# 使用Transformer与超图搜索策略预测逆合成路线

## One Sentence

该研究提出了一种超图的合成路线搜索策略，批判性地并提出了4个新的用于评估单步逆合成的指标。

## 引言

有机化学领域一直在发展，从关注合成复杂的天然产物到理解分子的功能与活性。在众多的任务中，设计合成路线是最为困难的之一，主要原因是逆合成需要确证大量的化学键断裂的方案、适合的构建模块以及适合的官能团保护策略。因此，不出意外的，自20世纪60年代起，计算机被广泛用于合成路线规划，并开发了许多计算机辅助工具。

在许多年内，基于规则（相似性）的方法是最为成功的计算机辅助方法。这些方法能有效的给出合成路线，但它们不是严格地从化学数据而是编码好的合成规则中学习信息。这些方法的主要的一个问题就是，它们需要人工的编码反应模板，这就决定了数据集的大小不会很大。

### AI驱动的化学的开端

虽然人类化学知识将在未来几年持续推动有机化学研究，但对当前趋势的仔细分析和基本外推原理的应用无可否认地表明，人们对使用人工智能 (AI) 架构的期望越来越高，它们可以模仿人类的化学直觉，并为全球所有实验室化学家提供帮助。

与基于规则的系统的同时期，广泛的 AI 方法被报导用于逆合成分析、反应结果预测和反应条件优化。 所有这些人工智能模型都取代了基于规则的方法，它们通过从大数据集中学习化学而无需人工干预来模仿人脑。

其中，公共数据的可用性使有机化学人工智能模型的广泛生产成为可能。 然而，文本挖掘提取过程产生的数据中包含的噪声严重阻碍了它们的发展。事实上，虽然基于规则的方法通过湿实验室实验证明了能够以较少的纯化步骤设计目标分子，从而节省时间和成本，但人工智能方法仍有很长的路要走。

在不同的 AI 方法中，将化学反应预测视为自然语言 (NL) 问题已成为一个主流的方法，在正向预测任务中，该方法达到SOTA (90%). 目前用不同类型的神经机器翻译架构将正向或逆向反应预测转换为翻译问题已成为主流手段， 将合成化学表示为一种语言的最显着优势之一是更大数据集的固有可扩展性，因为它避免了繁琐的操作，例如需要人类分配反应中心。Molecular Transformer 架构是目前最流行的将化学视为一种语言的方法，其训练好的模型为IBM的RXN for Chemistry平台提供服务。

### 基于Transformer的逆合成方法的现状

在基于Transformer的正向合成（正合成）预测方法提出后，该架构很快地被用于预测逆合成任务。Zheng等人提出了一种去模板的自纠正的逆合成预测器，该模型在一个小数据库(USTPO-50K)上的Top-1的ACC达到了43.7%，并且经过校正器后，无效分子的比率从12.1%降低到了0.7%; 然而，在本研究的前向研究中，无需任何校正器，无效分子的比率即可达到0.5% (**注：本人认为和训练的样本量有关)**; Karpov等人在同样的数据库上的Top-1的ACC可以达到42.7%; Lin等人结合模特卡洛搜素算法与单步的逆合成预测模型实现了多步的逆合成预测，其中单步的预测器在相同数据库的Top-1的ACC达到了43.1%，并且在一个10倍大的训练集上达到了54.1%; Duan等人通过增加训练的batch\_size, 使得在USTPO数据库上的Top-1 ACC达到了54.1%; 随后，相同的相同的架构在Top-1 ACC达到了43.8%，与先前报导的三个模型表现一致，但显著低于Duan等人的模型，但值得关注的是该模型是在一个专利数据集上训练的，在该数据集中，反应物之间的相似性最高达到了0.75，大约是USTPO的两倍。虽然在专利数据集上，模型在训练集和测试集都能达到一个不错的表现，但实际是否在特定化学信息上存在过拟合的问题是存疑的。近期，一个图提升Transformer(graph henced transformer)模型被报导在Top-1 ACC上达到了44.9%，并且可以提供更加丰富的反应物建议，除此之外，与之前的研究相比，没有很大的提升。

除了Lin等人的研究，所有的Transformer逆合成模型都**只是单步预测的。没有任何模型尝试预测试剂、催化剂与溶剂状态，只能预测反应物**。

在这项工作中，使用了超图的搜索策略来扩展分子Transformer架构。与其他工作相比，该工作预测了每一步的逆反应，并且**将试剂也视为反应前体**(图1)。此外，批判了现有评估指标(Top-1 ACC)的可信度，并基于正向合成预测器与分类预测器**提出了4个新的用于评估单步逆合成的指标**。这些指标符合人类专家在实际合成任务中关注的重要信息。

