Energy-based View of Retrosynthesis

Ruoxi Sun[†], Hanjun Dai[††], Li Li[†], Steven Kearnes[†], and Bo Dai[††]

[†]Google Research, Applied Science [††]Google Brain
{ruoxis, hadai, leeley, kearnes, bodai}@google.com

Abstract

Retrosynthesis—the process of identifying a set of reactants to synthesize a target molecule—is of vital importance to material design and drug discovery. Existing machine learning approaches based on language models and graph neural networks have achieved encouraging results. In this paper, we propose a framework that unifies sequence- and graph-based methods as energy-based models (EBMs) with different energy functions. This unified perspective provides critical insights about EBM variants through a comprehensive assessment of performance. Additionally, we present a novel "dual" variant within the framework that performs consistent training over Bayesian forward- and backward-prediction by constraining the agreement between the two directions. This model improves state-of-the-art performance for template-free approaches where the reaction type is unknown.

1 Introduction

Retrosynthesis is a critical problem in organic chemistry and drug discovery [1–4]. As the reverse process of chemical synthesis, the goal of retrosynthesis is to find the set of reactants that can synthesize the provided target via chemical reactions (Figure 1). The search space of theoretical feasible reactant candidates is enormous; hence, smart design of algorithms is required such that the model has the expression power to learn chemical rules while maintaining computational efficiency.

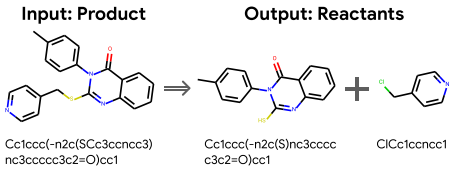

Figure 1: Retrosynthesis and SMILES.

Recent machine learning applications for retrosynthesis, including sequence- and graph-based models, have made significant progress. Sequence-based models treat molecules as one-dimensional token sequences (SMILES [5], bottom of Figure 1) and formulate retrosynthesis as a sequence-to-sequence problem, where recent advances in neural machine translation [6–9] can be applied. LSTM-based encoder–decoder frameworks and, more recently, transformer-based approaches have achieved promising results [9–12]. Graph-based models, on the other hand, have a natural representation of human-interpretable molecular graphs, where chemical rules are easily applied. Graph-based approaches that perform graph matching with chemical rules ("templates"; see the definition below) or reaction centers have reached encouraging results [13, 14].

Our goal here is to provide a unified view of both sequence- and graph-based retrosynthesis models using an energy-based model (EBM) framework. Within the framework, both types of models can be formulated as different EBM variants by instantiating the energy score functions into specific forms. A unified view is critical to provide insights into different EBM variants, as it's easy to extract commonalities and differences between EBM variants, understand strengths and limitations in designing models, compare the complexity of learning or inference, and inspire novel EBM variants. Note that here we are focused on one-step retrosynthesis, instead of multi-step planning; the design of the former case can be recursively applied to the latter. To summarize our contributions:

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

- We propose a unified EBM based framework that integrates sequence- and graph-based models for retrosynthesis.
- Based on this unified framework, we propose a novel Dual EBM variant that performs consistent training over forward and reverse prediction directions.
- We provide comprehensive empirical studies on multiple EBM variants and show that our proposed Dual model improves the state-of-the-art accuracy by 9.6% for template-free and 2.7% for template-based approaches.

