucidation of fundamental chemical processes and further analysis of complex chemical networks such as prediction of product yields or pollutant formation.[29] Thus,

taking the reaction mechanisms into account in molecule generation will not only improve the reliability of generated molecules but also help in subsequent synthesis. Recently, *ab initio* molecular dynamics (AIMD) simulation was developed to explore the reaction space of molecules without any prior knowledge, coupled with automatic analysis and refinement methods to build a quantitatively accurate reaction network.[30] The application of this method to Urey-Miller chemistry resulted in the formation of amino acids and other products from hydrogen, ammonia, methane, carbon monoxide and water, which shows the significant strengths of chemical space exploration. However, despite the GPU acceleration, the computational cost is still very high[31]. In addition, the enumeration with predefined reaction types[32] and sampling of the carbocation potential energy surface by an artificial force induced reaction (AFIR) method[33] were utilized to explore the reaction space of several mono- and sesqui-terpenoid natural products, respectively. However, the computational cost is also enormous and thus has limited their application. A recent alternative approach[34] was described that explores the reaction space with metadynamics simulations based on the tight-binding quantum chemistry method GFN2-xTB[35]. A so-called RMSD-PP was employed to find the reaction path and transition states (TSs) just by applying two bias potentials that "pushes" the molecule away from the substrate and "pulls" the molecule towards the product. It has been proved that this approach is accurate and fast enough to automatically identify promising products as well as TSs from substrates in different reaction types[36,37].

Although organic synthesis has become a highly powerful art in creating new molecules, the organic reaction rules are not suitable for generation or synthesis of the majority of natural products due to structural complexity[38]. In nature, enzymes use limited reaction rules to create many natural products, and the reaction types involved are far more than the known reaction types in organic small molecule synthesis. An example are the terpenoids, which are the largest family of natural products and a major source of drug discovery.[39] They are some of the most synthetically challenging structures due to their complex and stereochemically-rich polycyclic ring systems[40,41]. Hence the biosynthesis-driven strategies including metabolic engineering[42], semisynthesis[43] and biomimetic synthesis[44] provide alternative methods for producing high-value terpenoids. As shown in Figure 1a, starting with the $C_{5n}$ isoprenoid diphosphates (n = 1, 2, 3, etc.), the biosynthesis of terpenoids is initiated by the cyclizations and carbocation rearrangements, which lead to the formation of terpenes. And further oxidation and optional post-decoration are wide spread in most of the species[45]. The diverse carbocation rearrangement and functionalization in post-decoration produces a variety of terpenoid structures.[46,47] In view of the "biosynthetic tree" of terpenoids (Figure 1a), can we learn from nature to explore the chemical space based on the knowledge of their biosynthetic mechanisms?

By considering the biosynthetic logic of terpenoid natural products, we introduce a bio-inspired strategy for the chemical space exploration of terpenoids (Figure 1b), named TeroGen. In the protocol of TeroGen, first metadynamics simulations were used to explore the carbocation rearrangement reactions (Reactor), where the reaction barrier and energy can be estimated rapidly. Then a Transformer and an encoder-decoder RNN neural network architecture were applied to find the decorating sites and functional groups respectively (Decorator).

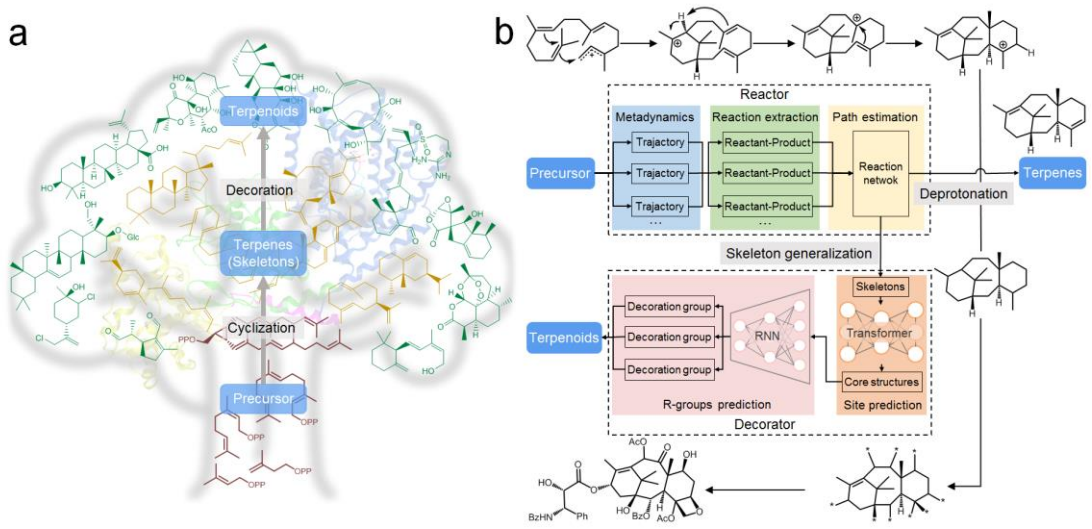


Figure 1. (**a**) The "biosynthetic tree" of terpenoids, from dozens of $C_{5n}$ isoprenoid diphosphate precursors (brown) to thousands of terpenes (yellow) and finally to the hundreds of thousands of terpenoids (green). (**b**) The workflow of TeroGen, which consists of Reactor (exploring carbocation reaction space) and Decorator (predicting the decorating sites as well as functional groups based on the carbocation skeletons and further assembling them into the final terpenoids). The biosynthesis of Taxol was taken as example for this protocol: OPP, diphosphate; Ac, acetyl; Glc, glucose; GGPP, geranylgeranyl diphosphate; Ph, phenyl; Bz, benzyl; "*", means the predicted sites.

