Doubly Stochastic Graph-based Non-autoregressive Reaction Prediction

Ziqiao Meng [1*] , **Peilin Zhao** [2†] , **Yang Yu** [2] and **Irwin King** [1†]

[1]The Chinese University of Hong Kong

[2]Tencent AI Lab

{zqmeng, king}@cse.cuhk.edu.hk, mazonzhao@tencent.com, kevinyyu@tencent.com

Abstract

Organic reaction prediction is a critical task in drug discovery. Recently, researchers have achieved non-autoregressive reaction prediction by modeling the redistribution of electrons, resulting in state-of-the-art top-1 accuracy, and enabling parallel sampling. However, the current non-autoregressive decoder does not satisfy two essential rules of electron redistribution modeling simultaneously: the electron-counting rule and the symmetry rule. This violation of the physical constraints of chemical reactions impairs model performance. In this work, we propose a new framework called *ReactionSink* that combines two doubly stochastic self-attention mappings to obtain electron redistribution predictions that follow both constraints. We further extend our solution to a general multi-head attention mechanism with augmented constraints. To achieve this, we apply Sinkhorn's algorithm to iteratively update self-attention mappings, which imposes doubly conservative constraints as additional informative priors on electron redistribution modeling. We theoretically demonstrate that our *ReactionSink* can simultaneously satisfy both rules, which the current decoder mechanism cannot do. Empirical results show that our approach consistently improves the predictive performance of non-autoregressive models and does not bring an unbearable additional computational cost.

1 Introduction

Reaction prediction is a crucial task in computational chemistry. With a reliable prediction model, chemists can verify the accuracy of retrosynthetic routes and potentially uncover new chemical reactions, which could significantly benefit the drug discovery industry. However, given a set of reactants, the number of potential products can be combinatorially large, necessitating automatic tools to narrow down the
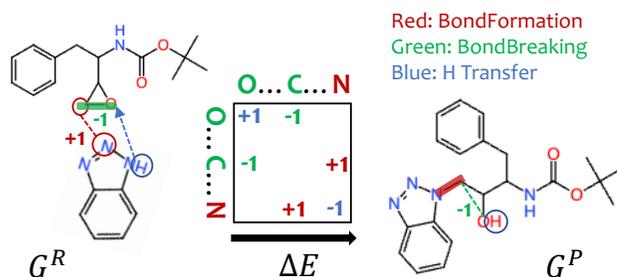

Figure 1: This figure illustrates chemical reaction from electron redistribution perspective. Red circles and arrows denote involved atoms that form new bond. Green arrow denotes the bond that breaks. Blue circles and arrows describe the hydrogen atom transfer. With electron redistribution ΔE, reactants G^R are transformed to products G^P such that $E^R + \Delta E = E^P$, where E^R and E^P are the adjacency matrices of reactants and products respectively. Note that E^R, E^P and ΔE does not consider Hydrogen atoms explicitly and ΔE records the change of attached hydrogen atoms in diagonal.

search space. Consequently, the use of deep learning techniques to automatically infer possible reaction products has become prevalent and essential.

Translation-based autoregressive models [Schwaller *et al.*, 2019; Tetko *et al.*, 2020; Tu and Coley, 2021; Lu and Zhang, 2022; Irwin *et al.*, 2022; Zhao *et al.*, 2022] have dominated the design of template-free end-to-end reaction models in recent years. These models represent molecules as sequences of characters, called SMILES [Weininger, 1988], and formulate the reaction prediction problem as a neural machine translation problem [Sutskever *et al.*, 2014; Vaswani *et al.*, 2017; Devlin *et al.*, 2019]. This class of models achieves great predictive performance and does not rely on pre-computed atom-mapping information. However, autoregressive models have two major shortcomings. *First*, autoregressive sampling is very inefficient because the model has to generate the final predictions step-by-step from intermediate parts. *Second*, training autoregressive models requires pre-defined generation orders, while the absolute generation order of molecules is ambiguous.

To avoid these issues, a non-autoregressive model called NERF [Bi *et al.*, 2021] has been proposed. It predicts the redistribution of electrons, which is an important observation

because chemical reactions can be reflected by bond reconnection, since the elements of the involved atoms remain unchanged. Bond reconnection inherently involves electron redistribution. For example, forming a single bond between a carbon atom and a nitrogen atom involves sharing a pair of electrons between them. Therefore, predicting electron redistribution ΔE is sufficient for inferring possible complete products. With a novel design for the electron redistribution decoder, NERF has achieved state-of-the-art top-1 accuracy and much faster sampling speed with parallel inference.

However, an important issue that has been neglected is that $\Delta \hat{E}$ must simultaneously follow the *electron counting rule* and the *symmetry rule* to better approximate ΔE. The chemical reaction that adheres to both of these rules is illustrated in Figure 1. We can clearly see that ΔE is row-wise conservative (conforming to the electron counting rule) and symmetric. Currently, the NERF decoder neglects to symmetrize $\Delta \hat{E}$, only ensuring that it follows the electron counting rule through combinations of *BondBreaking* and *BondFormation* self-attention matrices. This leads to inexact modeling of ΔE and impairs model performance. The predicted product adjacency matrix $\hat{E}^P$ is forced to be symmetric, but this operation implicitly corrupts the row-wise conservatism of $\Delta \hat{E}$.

