利用本地模板检索进行逆合成预测

Shufang Xie[1] , Rui Yan[1] *, Junliang Guo[2] , Yingce Xia[3] , Lijun Wu[3] , Tao Qin[3]

[1]中国人民大学高岭人工智能学院大数据管理与分析方法北京市重点实验室[2]
Microsoft Reserarch Aisa
[3]微软人工智能科学研究中心

{shufangxie,ruiyan}@ruc.edu.cn，{junliangguo,yingce.xia,lijuwu,taoqin}@microsoft.com

摘要

逆合成可以预测特定焦油分子的反应物，是药物发现的一项重要任务。近年来，基于机器学习的逆合成方法取得了可喜的成果。在这项工作中，我们介绍了一种局部反应模板检索方法 RetroKNN，以进一步提高基于模板的非参数检索系统的性能。我们首先建立了一个原子模板库和键模板库，其中包含了训练数据中的局部模板，然后在推理过程中使用 k-nearest-neighbor (KNN) 搜索从这些模板中检索。检索到的模板与神经网络预测相结合，作为最终输出。此外，我们还提出了一种轻量级适配器，用于在结合神经网络和 KNN 预测时根据隐藏表示和检索模板调整权重。我们在两个广泛使用的基准（USPTO-50K 和 USPTO-MIT）上进行了全面实验。特别是在前 1 位的准确率方面，我们在 USPTO-50K 数据集上提高了 7.1%，在 USPTO-MIT 数据集上提高了 12.0%。这些结果证明了我们方法的有效性。

1 导言

逆合成（预测给定产物分子的反应物）是药物研发的一项基本任务。传统方法严重依赖化学家的经验和启发式方法（Corey，1991 年）。最近，有人提出了基于机器学习的方法来辅助化学家，并取得了可喜的成果（Dong 等，2021 年）。典型的方法包括直接预测反应物的无模板方法和首先预测反应模板然后根据模板获得反应物的基于模板的方法。对于这些不同的方法，一个共同的研究挑战是有效地模拟这一任务的特殊属性。

如图 1 所示，化学再作用的一个关键特性是它与目标分子局部结构的改变密切相关，如取代一个功能基团或断开一个键。因此，最近的许多研究都集中于更好地模拟分子的局部结构（Chen 和 Jung，2021 年；Somnath 等，2021 年）。去

*通讯作者：Rui Yan (ruiyan@ruc.edu.cn).

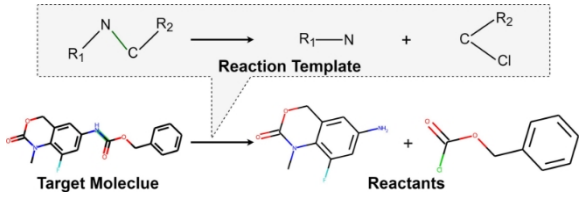


图 1：逆合成示意图，左侧为目标分子，右侧为两个反应物。呼号内是将碳-氮键分成两部分的反应模板。

尽管取得了令人鼓舞的成果，但我们注意到，仅靠神经网络工程来学习所有反应模式，尤其是罕见模板的反应模式，仍然十分困难。

因此，我们引入了一种基于非参数检索的方法，为预测提供具体指导。具体来说，我们使用一种局部模板检索方法，即 KNN（k- nearest-neighbor）方法，来提供额外的预判定，以提高预测精度。按照 LocalRetro（Chen 和 Jung，2021 年）的方法，我们首先利用训练有素的图神经网络（GNN）进行逆合成任务，然后离线构建包含反应模板的原子模板和键模板存储（第 2.1 节）。在存储构建阶段，我们遍历训练数据中的所有目标分子，并将每个原子和每个键的模板添加到相应的存储中。这些模板是由 GNN 提取的隐藏表示法生成的。在推理过程中，对于给定的新目标分子，我们首先使用原始 GNN 提取隐藏表示以及原始 GNN 预测模板。然后，我们使用隐藏表征搜索这两个存储空间，以检索与查询相似的本地模板。GNN 预测模板和 KNN 检索模板会以不同的权重合并，形成最终输出。

结合 GNN 和 KNN 预测是上述过程中的一个关键因素。传统的方法是使用固定参数来汇总所有反应的 预测值，这可能是次优的，会损害模型的泛化能力（Zheng 等，2021 年）。由于每个预测可能有不同的置信度，因此为每个反应自适应地分配权重将是有益的。

因此，我们采用轻量级适配器，根据 GNN 表示和检索结果来预测这些值（第 4.1 节）。因此，我们采用了一个轻量级适配器，根据 GNN 表示和检索结果来预测这些值。适配器网络结构简单，只需使用少量样本进行训练。虽然适配器会带来一些额外的成本，但它可以帮助有效提高模型的性能。

总之，我们有两方面的贡献：

- 我们提出了 RetroKNN 方法，这是一种通过非参数 KNN 方法的局部模板检索提高逆合成预测性能的新方法。
- 我们提出了一种轻量级元网络，用于在结合 GNN 和 KNN 预测时自适应地控制权重。

我们在两个广泛使用的基准上进行了实验：USPTO-50K 和 USPTO-MIT。这些数据集包含从美国专利商标局（USPTO）文献中提取的有机反应。在 USPTO- 50K 数据集上，我们将前 1 名的准确率从 53.4 分提高到 57.2 分（相对提高 7.1%），达到了新的先进水平。同时，在 USPTO-MIT 数据集上，我们将前 1 名的准确率从 54.1 分提高到 60.6 分（相对提高 12.0%）。此外，我们的方法在零镜头和少镜头数据集上也取得了可喜的成果，这些数据集对于传统的基于模板的方法来说是具有挑战性的，但对于这一研究领域却是必不可少的。这些结果证明了我们方法的有效性。

2 方法

2.1 序言

我们将分子表示为一个图 G（V，E），其中 V 是节点集，E 是键集。给定目标分子 M 作为输入，逆合成预测任务是生成作为 M 反应物的分子集 R。我们不直接预测 R，而是按照 LocalRetro（Chen 和 Jung，2021 年）的方法，预测反应中心 c 的局部反应模板 t 并将 (t, c) 应用于分子 M。更具体地说，根据 c 是原子还是键，t 可分为两种类型：原子模板 $t \in T_a$ 和键模板 $t \in T_b$。我们还假设有一个训练集 D_{train}、一个评估集 D_{val} 和一个测试集 D_{test}。每个数据分块都包含目标和相应的反应物，它们是为 $D = \{(M_i, t_i, c_i, R_i)\}_{i=1}^{|D|}$，其中 c_i 是 reaction M_i 应用模板 t_i 和 $|D|$ 是数据 D 的大小。