## Results

**Validation of Reactor**. To investigate the performance of our carbocation Reactor, four precursors of sesquiterpenoids were selected as the initial structures. A reaction network (Figure 2a) is ultimately mapped out after two rounds of metadynamics simulations in which 6631 unique reactions and 4767 carbocations were generated. 93.5% of reactions had reasonable barriers (here defined as <30 kcal/mol as shown in Figure 2b, and it was 90% if the threshold value is set at 25 kcal/mol), hence most of those reaction pathways are kinetically feasible. Meanwhile, 83.5% out of the 4767

carbocations are yielded by means of an exothermic reaction or endothermic process with less than 10 kcal/mol (also shown in Figure 2b and it was 89.4% if 15 kcal/mol is defined as the cutoff). Considering that those carbocations are serving as intermediates and will ultimately transform into neutral terpene products or ongoing further decoration, heat release is reasonable and the complete reaction pathway is plausibly thermodynamically favorable. In this sense, the reactions explored by the Reactor are mostly accessible with rational energetics characteristics in thermodynamics and kinetics.

In addition, all of the plausible carbocation intermediates directly connected to the bisabolyl cation as proposed by Tantillo[48] could be rapidly sampled by the Reactor in its first round. After two rounds, the Reactor covered all the rearrangements, intermediates and products. In addition, 41 terms of reactions were selected for further validation with high-level DFT calculation results[48,49]. The thermodynamic feasibility for different reaction styles for those reactions were investigated, as summarized in Figure 2c, the exothermicity and endothermicity could be perfectly reproduced for the widely existing cyclizations and also as well for H-shift and alkyl-transfer. Finally, it should be noted that the reaction energetics predicted by Reactor are not suitable for quantitative criteria as a semiempirical DFT method (GFN2-xTB) was used in Reactor. Nevertheless, considering the powerful sampling ability and the predicted reaction energetics are basically qualitatively consistent with the high-precision method. The Reactor is reliable and powerful for the exploration of reaction space triggered by a reactive carbocation, which is ubiquitous and fundamental in the biosynthesis of cyclic terpenes and diverse carbon skeletons.