In this work, we propose a novel framework called *ReactionSink* to refine the decoder. The framework applies Sinkhorn's algorithm to iteratively update the *BondBreaking* and *BondFormation* self-attention matrices. With a sufficient number of iterations, each self-attention matrix converges to a doubly stochastic matrix. To generalize this solution to multi-head attention mechanisms, we impose augmented constraints on the Sinkhorn normalization operation. We theoretically show that the NERF decoder leads to an implicit contradiction in preserving both physical constraints. In contrast, doubly stochastic self-attention mappings lead to valid $\Delta \hat{E}$, which preserve both rules simultaneously and are not corrupted by the symmetrization on $\hat{E}^P$. We further discuss the implicit connections between the reaction prediction problem and the discrete optimal transport problem to validate the rationality of imposing the doubly stochasticity constraint on the decoder as an information prior. Comprehensive empirical results on the open benchmark dataset USPTO-MIT demonstrate that our approach consistently outperforms baseline non-autoregressive reaction prediction models. Our contributions can be summarized as follows:

- We introduce two chemical rules that the predicted electron redistribution $\Delta \hat{E}$ should follow and theoretically show that the current non-autoregressive decoder inevitably violates either of these chemical rules;

- We propose a novel decoder design that conforms to both chemical rules through leveraging combinations of doubly stochastic self-attention mappings and we generalize our approach to multi-head attention mechanism;

- We provide some new interpretations on electron redistribution modeling, which facilitates further research on potential connections between the reaction prediction problem and the optimal transport problem.

2 Related Work

Reaction Prediction. In the early years, template-based methods were widely adopted for reaction prediction. Experts deduced possible products with the help of concluded reaction templates. Although these methods are reliable, they suffer from poor generalization. Quantum computation methods [Dral, 2020; Lam *et al.*, 2020; Raucci *et al.*, 2022] have also been adopted for reaction predictions, but they suffer from low computation speeds, despite being template-free. With the rise of deep learning, various deep learning architectures have been utilized in reaction modeling. Models that learn neural matching between reactions and templates have been proposed [Segler and Waller, 2017; Segler and Waller, 2016a]. WLDN [Jin *et al.*, 2017] proposes a template-free model that splits reaction prediction into a reaction center identification stage and candidate ranking stage. MEGAN and GTPN [Sacha *et al.*, 2020; Do *et al.*, 2019] model reactions as a sequence of graph edits, while Electro [Bradshaw *et al.*, 2019] regards reactions as a sequence of electron transfers. A major class of models are translation-based reaction prediction models [Schwaller *et al.*, 2019; Tetko *et al.*, 2020; Tu and Coley, 2021; Lu and Zhang, 2022; Irwin *et al.*, 2022; Zhao *et al.*, 2022], which mainly transfer techniques from language modeling to reaction modeling. NERF [Bi *et al.*, 2021] refines the decoder to model electron redistribution and achieves non-autoregressive modeling with state-of-the-art top-1 accuracy and parallel inference.

Sinkhorn and Attention. The Sinkhorn algorithm [Sinkhorn, 1966; Cuturi, 2013; Peyré and Cuturi, 2019] is a well-studied tool for approximating solutions to the optimal transport problem. For instance, a new set pooling algorithm is presented in [Mialon *et al.*, 2021], which embeds the optimal transport plan between input sets and reference sets, and then uses the Sinkhorn algorithm as a fast solver. Sinkhorn normalization is presented in [Marshall and Coifman, 2019; Wormell and Reich, 2021], and its convergence properties to a doubly stochastic matrix have been theoretically validated. SparseMAP [Niculae *et al.*, 2018] uses doubly stochastic attention matrices in LSTM-based encoder-decoder networks. Sinkhorn updates are also applied for differentiable ranking over internal representations in [Adams and Zemel, 2011]. Gumbel-Sinkhorn [Mena *et al.*, 2018] applies the Sinkhorn algorithm to learn stochastic maximization over permutations. The Sinkhorn algorithm is used to iteratively generate doubly stochastic matrices for approximating smooth filters in [Milanfar, 2013]. It has also been applied to refine self-attention mechanisms in [Vaswani *et al.*, 2017]. The Sparse Sinkhorn Transformer [Tay *et al.*, 2020] learns sparse self-attention by introducing a sorting network that generates a doubly stochastic matrix to permute input sequence elements. Sinkformer [Sander *et al.*, 2022] applies the Sinkhorn algorithm to derive doubly stochastic transformers.

3 Proposed Methods

This section introduces the learning framework of our new non-autoregressive reaction prediction model, *ReactionSink*. We first describe its problem formulation and general training
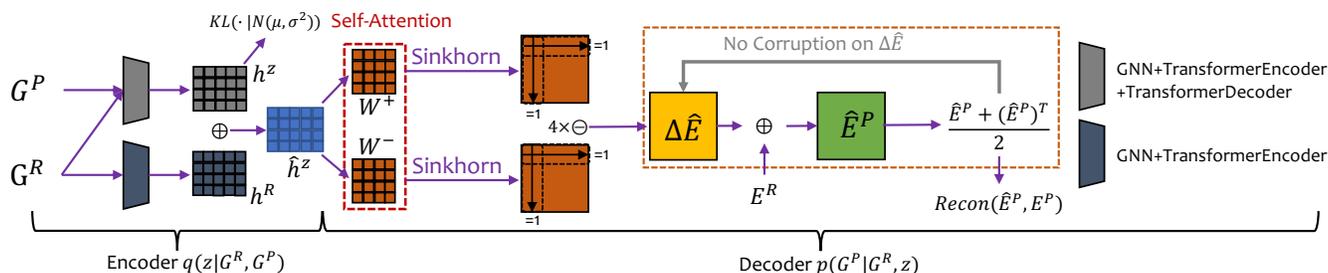
Figure 2: This figure illustrates the training procedure of *ReactionSink*. It adopts the conditional variational autoencoder framework with KL-divergence loss and reconstruction loss $||E^P - \hat{E}^P||_2^2$. Red rectangles highlight *BondFormation* and *BondBreaking* self-attention mappings. Orange rectangle highlights the process we aim to analyze. $\ominus$ denotes the element-wise subtraction, $\oplus$ denotes the element-wise plus. With Sinkhorn algorithm applied on *BondFormation* and *BondBreaking*, $\Delta\hat{E}$ becomes doubly conservative and would not be corrupted by symmetrization on $\hat{E}^P$. "4×" denotes the multi-head attention with 4 heads.