同时，我们假设在 D_{train} 存在。在不失一般性的前提下，我们将 GNN 分成两部分

算法 1：存储构建算法

输入：训练数据 D_{train}。
输入特征提取器 **f** .
输出：原子存储空间 S_A 和键存储空间 S_B。

```
让 S_A := ∅, S_B := ∅ ; // 初始化。
对于 (M、t、c、R) ∈ D_train 做
    让 V 表示 M 的节点集;
    让 E 表示 M 的边集;
    for v∈ V ; // 循环每个节点。
     做
        让 h_v := f (v|M );
        if v == c then
            让 S_A := S_A ∪ {(h_v , t)};
        否则
            让 S_A := S_A ∪ {(h_v , 0)};
        结束
    结束
    for e∈ E ; // 循环每条边。
     做
        让 h_e := f (e|M );
        if e == c then
            让 S_B := S_B ∪ {(h_e , t)};
        其他
            让 S_B := S_B ∪ {(h_e , 0)};
        结束
    结束
结束
返回 S_A , S_B
```

从 D_{train} 计算出的值对，其构建过程详见算法 1。

在该算法中，第一步是初始化原子存储空间 S_A 和键存储空间 S_B 为空集。接下来，对于训练数据 D_{train} 中的每个反应，我们分别在第 5 至 13 行和第 14 至 22 行遍历目标分子 M 的所有节点 $v \in V$ 和所有边 $e \in E$。对于每个节点 v，如果它是 $_{vA}$否则，我们会添加一个特殊标记 **0**，表示此处不使用模板。

同样，对于每条边 e，我们会添加 (h_e, t) 或 $(h_e, \mathbf{0})$ 到键存储空间 S_B。最后，我们得到原子存储空间 S_A 和债券存储 S_B。

特征提取器 **f** 将分子图 $G(V, E)$ 作为输入，并为每个节

点 v∈V 输出隐藏表示 h_v ，为每个边 e∈E 输出隐藏表示 $\mathbf{h}_e$ 。预测头 h 对 h_v 和 h_e 进行处理，以分别预测模板集 T_a 和 T_b 上的概率分布。

2.2 商店建设

我们的方法使用两个数据存储 S_A 和 S_B ，其中包含原子和键的信息。在推理之前，这两个存储空间都是离线结构化的。存储区内有关键

2.3 推理方法

推理过程概览见图 2。推理时，给定一个新的目标分子 M，我们首先计算隐藏表征 h_v 、h_e 和模板概率。能力 $P_{\mathrm{GNN}}(t_a \mid M, a)$，$P_{\mathrm{GNN}}(t_b \mid M, b)$，分别针对每个原子 a 和键 b，[1] 。接下来，我们将检索每个节点和边，可写成

[1]为简化符号，我们尽可能省略节点和边 ID 的下标。

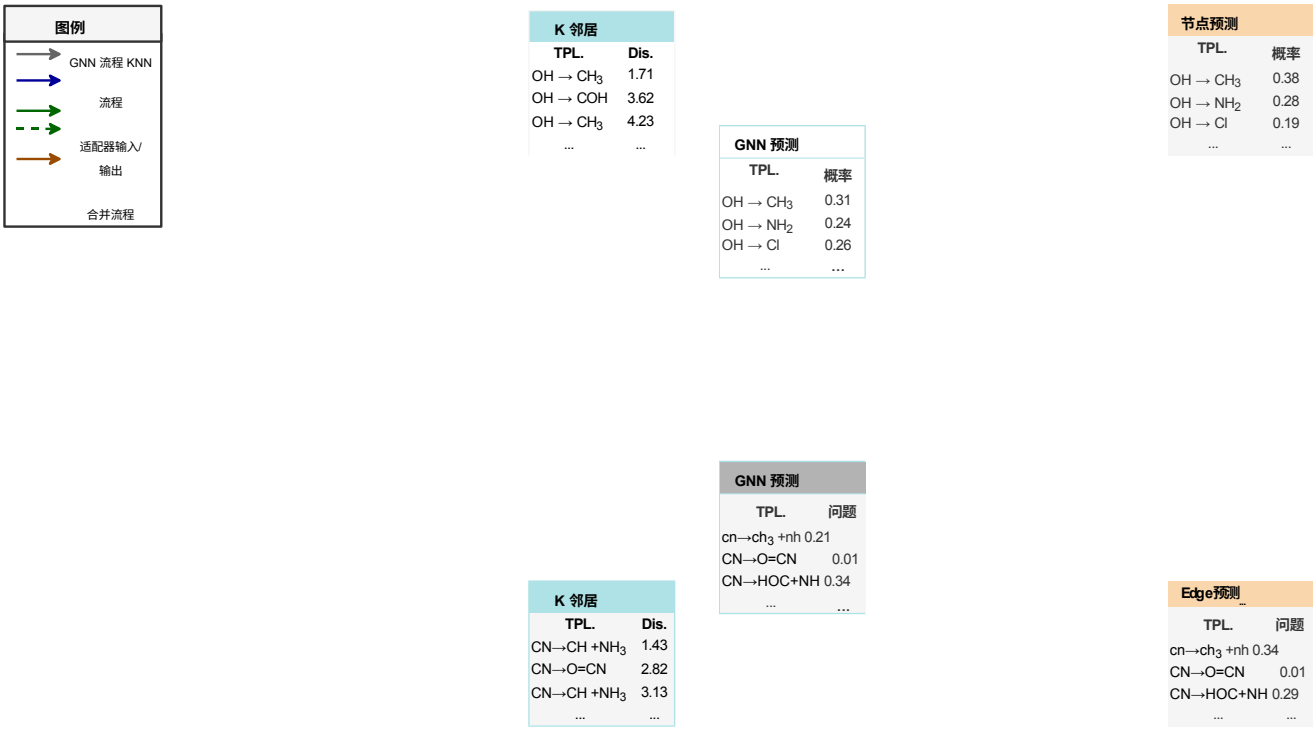


图 2：左中为 RetroKNN 对目标分子的说明。上半部分和下半部分分别显示了一个原子和键检索的示例。灰色、蓝色、绿色和棕色线条分别表示 GNN 预测、KNN 预测、适配器输入/输出和合并过程。粉色表格表示所有预测的最终输出结果。

$$P_{\mathrm{KNN}}(\mathrm{ta}|\mathrm{M}, a) \propto \sum_{(h_i, t_i)\in \mathrm{N}} \mathbb{I}_{a=t_i} \exp\left(\frac{-(\text{哈}, \text{嘀})}{\text{电讯}}\right), \quad (1)$$

$$P_{\mathrm{KNN}}(t_b|M, b) \propto \sum_{(h_i, t_i)\in \mathrm{N}_b} \mathbb{I}_{b=t_i} \exp\left(\frac{-d(h_b, h_i)}{T_B}\right). \quad (2)$$

