Generation of novel Diels-Alder reactions using a generative adversarial network

Sheng Li, ‡ Xinqiao Wang, ‡‡ Yejian Wu, ‡ Hongliang Duan ‡ and Lan Tang †‡

Deep learning has enormous potential in the chemical and pharmaceutical fields. Among these, Generative Adversarial Network, as an excellent generative model, has shown its remarkable performance in the field of molecular generation, but it has few applications in organic chemistry. Therefore, we attempt to apply GAN as a generative model for the task of reaction generation to expand the application of GAN in chemistry. In this work, we used the MaskGAN model trained with 14092 Diels-Alder reactions, and we finally generated 1441 novel Diels-Alder reactions that learn reaction rules in-depth, which demonstrates that reaction generation can be used in the field of chemistry, and helps chemists explore novel reactions.

Introduction

Organic chemistry has played a significant role in human history for hundreds of years since the synthesis of chemicals such as drugs or materials was inseparable from the development of organic reactions. However, as can be seen, the exploration of traditional chemical reactions remains an ongoing challenge due to the prolonged experimental time, exorbitant experimental cost and low rate of success. Fortunately, the advent of artificial intelligence (AI) offered new solutions for organic chemistry.¹

With the constant development of computer techniques, AI has achieved remarkable results in the fields of retrosynthetic prediction and reaction prediction in the past few years.²,³ As a specific example, Zheng et al. applied the Transformer model to develop a template-free self-corrected retrosynthesis predictor (SCRIP) to predict retrosynthesis reactions and displayed accuracy of 59.0% on a standard benchmark data set.²⁴ Wang et al proposed a method that uses transfer learning to improve the accuracy of the transformer model (94.9%), which is higher than the accuracy of the transformer-baseline model (66.3%).⁵ Furthermore, the application of deep neural networks in the field of pharmaceutical chemistry, such as drug molecular generation and toxicity risk assessment, has also received extensive attention recently.⁶,⁷ Lee et al. applied the generative adversarial network (GAN) to de novo molecular design and demonstrated outstanding performance in the five distribution learning benchmarks in the GuacaMol framework.⁸ The remarkable achievements of generative models in molecular generation inspired the chemists, both Bort et al. and Wang et al. obtained the reactions of their interest with reaction generative models.⁹,¹⁰

Generative models are an important class of models in machine learning in which we can generate new data that is not included in the training dataset and have shown its enormous potential in the field of image,¹¹ text¹² and sound generation¹³ in the past few years. Among these, GAN is an extremely popular deep learning model, it was originally proposed by Goodfellow et al. as a new framework for estimating generative models in adversarial processes.¹⁴ GAN uses two adversarial networks, a generator that captures the distribution of the data, and a discriminator that estimates the probability that samples come from the training data. They compete until the discriminator cannot distinguish between the real data and the generated data from the generator. Compared with other models, GAN has shown its superiority in generating more realistic images by this means.¹⁵ However, the application of GAN in the chemical field was limited due to the discreteness of SMILES strings that replace the molecular structures as inputs. Scientists then applied a policy gradient-based reinforcement learning approach proposed by Sutton et al. on GAN to provide feedback on information to unfreeze its restrictions.¹⁶ Lin et al. used generative adversarial networks to fulfill molecular de novo design, dimensionality reduction, and de novo peptide and protein design.¹⁷ Maziarka et al. proposed mol-CycleGAN which was a cycler-based molecular optimization model. This method could generate optimized compounds with the desired properties which were similar in structure to the originally given molecules, and further achieved more remarkable results than the previous model.¹⁸ Prykhodko et al. proposed a new deep learning...
Fig. 1 Flowchart for generating the Diels-Alder reactions with GAN. The real training data and the data that generated during the training process of generator, are simultaneously import into the discriminator for training, then the results are fed back to the generator for further training.