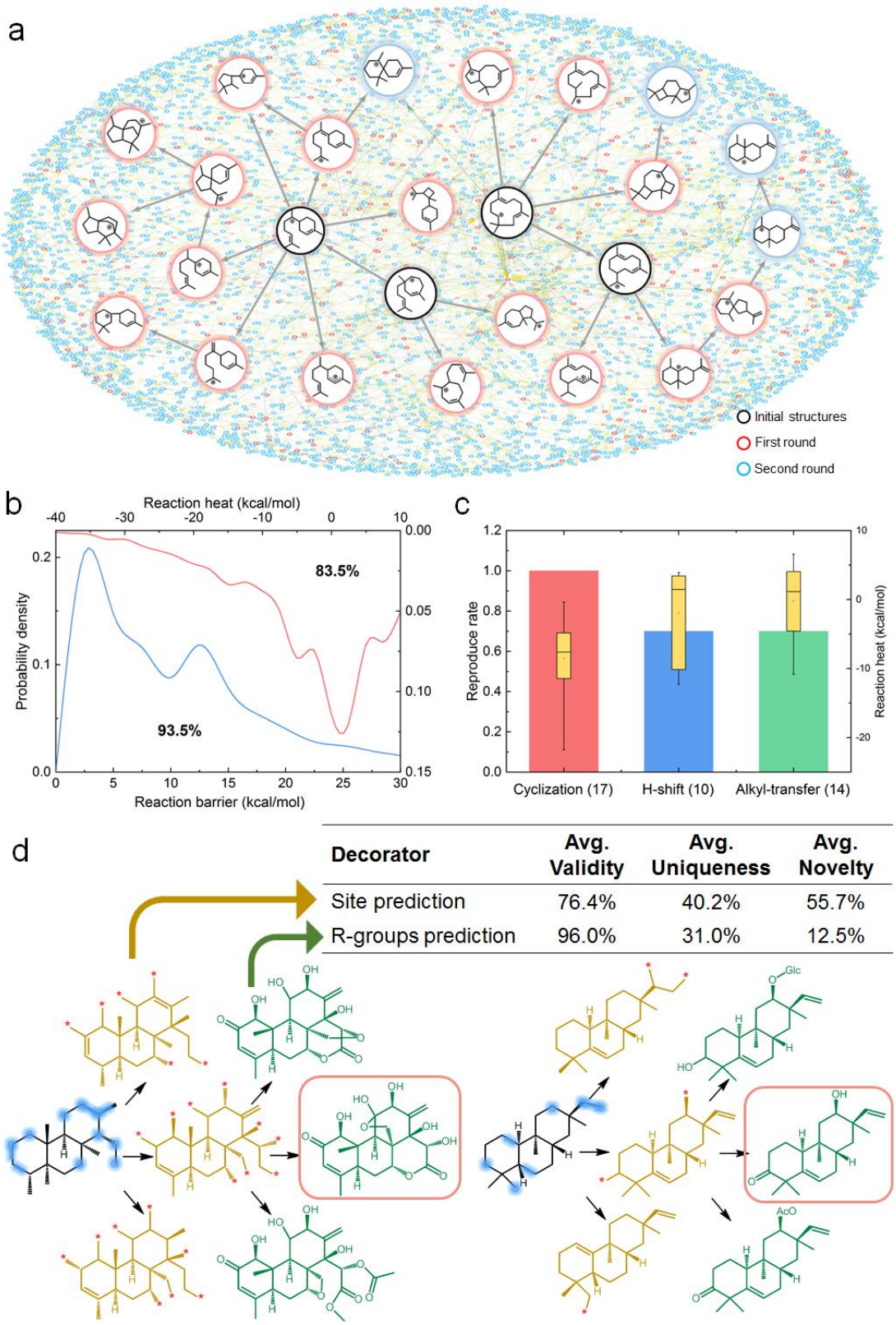


Figure 2. Validation of Reactor (a/b/c) and Decorator (d). (**a**) The predicted reaction network originating from four typical precursors of sesquiterpenoids as initial structures. The nodes and edges represent the carbocations and reactions, respectively. Several intermediates or products are highlighted for showing detailed structures, and a more

complicated network would be obtained if the number of rounds is increased in the "Reactor". (**b**) The reaction barriers and heat release for all reactions captured in the network. (**c**) The reproduction rate (histogram) and distribution (boxplot) of the exothermicity and endothermicity for different reaction styles, the high-level DFT (mPW1PW91/6-31+G(d,p)//B3LYP/6-31+G(d,p)) calculation results are extensively reported by Tantillo[48,49]. (**d**) The performance of Decorator for the modification-site and R-groups prediction and two examples taken for representing the workflow of core structures and terpenoids generation from two similar $C_{20}$-skeletons in the test set. All of the potential decoration sites and bonds were highlighted with blue and the points for linking substitute were noted by asterisks. The experimentally discovered terpenoids were shown in red box.

**Validation of Decorator.** After a series of data collection and notation, as well as model pre-optimization, the ensemble of canonical and mixed Transformer model was then used as a "site prediction" strategy, and the encoder-decoder RNN neural network architecture was employed for the "R-group prediction" strategy. As summarized in Figure 2d, the accuracy, validity, uniqueness, and novelty were used to estimate the performance of the Decorator. For site prediction only the outputs sharing the same carbon skeleton with the input were considered to be valid. And for the R-groups prediction, the validity was measured from the percent of the completely decorated output. The uniqueness was the ratio of nonredundant output to valid output. In site prediction, the output that contains at least one decorating site that is not in the known structure of terpenoids was defined as a novel output. While in the R-groups prediction, the output R-groups that contain at least one R-group that is not present in the training set was defined as novel output. The novelty was calculated as the ratio of a novel output to unique output.

Based on these criteria, most of the outputs (76.4%) of site prediction were valid SMILES and 40% of them were unique. Therefore, the model can reconstruct the core skeleton structures and find the correct decorating sites as validated by the known structures. More importantly, novel, potential decoration sites would also be found, as shown by the average percentage (55.7%) of at least one undiscovered modification site in each skeleton. Thus the exploration ability of new decoration sites is warranted by this ensemble model established by the Transformer architecture. For the R-groups prediction, the sampling process yielded a total of 6287 decorated molecules from 300 core structures, and 199 (66.3%) of them were found to be validated natural molecules. 96.0% of the output were completely decorated molecules, in which the average uniqueness is 31.0%. And 769 (12.5%) of the molecules contain at least one R-group not presented in the training set. This indicates that the end-to-end model of Decorator has learned the features of functional groups in the natural terpenoids while in the

meantime, it brings the creativity of R-groups, which is important to the expansion of chemical space. For example, several outputs derived from the two diterpenoid carbon skeletons predicted by of Decorator are presented in Figure 2d. We can see that the Decorator can distinguish between these two skeletons in spite of their high similarity, and not only the "optimal" decorating sites are detected but a series of R-group substituted terpenoids derivatives includes ones discovered in nature. All these results establish that the Decorator can produce terpenoid-like structures by learning the manner of natural modification.

By combining the Reactor and Decorator, the terpenoids exploration strategy named TeroGen is established as summarized in Figure 1b. It should be emphasized that TeroGen is not a rule-based approach but a physically based (metadynamics with semiempirical DFT) and learning-based (Transformer for site prediction and an encoder-decoder RNN neural network for R-group functionalization) hybrid strategy. Nevertheless, it can be used to greatly expand the chemical space of terpenoids following the rules of terpenoids biogenesis, since the principle for those method/models employed in TeroGen aims to closely follow the biosynthetic rules of terpenoids and its modifications. Next, its performance for sesterterenoids is discussed.