在等式 (1, 2) 中，N_a , N_b 是由 S_A , S_B 重新划分的候选集，I 是指标函数，只有当条件（即 $t_a = t_i$ 或 $t_b = t_i$ ）为 1 时才输出 1。

满足，而 T_A , T_B 是软最大温度。同时，d(-, -)是距离函数，用于测量"-"和"-"之间的距离。

h_i 与 h_v 或 h_e 之间的相似度。换句话说，$P_{\mathrm{KNN}}(t_a|M, a)$ 与模板为 t_a 的邻域权重之和成正比。

最后，我们合并 GNN 输出和 KNN 输出插值因子 λ，即

$$P(\mathrm{ta}|\mathrm{M}, a) = \lambda_a P_{\mathrm{GNN}}(\mathrm{ta}|\mathrm{M}, a) + (1-\lambda_a) P_{\mathrm{KNN}}(\mathrm{ta}|\mathrm{M}, a), \quad (3)$$

$$P(t_b|M, b) = \lambda_b P_{\mathrm{GNN}}(t_b|\mathrm{M}, \mathrm{b}) + (1-\lambda_b) P_{\mathrm{KNN}}(t_b|M, b). \quad (4)$$

在公式 (1)-(4) 中，温度 T_A , $T_B \in \mathrm{R}^+$ 和插值因子 λ , $\lambda \in [0, 1]$ 是通过

详情见第 2.4 节。

2.4 适配器网络

为了自适应地为每个原子和化学键选择 T_A 、 T_B 、 λ_a 和 λ_b ，我们设计了一个轻量级网络来预测这些结果。值。适配器的输入是隐藏表示 h_v , h_e 来自 GNN 侧，距离列表 $d(h_v, h_i)$, $d(h_e, h_i)$ 来自 KNN 方面。

我们的网络结构使用单层 GNN，然后是几个完全连接（FC）层。我们使用带边缘特征的图同构网络（GIN）（Hu 等人，2019 年）层来捕捉节点特征 h_v 和边缘特征 h_e ，其公式为：

$$h_v^{(g)} = W_{\mathrm{vg}}\left((1+\epsilon)h_v + \sum_{e\in E(v)} \mathrm{ReLU}(h_v + h_e)\right) + b_{\mathrm{vg}}, \quad (5)$$

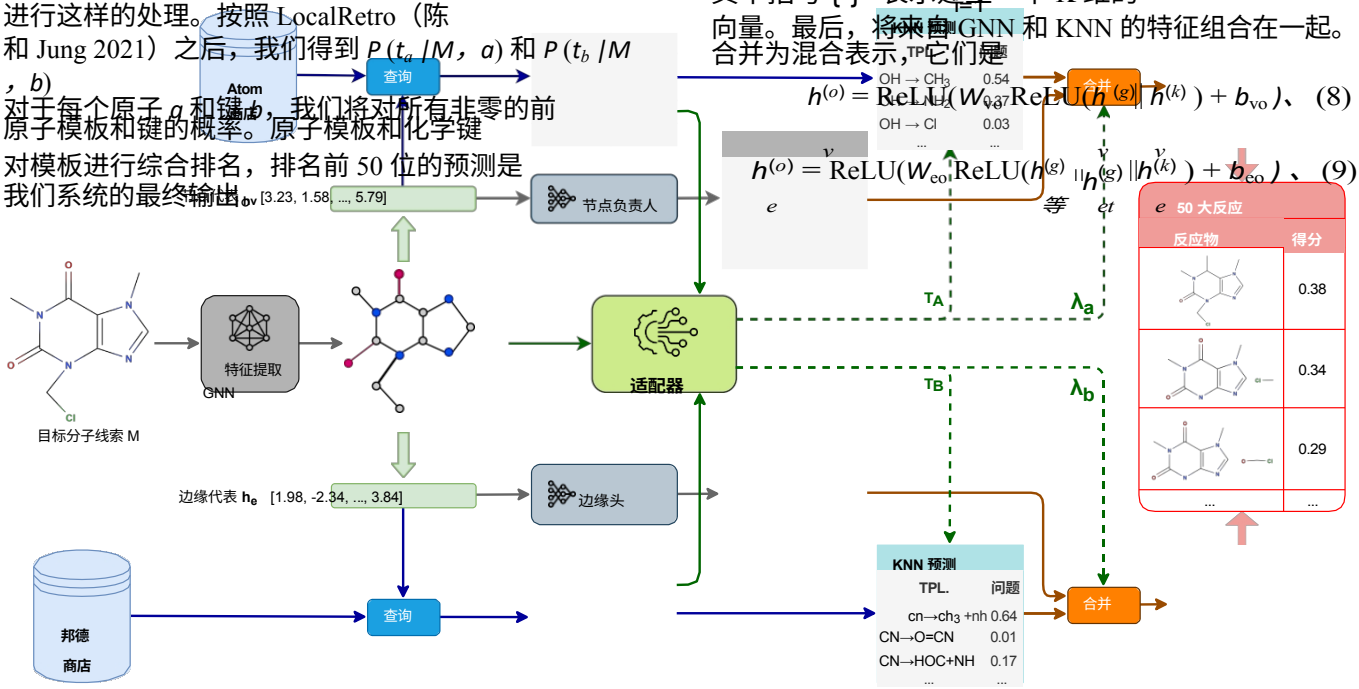
其中，$h^{(g)}$ 是输出，ϵ 和 W 是 GIN 的可学习参数，$E(v)$ 是 v 周围的边缘集。同时，我们使用 FC 层将 KNN 距离投射到