process. Then we introduce its encoder architecture, which obtains the conditional reaction latent embedding. Most importantly, we present our novel decoder mechanism that leverages doubly stochastic self-attention matrices to obtain ΔE. In the theoretical analysis section, we mathematically compare our decoder with the NERF [Bi *et al.*, 2021] decoder to demonstrate the advantages of our approach. Finally, we briefly discuss the new insights brought by *ReactionSink*.

Problem Formulation. A chemical reaction is represented as a pair (G^R, G^P), where $G^R = (V, E^R, f^R)$ denotes the set of reactants and $G^P = (V, E^P, f^P)$ denotes the set of products, V is the set of $|V|$ atoms, E^R and E^P denote the adjacency matrix of G^R and G^P respectively, f^R and f^P denote the atom feature matrix of G^R and G^P respectively. Following the atom-mapping principle, reactants and products share the same set of atoms V. Unlike most of the papers representing E as a 3D tensor, where $E \in \mathbb{R}^{|V|\times|V|}$ is a 2D matrix and E_{ij} denotes the number of shared electrons between atom i and atom j (i.e. single bond is 1, double bond is 2). For special aromatic bond, we denote it as $E_{ij} = 1.5$. The goal of reaction prediction is to predict the set of products G^P given a set of reactants G^R.

Following NERF [Bi *et al.*, 2021], we also adopt the conditional variational autoencoder [Sohn *et al.*, 2015] (CVAE) architecture to approximate $p(G^P|G^R)$ by introducing a latent variable z. Instead of directly maximizing the log-likelihood $\log p(G^P|G^R)$, CVAE is to maximize its evidence-lower bound (ELBO):

$$
\begin{aligned}
ELBO = &\mathbb{E}_{q(z|G^P,G^R)}[\log p(G^P|G^R, z)] - \\
&\text{KL}(q(z|G^P,G^R)||p(z|G^R)) \leq \log p(G^P|G^R),
\end{aligned} \tag{1}
$$

where $q(z|G^P, G^R)$ is the reaction encoder with reaction (G^R, G^P) as input and low-dimensional embedding $\mathbf{h}^z$ as output, $p(G^P|G^R, z)$ is product decoder with reactants G^R and latent embedding $\mathbf{h}^z$ as input, $p(z|G^R)$ denotes the prior distribution of latent variable z. KL term is minimizing the gap between $q(z|G^P, G^R)$ and $p(z|G^R)$. In real implementations, the backbone network architectures of the encoder $q(z|G^P, G^R)$ are graph neural networks (GNNs) and transformers. GNNs capture local interactions within each

molecule and transformers capture global interactions across different molecules. With this architecture, G^R and G^P will be projected to reactants embedding $\mathbf{h}^R$ and products embedding $\mathbf{h}^P$ respectively. The cross attention layer is used for mapping $\mathbf{h}^R$ to latent $\mathbf{h}^z$ with $\mathbf{h}^P$ as teacher forcing during training. Finally, the latent conditional reaction embedding is derived such that $\hat{\mathbf{h}}^z = \mathbf{h}^R + \mathbf{h}^z$. Note that similar reaction embedding networks are widely adopted in many related works as long as transformer-based reaction encoder architectures are used. Please refer the detailed architectures in Appendix 1 and the general architectures of *ReactionSink* is illustrated in Figure 2.

3.1 Decoder with Physical Constraints

The core mechanism of the decoder $p(G^P|G^R, z)$ is first decoding electron redistribution matrix $\Delta\hat{E}$ from conditional latent $\hat{\mathbf{h}}^z$, and then add it to E^R to predict $\hat{E}^P$, i.e., $\hat{E}^P = E^R + \Delta\hat{E}$. Before presenting *ReactionSink* decoder in details, we first introduce two important physical rules, the *electron-counting* rule and the *symmetry* rule, which should be followed by the predicted $\Delta\hat{E}$. However, these rules are not explicitly stated in the previous literature. We will mathematically characterize these two rules as matrix properties.

Electron Counting Rule. This is an important chemical rule of thumb: the octet rule. It states that main-group elements (carbon, nitrogen, oxygen, halogens) tend to bond in a way that each atom has eight electrons in its valence shell. Following this rule, electron redistribution of each atom is conservative, allowing it to maintain eight electrons in its valence shell. For example, when a carbon atom forms a new bond with one atom, it tends to break an existing bond with another atom in order to maintain eight electrons in its valence shell. This empirical rule leads to an important mathematical property of ΔE, which is **row-wise conservative** (i.e., $\sum_j \Delta E_{ij} = 0$ for each atom i). Note that this rule is more informative than the charge conservation rule, which can be mathematically expressed as $\sum_{i,j} \Delta E_{ij} = 0$.

Symmetry Rule. The ground-truth ΔE should clearly be symmetric. For example, if the edge number between atom i and atom j increases by 1, then the edge number between

atom j and atom i should also increase by 1. To better approximate ΔE, $\Delta \hat{E}$ should also be symmetric.