Choosing an applicable dataset for model to learn is a factor that affects the model’s performance. We chose the Diels-Alder reaction dataset as our training set in this paper, since the Diels-Alder reaction has become one of the most powerful organic reactions that are widely used synthetic tools in both drug and material synthesis. The Diels-Alder cycloaddition reaction is known as the cyclization of a diene and alkene to form a cyclohexene derivative, which was discovered by O. Diels and K. Alder while they established the structure of the cycloadduct of p-quinone and cyclopentadiene in 1928. And since its discovery, the Diels-Alder reactions as the classical reactions contains sufficient amount of data, which is another reason that we chose Diels-Alder reaction dataset. The key step in the synthesis of rubrolone aglycon’s seven-membered C-ring by Boger et al. in the laboratory was the intermolecular Diels-Alder reaction of an electron-rich diene with the very strained dienophile. Meanwhile, the excellent yield of cycloaddition is 97% and the products have complete enantioselectivity. During Diels-Alder reaction progress, the reactants close together and interact with each other to form a cyclic transition state, which is then gradually transformed into a product molecule.

In this study, we explored generating novel Diels-Alder reactions with MaskGAN which is composed of the generator network, discriminator network and critic network. Among this, the generator network uses sequence-to-sequence model with attention mechanism. We constructed a training dataset with Diels-Alder reactions from the Reaxys database, converted the Diels-Alder reaction in the dataset into SMILES strings and then imported it into the MaskGAN model for training and reaction generation. We compared the generated reactions to the Diels-Alder reaction datasets to confirm that the generated reactions were novel. We ended up with 1441 brand new Diels-Alder reactions. We try to combine the discovery of new reactions with artificial intelligence and look forward to contributing to speeding up the reaction discovery process and it could be useful in predicting outcomes of organic reactions.

Method

Dataset
To train MaskGAN for generating Diels-Alder reactions, we create a dataset of the Diels-Alder reactions. After downloading the reactions on the "Reaxys" database with the search keyword "Diels-Alder Reaction", we deleted duplicate reactions and invalid reactions (reactions with empty reactants or products, and reactions with reactants equal to products), and used the "RDKit" templates to screen the correct Diels-Alder reaction and finally, a dataset of 14,092 reactions was assembled. We split the dataset into a training set and validation set with a ratio of 8:2.

Model
Initially, GAN was designed to output differentiable values, so it is difficult for GAN to generate discrete language. In response to this challenge, MaskGAN, an actor-critic conditional GAN was introduced by filling in missing text conditioned on the context to generate higher quality samples. MaskGAN is composed of the generator network, discriminator network and critic network. Among this, the generator uses a sequence-to-sequence architecture with an architecture for de novo molecular design, LatentGAN, which combined autoencoders and generative adversarial networks. Therefore, we try to use GAN to achieve novel reaction generation to advance the progress of chemical reactions. Fig.1 shows Flowchart for generating the Diels-Alder reactions with GAN.
attention mechanism. The actual implementation of seq2seq in MaskGAN take the form of LSTM, and it contains an encoder structure and a decoder structure. The encoder processes each element in the input sequence, compiling the captured information into context vector. After processing the entire input sequence, the encoder sends the context vector to the decoder, which starts producing the output sequence item-by-item, eventually producing the entire sentence. The structure used by the discriminator is the same as the generator, both in the form of seq2seq, except that a scalar probability is output at each time step. In addition, for converging more rapidly, the critic in MaskGAN helps the generator reduce the high variance of the gradient updates in a high action-space environment. It enables a much more stable training procedure.

Result and discussion
To explore novel reactions with our model, we imported 14092 Diels-Alder reactions obtained from the database into the model in the form of SMILES strings. Among these, 80% of the Diels-Alder reactions were applied to train the model while remainders for validation. However, the reaction may still be chemically invalid, although the SMILES strings for both reactants and products were valid. Therefore, we removed chemically meaningless reactions through the screening with reaction templates in the RDKit module and the judgments of chemists with professional knowledge. We finally obtained 1441 novel Diels-Alder reactions after removing the reactions which repeated in the training set from generated set (in Fig.2). And Fig. 3 shows the reasonable reactions representatively selected from the generated set.

