化工流程图机器学习的数据增强

Lukas Schulze Balhorn[a]、Edwin Hirtreiter[a]、Lynn Luderer[a]、
Artur M. Schweidtmann[a]，*
[a]Process Intelligence Research，化学工程系，代尔夫特理工大学，Van der Maasweg 9，
代尔夫特 2629 HZ，荷兰*通讯作者。电子邮箱：*a.schweidtmann@tudelft.nl*

摘要
人工智能在加速化学过程的设计与工程开发方面具有巨大潜力。近期，我们已证明，基于变换器的语言模型能够利用 SFILES 2.0 字符串表示法对化工流程图进行自动补全。此外，我们还证明，语言翻译模型可用于将工艺流程图（PFDs）转换为管道及仪表流程图（P&IDs）。然而，人工智能方法需要大数据，而目前流程图数据较为有限。为应对数据有限这一挑战，我们提出了一种针对以 SFILES 2.0 表示法呈现的流程图数据的新型数据增强方法。我们证明，所提出的数据增强方法能够提升基于人工智能的过程设计模型的性能。在我们的案例研究中，流程图数据增强使流程图自动补全模型的预测不确定性降低了 14.7%。未来，我们的流程图数据增强方法还可用于其他基于 SFILES 表示法、针对化工流程图的机器学习算法。
关键词：数据增强、流程图自动补全、SFILES、变换器

1. 引言
流程图拓扑的设计是过程综合早期的一项重要工作。这一步包括为化学过程选择并排列单元操作。人工智能（AI）方法有望从以往的流程图中学习，并为过程开发中的工程师提供支持（Hirtreiter 等人，2022；Oeing 等人，2022；Schweidtmann，2022；Vogel 等，2023）。例如，Vogel 等（2023）提出了一种用于流程图自动补全的算法。该自动补全算法受到自然语言处理（NLP）中文本自动补全的启发，后者基于生成式变换器模型（Radford 等人，2019）。此外，Hirtreiter 等人（2022）指出，对工艺流程图（PFDs）中控制结构要素的预测可以视为工艺流程图与管道及仪表流程图（P&IDs）之间的翻译任务。因此，他们采用了常用于不同语言文本翻译的序列到序列变换器架构。这些流程图变换器依赖于机器可读流程图表示。

为了以机器可读的格式表示流程图，我们将其描绘为图，或使用唯一的（即规范的）SFILES 2.0 字符串来表示（Vogel 等人，2022b）。一般而言，流程图是化学过程的示意图。化学工程师使用流程图来交流、规划、操作、模拟和构建这些过程。图1展示了一个流程图示例。

一种直观的流程图表示方式是将其表示为图，其中单元操作为节点，物流连接为有向边。除了图表示法，流程图还可以表示为字符串。D'Anterroches（2005）提出了简化流程图输入行录入系统（SFILES）记法；我们最近在（Vogel 等人，2022b）中对其进行了扩展，纳入了控制结构和其他特性。创建 SFILES 时，我们从输入节点出发，沿物流方向遍历该图，直到到达产品节点或循环流。如果物流在某个节点处分支，即该节点为分流器，我们就需要决定先沿哪条物流前进。为确定线性字符串中各分支的顺序，SFILES 算法会为每个节点分配一个唯一的秩。图1所示流程图对应的 SFILES 字符串如下：

(raw)(hex){1}(r)<&|(raw)(pp)&|(mix)<1(v)(dist)[{tout}(prod)]{bout}(splt)1(prod)n|(raw)(hex){1}(prod)。

虽然人工智能模型需要大量训练数据，但机器可读流程图数据通常十分有限。数据有限的原因在于，流程图主要以图像形式呈现，因此无法被机器读取（Schweidtmann，2022）。最近，我们提出了在文献中自动查找流程图图像（Schulze Balhorn 等人，2022），并通过计算机视觉使其实现机器可读（Theisen 等人，2023）。然而，这些方法需要大规模实施。此外，由于公司的知识产权保护，大多数流程图并未公开。包含机器可读流程图的化学过程数据集十分罕见，而且通常仅包含几十个流程图（Hirtreiter 等人，2022；Oeing 等人，2022；Vogel等，2023；Zhang 等人，2018；Zheng 等人，2022）。然而，变换器等人工智能方法通常需要大量训练数据。