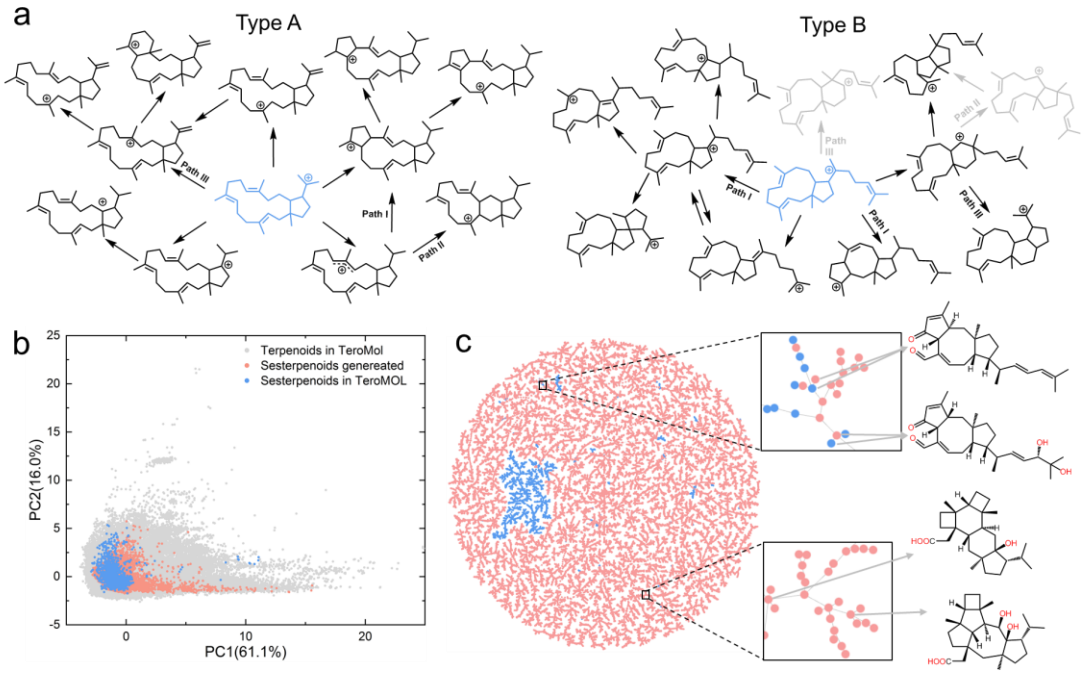


Figure 3. (**a**) Some typical carbocations generated by TeroGen. The modelling is starting from the two structures (blue) which were thought to be the key intermediates distinguished by the type A and type B cyclization pathways for sesterterenoids. The already proposed paths are marked, and those not sampled by Reactor (2 rounds) are colored gray. (**b**) The chemical space overlap for the terpenoids collected in TeroMOL and generated sesterterpenoids by TeroGen in this work. (**c**) The chemical space of all sesterterpenoids (blue) in TeroMOL and the generated ones (red) by TeroGen.

**Sesterterpenoids generation.** The TeroGen protocol (Figure 1b) was then used to explore the subspace of terpenoids, sesterterpenoids (with $C_{25}$ skeleton), whose number is only about 2500. Sesterterpenoids consist of are many fewer examples than the diterpenoids (with a $C_{20}$ skeleton) and triterpenoids (with a $C_{30}$ skeleton), both of which have about 40000 structures. Over the years, the bifunctional di-/sester-terpene synthases called cyclopentane-forming terpene synthases (CPF-TSs)[50] have caught much attention since they produce many chiral, congested and structurally complex polycyclic molecules. In the early stage of the ring construction, cation-mediated double annulation gives a 5/15 (type A) or a 5/11 bicyclic intermediate (type B).[51] In this work, first with Reactor, these two well validated intermediates were selected as the starting point to explore the carbocation reaction space, with 1374 carbocations and 2709 reactions output. A large number of carbocation rearrangements were detected during the metadynamics simulations including those already proposed mechanisms reviewed by Oikawa and Minami et al.[50] (Figure 3a). For the intermediates not sampled by Reactor, some of them may skipped in the concerted but asynchronous reactions to avoid the high energy secondary carbocations, which is proposed in previous gas phase calculation done by Hess and Tantillo[52,53], such as the path II of type B cyclization for sesterterpenoids. Secondly, followed by the Decorator, 3553 sesterterpenes were obtained by deprotonation and a total of 1716 carbon skeletons were generalized for further decoration. Since the number of sesterterpenoids in the training data is rather small, the validity of ensemble model in site prediction was low (the valid core structures were predicted for about 600 skeletons). To better explore the chemical space of sesterterpenoids, the mix model was used for site prediction, by which more than 1500 skeletons could be predicted with about 7000 valid core structures (top 5 are outputted). Eventually a total of 34439 sesterterpenoids were generated by the Decorator, with only about 50 of them existing in the TeroMOL database. This also indicated us that for site prediction, the ensemble model prefers the skeletons with more training information. Though it cannot guarantee the validity of the rare or novel skeletons, for which the mixed model might be better.

Furthermore, 11 kinds of physicochemical properties were calculated as in our previous work[39] and principal component analysis (PCA) was used to visualize the chemical space of the generated structures (Figure 3b). The result shows that the chemical space of generated sesterterpenoids are covered by the existing terpenoids and larger than the existing sesterterpenoids. Finally, our bio-inspired exploration protocol, TeroGen, has greatly expanded the chemical space of sesterterpenoids, generating 34439 sesterterpenoids. To obtain a clearer visualization of such huge molecules, the structure diversity was evaluated by Tree Maps (TMAPs)[54], which cluster molecules according to similarity of their fingerprints, as shown in Figure 3c. Obviously, the

generated sesterterpenoids varies greatly from the existing ones, in spite of a small overlap. This is mainly because that most of the existing sesterterpenoids, such as those with the linear skeletons and the scalarane-type tetracarbocyclic skeletons, do not belong to products of CPF-TSs. The generated sesterterpenoids not only contain varied carbon skeletons but also have been decorated with a series of R-groups. Considering that sesterterpenoids only occupy less than 1.6% of all well-known terpenoids in nature based on our TeroKit webserver (**http://terokit.qmclab.com/**), it means that only a rather limited number of sesterterpenoids exist in our training set in this work. Nevertheless, our TeroGen shows super efficiency on its prediction ability and extreme overlap of chemical space (namely coverage ability), thus a high portability to other types of terpenoids by TeroGen is expected.

