面向大规模相对自由能扰动的更快的先导优化映射算法

Kairi Furui东京工业大学计算机学院日本神奈川县
furui@li.c.titech.ac.jp

Masahito Ohue东京工业大学计算机学院日本神奈川县
ohue@c.titech.ac.jp

arXiv:2304.04713v1 [q-bio.BM] 10 Apr 2023

摘要—**近年来，自由能扰动（FEP）计算作为支持药物发现的工具，日益受到关注。先导化合物优化映射器（Lomap）被提出作为一种高效计算配体之间相对自由能的算法。然而，Lomap 需要检查 FEP 图中的每条边是否可移除，这就必须对所有边进行约束检查。因此，传统 Lomap 需要耗费大量计算时间：对于涉及数百个化合物的情况，至少需要数小时；而对于边数超过数万的情况，则不具备实用性。本研究旨在降低 Lomap 的计算成本，使其能够构建数百个化合物的 FEP 图。通过使用分块检查过程，同时检查尽可能多的边的约束，我们可以将所需约束检查的总次数从依赖边数降低至依赖节点数。此外，输出图与使用传统 Lomap 获得的图等价，因此可以直接用我们的方法替换原始 Lomap。经过改进后，该方法的执行速度比原始 Lomap 快数十至数百倍。**

索引词—**自由能扰动（FEP）、Lomap、计算药物发现、加速、FastLomap**
常规研究论文

I. 引言

在药物发现的早期阶段，需要通过命中到先导优化，从与目标受体弱结合的命中化合物中筛选出具有足够亲和力和理想药理性质的先导化合物 [1]。炼金术自由能扰动（FEP）计算 [2]–[4]是一种广泛使用的先导化合物优化工具，因为它采用基于扰动的方法，通过分子动力学模拟计算候选化合物与其目标受体之间精确的相对结合自由能 [5], [6]。

闭合热力学循环对于减少相对结合自由能计算中的误差至关重要 [7]–[9]。热力学闭合路径的相对自由能之和理论上为零。闭合循环中的误差称为循环闭合误差或滞后现象 [7]。基于闭合循环减少误差的循环闭合方法，相较于扰动计算可提供更高的预测准确度。因此，为确保 FEP计算的准确性，FEP 图必须包含计算资源允许的尽可能多的闭合循环。图1展示了一个由 Wang等人 [2] 使用 CressetFlare 的 FEP 应用程序 [3],[10]运行的 TYK2 靶点基准系列 FEP 图示例。

首先，最常见的 FEP图是星形图，它将所有化合物仅连接到一个参考化合物。不需要特定的图构建算法，且仅执行 $N-1$ 次扰动计算。然而，由于不存在单一的闭合热力学循环，因此无法保证 FEP计算的准确性。PyAutoFEP [11]，一个用于自动执行 FEP计算的开源工具，提供了使用蚁群优化构建轮辐状图和最优图的实现。

DiffNet [8] 为优化误差的协方差矩阵提供了理论范式。HiMap [9] 也基于协方差矩阵的优化提供了统计学上的最优设计，与DiffNet类似。

Lomap [7] 是用于计算相对自由能的自动化规划算法。Lomap根据多个适用于相对自由能计算的设计标准，计算化合物对之间的评分，并基于这些评分构建FEP图。在此，Lomap基于所有分子都必须至少属于一个闭合循环这一标准构建FEP图。然而，Lomap只能处理几十种化合物的FEP图生成，并且必须考虑数百次FEP计算。

最近，主动学习 [12]–[14],一种用于规划先导化合物优化实验的机器学习方法，已被应用于相对结合自由能计算 [15]–[17]。主动学习涉及数百或数千次大规模FEP计算。然而，对于包含数百种化合物的FEP 图，传统的Lomap计算会产生巨大的计算成本。因此，将Lomap图直接应用于主动学习场景需要付出大量努力。采用FEP的传统主动学习方法只能尝试执行不包含闭合循环的星形 FEP计算 [16], [17] 或对有限数量的化合物执行更详细的计算 [15]。