We have shown that $\Delta \hat{E}$ should preserve two mathematical properties which are $\sum_j \Delta \hat{E}_{ij} = 0$ for any j and $\Delta \hat{E}_{ij} = \Delta \hat{E}_{ji}$ for any i, j. This can easily come up with a new conclusion that $\Delta \hat{E}$ should also be **column-wise conservative** such that $\sum_i \Delta \hat{E}_{ij} = 0$ for any j. Therefore, $\Delta \hat{E}$ should be both row-wise conservative and column-wise conservative in the strict sense, which is **doubly conservative**. Inspired by NERF decoder that leverages the combinations of row-wise stochastic self-attention mappings to achieve row-conservative $\Delta \hat{E}$, we achieve doubly conservative $\Delta \hat{E}$ by using combinations of two **doubly stochastic self-attention mappings**. Specifically, we decode two self-attention mappings *BondFormation* W^+ and *BondBreaking* W^- from the conditional latent $\hat{\mathbf{h}}^z$. *BondFormation* W^+ is to predict bond addition and *BondBreaking* W^- is to predict bond removal. The detailed self-attention decoding mechanism is shown in Appendix 2. Then the combinations of W^+ and W^- should satisfy the following equations:

$$\begin{aligned} (\sum_j W^+_{ij} - \sum_j W^-_{ij}) = \sum_j \Delta \hat{E}_{ij} = 0, \text{ for any i,} \\ (\sum_i W^+_{ij} - \sum_i W^-_{ij}) = \sum_i \Delta \hat{E}_{ij} = 0, \text{ for any j.} \end{aligned} \quad (2)$$

To satisfy the above equations, we force W^+ and W^- to be doubly stochastic such that:

$$\begin{aligned} \sum_j W^+_{ij} = 1, \sum_j W^-_{ij} = 1 \text{ for any i,} \\ \sum_i W^+_{ij} = 1, \sum_i W^-_{ij} = 1 \text{ for any j.} \end{aligned} \quad (3)$$

Note that forcing W^+ and W^- to be doubly stochastic is probably not the only solution to Eq 2. But stochasticity (sum to 1) is the most straightforward solution and more compatible with normalization operations. Therefore, if W^+ and W^- can be doubly stochastic matrices as stated in Eq 3, their combinations can lead to a doubly conservative $\Delta \hat{E}$. Then we can apply symmetrization operation on $\Delta \hat{E}$ such that $\frac{\Delta \hat{E} + (\Delta \hat{E})^T}{2}$, which is averaging predictions of $\Delta \hat{E}_{ij}$ and $\Delta \hat{E}_{ji}$ for each i, j. Interestingly, symmetrization in this case would not corrupt the conservative of $\Delta \hat{E}$. We will theoretically validate these claims in the next section.

Unfortunately, decoding doubly stochastic self-attention mappings from conditional latent variables is non-trivial. Recall that the vanilla self-attention mapping is obtained as `SoftMax`$(\exp(QK^T))$, where Q and K are queries and keys linearly projected from latent variable $\hat{\mathbf{h}}^z$. Note that the `SoftMax` operator is basically a row-wise normalization technique that forces self-attention mappings to be row-wise stochastic, while it is not necessarily column-wise stochastic. Therefore, a specific operator is required to extend *BondFormation* W^+ and *BondBreaking* W^- to doubly stochastic matrices. Furthermore, this operator should be differentiable so that it can be naturally merged into decoder neural networks.

3.2 Sinkhorn's Algorithm

An iterative algorithm to convert positive matrix to doubly stochastic matrix is Sinkhorn's algorithm. The main idea of Sinkhorn's algorithm is alternatively applying row-wise normalization and column-wise normalization on the given positive matrix. Given a positive square matrix $X \in \mathbb{R}^{N \times N}$, we define the Sinkhorn operator as follows:

$$\begin{aligned} S^0(X) &= \exp(X), \\ S^l(X) &= T_c(T_r(S^{l-1}(X))), \quad (4) \\ S^\infty(X) &= \lim_{l \to \infty} S^l(X), \end{aligned}$$

where T_r denotes the row-wise normalization and T_c denotes the column-wise normalization. They are formally described as follows:

$$\begin{aligned} T_r(X) &= X \oslash (X \mathbf{1}_N \mathbf{1}_N^T), \\ T_c(X) &= X \oslash (\mathbf{1}_N \mathbf{1}_N^T X), \end{aligned} \quad (5)$$

where $\oslash$ denotes the element-wise division operator and $\mathbf{1}$ denotes the all one vector. We only need to apply the above normalization operations alternatively with enough number of iterations. In real implementation, to stabilize the updating process, we put the computation of Eq 5 in log domain. Details of this implementation are shown in the Appendix 6.

Sinkhorn [Sinkhorn, 1966] proves that $S^\infty(X)$ belongs to the Birkhoff polytope, the set of doubly stochastic matrices, such that $S^\infty(X)\mathbf{1}_N = \mathbf{1}_N$ and $S^\infty(X)^T\mathbf{1}_N = \mathbf{1}_N$. Hence, applying this Sinkhorn operator on W^+ and W^- separately will make them converge to doubly stochastic matrices. This Sinkhorn operator can be easily coupled with the self-attention mechanism as an additional layer. In addition, it can be perfectly suited for backpropagation updates of neural networks. Specifically, this Sinkhorn operator can be stacked after the vanilla self-attention, since the vanilla self-attention mapping can be regarded as $T_r(\exp(QK^T))$, which is equivalent to the initialization step plus the row-wise normalization step in Eq 4. Therefore, in real implementations, we start with the column-wise normalization T_c after the `SoftMax` layer in the vanilla self-attention.