2 Retrosynthesis Model

2.1 EBM framework for retrosynthesis

An energy-based model [15–17] defines the distribution using an energy function. Without loss of generality, we define the joint distribution of product and reactants as follows:

$$p_\theta(X, y) = \frac{\exp(-E_\theta(X, y))}{Z(\theta)} \tag{1}$$

where the partition function $Z(\theta) = \sum_y \sum_X \exp(-E_\theta(X, y))$ is a normalization constant to ensure a valid probability distribution. EBMs are well known for flexibility. By instantiating the energy score $E(\theta)$ with different designs and extra normalization conditions, EBMs can be used to unify many existing probabilistic models, including some directed graphical models like the autoregressive models. It is also easy to obtain arbitrary conditioning with different partition functions. For example, the predictive model for a chemical reaction can be obtained by $p_\theta(y|X) = \frac{\exp(-E_\theta(X,y))}{\sum_{y'} \exp(-E_\theta(X,y'))}$. However, learning EBMs with maximum likelihood estimation (MLE) is notoriously difficult in general, as the partition function $Z(\theta)$ is generally intractable. We will discuss trade-offs between capacity and learning tractability in detail (see below). Overall, the proposed framework works as follows: (1) design and train an energy score function E_θ, and (2) use E_θ for inference in retrosynthesis.

Inference with EBM for retrosynthesis: With the trained E_{θ^*}, inference identifies the best X that minimizes the energy function for given y^{test}, i.e. $X^{\text{test}} = \arg\min_{X \in \mathscr{X}} E_{\theta^*}(X, y^{\text{test}})$. Directly solving the above minimization is again intractable, but the energy function can generally be used for ranking. Let R denote the rank of candidate X_i for the given y^{test}

$$\{R(X_1) < R(X_2) \iff E_{\theta^*}(X_1, y^{\text{test}}) < E_{\theta^*}(X_2, y^{\text{test}})\} \tag{2}$$

One can use either template-based or template-free method to come up with initial proposals for ranking, as follows.

2.1R.1 Template-based Ranking (TB). Templates can be used to extract a list of proposed reactant candidates by using templates. We use $\mathscr{T}$ to define the set of available templates. Recall $T := t_y \to t_X$. Here we overload the notation to define a *template operator* $T(\cdot) : M \mapsto \mathscr{X}$ which takes a product as input, and returns a set of candidate reactant sets. Specifically, $T(\cdot)$ works as follows: enumerate all the templates with product-subgraph t_y matching with the given product y and define $S(y) = \{T : t_y \in y,\ \forall T \in \mathscr{T}\}$; then reconstruct the reactant candidates by instantiating reactant-subgraphs of the matched templates $R = \{X : t_x \in X,\ \forall T \in S(y)\}$.

2.1R.2 Template-free Ranking (TF). In this paper, *template-free ranking* makes proposals using the learned structure prediction model. We use a simple autoregressive form for $p(X|y)$, which can draw the top K most likely samples by beam search from this distribution.

2.2 Sequence based Models

In this section, we describe several sequence-based energy function designs. We first define the notation. Given a molecule x, we denote its SMILES representation as $s(x)$. We use superscript $s(x)^{(i)}$ to denote the character at i-th position of the SMILES string. For simplicity, we use $x^{(i)}$ when possible. The SMILES representation of a molecule set X, denoted as $s(X)$, is an ordered concatenation of $s(x)$ for every x in X with "." in between. For simplicity of notation, we use $X^{(i)}$ as the short form of $s(X)^{(i)}$ to denote the i-th position of this concatenated SMILES.

2.2.1 Full energy-based model

We start by proposing a most flexible model that imposes the minimum restrictions on design of E_θ. All the variants proposed in Sec 2 are special instantiations of this model.

The conditional probability of X given y is given as follows:

$$p(X|y) = \frac{\exp\left(-E_\theta(X, y)\right)}{\sum_{X' \in \mathscr{P}(M)} \exp\left(-E_\theta(X', y)\right)} \propto \exp(-E_\theta(X, y)) \tag{3}$$

Here the energy function $E_\theta : \mathscr{P}(M) \times M \mapsto \mathbb{R}$ takes a molecule and a molecule set as input, and outputs a scalar value. $\mathscr{P}(\cdot)$ represents the power set. Due to the intractability of the partition function, we focus on the following three general ways for training.