**Discussion**

The exploration ability of our bio-inspired strategy TeroGen is investigated above. In addition, the network map could be further constructed to decipher the correlation of this chemical space with the reaction network. That is, the biosynthesis pathway can be traceable, especially for the Reactor, where the plausible carbocation rearrangement reactions will provide insights into the biosynthetic mechanism of terpenoid skeletons. Although the semi-empirical method GFN2-xTB used here is not as accurate as the DFT methods, it is a good approximation and an initial point for further refinement by DFT methods. Take the application on the humulyl cation ($C_{15}H_{25}^{+}$), sampled by the Reactor. Twenty-six carbocations directly connected to the humulyl cation were obtained in the first round, and a total of 1979 carbocations were generated after two rounds of simulation, which cost about 7 days running with two 12-core CPUs (Xeon E5-2609 1.70GHz). In a previous work[33], the reaction space of humulyl cation was also investigated by DFT methods with the AFIR strategy. A total of eight accessible carbocations (without regard to stereochemistry) were located by the DFT methods, which are predicted to be formed through pathways with no individual steps having barriers greater than 20 kcal/mol. Here we sampled all of these eight carbocations (highlighted with blue circles, although some of them have different stereochemistry), seven of which were also predicted to be accessible (Figure 4) and except one (**L**) formed via an alkenyl carbocation intermediate (**K**). For the competitive pathways between **B** to **C** and **B** to **D**, the DFT calculation with AFIR[33] revealed that the barrier of the former is lower than the latter and both of them are significantly exothermic, which is consistent with our results. In additional, all the unlikely carbocations that are predicted to be formed through the pathways with at least one step having barriers greater than 20 kcal/mol in the DFT calculations[33] were not sampled by our Reactor except **Z**, which is also unlikely to form since the pathway contains an endothermic

step (**X** to **Y**) with barrier being 39.3 kcal/mol. All these results demonstrated that the Reactor can sample more feasible reaction pathways efficiently and the landscape of the carbocation potential energy surfaces are mostly in agreement qualitative agreement with those calculated by DFT. Considering such a low computational cost, Reactor can be used for high-throughput chemical reaction space exploration of terpenoids, while other available methods cannot be.

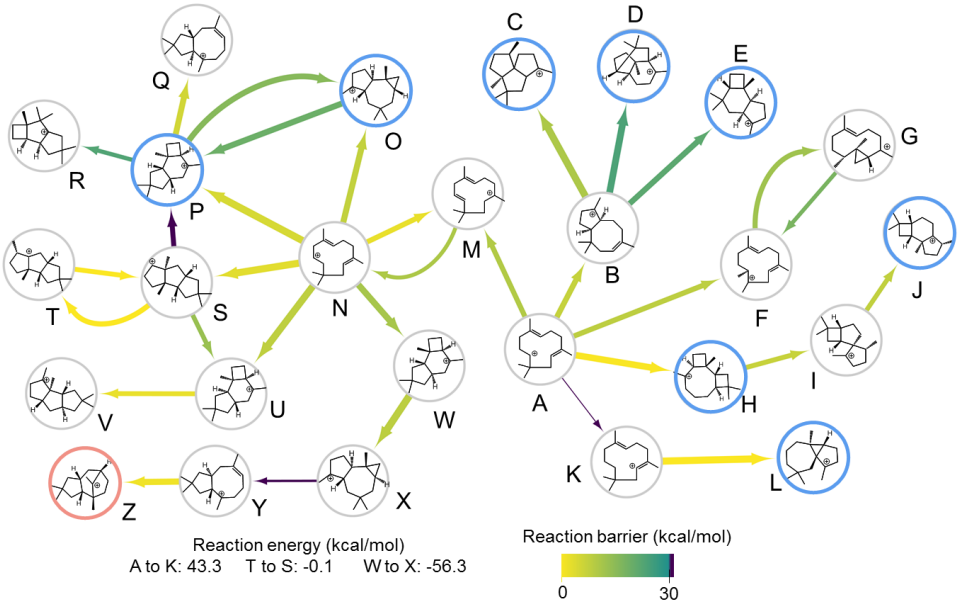


Figure 4. The selected reaction pathways in the complicated reaction network around the humulyl cation predicted by TeroGen (Reactor). The reaction barrier and energy (heat release) were represented by the color and width of edge, respectively. The reaction energies with highest (**A** to **K**), lowest (**W** to **X**) as well as the closest to 0 (**T** to **S**) were shown for reference.

Another aspect should be noted is that dividing the terpenoid biosynthesis into two relatively independent stage (cyclization and decoration) is idealistic although it can be correct for a considerable number of terpenoids. One exception is that deprotonation is not the only way of carbocation quenching. The other way was trapping by nucleophiles (usually the water) also exists, which leads to the formation of terpene alcohols[46]. Yet it is probably not a major problem since the hydroxyl can be added by Decorator. The other and more complicated case is that oxidation is sometime accompanied by further rearrangement[55], which could lead to the change of the initial carbon skeletons. In these circumstances, the skeletons cannot be sampled by the Reactor, and equally, these terpenoids cannot be generated by Decorator. Perhaps further metadynamics simulation of the oxidization starting from terpenes but not carbocations would be complementary to solve this issue, while obviously it will be much more complicated and

computationally expensive. And with the development of machine learning (ML) potential[56,57], ML-based simulations featuring both high speed and high accuracy could be an alternative to exploring the reaction network.