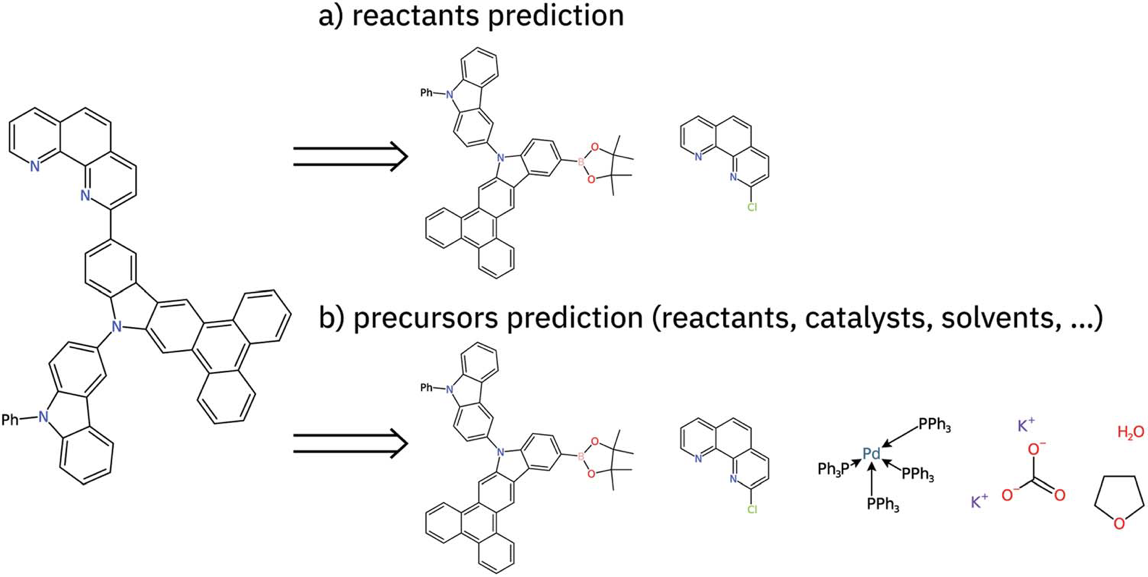


图1. a)只预测“反应物”; b)前体预测，包括反应物、催化剂、溶剂...

在该工作中，通过在超图上使用Beam Search来搜索最优的合成路线。其中，超图通过动态的方法来构建——基于一种类贝叶斯概率以及SCScore来筛选添加的节点。该筛选策略允许规避潜在的选择性陷阱，惩罚比目标复杂度更高的非选择性反应和前体。该研究奖将逆合成树的质量与前向预测模型的概率分布联系起来，并提出使用 Jensen-Shannon散度(JS散度, JSD)来表征分布的相似性。该研究首先提出了一种系统地提高多步逆合成工具质量的方法。

最后，这项工作通过回顾几个逆合成问题，批判性地评估了整个 AI 框架。结果表明，**在单步逆合成预测的指标子集上达到高性能对多步框架中是无益的**。 此外，该工作还证明，新定义的指标提供了评估端到端的评估方案，从而仅关注单步预测模型的质量。

## 材料与方法

### 评估单步逆合成预测的指标

评估逆合成路线是人类专家的一个任务，不幸的是，在大量的可用示例中去缩小范围是十分枯燥且无聊的。因此，让不同的模型生成统计学上相关的模型是是一项艰巨的挑战。通过与人类专家的类比，该研究提出使用一个正向预测模型与一个反应类型分类模型来评估逆合成预测模型的质量。

在这里简述一下辅助信息中的这两个模型的相关信息。正向预测模型使用Transformer架构，训练数据为Pistachio3.0. 正向预测模型有两种预测模式。即对于给定的反应前体 (precursor) 输出概率最大的产物或者用过Beam Search输出Top-N个产物，在该研究中Beam Width设为3; 对于反应分类器，同样使用Transformer架构，该模型由四个Encoder Layer和1个Decoder Layer组成。主要区别在于输入为完整的反应串(precursors到products)，输出为NameRXN 的划分反应的类型。反应类型字符由三个对应于超类、类/类别和命名反应的数字组成。 有关反应类别的更多详细信息，请参见[此处](https://doi.org/10.1021/acs.jmedchem.6b00153)。该工作中使用的分类模型与 NameRXN 工具中的93.8% 反应匹配的类别相同。(NameRXN为商用APP)

其中，正向预测器用来评估单步的逆合成器预测出的前体，通过正向反应的可能性，并且通过反应类型预测器得出反应类型。在该研究中，认为如果一个逆合成预测出的前体通过正反应预测，能得到原来的分子，则认为该逆合成预测是有效的(图2)。该研究提出4个指标(往返准确率、覆盖率、类别多样性与JS散度)来**彻底地(thoroughly)**评估逆合成预测模型。