一种有前景的解决人工智能训练数据有限问题的方法是数据增强。数据增强基于这样一种理念：通过对现有原始训练数据进行掩码、修改和扰动，生成额外的人工训练数据。例如，在计算机视觉领域，数据增强是一种成熟的方法，可通过添加已有图像的修改副本来提升模型性能（Shorten 和 Khoshgoftaar，2019）。然而，目前还没有针对流程图的数据增强方法。因此，我们需要一种增强方法，通过添加已有流程图的修改副本，缓解可用流程图数据有限的问题。

我们提出了一种针对过程流程图的新型数据增强方法。具体而言，我们采用一种针对 SFILES 的基于文本的增强方法，其灵感来自对分子 SMILES 字符串的增强（Bjerrum，2017）。我们的方法是在生成 SFILES 时随机化分支决策。我们以 Vogel 等人（2023）提出的流程图自动补全模型为背景，展示所提出的流程图增强方法。

2. 数据增强方法

为了扩充流程图数据集，我们修改 SFILES 生成算法中的分支决策，从而生成多个表示同一流程图的 SFILES 字符串。在确定规范 SFILES 时，通过为图中的每个节点分配唯一的秩来作出分支决策。因此，规范 SFILES 是从流程图图结构到字符串的唯一映射。生成增强 SFILES 时，分支决策是随机作出的，因此得到的是非规范形式。二者的区别在于线性字符串中分支的顺序不同。生成的增强 SFILES 包含与规范 SFILES 完全相同的信息，因此描述的是同一过程流程图。因此，所有增强结果都可以转换回原始规范 SFILES。图1展示了示例流程图的数据增强过程。

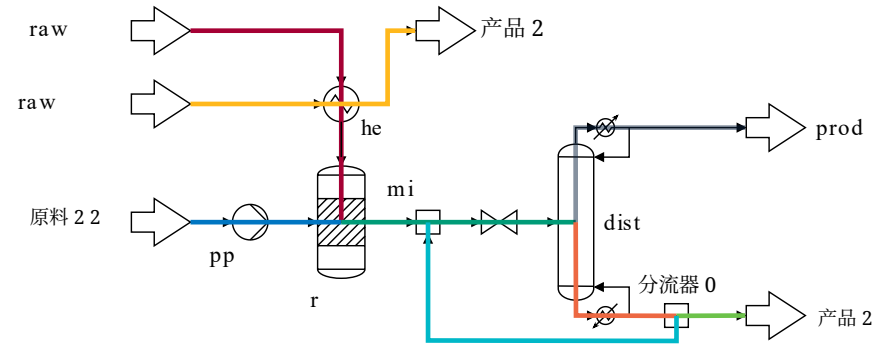



图1：图1所示流程图的数据增强示例。节点秩标注在节点名称后的方括号内。

数据增强过程中，仅 SFILES 表示的唯一性会丢失，而完整的流程图拓扑信息得以保留。值得注意的是，数据增强不会改变输入分支的顺序。否则，包含非连通子图的流程图将无法转换回原始规范形式。例如，对于图1示例流程图中的(raw)(hex){1}(prod)，这一点对于带有热集成的独立过程尤为重要。

由单个流程图图结构推导出的增强SFILES数量是有限的。具体而言，潜在增强SFILES的数量取决于给定流程图中可用分支点的数量；分支点越多，数据增强的可能性就呈指数增长。例如，图1中的流程图包含两个分支点。这会产生三种增强SFILES表示，总计四种SFILES表示。

本例中的两个分支点分别是蒸馏塔 dist-1 和蒸馏塔后的分流器 splt-1。在第一个分支点处，我们交换塔顶和塔底出口流股的顺序。第二次增强仅影响塔底产品流股 prod-2。不包含单元操作的循环流总是直接出现在分流单元之后，在此情况下，这会使 (splt-1)1 在所有情形中均出现。当产品分支先于循环分支被访问时，会在 prod-2 外使用额外的一对方括号。为使流程图数据增强方法能够公开获取，我们将其纳入公开的 SFILES 2.0 GitHub 代码库（Vogel等，2022a）。

3. 案例研究与结果

3.1. 数据与数据增强 我们使用两个 SFILES 数据集，它们

由（Vogel等，2023）创建。首先，我们使用他们提出的流程图生成器来生成一个大规模的约 8,000 个数据集

人工流程图。在这里，我们可以灵活调整数据集的规模。其次，我们使用由 223 个 Aspen 和 DWSIM 化工过程模拟构成的数据集，并将其称为真实数据集。在开始训练运行之前，我们将每个数据集拆分为训练数据集、验证数据集和测试数据集。对于这两个训练数据集，我们为每个 SFILES 最多创建五个增强样本，使数据集规模大致扩大五倍。通过将增强样本数量限制为五个，我们确保较大的流程图不会占据过高比例。