In summary, we propose a bio-inspired terpenoids exploration strategy using metadynamics simulations and deep learning according to the characteristics of two key stages in terpenoid biogenesis. The carbocation reactions are sampled by the Reactor to explore the covered space, then a compound slicing algorithm was developed to decompose natural terpenoids into core skeletons and decoration groups, which help to construct the Decorator models. By combining the physical-model-based Reactor and data-learning-based Decorator, TeroGen protocol will generate diverse terpenoid structures in the manner of biosynthesis. In practices, TeroGen not only provides an efficient, intuitively accessible strategy to map out and clearly visualize the cryptic chemical space difference between the known and generated terpenoids, but also could map out the generated chemicals ensembled in a reaction pathway network which obeys the general biogenesis rules for terpenoids. Therefore, it is very promising for user's personalized aim to navigate and analyze the detailed reaction space of terpenoids of practical use, such as to interpret the biosynthetic mechanism of existing terpenoids, to provide primary inspiration for the synthesis strategies of terpenoids, to supplement the clues for discovering plausible biosynthetic pathways to produce high-value terpenoids for heterologous biosynthesis, to define the known subspace of terpenoids natural products and to explore where the boundaries of terpenoid space are.

## Methods

**Reactor.** As shown in Figure 1b, the Reactor consists of parallel metadynamics simulations carried out using the *xtb*[34,58],in which the RMSD-PP plugin (with an optimized parameter set) was used to calculated the reaction paths, transition states as well as the reaction barriers and energies. An in-house script was used to extracted the reactions along the trajectory by detecting the location of positive charge and change of bond order. After the exploration, all carbocations were deprotonated exhaustivity to form terpenes or generalized to generate the carbon skeleton.

**Decorator**. The Decorator consisted of two models that predict the decorating sites and R-groups, both of which were constructed by Pytorch[59] with molecules represented by SMILES. The Transformer architecture used in site prediction was provided by OpenNMT[60,61]. The structures from the terpenoids database, TeroMOL, was preprocessed to train and test the Decorator model.

## Author contributions

R.W. designed and supervised the whole research. T.Z, R.W. and B.A.H. wrote the manuscript. T.Z. and F.Z. contributed concept and implementation. All authors contributed to the interpretation of results. All authors reviewed and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Reference


1. Bohacek, R.S., McMartin, C., and Guida, W.C. (1996). The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev *16*, 3-50. 10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6.
2. Kirkpatrick, P., and Ellis, C. (2004). Chemical space. Nature *432*, 823-823. 10.1038/432823a.
3. Walters, W.P. (2019). Virtual Chemical Libraries. J. Med. Chem. *62*, 1116-1124. 10.1021/acs.jmedchem.8b01048.
4. Sorokina, M., Merseburger, P., Rajan, K., Yirik, M.A., and Steinbeck, C. (2021). COCONUT online: Collection of Open Natural Products database. J. Cheminf. *13*, 2. 10.1186/s13321-020-00478-9.
5. Pye, C.R., Bertin, M.J., Lokey, R.S., Gerwick, W.H., and Linington, R.G. (2017). Retrospective analysis of natural products provides insights for future discovery trends. Proc. Natl. Acad. Sci. U. S. A. *114*, 5601-5606. 10.1073/pnas.1614680114.
6. Rodrigues, T., Reker, D., Schneider, P., and Schneider, G. (2016). Counting on natural products for drug design. Nat. Chem. *8*, 531-541. 10.1038/nchem.2479.
7. Newman, D.J., and Cragg, G.M. (2020). Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. J. Nat. Prod. *83*, 770-803. 10.1021/acs.jnatprod.9b01285.
8. Ertl P., S.A. (2008). Cheminformatics analysis of natural products: Lessons from nature inspiring the design of new drugs. Prog. Drug Res. *66*, 217-235. 10.1007/978-3-7643-8595-8_4.
9. Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J., Kapono, C.A., Luzzatto-Knaan, T., et al. (2016). Sharing and community curation of mass spectrometry data with Global Natural

Products Social Molecular Networking. Nat. Biotechnol. *34*, 828-837. 10.1038/nbt.3597.

10. Tabudravu, J.N., Pellissier, L., Smith, A.J., Subko, K., Autréau, C., Feussner, K., Hardy, D., Butler, D., Kidd, R., Milton, E.J., et al. (2019). LC-HRMS-Database Screening Metrics for Rapid Prioritization of Samples to Accelerate the Discovery of Structurally New Natural Products. J. Nat. Prod. *82*, 211-220. 10.1021/acs.jnatprod.8b00575.

11. Metelev, M., Osterman, I.A., Ghilarov, D., Khabibullina, N.F., Yakimov, A., Shabalin, K., Utkina, I., Travin, D.Y., Komarova, E.S., Serebryakova, M., et al. (2017). Klebsazolicin inhibits 70S ribosome by obstructing the peptide exit tunnel. Nat. Chem. Biol. *13*, 1129-1136. 10.1038/nchembio.2462.

12. Wang, S., Dong, G., and Sheng, C. (2019). Structural Simplification of Natural Products. Chem. Rev. *119*, 4180-4220. 10.1021/acs.chemrev.8b00504.

13. Liu, J., Flegel, J., Otte, F., Pahl, A., Sievers, S., Strohmann, C., and Waldmann, H. (2021). Combination of Pseudo-Natural Product Design and Formal Natural Product Ring Distortion Yields Stereochemically and Biologically Diverse Pseudo-Sesquiterpenoid Alkaloids. Angew. Chem. Int. Ed. Engl. *60*, 21384-21395. 10.1002/anie.202106654.