提取的特征可表述为

$$h_v^{(k)} = W_{\mathrm{vk}}(\{d(h_v, h_i)\}_{i=1}^{K}) + b_{\mathrm{vk}}, \quad (6)$$

$$h_e^{(k)} = W_{\mathrm{ek}}(\{d(h_e, h_i)\}_{i=1}^{K}) + b_{\mathrm{ek}}, \quad (7)$$

在图 2 中，我们仅以一个节点和一个键的重估为例进行说明，但在实际操作中，我们会对所有原子和键进行这样的处理。按照 LocalRetro（陈和 Jung 2021）之后，我们得到 $P(t_a|M, a)$ 和 $P(t_b|M, b)$ 对于每个原子 a 和键 b，我们将对所有非零的前原子模板和键的概率。原子模板和化学键对模板进行综合排名，排名前 50 位的预测是我们系统的最终输出。



$$e \qquad i=1$$

其中括号 $\{\cdot\}^K$ 表示建立一个 K 维的向量。最后，将来自 GNN 和 KNN 的特征组合在一起。合并为混合表示，它们是

$$h^{(o)} = \mathrm{ReLU}(W_{\mathrm{vo}}\mathrm{ReLU}(h^{(g)} \| h^{(k)}) + b_{\mathrm{vo}}), \quad (8)$$

$$h^{(o)} = \mathrm{ReLU}(W_{\mathrm{eo}}\mathrm{ReLU}(h^{(g)} \| h^{(k)}) + b_{\mathrm{eo}}), \quad (9)$$

$$v \qquad v \qquad v$$

$$e \qquad \text{等} \qquad et$$

其中‖表示张量连接，*es* 和 *et* 表示边 *e* 的起点和终点节点。

T_A 、λ_a 由 $h^{(o)}$ 预测，T_B、λ_b 由 h 预测。$^{(o)}$。我们还使用了 sigmoid 函数。^{e} (σ)和 以保证 λ_a， $\in$（0，1），并夹住 T_A ，T_B (σ) λ_b 范围 [1, 100]。形式上，我们有

$$T_A = \max(1, \min(100, W_{ta} h^{(o)} + b_{ta}))、 \quad (10)$$

$$\lambda_a = \sigma(W_{la} h^{(o)} + b_{la})、 \quad (11)$$

$$T_B = \max(1, \min(100, W_{tb} h^{(o)} + b_{tb}, 1, 100))、 \quad (12)$$

$$\lambda_b = \sigma(W_{lb} h^{(o)} + b_{lb}). \quad (13)$$

因为这里使用的所有公式都是可微分的，所以我们用梯度 decent 来优化适配器参数 W，以最小化模板分类损失

$$L_M = -\frac{1}{|V|}\sum_{a\in V} \log P(\hat{t}_a|M, a) - \frac{1}{|E|}\sum_{b\in E} \log P(t_b|M, b)、 \quad (14)$$

$P(\hat{t}_a|M), P(\hat{t}_b|M)$ 由公式 (3) 和公式 (4) 计算。$\hat{t}_a$，$\hat{t}_b$ 是地面实况模板。

3 实验

3.1 实验设置

数据。我们的实验基于从美国专利商标局（USPTO）文献中提取的化学反应。我们使用了两个版本的 USPTO 基准：USPTO-50K（Coley 等人，2017 年）和 USPTO-MIT（Jin 等人，2017 年）。USPTO-50K 包含 50k 个化学反应，分为 40k/5k/5k 个反应，分别作为训练、验证和测试。同时，USPTO-MIT 包含约 479k 个反应，分区为 409k/40k/30k。所有分区均与上一版本相同。

许多作品（Coley 等人，2017 年；Jin 等人，2017 年）使公平比较。我们还使用 Chen 和 Jung（2021 年）的预处理脚本从这些反应中提取反应模板，从而在 USPTO-50K 和 USPTO-MIT 中分别获得 658 和 20,221 个反应模板到 32。对于适配器网络，我们使用了与骨干 GNN 相同的隐藏层。适配器也使用学习率为 0.001 的 Adam 优化器进行训练。考虑到数据大小的差异，我们对适配器进行了 10 次训练。

在美国专利商标局 50K 数据集和美国专利商标局麻省理工学院数据集的验证集上分别进行了两个历元和两个历元的测试。适配器

在测试中使用了验证损失最大的

评估和基线 按照之前的工作，我们的系统将预测每个目标分子的前 50 个结果，并根据 Chen 和 Jung（2021 年）的脚本报告 K=1、3、5、10 和 50 的前 K 准确率。我们还使用了近年来具有代表性的基线系统，包括

- 基于模板的方法：retrosim（Coley 等人，2017 年）、neuralsym（Segler 和 Waller，2017 年）、GLN（Dai 等人，2020 年）、Hopfield（Seidl 等人，2021 年）和 LocalRetro（Chen 和 Jung，2021 年）；
- 基于半模板的方法：G2Gs（Shi 等人，2021 年）、RetroXpert (Yan et al. 2020) 和 GraphRtro (Somnath et al. 2020)。
- 无温度的方法：Transformer (Lin et al. 2020)、MEGAN (Sacha et al. 2021)、Chemformer (Irwin et al. 2020)、Transformer (Lin et al. 2020)。

2021）、GTA（Seo 等人，2021）和 DualTF（Sun 等人，2021）。

3.2 主要成果

当反应类型未知时，USPTO-50K 基准的实验结果如表 1 所示；当反应类型已知时，实验结果如表 2 所 示。 同时，USPTO-MIT 基准的结果见表 3。在这些表格中，我们按准确率最高的 1 个系统对所有系统进行了排序，并通过填写循环符号标记了它们的类型。我们的方法（RetroKNN）位于最后一行，并用粗体标出。

比较这些准确率数字，我们可以发现，我们的方法在很大程度上优于基线系统。当反应类型未知时，我们取得了 57.2 分的最高-1 准确率，并将 LocalRetro 的骨干结果提高了 3.8 分，相对提高了 7.1%。

方法	TPL.	K = 1	3	5	10	50
回顾	●	37.3	54.7	63.3	74.1	85.3
神经元	●	44.4	65.3	72.4	78.9	83.1
梅根	○	48.1	70.7	78.4	86.1	93.2

。

实施细节。 我们沿用相同的模式，即
在此基础上，我们可以将其视为 LocalRetro（Chen 和 Jung，2021 年），以建立
骨干 GNN 模型。特征提取器 **f** 是一个 6-