<table>
<thead>
<tr>
<th>Components in the reactions</th>
<th>Valid molecular</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>dienes</td>
<td>10000</td>
</tr>
<tr>
<td>dienophiles</td>
<td>10000</td>
</tr>
<tr>
<td>products</td>
<td>10000</td>
</tr>
</tbody>
</table>

Table 1 The validity of components in the generated set

To further explore the generated reactions, firstly, we analyzed the novel reactions at the molecular level. Table 1 shows the amounts and ratios of valid molecules out of 10000 generated molecules for each component in the reactions, and Table 2 shows the amounts and ratios of reactants and products calculated with the different metrics in the generated set.

Validity. The validity is calculated as the ratio of the valid molecules to generated molecules. The ratios of dienes, dienophiles and products components are 70.1%, 74.8% and 30.5%, respectively. Our model exhibits outstanding performance for the validity of dienes and dienophiles, however, our model gives a mediocre performance in products, and there is greater space for optimization.

Uniqueness. Here, we computed the uniqueness as the ratio of unique molecules to valid molecules in generated set. As shown in Table 2, the ratios of products, dienes and dienophiles components are 97.0%, 42.8% and 62.0%, respectively. We speculate that the reason for the higher ratio of products to reactants is the complexity of the structure of the products. Uniqueness checks that the model does not produce only a few typical molecules. and our model gives an impressive performance in this metric. In particular, it exhibited a high ratio of generating unique products.

Novelty. We counted the amount of novel molecules of

<table>
<thead>
<tr>
<th>Components in the reactions</th>
<th>Unique molecular</th>
<th>Novel molecular</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Amount</td>
<td>Rate</td>
</tr>
<tr>
<td>dienes</td>
<td>661</td>
<td>42.8%</td>
</tr>
<tr>
<td>dienophiles</td>
<td>825</td>
<td>62.0%</td>
</tr>
<tr>
<td>products</td>
<td>1394</td>
<td>97.0%</td>
</tr>
</tbody>
</table>

Table 2 The uniqueness and novelty of components in the generated set

J. Name., 2013, 00, 1-3
each component in the reactions. Novelty shows the ratio of generated molecules not in training set and unique molecules in generated set. The amount of novel dienophiles is 628 while the ratio is 76.1%, which indicates that the innovation of the novel reactions generated by our model mainly focused on the novel dienophiles, rather than simply replacing different reactants to produce the same products as the products in training set. Meanwhile, the ratios of dienes and products components are 68.3% and 74.2%. Although the proportions of our novel dienes are mediocre, since reactants can combine in pairs, they provide sufficient possibilities for innovation to create brand new products without the need for a large number of novel reactants.

Interestingly, as shown in Fig. 4, we observed the situation that the dienophile components chose to react with the two double bonds with the s-cis conformation when the simultaneous appearance of three carbon-carbon double bonds in reactant, and 98.4% of reactions conform to this rule in generated set. Our model achieves outstanding performance on this metric, and this proves that our model has learned this rule. It is due to the s-cis conformation being more favorable in the formation of the transition state during the Diels-Alder reaction. Therefore, dienes that are permanently in the s-trans conformation and cannot adopt the s-cis conformation will not do the Diels-Alder reaction at all. These dienes cannot react with dienophile in Diels-Alder reactions since the two ends of the dienes cannot get close enough and, at the least, the new six-membered ring of products would form an impossible trans double bond. On the contrary, dienes that are permanently in the s-cis conformation, such as cyclic dienes, exceptionally favor Diels-Alder reactions.