2.2.2 Ordered sequential model

As the full energy-based model in the previous section relies on templates for training and doesn't explicitly exploit the dependency between positions in a sequence, one can use an *ordered sequential model*, which performs forward auto-regressive factorization of the input sequence [8, 9, 18].

$$P(X|y) = p(X^{(1)}, X^{(2)}, \ldots, X^{(|s(X)|)}|y) = p_\theta(X^{(1)}|y) \prod_{i=2}^{|s(X)|} p_\theta(X^{(i)}|X^{(1:i-1)}, y) \tag{4}$$

where conditional probability $p(X^{(i)}|X^{(1:i-1)}, y)$ is parameterized by transformer $h_\theta(p, q) : S^{|p|} \times S^{|q|} \mapsto R^{|S|}$, and S is the vocabulary size for chemistry symbols like atoms, charges, *etc.*

$$P(X|y) = \exp\left(\sum_{i=1}^{|s(X)|} \log p_\theta(X^{(i)}|X^{(1:i-1)}, y)\right) = \exp\left(\sum_{i=1}^{|s(X)|} \log \frac{\exp\left(h_\theta(X^{(1:i-1)}, y)^\top e(X^{(i)})\right)}{\sum_{c \in S} \exp\left(h_\theta(X^{(1:i-1)}, y)^\top e(c)\right)}\right) \tag{5}$$

where $e(c)$ is a one-hot vector with dimension c set to 1. This choice of $h_\theta(p|q)$ enables efficient computing of the denominator of Eq. (5) by outputting a vector with length equal to $|S|$ to indicate logits (unnormalized log probability) for each value in the vocabulary. Directly using MLE is feasible for training this model.

2.2.3 Perturbed sequential model

In contrast to the ordered sequential model that factorizes the sequence in one direction, we adapt a method from XLNet [19], which uses a perturbed sequential model to achieve stochastic bidirectional factorization. In particular, the model permutes the factorization order (while maintaining position encoding of the original order) that is used in the forward auto-regressive model.

$$P(X|y, z) = p(X^{(z_1)}, X^{(z_2)}, \ldots, X^{(z_{|s(X)|})}|y) = \prod_{i=1}^{|s(X)|} p_\theta(X^{(z_i)}|X^{(z_1:z_{i-1})}, y) \tag{6}$$

where the permutation order z is a permutation of the original order sequence $zo = [1, 2, \ldots, |X|]$ and z_i denotes the i-th element of permutation z. Here z is treated as hidden variable.

During training, permutation order z is randomly sampled and uses the following training objective:

$$P(X|y) \approx \exp\left(\mathbb{E}_{z \sim Z_{|s(x)|}}\left[\sum_{i=1}^{|X|} \log p_\theta(X^{(z_i)}|z_i, X^{(z_1:z_{i-1})}, y)\right]\right) \tag{7}$$

and the corresponding parameterization:

$$p_\theta(X^{(z_i)}|z_i, X^{(z_1:z_{i-1})}y) = \log \frac{\exp\left(h(X^{(z_1:z_{i-1})}, z_i, y)^\top e(X_{z_i})\right)}{\sum_{c \in S} \exp\left(h(X^{(z_1:z_{i-1})}, z_i, y)^\top e(c)\right)} \tag{8}$$

where z_i encodes which position index in the permutation order to predict next, implemented by a second position attention (in addition to the primary context attention).

Eq. (7) is actually a lower bound of the latent variable model, due to Jensen's inequality. However, we focus on this model design for simplicity of permuting order in training to avoid difficult posterior inference. The benefit of such a model is that it has seen information from both directions during training via random permutation order. During testing, we use original order zo to compute $p(X^{\text{test}}|y^{\text{test}})$. With the lower-bound approximation, the direct MLE is feasible for model training.