Currently, each entry of $\Delta \hat{E}$ is in range $[-1, 1]$ while edge number changes more than 1 in many cases. Therefore, we apply multi-head attention mechanism such that $\Delta \hat{E} = \sum_{d=1}^4 W^{+d} - \sum_{d=1}^4 W^{-d}$ with $d = 4$ attention heads. In this way, $\Delta \hat{E}_{ij}$ is in range $[-4, 4]$, which could cover enough support of edge number changes of chemical reactions. We only need to slightly change Eq 3 by increasing conservation constraints to 4 and impose the augmented doubly conservative constraint on the sum of attention heads $\sum_{d=1}^4 W^{+d}$ and $\sum_{d=1}^4 W^{-d}$ respectively as follows:

$$\begin{aligned} \sum_{d=1}^4 \sum_j W^{+d}_{ij} = 4, \sum_{d=1}^4 \sum_j W^{-d}_{ij} = 4 \text{ for any i,} \\ \sum_{d=1}^4 \sum_i W^{+d}_{ij} = 4, \sum_{d=1}^4 \sum_i W^{-d}_{ij} = 4 \text{ for any j.} \end{aligned} \quad (6)$$

For applying the Sinkhorn normalization operator to the multi-head attention mechanism, we only need to impose the

augmented constraints to both row-wise normalization and column-wise normalization such that:

$$T_r(X) = 4X \oslash (X\mathbf{1}_N\mathbf{1}_N^T),$$
$$T_c(X) = 4X \oslash (\mathbf{1}_N\mathbf{1}_N^T X). \qquad (7)$$

To conclude, with Sinkhorn's algorithm, we can efficiently convert W^+ and W^- to doubly stochastic matrices, to finally achieve symmetric doubly conservative $\Delta\hat{E}$.

4 Theoretical Analysis

In this section, we first theoretically show that the NERF decoder cannot output valid $\Delta\hat{E}$ that follows the previously mentioned two physical constraints. In comparison, we mathematically validate that our *ReactionSink* decoder can generate valid $\Delta\hat{E}$ and the symmetry is not contradictory to the conservative.

In NERF decoder, it generates two vanilla self-attention mappings W^+ and W^- based on conditional latent for *Bond-Formation* and *BondBreaking* respectively and approximates $\Delta\hat{E}$ through subtraction with multi-head attention such that $\Delta\hat{E} = \sum_{d=1}^{4} W^{+d} - \sum_{d=1}^{4} W^{-d}$. Then it applies symmetrization on the final predicted $\hat{E}^P = E^R + \Delta\hat{E}$ such that $\frac{E^P + (E^P)^T}{2}$. Therefore, we begin to verify that $\Delta\hat{E}$ has also been implicitly symmetrized by NERF decoder, which is the following lemma:

Lemma 4.1 *Symmetrization on the final predicted product adjacency matrix $\hat{E}^P$ is equivalent to implicit symmetrization on the predicted electron redistribution $\Delta\hat{E}$.*

The proof procedures are shown in Appendix 2.

Then we argue that NERF decoder design cannot satisfy the two proposed physical constraints simultaneously. Specifically, the symmetrization operation on $\hat{E}^P$ would result in that $\Delta\hat{E}$ violates the first rule. Since we have already shown that the symmetrization of $\hat{E}^P$ is equivalent to the implicit symmetrization of $\Delta\hat{E}$ in lemma 4.1, we only need to further prove the following lemma:

Lemma 4.2 *Implicit symmetrization on $\Delta\hat{E}$ would result in that $\Delta\hat{E}$ violates the electron counting rule such that $\sum_j \Delta\hat{E}_{ij} \neq 0$ for some i.*

The proof procedures are shown in Appendix 3.

To summarize, electron redistribution predictions generated by NERF cannot satisfy both constraints simultaneously. Implicit symmetrization operation applied on $\Delta\hat{E}$ makes it violate the first rule. Otherwise it will violate the second rule if no symmetrization operation is adopted. In short, NERF decoder actually leads to a contradiction between two physical constraints.

In comparison, with doubly stochastic self-attention mapping W^+ and W^-, *ReactionSink* can conform to both physical rules concurrently. We can prove the following theorem:

Theorem 4.3 *If the self-attention mappings for bondformation W^+ and bondbreaking W^- are doubly stochastic, then $\Delta\hat{E}$ would be guaranteed to be doubly conservative and its*

doubly conservation would not be corrupted by the symmetrization of $\hat{E}^P$.

The proof procedures are shown in Appendix 4.

To extend the above conclusion to multi-head attention mechanism. The following corollary can be easily proved:

Corollary 4.3.1 *If the multi-head self-attention mechanisms for bondformation $\sum_{d=1}^{D} W^{+d}$ and bondbreaking $\sum_{d=1}^{D} W^{-d}$ are doubly conservative with D-sum constraints, then $\Delta\hat{E}$ would be guaranteed to be doubly conservative and its doubly conservation would not be corrupted by the symmetrization of $\hat{E}^P$.*

The proof procedures are shown in Appendix 5.

5 Further Discussions

Our novel framework, *ReactionSink*, provides new insights for the non-autoregressive reaction prediction problem. Doubly stochastic matrices usually have special meaning in mathematics, with the most well-known being the permutation matrix. This suggests that electron redistribution can be formulated as a learned permutation of electrons. Note that permuting a matrix is a multiplication operation, which is different from the current formulation. The permutation formulation is intuitively closer to the discrete nature of electron redistribution, but the detailed formulation method requires further research. More generally, reaction prediction can be connected with the optimal transport (OT) problem. The permutation problem is the most elementary discrete OT problem. Specifically, electron redistribution can also be regarded as allocating electrons to lower the energy of the entire reaction system, which is implicitly connected to the discrete OT problem. Our work presents these new insights to inspire future research, although they are still in their nascent stages and have not been revealed by existing research.