层 MPNN（Gilmer 等人，2017 年），然后是单个 GRA
层（Chen 和 Jung，2021 年），共 8 个头。我们使用隐藏
维数为 320，滤除率为 0.2。原子和键的内
该特征由 DGL-LifeSci（Li 等人，2021 年）提取。
预测头 h 由两个密集层组成，每个密集层都有 ReLU 数据
训练器的学习率为 0.001，共训练 50 次。此外，当 5
激活。骨干网模型由亚当-斯蒂文斯（Adam Opti-Steven）
个历元的验证损失没有改善时，我们会提前停止训练
优化。

。背骨的配置与 Chen 和 Jung（2021 年）相同。

KNN 的实现基于 faiss（John-son、Douze 和 Je´gou
2019）库和用于快速嵌入搜索的 IndexIVFPQ in-dex
，KNN 的 K 设置为

G2Gs	⬭❨	48.9	67.6	72.5	75.5	-
RetroXpert	⬭❨	50.4	61.1	62.3	63.4	64.0
一般临时人员	○	51.1	67.6	67.8	81.6	-
霍普菲尔德	●	51.8	74.6	81.2	88.1	94.0
GLN	●	52.5	69,0	75.6	83.7	92.4
本地零售	●	53.4	77.5	85.9	92.4	97.7
双 TF	○	53.6	70.7	74.6	77.0	-
GraphRetro	⬭❨	53.7	68.3	72.2	75.5	-
RetroKNN	●	**57.2**	**78.9**	**86.4**	**92.7**	**98.1**
Chemformer	○	54.3	-	62.3	63.0	-

表 1：当反应类型未知时，USPTO-50K 数据集的 Top-K 精确匹配准确率。●、⬭ （和
表示基于模板、半模板和无模板、
分别为系统按最高精度排序。

方法	TPL.	K = 1	3	5	10	50
回顾	●	52.9	73.8	81.2	88.1	-
神经元	●	55.3	76.0	81.4	85.1	-
梅根	○	60.7	82.0	87.5	91.6	95.3
G2Gs	⬭❨	61.0	81.3	86.0	88.7	-
RetroXpert	⬭❨	62.1	75.8	78.5	80.9	-
GraphRetro	⬭❨	63.9	81.5	85.2	88.1	-
本地复古	●	63.9	86.8	92.4	96.3	97.9
GLN	●	64.2	79.1	85.2	90.0	93.2
双 TF	○	65.7	81.9	84.7	85.9	-
RetroKNN	●	**66.7**	**88.2**	**93.6**	**96.6**	**98.4**

表 2：当给出反应类型时，USPTO-50K 数据集的 Top-K 精确匹配准确率。表中的●、⬭❨和◯de 分别表示基于模板、半模板和无模板，再加上
非常明显。系统按最高精度-1 排序。

方法	TPL.	K = 1	3	5	10	50
Seq2Seq	○	46.9	61.6	66.3	70.8	-
神经元	●	47.8	67.9	74.1	80.2	-
变压器	○	54.1	71.8	76.9	81.8	-
本地零售	●	54.1	73.7	79.4	84.4	90.4
RetroKNN	●	**60.6**	**77.1**	**82.3**	**87.3**	**92.9**

表 3：USPTO-MIT 数据集的 Top-K 精确匹配准确率。●和◯表示基于模板的方法和无模板方法。系统按 Top-1 准确率排序。

当给出反应类型时，我们也将前 1 名的准确率提高了 2.8 个百分点，从 63.9 提高到 66.7。同时，在美国专利商标局-麻省理工学院（USPTO-MIT）上，我们的方法显示出 60.6 分的最高-1 准确率，提高了 6.5 分或 12% 的相对收益。更重要的是，这些 Top-1 准确率也优于其他强大的基线和最先进的方法，证明了我们方法的有效性。

同时，在美国专利商标局-50K 中，当再行为类型未知时，我们的前 3 名准确率为 78.9 分，准确率为 86.4 分，也远高于基线。至于前 10 名和前 50 名的准确率，我们的准确率分别为 92.7 分和 98.1 分。考虑到准确率已经非常高，我们的改进仍然非常显著。

总之，局部模板重试法有效地提高了逆合成预测的准确性。

4 研究与分析

4.1 案例研究

检索案例研究。为了更好地了解我们能否通过隐藏表征检索到有用的反应，我们对 USPTO-50K 数据集进行了案例研究，结果如图 3 所示。我们首先从数据中选择一个原子-模板再反应和第一个键-模板反应。接着，我们通过相应的原子和键查询原子和键存储。最后，对于每个检索到的模板，我们都会显示训练数据中的原始目标分子，其中反应原子/键以绿色背景高亮显示。

图中第一行和第二行分别展示了键模板反应和原子模板反应。在每一行中，我们首先显示了反应的目标分子 M，然后显示了 M 的五个邻域。从这些案例中，我们可以发现隐藏表征所检索到的邻域可以有效地捕捉分子的局部结构。例如，在边缘模板反应中，碳-氮键检索到了所有邻域。此外，所有碳原子都被双键中的氧（=O）和三氟化碳（-CF3）包围，所有氮原子都与芳香环相连。同时，在节点模板反应中，所有检索到的原子都是与苯基相连的氧原子。总之，利用隐藏表征检索分子是有效的，因为它能很好地捕捉局部结构。因此，我们可以利用检索到的模板来提高预测精度。

适配器案例研究。我们在表 4 中展示了三个具有代表性的适配器效果案例。在每一行中，我们显示了目标分子和地面实况模板 id，然后是适配器输出的 λ 和 T，最后是 GNN 预测值和 KNN 检索到的邻居。当第一行的 GNN 预测准确时，适配器会生成较高的λ 值（如 0.96），从而使 GNN 输出具有较高权重。然而，当情况并非如此时（第二行和第三行），λ *值*往往较低（如 0.14），这就赋予了 KNN 预测更高的权重。同时，当只有 N1 有正确预测时（第二行），适配器倾向于输出较小的 T（如 7.89），以进行尖锐的分配，给予 N1 的预测更多权重。相反（第三行），适配器倾向于输出一个较大的值（如 19.36），以便让更多的邻居为最终输出做出贡献。此外，我们的统计数据显示，当 $\lambda < 0.5$ 时，GNN 和 KNN 的准确率分别为 46.9% 和 69.2%，这表明 KNN 与 GNN 预测是互补的。