2.2.4 Bidirectional model

An alternative way to achieve bidirectional context conditioning is the denoising auto-encoding model. We adapt *bidirectional model* from BERT [20] to our application. The conditional probability $p(X|y)$

is factorized into product of conditional distributions of one single random variable on the others.

$$p(X|y) \approx \exp\left(\sum_{i=1}^{|s(X)|} \log p_\theta(X^{(i)}|X^{\neg i}, y)\right) = \exp\left(\sum_{i}^{|s(X)|} \log \frac{\exp\left(h_\theta(X^{\neg i}, y)^\top \vec{e}(X^{(i)})\right)}{\sum_{c\in S}\exp(h_\theta(X^{\neg i}, y)^\top \vec{e}(c))}\right) \quad (9)$$

where h and e are the same as in Eq. (5). As presented in [21], although the model is similar to MRF [22], the marginal of each dimension in Eq. (9) does not have a simple form as in BERT training objective. This may result in mismatch between model and learning objective.

2.2.5 Dual model

Retrosynthesis and reaction prediction are a pair of mutual reversible processes that factorize the joint distribution in different orders, where reaction prediction is "forward direction" – $p(y|X)$) and retrosynthesis is the "backward direction" – $p(X|y)$. With additional prior modeling, the joint probability $p(X, y)$ factorizes to either $p(X|y)p(y)$ or $p(y|X)p(X)$. Based on this, we propose a training framework, which leverages the duality of the forward and backward directions, and performs consistent training between the two directions to bridge the divergence. In this case, the energy based model is defined as:

$$p(X|y) \propto \exp\big(\log p(X) + \log p(y|X) + \log p(X|y)\big) \quad (10)$$

The duality of reversible processes has also demonstrated its advantage in other applications [23–26]. We here provide a different learning method that is more practical for retrosynthesis task. Specifically, our consistent training is achieved by minimizing the dual loss, where the dual constrains in the equation below are imposed to penalize KL divergence of the two directions, *i.e.*, KL(backward|forward). For simplicity we fix the backward probability, and therefore entropy $H(\text{backward})$ is dropped.

$$\ell_{\text{dual}} = -\Bigg(\underbrace{\widehat{E}[\log p(X) + \log p(y|X)]}_{\text{forward direction}} + \underbrace{\beta\widehat{E_y}E_{X|y}[\log p(X) + \log p(y|X)]}_{\text{dual constraint}} + \underbrace{\widehat{E}[\log p(X|y)]}_{\text{backward direction}}\Bigg)$$

where $\widehat{E}[\cdot]$ indicates expectation of empirical data distribution $\hat{p}(X, y)$. As each term in above equation only involves with likelihood evaluation, we model $p(X|y)$, $p(X)$ and $p(y|X)$ as autoregressive models for simplicity. With such design, sampling from $p(X|y)$ is also tractable for dual constraint optimization, with the estimation of $\hat{E}_y[\cdot]$ using empirical data.

3 Experiments

We first present the evaluation of our best EBM variant against existing methods for both template-based and template-free approaches in Table 1, then we provide comprehensive study on different variants of sequence-based EBMs in Table 2. We provide Template free evaluation in Table 3.

Table 1: **Top K exact match accuracy** of existing methods

Category	Model	Reaction type unknown				Reaction type known			
		top1	top3	top5	top10	top1	top3	top5	top10
TB	retrosim [27]	37.3	54.7	63.3	74.1	52.9	73.8	81.2	88.1
	NeuralSym [28]	44.4	65.3	72.4	78.9	55.3	76.0	81.4	85.1
	GLN [13]	52.5	69.0	75.6	83.7	64.2	79.1	85.2	90.0
	Dual-TB (Ours)	**55.2**	**74.6**	**80.5**	**86.9**	**67.7**	**84.8**	**88.9**	**92.0**
Semi-TB	G2Gs [14]	48.9	67.6	72.5	75.5	61.0	81.3	86.0	88.7
TF	LSTM [10]	-	-	-	-	37.4	52.4	57.0	61.7
	Transformer [12]	43.7	60.0	65.2	68.7	59.0	74.8	78.1	81.1
	Dual-TF (Ours)	**53.3**	**69.7**	**73.0**	**75.0**	**65.7**	**81.9**	**84.7**	**85.9**

*Dual-TB/TF: Dual model with template-based or -free ranking.