6 Experiments

Dataset. Following previous work, we evaluate our approach on the open public benchmark dataset USPTO-MIT [Jin *et al.*, 2017], which contains 479K reactions filtered by removing duplicates and erroneous reactions from Lowe's original data [Lowe, 2012]. An exisitng work discovers that about 0.3% of reactions in USPTO-479K do not satisfy the non-autoregressive learning settings for various reasons [Bi *et al.*, 2021]. Following this work [Bi *et al.*, 2021], we also filter 0.3% reactions from dataset and deduct 0.3% from the top-k accuracies of our model correspondingly.

Experimental Setting. We conduct experiments on three different splits of reaction prediction, which are random split, tanimoto-0.4 split and tanimoto-0.6 split. Random split is adopted by most of the previous work. Scaffold splits, tanimoto-0.4 split and tanimoto-0.6 split, are adopted by [Kovacs *et al.*, 2021] to test the generalization of reaction model with larger distribution shift between training and testing. Tanimoto similarity is measuring whether two reactions are structurally similar. Higher tanimoto index indicates that two reactions are more similar and otherwise two reactions are

Category	Model	Top-1	Top-2	Top-3	Top-5	Top-10	Parallel	End-to-end
Template-based	Symbolic†	90.4	93.2	94.1	95.0	-	✓	×
Two-stage	WLDN†	79.6	-	87.7	89.2	-	✓	×
Autoregressive	GTPN†	83.2	-	86.0	86.5	-	×	✓
	MT-base†	88.8	92.6	93.7	94.4	94.9	×	✓
	MEGAN†	89.3	92.7	94.4	95.6	95.4	×	✓
	MT†	90.4	93.7	**94.6**	95.3	-	×	✓
	Chemformer†	91.3	-	-	93.7	94.0	×	✓
	Sub-reaction†	91.0	-	94.5	95.7	-	×	✓
	Graph2Smiles†	90.3	-	94.0	94.8	95.3	×	✓
	AT $\times 100$†	90.6	**94.4**	-	**96.1**	-	×	✓
Non-autoregressive	NERF	90.7	92.3	93.3	93.7	94.0	✓	✓
	ReactionSink	**91.3**	93.3	94.0	94.5	94.9	✓	✓

Table 1: Top-K Accuracy % on USPTO-479K with original random split. Best results are bolded. † indicates that the reported results are copied from the corresponding published papers. We also show the category of each model to reflect its property. "Parallel" indicates whether the model is capable of parallel inference. "End-to-end" indicates whether the model can make end-to-end inference.

dissimilar with each other (e.g. Tanimoto-0.4 has larger distribution gap between training and testing). For original random split, the training set, validation set and testing set follows split ratio 409K:30K:40K. For scaffold split, the split ratio is 392K:30K:50K.

Comparison Baselines. We compare *ReactionSink* with the following baseline models: **WLDN** [Jin *et al.*, 2017] is a two-stage model, which firstly predicts reaction centers and then ranks enumerated products; **GTPN** [Do *et al.*, 2019] models reaction predictions as a series of graph transformations and use policy networks to learn the transformations; **MT-base** [Schwaller *et al.*, 2019] is a transformer-based autoregressive modeling with both input and output in SMILES sequence format; **MEGAN** [Sacha *et al.*, 2020] models reaction predictions as series of graph edit operations and generates edit sequences in an autoregressive manner; **MT** [Schwaller *et al.*, 2019] is MT-base models with data augmentation techniques applied to SMILES sequence input; **Symbolic** [Qian *et al.*, 2020] introduces the chemical rules to reaction modeling using symbolic inference; **Chemformer** [Irwin *et al.*, 2022] leverages molecular SMILES encoder pretrained on 100M molecular datasets with three self-supervised tasks; **Graph2Smiles** [Tu and Coley, 2021] leverages the similar backbone network of **NERF** [Bi *et al.*, 2021]; **AT×100** [Tetko *et al.*, 2020] leverages molecular transformer with ×100 SMILES augmentations; **Sub-reaction** [Zhao *et al.*, 2022] leverages motif tree to achieve substructure-aware reaction prediction; **NERF** [Bi *et al.*, 2021] uses CVAE to achieve non-autoregressive modeling;

Model Configuration and Reproducibility Setting. The major encoder and decoder architectures are following NERF [Bi *et al.*, 2021]. To ensure fair comparison with major baseline models Molecular Transformer and NERF, we set the number of transformer encoder layers and transformer decoder layers (cross-attention layer) to be 4, the same as previous work. And we set the dimension of latent embedding

to be 256. For multi-head attention decoder, *BondFormation* and *BondBreaking* both have 4 attention heads. The model is optimized using Adam optimizer [Kingma and Ba, 2015] at learning rate 10^{-4} with linear warm-up and linear learning rate decay. The number of iterations l of Sinkhorn normalization is also an hyperparameter to fine-tune. Larger l indicates more rounds of Sinkhorn normalization. Finally, we train our *ReactionSink* for 100 epochs with a batch size of 128 using 8 Nvidia V100 GPUs in this work.

Evaluation Metrics. Following the convention of previous work, we adopt the top-k accuracies to evaluate the performance of all the compared algorithms. The top-k accuracy is the percentage of reactions that have the ground-truth product in the set of the top-k predicted molecules. As long as the set of predicted products contains the ground-truth main product, then the prediction would be counted as a correct one. In this work, the value of k is set as 5 different values: $\{1, 2, 3, 5, 10\}$.