4.2 零镜头和少镜头研究

我们将 USPTO-50K 数据集修改为零镜头和少镜头版本，以研究我们方法的领域适应能力。具体来说，在 USPTO-50K 数据中，每个反应的反应类别都在 1 到 10 类中。为了建立零次搜索数据，我们对训练数据和验证数据进行过滤，将反应类别为 6 至 10 的所有反应重新移动，只保留反应类别为 1 至 5 的反应。同样，为了建立少量反应数据，我们只保留 10%的反应类别为 6 至 10 的反应。最后，我们用 LocalRetro 基线和 RetroKNN 方法评估了这些新数据的性能。结果汇总如图 4 所示。

从这些图中，我们注意到，对于传统的基于模板的方法来说，"零镜头 "是一个具有挑战性的设置，这也是这类方法的一个众所周知的缺点。然而，当与 KNN 结合时，我们的系统就能产生有意义的 结果。例如，在第 8 类反应中，RetroKNN 在零镜头数据中的前 5 名准确率为 6.1 分，前 10 名准确率为 9.8 分。由于在训练过程中可以获得少量示例，因此少发设置比零发设置更容易。尽管如此，RetroKNN 在所有反应类型上的表现都优于基线。平均而言，RetroKNN 将前 5 名的准确率提高了 8.56 分，将前 10 名的准确率提高了 5.64 分。这些结果表明，我们的方法也可以

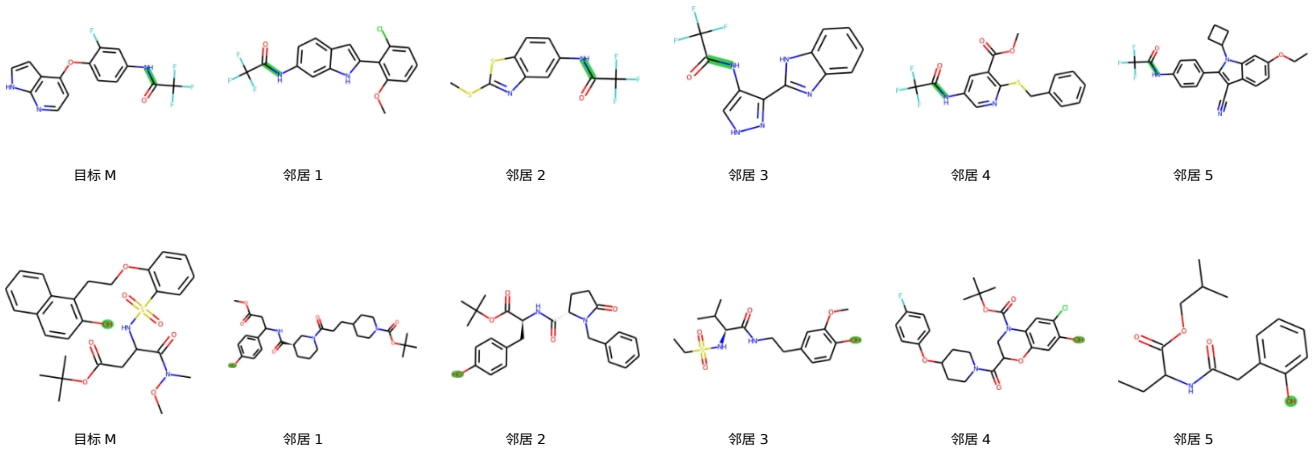
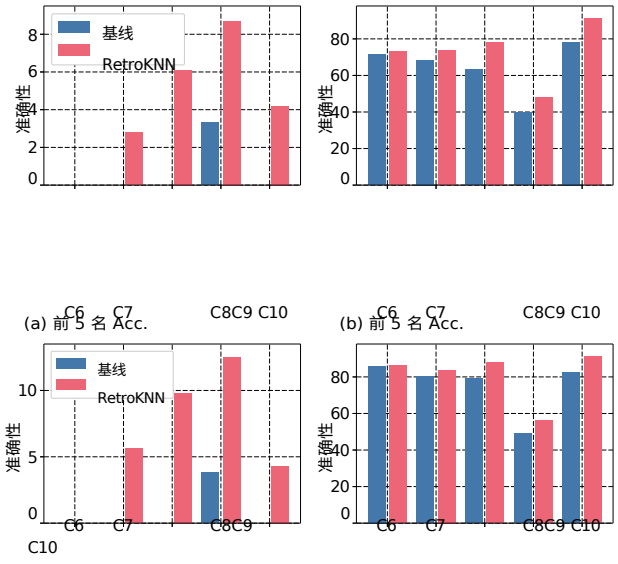


图 3：检索分子的案例研究。绿色背景突出显示了检索中使用的化学键和原子。第一列为目标分子，其余为训练数据中的五个邻近目标。

目标分子	GT。	λ	T	GNN	N1	N2	N3	N4	N5
Cc1ccc(-c2cccnc2C#N)cc1	**b542**	0.96	21.42	**b542**	b519 (67.51)	b519 (77.35)	b519 (77.35)	b519 (77.35)	b0 (104.00)
CCOc1ccc(C[C@H](NC(=O)C(F)(F)F)C(=O)O)cc1	**b524**	0.14	7.89	b495	**b524** (22.79)	b523 (33.84)	b495 (67.3)	b495 (76.21)	b495 (76.55)
CC1(C)CC(=O)N(Cc2ccccc2)c2ccc(C#Cc3ccc(C(=O)O)cc3)cc21	**a121**	0.02	19.36	a124	**a121** (34.41)	**a121** (57.3)	**a121** (58.4)	a0 (59.91)	a0 (61.17)

表 4：参数 T 和 λ *的*案例研究。GT.表示地面实况模板 id，GNN 表示 GNN 预测值，N1 至 N5 表示五个邻居。模板 id 的前缀 a、b 表示原子或键模板。我们在模板 id 下方的括号中显示了每个邻居的距离。正确的预测以粗体标出。



提高零/少镜头数据的性能，这是该领域的重要应用场景。

4.3 消融研究

我们对美国专利商标局 50K 数据集进行了消融研究，以研究不同组件的贡献，结果如表 5 所示。我们显示了
通过对不同系统进行比较，得出表格。系统(1)是
不使用 KNN 的 LocalRetro 基线达到了
准确率为 53.4 分。在系统(2)中，我们在不使用适配器的情况下添加了 KNN。为了找到最优参数，我们在 T∈{1, 5, 25, 50} 和 λ∈{0.1, 0.3, 0.5, 0.7, 0.9} 的条件下进行了全面的网格搜索，共搜索到 20 个网格，这些网格中的每一个都是最优的。
组合。我们通过验证损失来选择参数