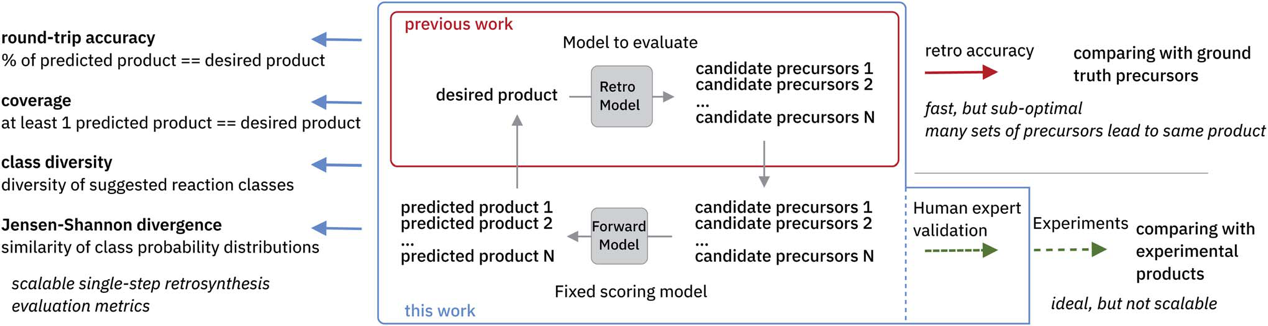


图2. 单步逆合成预测模型评估大纲

往返准确率 (round-trip accuracy) 用来量化多少的逆合成建议是有效的。该指标高度依赖于Beam Seach中的Beam Width，因为在Beam Width较多的情况下，由于建议质量较低，可能会导致有效建议的百分比降低。

覆盖率 (coverage) 用来**定量分析逆合成模型产生至少一种有效前体的分子的数量**。有了这个指标，可以防止奖励模型只为少数反应产生许多有效的前体。这种行为可能会导致相对较高的往返精度，但会导致较小的率。以保证逆合成模型能够为各种目标分子提供有效的建议。

类别多样性 (class diversity) 是对覆盖范围的补充，它不是与目标分子相关，而是**计算由逆合成模型在分类时预测的不同反应超类的数量**。即，对于一组给定的前体，能预测出几种的有效反应的类型。单步逆合成模型应该预测各种各样的断开策略，**这意味着生成导致相同产物的前体，相应的反应可能属于不同的反应类别**。允许多种不同的断开策略有利于最佳路径搜索，并且在目标分子包含多个官能团时是必不可少的。

最后是JS散度 (Jensen–Shannon divergence), 该指标用于评估12个不同反应超类, , 的高于0.5的似然分布之间的相似性。计算公式为：

其中表示概率分布，表示分布的信息熵。

为了计算JS散度，该研究将单步逆合成反应划分到某一超类，并使用正向预测模型的似然概率为每个类构建似然密度函数，然后通过JS散度计算方程中的熵函数对其进行归一化。该指标对于评估一系列逆合成步骤的质量至关重要。**拥有一个具有难以区分的似然分布的模型相当于让人类专家偏爱一些特定的反应类别而不是其他类别**。这将导致模型偏向于那些具有显性似然分布的类。虽然我们希望有一个峰值分布，因为这是模型从数据中学习的明显标志，但也希望所有的似然分布均等地达到峰值。JS散度 (1/JSD) 的倒数是不同超类之间似然分布相似性的度量，该研究使用此参数作为**有效度量所有可能的预测反应类之间的似然分布的均匀性质**。(十分幸运地，这四个指标作者已给出[轮子](https://gist.github.com/pschwllr/fb795e5384817c8b79089902bf4d0920)，极大程度上地辅助了我理解)。

### 超图的探索

逆合成搜索树相当于一个有向无环的超图 (hyper-graph), 即由超弧(hyper-arc)与结点构成的图。与一般的图的差异在于，超弧可以同时连接多个结点。这恰好与逆合成预测任务相似: 如果一个结点代表目标分子，那么超弧连接的不同结点则代表所有可能的发生反应的分子。超弧有一个天然的方向，决定反应是向前还是向后 (图3)。

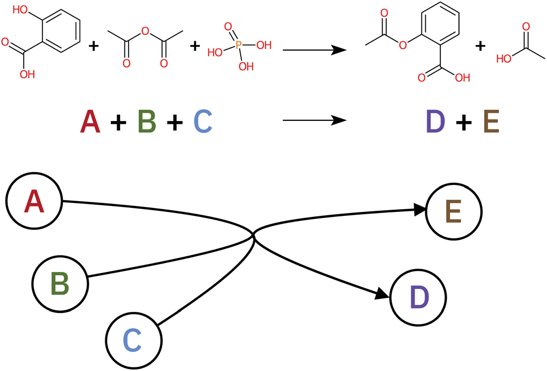


图3. 一般的反应（图片顶部）可以表示为超图。参与反应的每个分子都成为超图中的一个节点，而将反应物和试剂连接到产物的超弧线代表反应箭头。

逆合成路线是无环的。这一要求使逆合成路线成为一棵超级树 (hyper-tree)，其中根结点是目标分子，叶子结点是可商购的起始材料(见图 4)。该研究使用 eMolecules工具提供的数据库来确定分子是否可用。