14. Renner, S., van Otterlo, W.A.L., Dominguez Seoane, M., Möcklinghoff, S., Hofmann, B., Wetzel, S., Schuffenhauer, A., Ertl, P., Oprea, T.I., Steinhilber, D., et al. (2009). Bioactivity-guided mapping and navigation of chemical space. Nat. Chem. Biol. *5*, 585-592. 10.1038/nchembio.188.

15. Yu, M.J. (2011). Natural Product-Like Virtual Libraries: Recursive Atom-Based Enumeration. J. Chem. Inf. Model. *51*, 541-557. 10.1021/ci1002087.

16. Koch, M., Duigou, T., Carbonell, P., and Faulon, J.-L. (2017). Molecular structures enumeration and virtual screening in the chemical space with RetroPath2.0. J. Cheminf. *9*, 64. 10.1186/s13321-017-0252-9.

17. Xu, Y., Lin, K., Wang, S., Wang, L., Cai, C., Song, C., Lai, L., and Pei, J. (2019). Deep learning for molecular generation. Future Medicinal Chemistry *11*, 567-597. 10.4155/fmc-2018-0358.

18. Elton, D.C., Boukouvalas, Z., Fuge, M.D., and Chung, P.W. (2019). Deep learning for molecular design—a review of the state of the art. Molecular Systems Design & Engineering *4*, 828-849. 10.1039/C9ME00039A.

19. Segler, M.H.S., Kogej, T., Tyrchan, C., and Waller, M.P. (2018). Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Cent. Sci. *4*, 120-131. 10.1021/acscentsci.7b00512.

20. De Cao, N., and Kipf, T. (2018). MolGAN: An implicit generative model for small molecular graphs. arXiv preprint arXiv:1805.11973.

21. Lim, J., Hwang, S.-Y., Moon, S., Kim, S., and Kim, W.Y. (2020). Scaffold-based molecular design with a graph generative model. Chem. Sci. *11*, 1153-1164. 10.1039/c9sc04503a.

22. Arus-Pous, J., Patronov, A., Bjerrum, E.J., Tyrchan, C., Reymond, J.L., Chen, H.M., and Engkvist, O. (2020). SMILES-based deep generative scaffold decorator for de-novo drug design. J. Cheminf. *12*. ARTN 38 10.1186/s13321-020-00441-8.

23. Zheng, S., Yan, X., Gu, Q., Yang, Y., Du, Y., Lu, Y., and Xu, J. (2019). QBMG: quasi-biogenic molecule generator with deep recurrent neural network. J. Cheminf. *11*, 5. 10.1186/s13321-019-0328-9.

24. Gao, W., and Coley, C.W. (2020). The Synthesizability of Molecules Proposed by Generative Models. J. Chem. Inf. Model. 10.1021/acs.jcim.0c00174.

25. Horwood, J., and Noutahi, E. (2020). Molecular Design in Synthetically Accessible Chemical Space via Deep Reinforcement Learning. ACS Omega *5*, 32984-32994. 10.1021/acsomega.0c04153.

26. Fialkova, V., Zhao, J., Papadopoulos, K., Engkvist, O., Bjerrum, E.J., Kogej, T., and Patronov, A. (2021). LibINVENT: Reaction-based Generative Scaffold Decoration for in Silico Library Design. J. Chem. Inf. Model. 10.1021/acs.jcim.1c00469.

27. Segler, M.H.S., and Waller, M.P. (2017). Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. Chem. - Eur. J *23*, 5966-5971. https://doi.org/10.1002/chem.201605499.

28. Unsleber, J.P., and Reiher, M. (2020). The Exploration of Chemical Reaction Networks. Annu. Rev. Phys. Chem. *71*, 121-142. 10.1146/annurev-physchem-071119-040123.

29. Dontgen, M., Przybylski-Freund, M.D., Kroger, L.C., Kopp, W.A., Ismail, A.E., and Leonhard, K. (2015). Automated discovery of reaction pathways, rate constants, and transition states using reactive molecular dynamics simulations. J Chem Theory Comput *11*, 2517-2524. 10.1021/acs.jctc.5b00201.

30. Wang, L.P., Titov, A., McGibbon, R., Liu, F., Pande, V.S., and Martinez, T.J. (2014). Discovering chemistry with an ab initio nanoreactor. Nat. Chem. *6*, 1044-1048. 10.1038/nchem.2099.

31. Dewyer, A.L., Argüelles, A.J., and Zimmerman, P.M. (2018). Methods for exploring reaction space in molecular systems. Wiley Interdisciplinary Reviews: Computational Molecular Science *8*. 10.1002/wcms.1354.

32. Tian, B., Poulter, C.D., and Jacobson, M.P. (2016). Defining the Product Chemical Space of Monoterpenoid Synthases. PLoS Comput. Biol. *12*, e1005053. 10.1371/journal.pcbi.1005053.

33. Isegawa, M., Maeda, S., Tantillo, D.J., and Morokuma, K. (2014). Predicting pathways for terpene formation from first principles - routes to known and new sesquiterpenes. Chem. Sci. *5*, 1555-1560. 10.1039/c3sc53293c.

34. Grimme, S. (2019). Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. Journal of Chemical Theory and Computation

*15*, 2847-2862. 10.1021/acs.jctc.9b00143.

35. Bannwarth, C., Ehlert, S., and Grimme, S. (2019). GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. Journal of Chemical Theory and Computation *15*, 1652-1671. 10.1021/acs.jctc.8b01176.

36. Rasmussen, M.H., and Jensen, J.H. (2021). Fast and automated identification of reactions with low barriers using meta-MD simulations.