Appendix

Table 2: **Top K accuracy of sequence variants**

		Reaction type unknown				Reaction type known			
Dataset	Models	Top 1	Top 3	Top 5	Top 10	Top 1	Top 3	Top 5	Top 10
USPTO 50k	Full model	39.5	63.5	73.0	83.8	55.0	79.9	86.3	**92.0**
	Ordered	47.0	67.4	75.4	83.1	60.9	80.9	85.8	90.2
	Perturbed	42.9	58.7	63.9	69.6	56.6	73.6	77.2	81.6
	Bidirectional	16.9	34.4	45.6	61.1	31.4	57.0	69.8	81.3
	Dual	**48.4**	**69.1**	**77.0**	**84.4**	**61.7**	**81.5**	**86.9**	91.1
Augmented USPTO 50k	Ordered	54.2	72.0	77.7	84.2	66.4	82.9	87.4	91.0
	Perturbed	47.3	64.6	70.4	75.8	64.2	79.8	83.3	86.4
	Bidirectional	23.5	43.7	54.3	69.5	41.9	66.3	75.6	84.6
	Dual	**55.2**	**74.6**	**80.5**	**86.9**	**67.7**	**84.8**	**88.9**	**92.0**

Table 3: **Template-free**: Translation Proposal and Dual Ranking

Type	Proposal						Re-rank				
	Proposal model	Top 1	Top 5	Top 10	Top 50	Top 100	Rank model	Top 1	Top 3	Top 5	Top 10
No	Ordered on UPSPTO	44.4	64.9	69.9	77.2	78.0	Dual trained on Aug USPTO	53.6	**70.7**	**74.6**	**77.0**
	Ordered on Aug USPTO	53.2	54.7	55.6	60.5	60.5		**54.5**	60.0	60.4	60.5
	-	-	-	-	-	-	SOTA (SCROP [12])	43.7	60.0	65.2	68.7
Yes	Ordered on USPTO	56.0	76.1	79.7	85.2	86.4	Dual trained on Aug USPTO	65.7	**81.9**	**84.7**	**85.9**
	Ordered on Aug USPTO	64.7	66.5	67.3	69.7	75.7		**66.2**	75.1	75.6	75.7
	-	-	-	-	-	-	SOTA (SCROP [12])	59.0	74.8	78.1	81.1

EBM framework

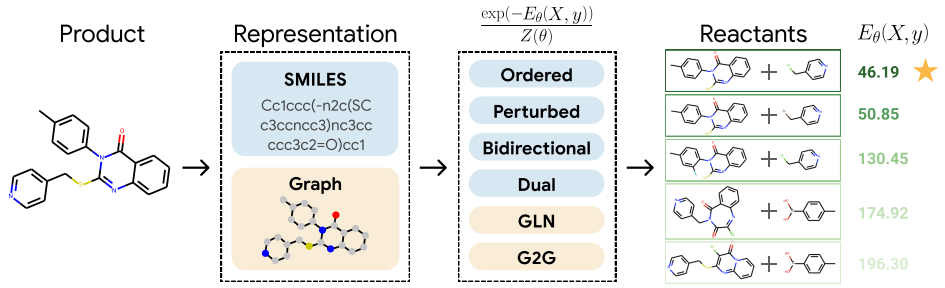

Figure 2: **EBM framework for retrosynthesis**. Given the product as input, works as follows: The product is the EBM framework (1) represents it as SMILES or a graph, (2) designs and trains the energy function E_θ, (3) ranks reactant candidates with the trained energy score $E_{\theta^\star}$, and (4) identifies the top K reactant candidates. The best candidate has the lowest energy score (denoted by a star). The list of reactant candidates is obtained via templates or directly from the trained model.