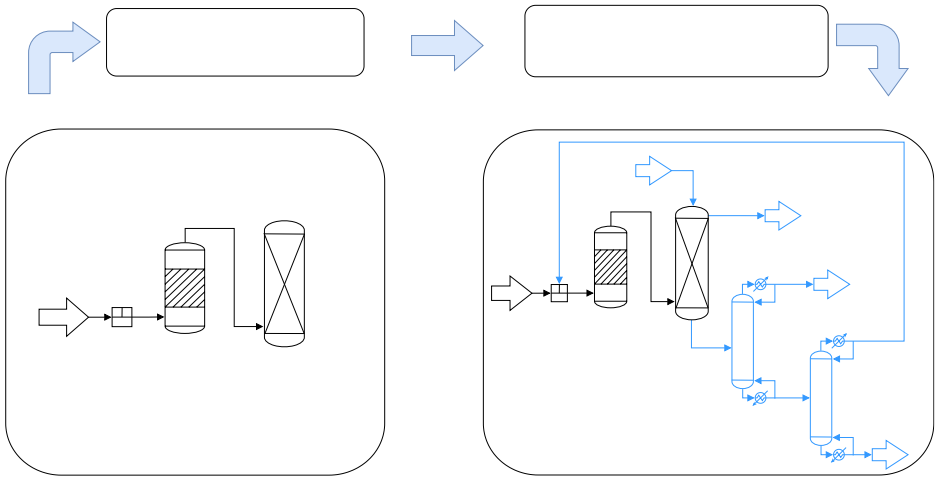

图2：流程图自动补全模型的预测示例。图改编自（Vogel等，2023）。

3.2. 流程图自动补全 我们考虑一个过程设计

案例研究，并使用（Vogel等，2023）提出的流程图自动补全模型。该模型旨在支持化学工程师设计新的过程拓扑结构。它可以为不完整的流程图建议下一个单元操作，类似于即时通讯应用中的句子补全。具体而言，给定一段 SFILES 开头，自动补全模型会通过迭代预测 SFILES 中的下一个构件（词元），来预测序列的后续内容（图2）。需要强调的是，该预测仅基于过程拓扑结构，不考虑操作点、组分和物料流。该自动补全模型采用仅解码器方法，基于 Transformer架构构建（Radford等人，2019）。由于 Transformer模型对数据需求量很大，因此使用生成的数据集对流程图自动补全模型进行预训练。随后，使用真实数据集对流程图自动补全模型进行微调。有关该模型的详细说明，请参阅（Vogel等，2023）。

为了研究数据增强的有效性，我们重新训练了来自（Vogel等，2023）的流程图自动补全模型。具体而言，我们分别使用增强和未增强的流程图数据训练模型。总体上，我们考察了三种不同的训练方案。首先，使用未增强的生成数据集训练模型，再使用真实数据集对该模型进行微调（i）。该模型用于复现（Vogel等，2023）的结果。其次，我们对真实数据进行增强，并用其微调预训练模型，得到模型（ii）。因此，我们只调整了微调过程。最后，我们在预训练和微调阶段均使用增强数据训练流程图自动补全模型

，得到模型（iii）。为公平比较采用和未采用数据增强的训练运行，我们使用相同的超参数。

3.3. 结果 我们使用困

惑度来衡量模型性能。困惑度描述了模型预测时的不确定性。因此，较低的困惑度更为理想。在此，较低的困惑度意味着模型更确信所建议的单元操作是合理的。困惑度是下一个词元预测的负平均对数似然的指数。它也等价于模型训练期间获得的交叉熵损失的指数。困惑度计算如下：

$$PP(T) = \exp\left(-\frac{1}{n}\sum_{i}^{n}\log P(t_i|t_{1:i-1})\right), \quad (1)$$

其中，$T = (t_1, \ldots, t_n)$ is a sequence of tokens and $P(t_i|t_{1:i-1})$描述了给定序列$t_{1:i-1}$时下一个词元t_i 的预测概率。

表1：数据增强的困惑度PP 结果。最佳测试结果以粗体显示。“Augm.”列表示训练数据是否经过增强。预训练和微调测试困惑度均在使用微调数据集完成模型训练后进行评估。

模型	***PP*** 预训练		***PP*** 微调				训练
	增强。	测试增强。	训练	验证	测试		**Time**
(i)	No	5.38	No	3.13	4.23	5.02	51 分钟
(ii)	No	6.33	Yes	3.32	4.12	4.69	57 分钟
(iii)	Yes	**5.16**	Yes	3.07	3.80	4.28	1小时31分钟