Sampling Top-k Predictions. Since *ReactionSink* follows CVAE architecture, multi-modal outputs are generated through sampling k different latent vectors $\hat{h}^z$ by increasing temperature values. To sample top-k predictions, we multiply a scalar temperature parameter t to the variance of standard Gaussian distribution, such that h^zs are sampled from different $N(\mathbf{0}, t\mathbf{I})$. Increasing the temperature value t would make the model predict different products. Specifically, we sample the first k predictions as our top-k predictions. Lower temperatures are set to output predictions with higher rank (e.g. prediction using $t = 1$ is treated as the top-1 prediction).

Main Experiments with Random Split and Tanimoto Splits. We conduct experiments under random split and tanimoto splits, which can be seen as a cross-validation process. Table 1 is the major benchmark adopted by previous work under random split conducted by WLDN [Jin *et al.*, 2017]. From this table, we can see that our method reaches the state-

Model Name	Top-1	Top-3	Top-5
WLDN5	75.9	86.2	88.8
MT	80.9	**88.2**	**89.6**
NERF	85.0	86.7	88.8
ReactionSink	**86.0**	87.2	89.3

Table 2: Top-K Accuracy % on USPTO-479K with Tanimoto Similarity < 0.6 (left)

Model Name	Top-1	Top-3	Top-5
WLDN5	69.3	80.9	84.1
MT	74.6	82.9	84.5
NERF	80.0	82.5	84.4
ReactionSink	**82.2**	**83.2**	**84.5**

Table 3: Top-K Accuracy % on USPTO-479K with Tanimoto Similarity < 0.4 (right)

Model Name	Wall-time	Latency	Speedup
Transformer (b=5)	9min	448ms	$1\times$
MEGAN (b=10)	31.5min	144ms	$0.29\times$
Symbolic	>7h	1130ms	$0.02\times$
NERF	20s	17ms	$27\times$
ReactionSink	24s	20ms	$26\times$

Table 4: Computation speedup (compared with Transformer)

Model Name	Top-1	Top-3	Top-5
ReactionSink (l=1)	90.9	93.2	93.7
ReactionSink (l=2)	91.0	93.3	94.0
ReactionSink (l=3)	**91.3**	93.7	94.4
ReactionSink (l=5)	**91.3**	94.0	94.5
ReactionSink (l=10)	91.3	94.0	94.5

Table 5: Ablation studies on the number of iterations l of Sinkhorn normalization

of-the-art top-1 accuracy over all baseline methods and consistently outperforms the non-autoregressive model NERF with top-k accuracies. Note that these results are evaluated on models without knowing reagents, which means these models do not know which reactant is reagent during inference stage. We can see that currently non-autoregressive models are performing worse than autoregressive models when k is higher. We conjecture that this may be caused by the simple sampling method for CVAE, or simple latent distribution assumption on uncertainty modeling (Gaussian distribution). In the future, it is worth to find better sampling method for CVAE, and explore stronger generative model than CVAE for reaction modeling. From Table 2 and Table 3, we can observe that non-autoregressive models have stronger generalization than autoregressive models. We conjecture that electron redistribution modeling proposed by NERF [Bi *et al.*, 2021] is closer to the nature of this problem. Under tanimoto splits, *ReactionSink* consistently improves the performance of non-autoregressive models in terms of top-k accuracies.

Ablation Studies on Sinkhorn Iteration. Since the Sinkhorn algorithm is also an iterative normalization process, we conduct ablation studies on the number of iterations l to check its effect on model performance. From table 5, we can see that the model performance is consistently improved when increasing l from 1 to 5. However, when l increases to 10, it does not bring further performance gains compared to $l = 3$. This demonstrates that the *Bondformation* and *Bond-Breaking* matrices are converted to doubly stochastic matrices with few iterations of the Sinkhorn normalization.

Computational Efficiency. The main concern about *ReactionSink* is its computational complexity, which is caused by the additional layers of Sinkhorn normalization. Fortunately, with a few iterations of Sinkhorn normalization, the self-attention matrices will converge to doubly stochastic matrices with negligible error. Therefore, the computational burden would not be significantly increased due to additional normalization. Following NERF, we report the wall-time and la-

tency of the inference model. Wall-time is the total time cost for inferring all testing samples, and latency is the inference time cost for a single testing sample. The detailed wall-time and latency computation standards are stated in NERF [Bi *et al.*, 2021]. From Table 4, we can see that the proposed method only adds a few additional seconds to Wall-time and a few more milliseconds to latency. This demonstrates that the additional normalization layers do not create a trade-off between accuracy and speed. Regarding the training process, additional Sinkhorn normalization only increases the training time by a few seconds for each epoch, which is almost negligible compared to the total training time. Overall, *ReactionSink* is more suitable for high-throughput reaction predictions than auto-regressive methods.

7 Conclusion

In this work, we first introduce two chemical rules that govern electron distribution and characterize their corresponding mathematical properties. We then identify that the current electron redistribution decoder does not maintain these two physical constraints simultaneously, and we theoretically prove our claims. To address this issue, we propose the *ReactionSink* architecture to extend the current self-attention mapping to doubly stochastic matrices. We also prove that the predicted electron distribution generated by our proposed methods better adheres to the physical constraints. Additionally, we establish the connections between electron redistribution and the learned optimal transport problem. This new perspective has the potential to lead to a new problem formulation for reaction prediction. Experimental results demonstrate that *ReactionSink* consistently outperforms the state-of-the-art non-autoregressive reaction prediction model and does not require expensive computational costs.

Acknowledgments

References

[Adams and Zemel, 2011] Ryan Prescott Adams and Richard S. Zemel. Ranking via sinkhorn propagation. *CoRR*, abs/1106.1925, 2011.

[Bi *et al.*, 2021] Hangrui Bi, Hengyi Wang, Chence Shi, Connor W. Coley, Jian Tang, and Hongyu Guo. Non-autoregressive electron redistribution modeling for reaction prediction. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 904–913. PMLR, 2021.