37. Rasmussen, M.H., and Jensen, J.H. (2020). Fast and automatic estimation of transition state structures using tight binding quantum chemical calculations. PeerJ Physical Chemistry *2*. 10.7717/peerj-pchem.15.

38. Kirschning, A., and Hahn, F. (2012). Merging chemical synthesis and biosynthesis: a new chapter in the total synthesis of natural products and natural product libraries. Angew. Chem. Int. Ed. Engl. *51*, 4012-4022. 10.1002/anie.201107386.

39. Zeng, T., Liu, Z., Liu, H., He, W., Tang, X., Xie, L., and Wu, R. (2019). Exploring Chemical and Biological Space of Terpenoids. J. Chem. Inf. Model. *59*, 3667-3678. 10.1021/acs.jcim.9b00443.

40. Maimone, T.J., and Baran, P.S. (2007). Modern synthetic efforts toward biologically active terpenes. Nat. Chem. Biol. *3*, 396-407. 10.1038/nchembio.2007.1.

41. Hung, K., Hu, X., and Maimone, T.J. (2018). Total synthesis of complex terpenoids employing radical cascade processes. Nat. Prod. Rep. *35*, 174-202. 10.1039/C7NP00065K.

42. Bian, G., Deng, Z., and Liu, T. (2017). Strategies for terpenoid overproduction and new terpenoid discovery. Curr. Opin. Biotechnol. *48*, 234-241. 10.1016/j.copbio.2017.07.002.

43. Paddon, C.J., and Keasling, J.D. (2014). Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. Nature Reviews Microbiology *12*, 355-367. 10.1038/nrmicro3240.

44. Chen, K., and Baran, P.S. (2009). Total synthesis of eudesmane terpenes by site-selective C–H oxidations. Nature *459*, 824-828. 10.1038/nature08043.

45. Bathe, U., and Tissier, A. (2019). Cytochrome P450 enzymes: A driving force of plant diterpene diversity. Phytochemistry *161*, 149-162. 10.1016/j.phytochem.2018.12.003.

46. Christianson, D.W. (2017). Structural and Chemical Biology of Terpenoid Cyclases. Chem. Rev. *117*, 11570-11648. 10.1021/acs.chemrev.7b00287.

47. Banerjee, A., and Hamberger, B. (2018). P450s controlling metabolic bifurcations in plant terpene specialized metabolism. Phytochem Rev *17*, 81-111. 10.1007/s11101-017-9530-4.

48. Hong, Y.J., and Tantillo, D.J. (2014). Branching out from the bisabolyl cation. Unifying mechanistic pathways to barbatene, bazzanene,

chamigrene, chamipinene, cumacrene, cuprenene, dunniene, isobazzanene, iso-gamma-bisabolene, isochamigrene, laurene, microbiotene, sesquithujene, sesquisabinene, thujopsene, trichodiene, and widdradiene sesquiterpenes. J. Am. Chem. Soc. *136*, 2450-2463. 10.1021/ja4106489.

49. Hong, Y.J., and Tantillo, D.J. (2009). Consequences of Conformational Preorganization in Sesquiterpene Biosynthesis: Theoretical Studies on the Formation of the Bisabolene, Curcumene, Acoradiene,Zizaene, Cedrene, Duprezianene, and Sesquithuriferol Sesquiterpenes. JACS *131*, 7999-8015.

50. Minami, A., Ozaki, T., Liu, C., and Oikawa, H. (2018). Cyclopentane-forming di/sesterterpene synthases: widely distributed enzymes in bacteria, fungi, and plants. Nat. Prod. Rep. *35*, 1330-1346. 10.1039/c8np00026c.

51. Ye, Y., Minami, A., Mandi, A., Liu, C., Taniguchi, T., Kuzuyama, T., Monde, K., Gomi, K., and Oikawa, H. (2015). Genome Mining for Sesterterpenes Using Bifunctional Terpene Synthases Reveals a Unified Intermediate of Di/Sesterterpenes. JACS *137*, 11846-11853. 10.1021/jacs.5b08319.

52. Tantillo, D.J. (2010). The carbocation continuum in terpene biosynthesis—where are the secondary cations? Chem. Soc. Rev. *39*, 2847-2854. 10.1039/B917107J.

53. Hess, B.A. (2002). Concomitant C-Ring Expansion and D-Ring Formation in Lanosterol Biosynthesis from Squalene without Violation of Markovnikov's Rule. JACS *124*, 10286-10287. 10.1021/ja026850r.

54. Probst, D., and Reymond, J.-L. (2018). A probabilistic molecular fingerprint for big data settings. J. Cheminf. *10*, 66. 10.1186/s13321-018-0321-8.

55. McCulley, C.H., and Tantillo, D.J. (2020). Predicting Rearrangement-Competent Terpenoid Oxidation Levels. JACS *142*, 6060-6065. 10.1021/jacs.9b12398.

56. Behler, J. (2016). Perspective: Machine learning potentials for atomistic simulations. J. Chem. Phys. *145*, 170901. 10.1063/1.4966192.

57. Ang, S.J., Wang, W., Schwalbe-Koda, D., Axelrod, S., and Gómez-Bombarelli, R. (2021). Active learning accelerates ab initio molecular dynamics on reactive energy surfaces. Chem *7*, 738-751. 10.1016/j.chempr.2020.12.009.

58. xtb, version 6.3.2. (2020). https://github.com/grimme-lab/xtb.

59. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., and Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems *32*, 8026-8037.

60. Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. Proceedings of ACL, 67-72.

61. OpenNMT-py. (2020). https://github.com/OpenNMT/OpenNMT-py.