不同训练运行的结果如表1所示。在微调方面，模型(ii)在测试集上的表现略优于模型(i)。我们认为，这一提升是因为微调数据规模有限（i），而数据增强能够使模型更加稳健（ii）。虽然采用数据增强后测试困惑度有所降低，但训练困惑度却更高。这表明，经过数据增强训练的模型不易过拟合，并且对未见数据的泛化能力更强。对预训练也采用数据增强（模型(iii)）后，我们看到模型在所有类别中的表现均为最佳，与模型(i)相比，微调困惑度降低了14.7%。我们得出结论：数据增强也能改善使用规模相对较大的生成数据集进行的预训练，从而提升微调效果。

总体而言，增强SFILES是有效的流程图表示，并且有助于流程图自动补全学习SFILES语法和化学过程结构。未来可以探究：对于预训练而言，生成更多人工数据、通过数据增强扩大数据集，还是结合这两种方法更有利。值得注意的是，在所有情况下，使用增强数据训练的模型，其训练困惑度和测试困惑度都更为接近。由于训练数据的方差更高，这些模型不易过拟合。

4. 结论

我们提出了一种化学流程图数据增强方法，即在图遍历过程中随机化分支决策，以生成 SFILES。由此，我们展示了一种增加可用于后续训练人工智能模型的流程图数据的方法。我们将增强SFILES应用于流程图自动补全问题，并表明这种增强在低数据量场景下能够提升模型性能。在

未来的研究中，我们计划将数据增强方法应用于更多自然语言处理模型，以提升其性能。长期目标应是通过从文献（Schweidtmann，2022；Schulze Balhorn 等人，2022）和公司中挖掘更多流程图来扩大流程图数据集，并在必要时将其数字化（Theisen 等人，2023）。结合数据增强与扩大数据集规模，将十分必要且最为有益。

参考文献

Bjerrum, E. J. (2017). 将SMILES枚举作为分子神经网络建模的数据增强方法。arXiv预印本 arXiv:1703.07076。d'Anterroches, L. (2005). 基于基团贡献法的过程流程图生成设计。[CAPEC], 丹麦技术大学化学工程系。Hirtreiter, E., Schulze Balhorn, L., & Schweidtmann, A. M. (2022). 探索利用人工智能自动生成管道与仪表图（P&IDs）。arXiv预印本 arXiv:2211.05583。Oeing, J., Welscher, W., Krink, N., Jansen, L., Henke, F., & Kockmann, N. (2022). 使用人工智能支持采用DEXPI标准绘制管道与仪表图。《数字化化学工程》，4，100038。Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). 语言模型是无监督多任务学习者。《OpenAI博客》，1(8)，9。Schulze Balhorn, L., Gao, Q., Goldstein, D., & Schweidtmanna, A. M. (2022). 利用深度卷积神经网络识别流程图。载于《计算机辅助化学工程》（第49卷，第1567-1572页）。爱思唯尔。Schweidtmann, A. M. (2022). 流程图挖掘。手稿，编写中。代尔夫特理工大学。Shorten, C., & Khoshgoftaar, T. M. (2019). 深度学习图像数据增强综述。《大数据期刊》，6(1)，1-48。Theisen, M. F., Flores, K. N., Schulze Balhorn, L., & Schweidtmann, A. M. (2023). 利用深度卷积神经网络实现化学过程流程图的数字化。《数字化化学工程》，6，100072。Vogel, G., Schulze Balhorn, L., Hirtreiter, E., & Schweidtmann, A. M. (2022a). Process-intelligence-research/sfiles2：V1.0.0（版本发布）。GitHub。https://github.com/process-intelligence-research/SFILES2 Vogel, G., Schulze Balhorn, L., Hirtreiter, E., & Schweidtmann, A. M. (2022b). SFILES 2.0：一种扩展的基于文本的流程图表示。arXiv预印本 arXiv:2208.00778。Vogel, G., Schulze Balhorn, L., & Schweidtmann, A. M. (2023). 从流程图中学习：用于流程图自动补全的生成式Transformer模型。《计算机与化学工程》，171，108162。Zhang, T., Sahinidis, N. V., & Siirola, J. J. (2019). 化工流程图中的模式识别。《AIChE期刊》，65(2)，592-603。Zheng, C., Chen, X., Zhang, T., Sahinidis, N. V., & Siirola, J. J. (2022). 通过多序列比对学习过程模式。《计算机与化学工程》，159，107676。