[Bradshaw *et al.*, 2019] John Bradshaw, Matt J. Kusner, Brooks Paige, Marwin H. S. Segler, and José Miguel Hernández-Lobato. A generative model for electron paths. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Light-speed computation of optimal transport. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2292–2300, 2013.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[Do *et al.*, 2019] Kien Do, Truyen Tran, and Svetha Venkatesh. Graph transformation policy network for chemical reaction prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 750–760. ACM, 2019.

[Dral, 2020] Pavlo Dral. Quantum chemistry in the age of machine learning. *Journal of Physical Chemistry Letters*, 11:2336–2347, 03 2020.

[Irwin *et al.*, 2022] Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, jan 2022.

[Jin *et al.*, 2017] Wengong Jin, Connor W. Coley, Regina Barzilay, and Tommi S. Jaakkola. Predicting organic reaction outcomes with weisfeiler-lehman network. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2607–2616, 2017.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Kovacs *et al.*, 2021] David Kovacs, William McCorkindale, and Alpha Lee. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nature Communications*, 12, 03 2021.

[Lam *et al.*, 2020] Yu-hong Lam, Yuriy Abramov, Ravi S. Ananthula, Jennifer M. Elward, Lori R. Hilden, Sten O. Nilsson Lill, Per-Ola Norrby, Antonio Ramirez, Edward C. Sherer, Jason Mustakis, and Gerald J. Tanoury. Applications of quantum chemistry in pharmaceutical process development: Current state and opportunities. *Organic Process Research & Development*, 24(8):1496–1507, 2020.

[Lowe, 2012] Daniel Mark Lowe. Extraction of chemical structures and reactions from the literature. *Apollo - University of Cambridge Repository*, 2012.

[Lu and Zhang, 2022] Jieyu Lu and Yingkai Zhang. Unified deep learning model for multitask reaction predictions with explanation. *J. Chem. Inf. Model.*, 62(6):1376–1387, 2022.

[Marshall and Coifman, 2019] Nicholas F Marshall and Ronald R Coifman. Manifold learning with bi-stochastic kernels. *IMA Journal of Applied Mathematics*, 84(3):455–482, 2019.

[Mena *et al.*, 2018] Gonzalo E. Mena, David Belanger, Scott W. Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[Mialon *et al.*, 2021] Grégoire Mialon, Dexiong Chen, Alexandre d'Aspremont, and Julien Mairal. A trainable optimal transport embedding for feature aggregation and its relationship to attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Milanfar, 2013] Peyman Milanfar. Symmetrizing smoothing filters. *SIAM J. Imaging Sci.*, 6(1):263–284, 2013.

[Niculae *et al.*, 2018] Vlad Niculae, André F. T. Martins, Mathieu Blondel, and Claire Cardie. Sparsemap: Differentiable sparse structured inference. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July*

10-15, 2018, volume 80 of *Proceedings of Machine Learning Research*, pages 3796–3805. PMLR, 2018.

[Peyré and Cuturi, 2019] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[Qian *et al.*, 2020] Wesley Wei Qian, Nathan T. Russell, Claire L. W. Simons, Yunan Luo, Martin D. Burke, and Jian Peng. Integrating deep neural networks and symbolic inference for organic reactivity prediction. *ChemRxiv*, 2020.

[Raucci *et al.*, 2022] Umberto Raucci, Valerio Rizzi, and Michele Parrinello. Discover, sample, and refine: Exploring chemistry with enhanced sampling techniques. *The Journal of Physical Chemistry Letters*, 13:1424–1430, 02 2022.

[Sacha *et al.*, 2020] Mikolaj Sacha, Mikolaj Blaz, Piotr Byrski, Pawel Wlodarczyk-Pruszynski, and Stanislaw Jastrzebski. Molecule edit graph attention network: Modeling chemical reactions as sequences of graph edits. *CoRR*, abs/2006.15426, 2020.

[Sander *et al.*, 2022] Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 3515–3530. PMLR, 2022.

[Schwaller *et al.*, 2019] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9), 2019.

[Segler and Waller, 2016a] Marwin H. S. Segler and Mark P. Waller. Modelling chemical reasoning to predict reactions. *Chemistry–A European Journal*, 23(25):6118–6128, 2016a.

[Segler and Waller, 2017] Marwin H. S. Segler and Mark P. Waller. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry – A European Journal*, 23(25):5966–5971, 2017.

[Sinkhorn, 1966] Richard Sinkhorn. A relationship between arbitrary positive matrices and stochastic matrices. *Canadian Journal of Mathematics*, 18:303–306, 1966.

[Sohn *et al.*, 2015] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3483–3491, 2015.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014.

[Tay *et al.*, 2020] Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. Sparse sinkhorn attention. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9438–9447. PMLR, 2020.

[Tetko *et al.*, 2020] Igor Tetko, Pavel Karpov, Ruud Deursen, and Guillaume Godin. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature Communications*, 11, 11 2020.

[Tu and Coley, 2021] Zhengkai Tu and Connor W. Coley. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *CoRR*, abs/2110.09681, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[Weininger, 1988] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, 1988.

[Wormell and Reich, 2021] Caroline L. Wormell and Sebastian Reich. Spectral convergence of diffusion maps: Improved error bounds and an alternative normalization. *SIAM J. Numer. Anal.*, 59(3):1687–1734, 2021.

[Zhao *et al.*, 2022] Ming Zhao, Lei Fang, Li Tan, Yves Lepage, and Jian-Guang Lou. Leveraging reaction-aware substructures for retrosynthesis and reaction prediction. *CoRR*, abs/2204.05919, 2022.