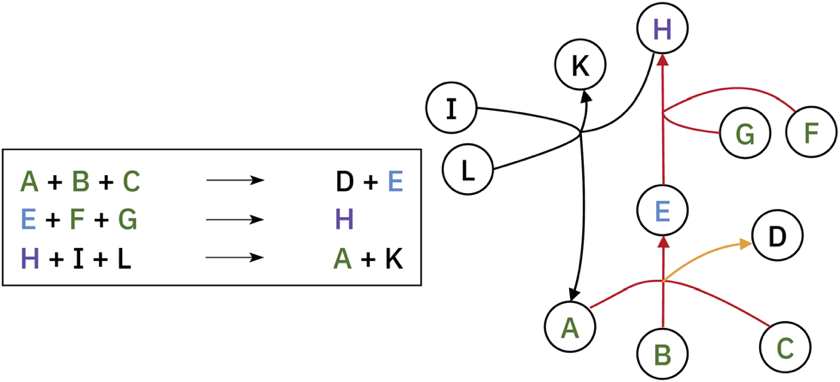


图4. 超图复杂性的示例。分子H是目标化合物(紫色标签)。红线代表从市售前体(以绿色突出显示)到目标分子的合成路径。黄线不影响H的逆合成，黑线的最后反应也不影响。

在整个化学空间的超图可用的情况下，详尽的搜索可能会揭示所有可能的合成途径，这些途径定义了从起始材料到生成目标分子的全过程。相反，该研究在这里动态构建超树：仅计算沿最有意义的逆合成扩展的节点和弧并将其添加到现有树中。逆合成探索使用基于SCScore(一种评估合成难易度的指标，类似于Scopy中的SAscore)的类贝叶斯概率来决定图扩展的方向，从而将树推向更简单的前体。在图 5 中，该研究展示了多步逆合成工作流程的示意图。为了阻止使用非选择性反应，该研究设定通过使用正向模型返回的反应可能性阈值来过滤单步逆合成预测。最后，过滤后的预测的似然性和 SCScore 被结合起来计算一个概率分数来对所有选项进行排序。 如果所有预测的前体都可在商业上获得，逆合成分析提供该选项作为可能的解决方案，并该搜索数是完备的。 如果没有，则使用前体作为初始目标分子重复整个循环，直到我们达到市售分子或特定逆合成步骤的最大数量。下面详细阐述一下整个搜索过程

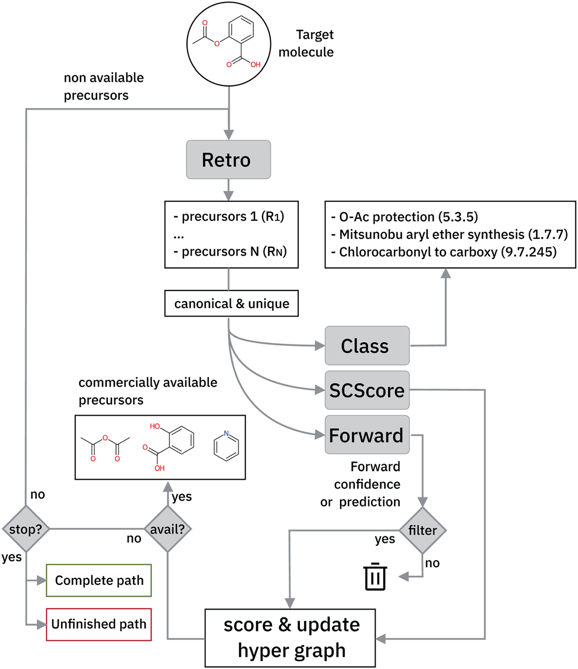
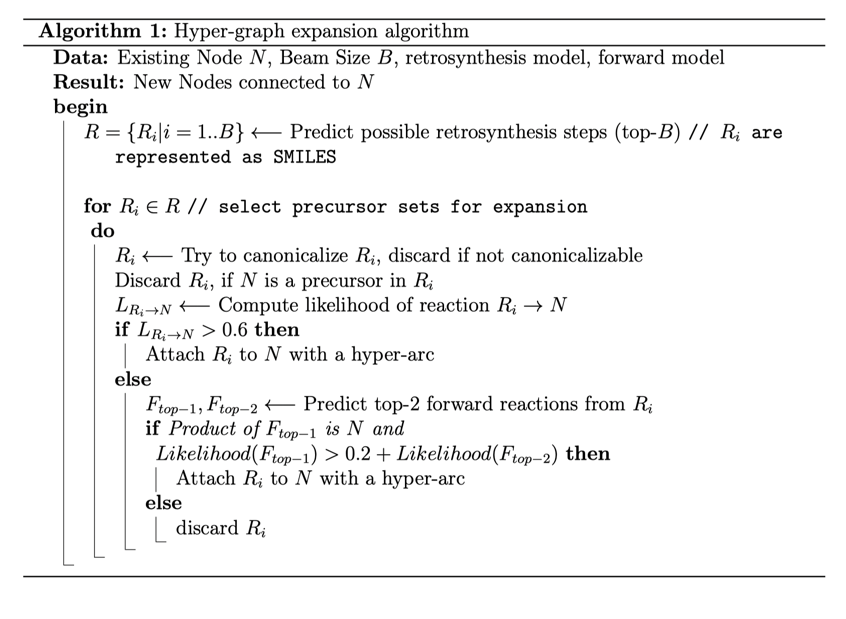