To find the relationship between the generated reactions and the Diels-Alder reaction of the training set, we preliminary analyzed the generated reactions. We applied MACCS molecular fingerprint combined with t-distributed Stochastic Neighbor Embedding (t-SNE) method to represent the distribution of reactants. MACCS fingerprint as a molecular qualitative descriptor contained various functional groups of the molecule, which is the high-dimensional data consisting of 166-dimensional molecular features and 1-dimensional placeholders. To visualize the MACCS molecular fingerprints of the reactants, we applied t-SNE as a dimensionality reduction technique. t-SNE is a variation on the Stochastic Neighbor Embedding (SNE) proposed by Matten et al., which is easier to optimize and equipped with significantly better by reducing the tendency of points to cluster in the center of the map. Fig. 5(A) shows the t-SNE plot of the MACCS fingerprint distribution of the novel dienophile components in generated set and the dienophile in training dataset. We observed that the distribution of training set covered the generated set well, which indicates that the dienophile components generated by model satisfy the features of reactants in Diels-Alder reactions though the reactants are novel. Meanwhile, we get a similar conclusion in Fig. 5(B) of the diene components, which is powerful evidence to prove that the generated reactions follow the feature distribution of the dataset.

Then, we further analyzed the generated reactions at the level of chemical transformation. Table 3 summarizes the results of the amounts and ratios for correct Diels-Alder reactions, unique reactions and novel reactions out of 10000 generated valid reactions and Table 4 shows the proportion of reactions that conform to the regioselectivity and stereospecificity of the reactants.
Diels-Alder reaction. We computed the correctness as the ratio of correct reactions that were screened by 'RDKit' templates to valid reactions in generated set. Uniqueness shows the ratio of the generated correct reactions from the model and reactions excluding duplicate reactions from the correct reactions, and the novelty is calculated as the ratio of molecules that are not included in the training dataset. The ratios of correctness, uniqueness and novelty are 50.4%, 40.6% and 21.4%, respectively. From these three metrics, we can see that our model presents a mediocre performance in novelty, we speculate that it is related to small data, and it is still a challenge in our task.

We then discuss the regioselectivity of the generated Diels-Alder reactions. As shown in Table 4, we observed a 100% probability that the model produces reactions with an ortho or para product, which is consistent with our knowledge of regioselectivity of Diels-Alder reactions, and our model achieves excellent performance on this metric. Regioselectivity, in a nutshell, implies that a reagent must choose where to react with one functional group that can react in two different places. The Diels-Alder reactions are highly regioselective since one of the carbon-carbon double bonds in the diene is much more likely to be attacked by the dienophile components as an electrophilic site due to the attachment of an electron-donating group. As shown in Fig. 6(A) and Fig. 6(B), the electron-donating group is located at the end of the diene, which contributes to the other end of diene being more electrophilic, and then the dienophile components attack the electrophilic site. In other words, the dienophile tends to attack the end of the diene to produce ortho products when the electron-donating group is located at another end of the diene. The situation has changed while the electron-donating group is in the middle of the diene in Fig. 6(C) and Fig. 6(D) The dienophile components attack the carbon-carbon double bond that is directly connected with the electron-donating group and produce the para product. Therefore, the Diels-Alder reaction is a cycloaddition with an aromatic transition state that is ortho and para directing, and the reactions generated by our model conform to this rule.

Stereoselectivity is how the group reacts with regard to the stereochemistry of the product. For the Diels-Alder reactions, the process of product formation mostly follows the endo rule, in other words, the electron-withdrawing group of the dienophile components and the newly formed carbon-carbon double bond in the middle of the old diene tend to be on the same side during the process, which called endo product. This
is due to the bonding interaction between the electron-
withdrawing group of the dienophile and the π bond formed at
the back of the diene contributing to an increased rate of endo
product formation. Therefore, endo products are preferred as
kinetic products in irreversible Diels–Alder reactions. In
reversible Diels–Alder reaction, however, the exo product is
formed instead since the exo product with less steric hindrance is more stable than the endo product. But we here
selected the irreversible Diels–Alder reactions for training set,
only, endo products are formed in most reactions. In terms of
the stereoselectivity of our generated novel reaction is shown in
Fig. 7(A) The asymmetric dienophile reacts with the cyclic
diene, we can observe that the carbonyl groups on the
dienophile and the newly formed double bond in the middle of
the old diene are on the same side, and the hydrogen atoms
are above the generated ring. Therefore, the product is an
endo product, which conforms to the stereoselectivity rule of
the Diels–Alder reaction, and proves that our model learns the
stereoselectivity of the Diels–Alder reaction.