最后得到 56.3 分的精　　确度。此外，在系统

(c) 零发数据的前 10 名加速度。　　(d) 少量射击数据的前 10 次加速。

图 4：零镜头（a、c）和少镜头（b、d）数据的前 5 名（a、b）和前 10 名（c、d）准确率（Acc.）C6 至 C10 列表示不同的反应类别。

tem ③，我们只为 T 添加适配器，并保持 λ 不变系统 ②。同样，我们只为 λ 添加适配器。系统④中。系统 ⑤是完整的 RetroKNN 模型。

将①系统与其他使用 KNN 的系统进行比较，我们可以发现，在这项任务中引入 KNN 可以有效地提高模型性能。这些数字表明

身份证	系统	准确性
①	基线	53.4
②	+ KNN	56.3
③	+ KNN，自适应 T	56.7
④	+ KNN，自适应 λ	56.8
⑤	+ KNN、自适应 T、自适应 λ	57.2

表 5：当反应类型未知时，对 USPTO-50K 数据集进行的消融研究。

#检索反应	1	4	8	16	32
准确性	55.6	57.4	57.1	56.9	57.2

表 6：对 KNN 检索反应数的研究。平均值

而将③④系统与②系统相比，我们注意到

加入 T 和 λ 适配器会有所帮助。最后，当这两个参数在系统⑤中进行自适应预测时，准确率可以提高到 57.2，这表明它们可以在系统⑤中起作用。

有效地结合在一起。因此，该系统需要所有组件。

4.4 检索到的模板尺寸

在表 6 中，我们展示了检索到的重新动作数量（即 KNN 的 K）对模型性能的影响。更具体地说，在 KNN 搜索中，我们设置了 K∈ [1、4、8、16、32]，然后分别训练适配器。最后我们在表中报告了前 1 位的准确率。

从这些结果中，我们首先发现，只增加一个检索模板（K=1）就能将准确率从 53.4 至 55.6。当 K≥4 时，精确度可进一步提高到 57 点左右。不会再有

当检索到更多反应时，预测结果不会有明显改善，收到更多模板也不会影响预测结果。我们认为这是因为已经有足够的信息来提高准确率，因为远离查询的模板对预测的贡献较小。

4.5 推理延迟

数据集	美国专利商标局-50K	美国专利商标局-麻省理工学院
|Dtrain|	40k	409k
$\|S\|_A$	1,039k	10,012k
$\|S\|_B$	2,241k	21,495k
无 KNN 时的延迟	2.71 ± 0.02 毫秒	3.51 ± 0.05 毫秒
延迟与 KNN	3.31 ± 0.09 毫秒	14.69 ± 0.29 毫秒

表 7：数据存储大小和推理延迟研究。

我们希望在今后的工作中加入这些技术。

5 相关工作

5.1 逆合成预测

逆合成预测是科学发现的一项重要任务，近年来取得了可喜的成果
年（Segler 和 Waller，2017 年；Liu 等，2017 年；Coley 等，2017 年）。

在表 7 中，我们研究了数据存储大小和推理延迟。最后两行显示的是推理过程中进行或不进行检索时的延迟时间，这些延迟时间是在一台配有英伟达 A100 GPU 的机器上测量的。每个延迟值（即每个反应的平均运行时间）都是通过十次独立运行测得的。在 USPTO-50K 数据集中，我们发现平均延迟从 2.71 毫秒增加到了 3.31 毫秒，每个反应约 0.6 毫秒。由于 USPTO-MIT 数据集比 USPTO- 50K 数据集大十倍左右，因此前者的延迟更为明显。不过，考虑到更精确的系统可以为化学家节省数小时甚至数天的时间，额外的 10 毫秒成本并不会真正阻碍这种方法的实际应用。最后，一些研究（He、Neubig 和 Berg- Kirkpatrick，2021 年；Meng 等，2021 年）表明，KNN

2017；Tetko 等人，2020；Irwin 等人，2021；Dai 等人，2020；Yan 等人，2020；Seidl 等人，2021；Chen 和 Jung，2021；Shi 等人，2021；Somnath 等人，2021；Wan 等人，2022）。一些研究还将检索机制用于这项任务。例如，Seidl 等人（2021 年）使用 Hopfield 网络选择模板，Lee 等人（2021 年）使用检索方法从数据库中获取分子。不同的是，我们是第一个在这项任务中结合深度学习和 KNN 检索的人。

5.2 检索方法

从数据存储或内存中检索以提高机器学习模型的性能是一个重要的重新搜索课题。SVM-KNN （Zhang 等人，2006 年）首次将 SVM 和 KNN 结合起来用于识别任务。此外，KNN-LM（Khandelwal 等人，2020 年）和 KNN-MT（Khan- delwal 等人，2021 年）在将 KNN 与 Transformer 网络相结合时也显示出了良好的效果。同时，He、Neubig 和 Berg-Kirkpatrick (2021)；Meng 等人 (2021) 研究了重取方法的速度，Zheng 等人 (2021) 研究了适应问题。然而，我们是第一个将 KNN 的强大功能与 GNN 结合起来，并将它们用于反合成任务的人。

6 结论

逆合成预测对于科学发现，尤其是药物发现和医疗保健至关重要。在这项工作中，我们提出了一种利用局部模板检索提高预测准确性的新方法。我们首先用训练数据和训练过的 GNN 建立原子和键存储，然后在推理过程中从这些存储中检索模板。检索到的模板与原始 GNN 预测相结合，形成最终输出。我们进一步利用轻量级适配器来自适应预测权重，以整合 GNN 预测和检索模板。我们大大提高了在两个广泛使用的基准（USPTO-50K 和 USPTO-MIT）上的预测性能，最高准确率分别达到 57.2 分和 60.6 分。这些结果证明了我们方法的有效性。

致谢

感谢匿名审稿人提出的宝贵意见。本研究得到了国家自然科学基金的资助（NSFC 批准号：62122089 和 61876196）、

北京市杰出青年科学基金项目，编号：BJWZYJH012019100020098。感谢中国人民大学 "双一流 "建设重大创新规划跨学科平台智能社会治理平台（BJJWZYJH012019100020098）的支持。我们还要感谢中国人民大学公共政策与决策研究室提供的支持和做出的贡献。阎锐由北京人工智能学会（BAAI）资助。

参考资料

Chen, S.; and Jung, Y. 2021.利用局部反应性和全局注意力进行深度逆合成反应预测。*JACS Au*.