图5. 多步合成的搜索策略流程图。

算法1为超图扩展策略的概述，其中给定起始节点，通过预测通过反应得到分子的前体来扩展超图。单步逆合成模型使用Beam Search来探索可能的断键方式，该算法保留前15组预测的前体(因此), 预测的 SMILES 被标准化并删除重复条目，此外任何无效的SMILES也将被删除(常规的处理方法了，嘻嘻). **剩余的前体组通过使用正向预测模型进一步地评估反应可行性和选择性**。

对于反应的可行性，该研究**只保留那些通过正向预测，其结果的Top-1与分子匹配的前体** 。这保证了在存在多个官能团的情况下，断开化学键会导致前体。但这是一个必要不充分条件，因为竞争性反应(Top-2集之后)可能会导致与所需目标不同的分子混合的现象。为了加强化学选择性，该研究**仅选择似然比随后的 top-2 大至少 0.2 的 top-1的预测的前体**。由于前体集的预测的似然总和为 1，任何高于 0.6 的预测似然自动满足上述要求并通过我们的过滤器。 这种过滤协议增加了沿逆合成路径的化学选择性反应的发生，惩罚了高度竞争的断开连接。(附上优雅的伪代码，感觉比文字直观太多了)



此外，具有相似断开化学键策略的前体被聚集在一起以降低树的复杂性。在同一簇内，与最高前向预测似然相关的前体被用作进一步树扩展的起始节点。 每个前体分子，除非已经出现在图中，否则将产生一个新节点，每个反应都会通过新的超弧将每个反应物连接到目标分子。

树中的每个超弧都用所谓的优化分数进行评估，来用于定义“最佳”逆合成路线。通过将路径中包含的所有超弧的分数乘积来表征逆合成途径的总分。 单个弧的得分定义为：

其中 表示某一步的逆合成路线的分数。 是由正向预测模型所得出的发生反应的可能性。 是分子的简单性得分。其计算公式为：

其中为分子的复杂性得分，越复杂值越大，取值范围为1～5.

## 结果与讨论

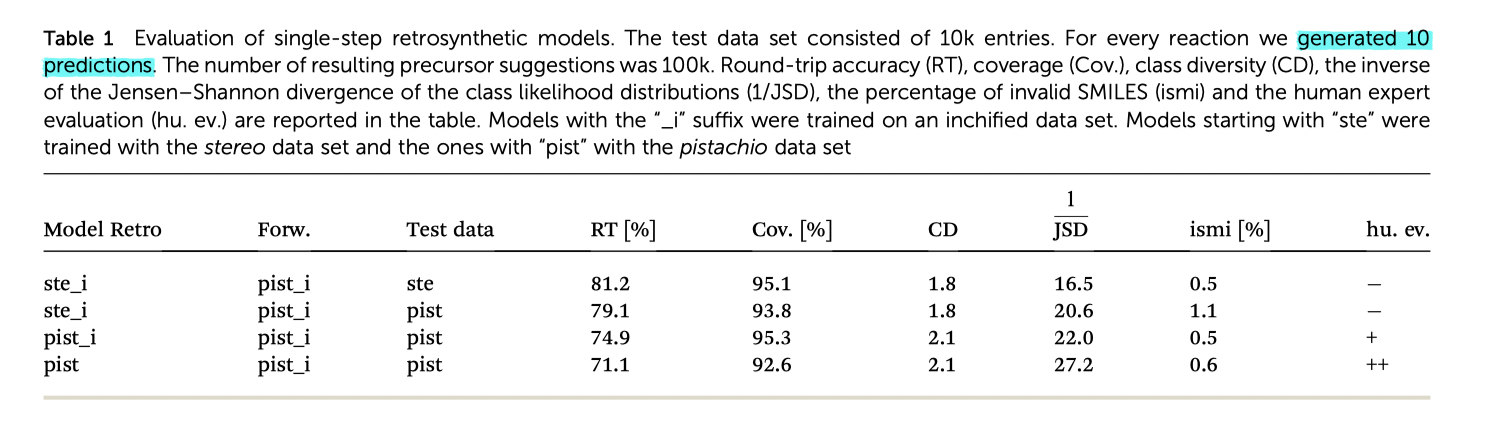
### 单步逆合成

Top-N ACC在之前的逆合成研究中，是评估单步预测模型质量的首选方法。 虽然这对于正向反应预测的评估是完全合理的，但它**在单步逆合成模型的背景下的使用具有误导性**，正如 Thakkar 等人所提出的，**Top-N ACC代表在逆合成预测模型的前 N 个建议中发现了真实的前体**。与正向预测模型相反，目标分子很少仅来自一组前体，**不同官能团往往存在允许存在多种可能的断开策略**，从而导致不同的反应物组以及可能的溶剂和催化剂。