Dual model framework

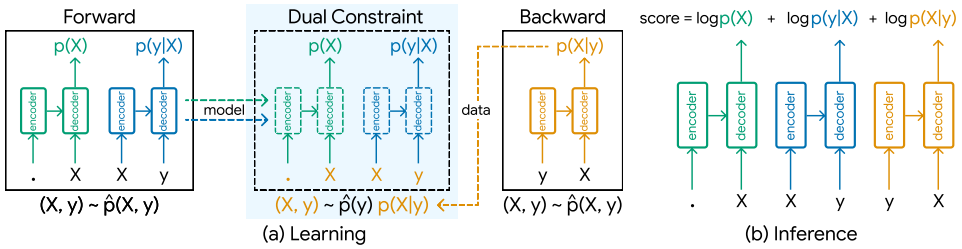

Figure 3: **Dual model. (a) Learning** consists of training three transformers: prior $p(X)$ (green), likelihood $p(y|X)$ (blue), and backward $p(X|y)$ (orange). Dual model penalizes the divergence between forward $p(X)p(y|X)$ and backward direction $p(y|X)$ with Dual constraint (highlighted). **(b) Inference** Given reactant candidates list, we rank them using Eq. (10).

Case study

Here we provide another case study showing with Dual model ranking (Sec 2.2.5), the accuracy improves upon translation proposal.

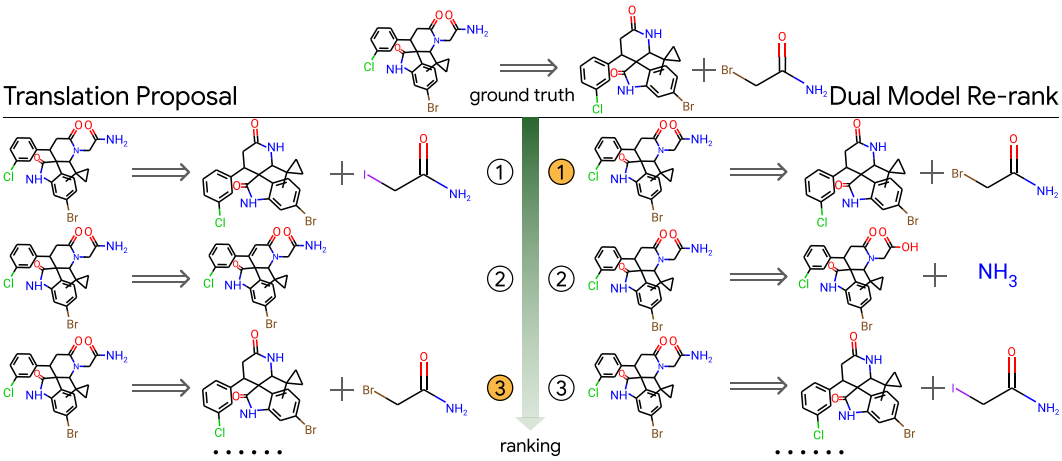

Figure 4: **Dual ranking improves upon translation proposal.**

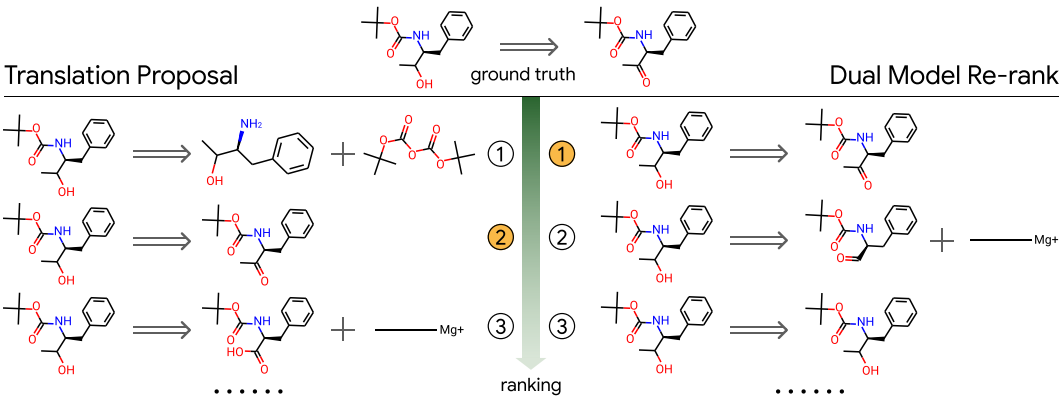

Figure 5: **Dual ranking improves upon translation proposal.** Another example in addition to Figure 4