We finally discuss the stereospecificity of the generated
Diels–Alder reactions. We besides found that the structure of
the product mostly depends on the reactants’ structure. As
shown in Fig. 7(B)(a), the product keeps the cis configuration
when the dienophile component with the cis configuration
reacts with the diene in the Diels–Alder reaction. While the
trans configuration is similarly reproduced from the dienophile
reactant to product in Fig. 7(B)(b) For dienophiles with trans
configuration, one of the functional groups is tucked under the
diene in the transition state and then, that functional group
appears underneath the ring when the product molecule is
formed and reproduced the trans configuration. The
configuration of the diene components, except the dienophile,
also has a great influence on the configuration of products. Fig.
7(B)(c) shows that both carbon-carbon double bonds of the
diene are cis conformation, thus the two hydrogen atoms are
below the newly formed six-membered ring. And Fig. 7(B)(d)
shows that the functional groups of dienes lie outside the
newly formed six-membered ring when the configurations of
the two carbon-carbon double bonds are trans. Therefore, the
product is faithfully reproduced if there is stereochemistry in
the dienophile, in other words, the Diels–Alder reaction is
stereospecific, as shown in Table 4, 76.6% of the reactions that
products with cis/trans isomerism conform to this rule.

Conclusion

In this study, we trained the MaskGAN model with 14,092
Diels–Alder reactions dataset from database and finally
generated 1441 novel Diels–Alder reactions. To figure out the
relationship between the generated reactions and Diels–Alder
reactions, we applied the Diels–Alder reaction templates of the
‘RDKit’ package and then artificially judged these novel
reactions. And we further analyzed the generated reactions from
the level of reaction components and reaction
mechanism and found that the generated reactions mostly
satisfy the features of Diels–Alder reactions, and our model
achieved excellent performance on two aspects:

regioselectivity and conforming to the rule that dienes with
the s-cis conformation are exceptionally good at Diels–Alder
reactions, which indicates that MaskGAN has a clear sense of
the intrinsic rules of the Diels–Alder reactions. The purpose of
our experiment is to try to generate novel reactions that
conform to mechanism by GAN. The task of reaction
generation, similar to other generation tasks, aims to de novo
reactions design without additional import. In this task, the
model is no longer limited by imports, which can provide
chemists with more ideas for novel reactions and helps
chemists explore novel reactions.

Author Contributions

These authors contributed equally: S.L. and X.W., H.D.
conceived the presented idea. S. L. trained models. S. L. and Y.
W. analyzed the data. X.W. and S.L. wrote the manuscript. All
authors discussed the results and approved the manuscript.

Conflicts of interest

The authors declare no conflict of interest.

Data availability

The training and validation used in our study are available from
https://github.com/hongliangduan/Generation-of-novel-Diels-
Alder-reaction-using-a-GAN-dataset. Source data are provided with
this paper.

Code availability

The code and the trained model are available from
https://github.com/hongliangduan/Generation-of-novel-Diels-
Alder-reaction-using-a-GAN-.git

Acknowledgements

This project was supported by the National Natural Science Fo
undation of China (No. 81903438), Natural Science Foundatio
n of Zhejiang Province (LD22H300004) and the Public welfare
project of Zhejiang Science and Technology Department [Grant
No.GF20H280010].

Notes and references

2545-2559.
2. S. Ishida, K. Terayama, R. Kojima, K. Takasu and Y. Okuno, J.
Model., 2019, 60, 47-55.
2020, 56, 9368-9371.


J. Gu, Y. Shen and B. Zhou, in CVPR. 2020.