通过分析Lowe 文本挖掘开源反应数据集的 USPTO 立体数据集和Pistachio 数据集的分析表明，分别有6%和14%的化合物至少有两个不同的集前体。 虽然这些数字仅反映了每个数据集中代表的有机化学，但可能断开的总数无疑更大。考虑到现有数据集的有限大小，很明显，在逆合成的背景下，**Top-N ACC 会励模型从数据集中检索预期答案的能力，而不是预测具有化学意义的前体的能力**。 因此，**Top-N ACC 并不是评估逆合成模型的适当指标。**

该研究对先前在单步逆合成模型中使用 Top-N ACC提出异议，并提出了四个新的不同指标（往返精度、覆盖率、类别多样性和 Jensen-Shannon 散度）。

在开发阶段，该研究使用两个不同的数据集训练了不同的基于Transformer架构的模型，一个完全基于开源数据(stereo)，另一个基于来自Pistachio的商用数据(pistachio)。在某些情况下会使用聚焦数据集(被标记为 \_i). 表 1 显示了逆合成模型的结果，使用修正的正向预测模型 (pistachio\_i) 在两个验证集(stereo和pistachio)上进行评估。其中，覆盖率(coverage)表示建议至少一个有效前体组的所需产品的百分比，在这方面stereo数据集稍微好一点，但所有模型组合都超过 90%，这是保证始终提供至少一种断开连接策略的可能性的重要要求。同样，类别多样性(class diversity)是在单个逆合成步骤中预测的不同反应类别的平均值，这两种模型具有可比性，pistachio模型的性能稍好一些。往返准确度(round-trip accuracy)，即在使用前向模型评估时导致初始目标的前体集的百分比，stereo比pistachio更好。尽管stereo逆合成模型在往返精度和覆盖范围方面比pistachio模型表现更好，但通过人类专家评估，使用该模型生成的合成路线质量较低，并且通常以一系列不合逻辑的保护/去保护步骤为特征(表 1 中的最后一列)。当我们详细分析人类如何处理逆合成问题时，这个明显的悖论就变得清晰起来。



解决逆合成问题需要仔细分析多种前体中的哪些可以更有效地产生所需的产品，如图 6，以5-bromo-2-methoxypyridine为例 。人类通过在脑海中列出和分析所有可能的断开位置并仅保留选项来解决这个问题，对于这些选项，相应的前体被认为最有选择性地产生目标分子。

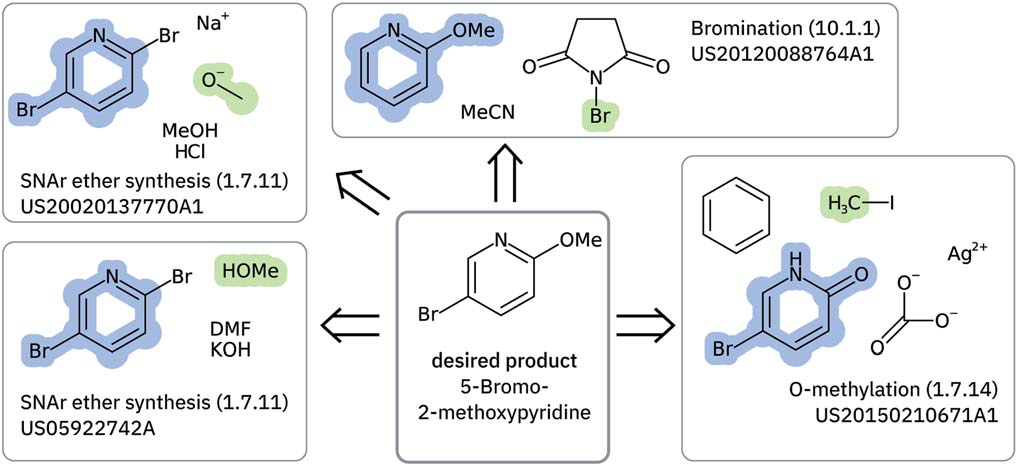


图6.

对于人类专家来说，总是找到至少一个断开位置（覆盖率）并确保相应的前体将选择性地导向原始目标（往返精度）是不够的。还有必要生成不同的断开策略样本以应对竞争性功能组反应性（类别多样性）。 此外，最重要的是，每个断开类都需要与所有其他类（Jensen-Shannon divergence，JSD）具有相似的概率分布。 继续与人类专家相比，如果一个人接触相同的反应类别多年，在路线规划中使用那些熟悉的方案会更频繁地出现，导致强烈偏向的逆合成。 因此，必须将单步逆合成模型中的任何偏差降至最低。