Coley, C. W.; Rogers, L.; Green, W. H.; and Jensen, K. F. 2017.基于分子相似性的计算机辅助逆合成。*ACS Central Science*, 3(12)：1237-1245.

Corey, E. J. 1991.*化学合成的逻辑》*。Ripol Classic.

Dai, H.; Li, C.; Coley, C. W.; Dai, B.; and Song, L. 2020.用条件图逻辑网络进行逆合成预测》，*arXiv:2001.01408*.

Dong, J.; Zhao, M.; Liu, Y.; Su, Y.; and Zeng, X. 2021.深度学习在逆合成规划中的应用：数据集、模型和工具。*生物信息学简报*，00（8 月）：1-15.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017.量子化学的神经信息传递。*国际机器学习会议*，1263-1272。PMLR.

He, J.; Neubig, G.; and Berg-Kirkpatrick, T. 2021.高效近邻语言模型。*EMNLP*.

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; and Leskovec, J. 2019.预训练图神经网络的策略》。ArXiv:1905.12265.

Irwin, R.; Dimitriadis, S.; He, J.; and Bjerrum, E. J. 2021.Chemformer：用于计算化学的预训练变换器。*机器学习：Science and Technology*, 3.

Jin, W.; Coley, C.; Barzilay, R.; and Jaakkola, T. 2017.用 Weisfeiler-lehman 网络预判有机反应结果。*神经信息处理系统进展》*，第 30 期。

Johnson, J.; Douze, M.; and Je´gou, H. 2019.使用 GPU 进行亿级规模的相似性搜索。*IEEE Transactions on Big Data*, 7(3)：535-547.

Khandelwal, U.; Fan, A.; Jurafsky, D.; Zettlemoyer, L.; and Lewis, M. 2021.最近邻机器翻译。*国际学习表征会议*。

Khandelwal, U.; Levy, O.; Jurafsky, D.; Zettlemoyer, L.; and Lewis, M. 2020.通过记忆实现泛化：最近邻语言模型。In *International Con- ference on Learning Representations (ICLR)*.

Lee, H.; Ahn, S.; Seo, S.-W.; Song, Y. Y.; Yang, E.; Hwang、
S.J.; and Shin, J. 2021.RetCL: A Selection-based Approach for Retrosynthesis via Contrastive Learning.In *IJCAI*, 2673-2679.

Li, M.; Zhou, J.; Hu, J.; Fan, W.; Zhang, Y.; Gu, Y.; and Karypis, G. 2021.DGL-LifeSci: An Open-Source Toolkit for Deep Learning on Graphs in Life Science.*ACS Omega*.

Lin, K.; Pei, J.; Lai, L.; and Xu, Y. 2020.使用无模板模型的自动再合成途径规划。*化学科学*。

Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; and Pande、

V.2017.利用神经序列到序列模型进行逆合成反应预测。*ACS Central Science*, 3(10)：1103-1113.

Meng, Y.; Li, X.; Zheng, X.; Wu, F.; Sun, X.; Zhang, T.; and Li, J. 2021. 快速近邻机器翻译》，arXiv:2105.14528.

Sacha, M.; Błaz, M.; Byrski, P.; Dabrowski-Tumanski, P.; Chrominski, M.; Loska, R.; Włodarczyk-Pruszynski, P.; and Jastrzebski, S. 2021.分子编辑图注意网：将化学反应建模为图编辑序列。*化学信息与建模期刊*》，61（7）：3273-3284.

Segler, M. H. S.; and Waller, M. P. 2017.用于逆合成和反应预测的神经符号机器学习。*Chemistry - A European Journal*, 23(25)：5966-5971.

Seidl, P.; Renz, P.; Dyubankova, N.; Neves, P.; Verhoeven, J.; Segler, M.; Wegner, J. K.; Hochreiter, S.; and Klambauer, G. 2021.用于少量和零次反应模板预测的现代 Hopfield 网络。*arXiv:2104.03279*.

Seo, S.-W.; Young Song, Y.; Yong Yang, J.; Bae, S.; Lee, H.; Shin, J.; Ju Hwang, S.; and Yang, E. 2021.GTA: Graph Truncated Attention for Retrosynthesis.*AAAI*.

Shi, C.; Xu, M.; Guo, H.; Zhang, M.; and Tang, J. 2021. 用于逆合成预测的图到图框架. *arXiv:2003.12725*.

Somnath, V. R.; Bunne, C.; Coley, C. W.; Krause, A.; and Barzilay, R. 2021.Learning Graph Models for Retrosynthe- sis Prediction. *ArXiv:2006.07038*.

Sun, R.; Dai, H.; Li, L.; Kearnes, S.; and Dai, B. 2021.通过基于能量的模式了解逆合成。*NeurIPS*, 9.

Tetko, I. V.; Karpov, P.; Van Deursen, R.; and Godin, G. 2020.用于直接和单步逆合成的最新增强型 NLP 变换器模型。*自然通讯*》，11（1）：5575.

Wan, Y.; Liao, B.; Hsieh, C.-Y.; and Zhang, S. 2022.Retro- former：推动可解释端到端 Retrosynthesis 变换器的极限。*arXiv:2201.12475*.

Yan, C.; Ding, Q.; Zhao, P.; Zheng, S.; Yang, J.; Yu, Y.; and Huang, J. 2020.Retroxpert：像化学家一样分解逆合成预词典。*神经信息处理系统进展*》，33: 11248-11258.

Zhang, H.; Berg, A. C.; Maire, M.; and Malik, J. 2006.SVM-KNN：鉴别性近邻分类

视觉类别识别。*2006 年 IEEE 计算机学会计算机视觉与模式识别会议（CVPR'06）*，第 2 卷，2126-2136 页。IEEE.

Zheng, X.; Zhang, Z.; Guo, J.; Huang, S.; Chen, B.; Luo, W.; and Chen, J. 2021.自适应近邻机器翻译。第 *59 届计算语言学协会年会暨第 11 届国际自然语言处理联合会议论文集（第 2 卷：短篇论文）*。