References

1. Corey, E. Robert Robinson lecture. Retrosynthetic thinking—essentials and examples. *Chemical Society Reviews* **17,** 111–133 (1988).
2. Corey, E. J. The logic of chemical synthesis: multistep synthesis of complex carbogenic molecules (Nobel Lecture). *Angewandte Chemie International Edition in English* **30,** 455–465 (1991).
3. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Accounts of chemical research* **51,** 1281–1289 (2018).
4. Szymkuć, S. *et al.* Computer-Assisted Synthetic Planning: The End of the Beginning. *Angewandte Chemie International Edition* **55,** 5904–5937 (2016).
5. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **28,** 31–36 (1988).
6. Cho, K. *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
7. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
8. Vaswani, A. *et al. Attention is all you need* in *Advances in neural information processing systems* (2017), 5998–6008.
9. Schwaller, P. *et al.* Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS central science* **5,** 1572–1583.
10. Liu, B. *et al.* Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science* **3,** 1103–1113 (2017).
11. Karpov, P., Godin, G. & Tetko, I. V. *A transformer model for retrosynthesis* in *International Conference on Artificial Neural Networks* (2019), 817–830.
12. Zheng, S., Rao, J., Zhang, Z., Xu, J. & Yang, Y. Predicting Retrosynthetic Reactions using Self-Corrected Transformer Neural Networks. *Journal of Chemical Information and Modeling* (2019).
13. Dai, H., Li, C., Coley, C., Dai, B. & Song, L. *Retrosynthesis Prediction with Conditional Graph Logic Network* in *Advances in Neural Information Processing Systems* (2019), 8870–8880.
14. Shi, C., Xu, M., Guo, H., Zhang, M. & Tang, J. A Graph to Graphs Framework for Retrosynthesis Prediction. *arXiv preprint arXiv:2003.12725* (2020).
15. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M. & Huang, F. A tutorial on energy-based learning. *Predicting structured data* **1** (2006).
16. Hinton, G. E. in *Neural networks: Tricks of the trade* 599–619 (Springer, 2012).
17. *Energy-Based Models (EBM) deeplearning tutorial,* http://deeplearning.net/tutorial/rbm.html.
18. Sutskever, I., Vinyals, O. & Le, Q. V. *Sequence to sequence learning with neural networks* in *Advances in neural information processing systems* (2014), 3104–3112.
19. Yang, Z. *et al. Xlnet: Generalized autoregressive pretraining for language understanding* in *Advances in neural information processing systems* (2019), 5754–5764.
20. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
21. Wang, A. & Cho, K. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094* (2019).
22. Kindermann, R. Markov random fields and their applications. *American mathematical society* (1980).
23. Wei, B., Li, G., Xia, X., Fu, Z. & Jin, Z. *Code Generation as a Dual Task of Code Summarization* in *Advances in Neural Information Processing Systems* (2019), 6559–6569.
24. He, D. *et al. Dual learning for machine translation* in *Advances in neural information processing systems* (2016), 820–828.
25. Xia, Y. *et al. Dual supervised learning* in *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (2017), 3789–3798.
26. Tang, D., Duan, N., Qin, T., Yan, Z. & Zhou, M. Question answering and question generation as dual tasks. *arXiv preprint arXiv:1706.02027* (2017).

27. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science* **3,** 1237–1245 (2017).
28. Segler, M. H. & Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal* **23,** 5966–5971 (2017).