为了评估单步模型的对于每种反应类型的偏好(bias)，该研究使预测在不同反应超类中划分的似然分布的 JSD，表 1 中将其报告为 1/JSD。这个数字越大，属于不同类别的反应的似然分布越相似(即对每种反应的偏好越小)，因此，在多步合成中，单个反应类别的主导性（较低偏差）越小。在图 7 显示了表 1 中不同模型的似然分布。所有分布都显示接近 1.0 的峰值，这清楚地表明模型学会了如何预测这些类别中的反应。由于stereo和pistachio数据集中立体化学反应的数据质量/数量较差，因此分辨率等级相对平坦。有趣的是，可以看到，对于立体模型，脱保护、还原和氧化反应的似然分布与使用相同模型生成的所有其他分布不同（并且通常更加峰值）。这种统计不平衡有利于那些反应类别，并解释了保护/脱保护或氧化/还原策略的不合逻辑循环的发生。虽然峰值分布是可取的，因为这是模型学习在精确类别中预测断开策略的结果，但十二个概率分布之间的差异 (JSD) 反映了内在偏差，这可能是由于数据集不平衡造成的。在报告的少数模型中，发现pistachio模型的相似度 (1/JSD) 得分最高。

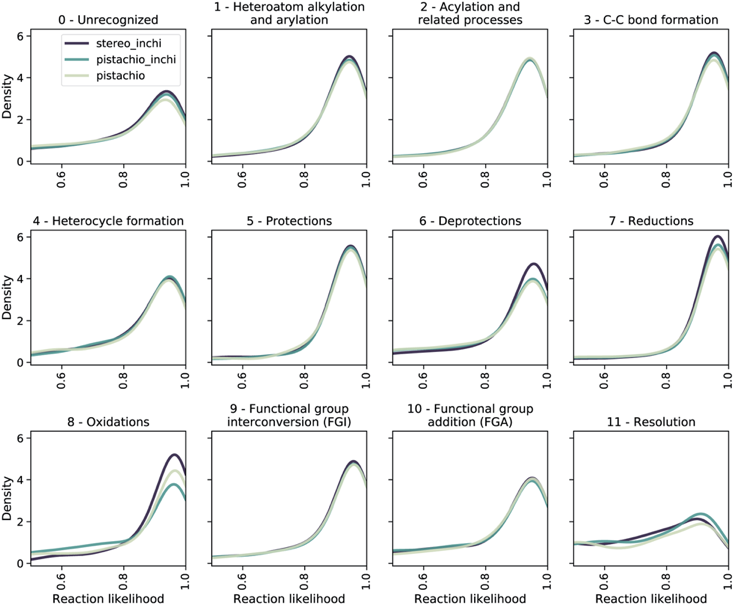


图7. 由前向模型(pistachio\_i)预测不同的逆合成预测模型预测前体的似然分布。只截取0.5～1.0

类别多样性和相似性分数需要识别每个预测的反应类别。 该研究使用了[这篇文章](https://chemrxiv.org/engage/chemrxiv/article-details/60c7531d469df45a9cf44cb5)中的基于Transformer的反应分类模型。在图 8 报告了 NameRXN 工具分类的真实情况、本研究的分类模型对真实反应预测的类分布。 观察到，该研究的类别预测模型所做的分类与 NameRXN 的分类一致，并且ACC达到了93.8%。

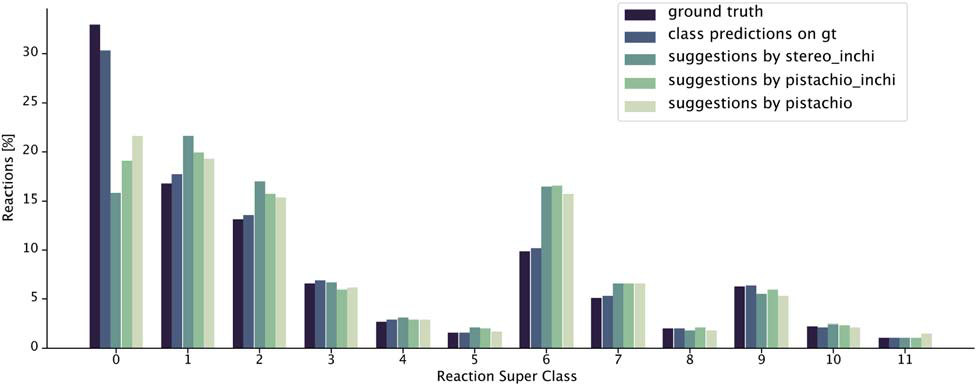


图8. 真实情况的反应超类分布、真实情况反应的预测超类和不同逆合成模型建议的反应的预测超类的分布。

### 路径预测的整体评估

在该部分中作者选用了8个分子(图9)，其中5个已有文献报导的合成路线，其中大部分都在预测路线中的前三名。并作出了分析。该部分有大量的有机机理分析，我难以理解就先跳过该部分。

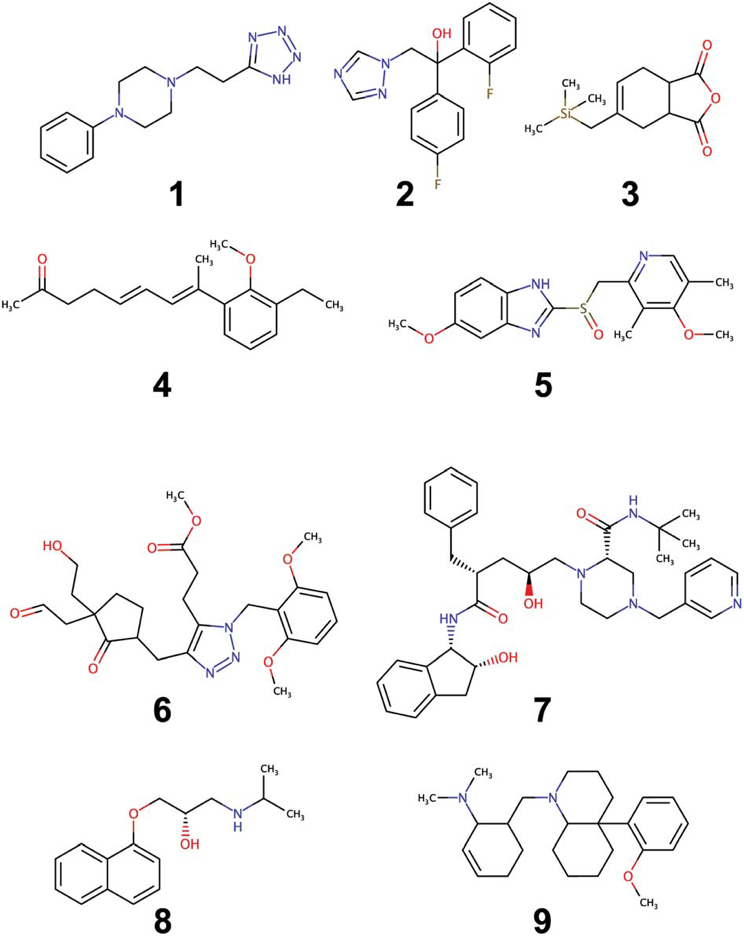


图9. 用于评估逆合成质量的一组分子。

## 结论

在这项工作中，展示了在Molecular Transformer架构上的扩展，并结合了超图探索策略，以在没有人工干预的情况下设计逆合成。该研究首次引入了预测反应物和试剂的单步逆合成模型。我们还引入了四个新指标(覆盖率、类别多样性、往返精度和 Jensen-Shannon 散度)，以提供对单步逆合成模型的全面评估。通过对可能的断开策略的超图进行波束搜索找到最佳合成途径，并允许规避潜在的选择性陷阱。超图是动态构建的，节点通过过滤策略，并基于类似贝叶斯的概率分数进一步扩展，直到识别出商业可用的构建块。此外，处理立体化学反应的能力不足是由于训练数据集质量差。通过专注于单步逆合成模型的性能而无需手动审查质量，提供了一种定义明确的策略来优化逆合成框架整个逆合成路线。该策略的一个关键作用是构建统计相关的训练数据集，以提高模型在不同类型的反应类别和断开连接中的置信度。

## 读后总结与启发

这篇文章首先犀利地批评了之前众多逆合成预测工作中使用的Top-N ACC这个评估指标，因为Top-N ACC只能评估模型是否能将一个目标化合物按照数据库中所记录的那样预测出其前体化合物，然而一个化合物往往不知具有一种合成方法，所以应当去评估逆合成模型全部的Top-N的潜在前体化合物，在这个背景下作者提出了往返准确率(round-trip)。为了防止奖励模型只为少数反应产生许多有效的前体，又继续提出了覆盖率(coverage)，接着为了评估模型预测不同类型的能力，提出了类别多样性(class diversoty)。最后，为了检测模型是否对某种类型有特定的偏好，提出了有JS散度评估模型。这些参数都可以用于我们之后的评估任务

此外，该文章还提出了基于超图(hyper-graph)的多步合成策略，从文章的结果看出该策略还是非常成功的。

最后，我有一些想法：

* 文章中虽然提出了往返准确率这种更好的评估参数，作者提出的想法仅仅只作为评估指标，没有真正参与到模型训练之中，对于一个分子有多种合成方法，使用交叉熵的模型是否能学到这里面的信息，我本人是存疑的，**是否需要提出一种新的loss**？
* 我们在超图的搜索策略中我们能**引入一些自定义的指标**，让预测出的化合物具有我们期望的性质；
* 我个人感觉，可以使用强化学习的方法来进行逆合成的研究。大致的实现流程为: (1)训练或找到一个预训练的分子(SMILES)的生成模型，同时训练一个正向反应的预测模型; (2) 使用超图的策略来定义模型的奖励，微调预训练的模型。在该设想中，只需要训练一个正向预测模型即可。目前的SMILES生成模型，均为生成单个分子的，似乎没有能生成多个分子的模型，可能还是得自行训练，这无疑是非常困难的，或者可以用Bert？我再找点相关资料，看看能不能用什么新的trick。