因此，简化 FEP 循环的引入至关重要，即使涉及数百种化合物，也能支持通过主动学习进行 FEP计算。

Lomap 计算耗时较长的主要原因是，它会检查

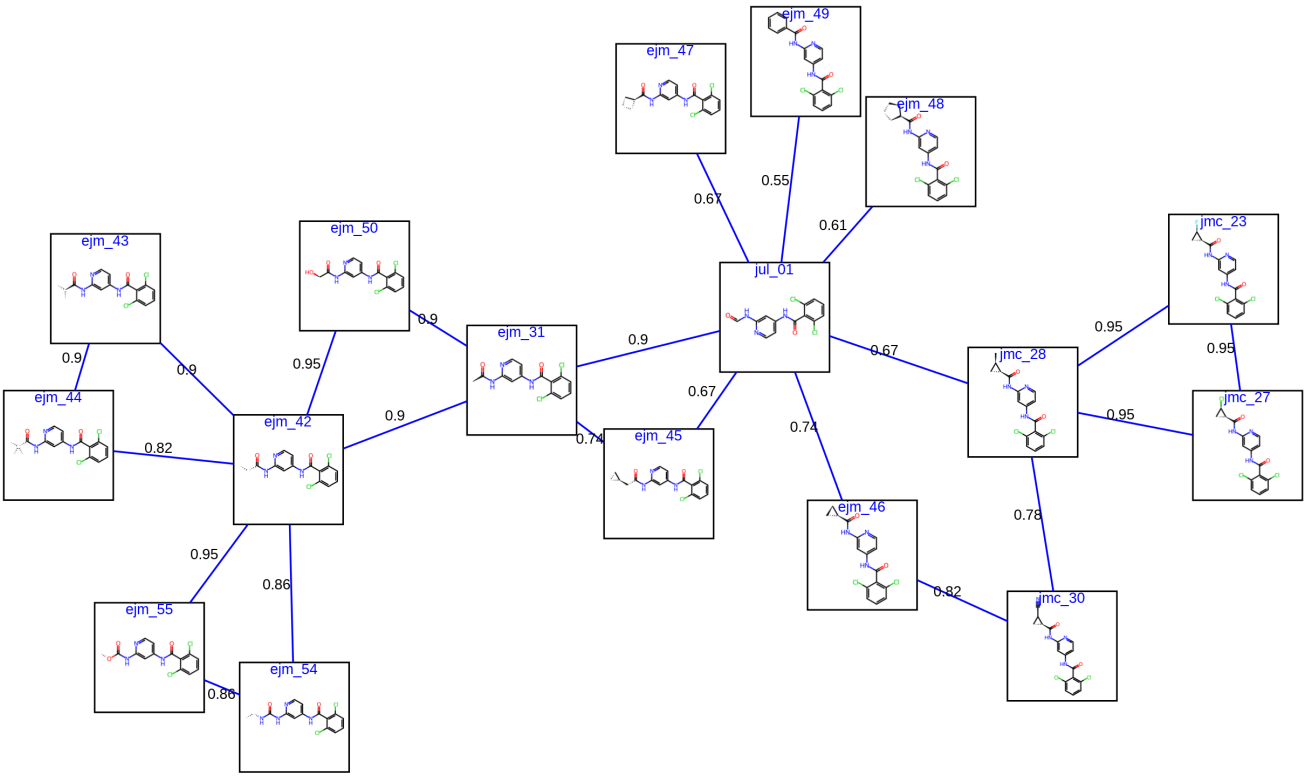


图1. Wang基准数据集上的 FEP 图示例（靶标：TYK2）。使用 Lomap 和 NetworkX 库绘制。

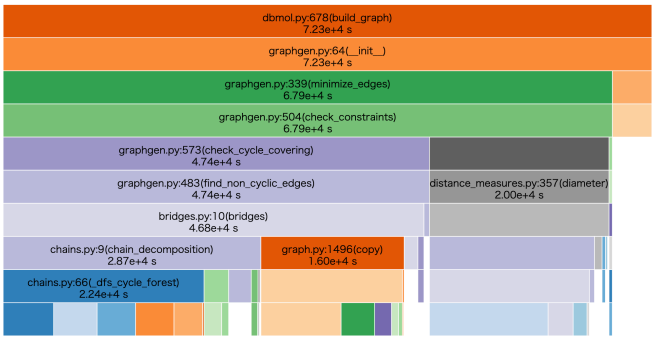


图2. 基线Lomap算法的 cProfile 性能分析结果（Thompson TYK2 数据集，$N = 1000$）。由 snakeviz 绘制。

初始图中每条边的约束。

图2 展示了使用 Lomap 算法为 1000 个节点生成 FEP 图的结果。build_graph 函数的大部分执行时间都耗费在 check constraints 函数上。_

在本文中，我们提出了一种改进的Lomap算法，该算法引入了分块检查过程，尽可能批量处理边，以使约束检查次数取决于节点数量。对于数百种化合物，所提方法构建 FEP 图的速度可提高数十至数百倍。此外，输出的 FEP 图与原始 Lomap 生成的图等效。

II. 方法

A. Lomap 算法

Lomap 是基于 RDKit 实现的开源工具（https://github.com/OpenFreeEnergy/Lomap），本研究以该实现为基础并对其进行了改进。本节介绍原始 Lomap 中使用的算法。Lomap 会生成一个无向图，用于确定将进行相对结合自由能计算的化合物对。FEP 图必须由原子间足够相似的化合物对所对应的边组成，才能计算出有效的自由能。链接分数是基于两个化合物之间相似度的指标。链接分数根据一对化合物的最大公共子结构（MCS），使用以下公式计算：

$$S = \exp[-\beta(N_A + N_B - 2N_{MCS})], \quad (1)$$

其中，β 是一个参数，而 N_A、N_B和 N_{MCS} 分别是两个输入化合物和MCS中的重原子数。这就是该扰动所需的插入和删除次数。该分数的取值范围为0到1，当两个化合物相同时，最大值为1。该分数包含一项校正，该校正取基于其是否适用于扰动计算而计算的其他分数的乘积，例如具有相同的净电荷并尽可能保留环。图生成的一个关键方面是，应移除链接分数较低的边，除非它们对闭合产生有利影响

周期。使用Lomap算法生成FEP图需要计算所有化合物之间的连接分数。在本研究中，每个化合物对的分数计算未进行加速，因为该过程可以并行化。此外，分数计算不纳入执行时间测量，因为分数是预先计算的。

Lomap 的 FEP图生成过程如下所述：
1) 为所有化合物对计算评分矩阵。2) 初始图仅由连接分数高于阈值的化合物对所对应的边组成。初始图可能被划分为多个子图。3) 对初始图的每个连通分量执行边移除过程。首先，按连接分数从高到低对该子图中的边排序。4) 从连接分数最低的边开始，评估移除该边后是否满足约束条件。如果满足，则移除该边；否则，保留该边。四项约束如下：

a) 子图是连通的。b) 不在环中的节点数量不变。可通过环中的桥数量是否保持不变来判断是否满足此约束。c) 子图的直径小于或等于阈值 MAXDIST。d)（如有必要）与已知活性配体之间的距离小于或等于阈值。

5）连接每个连通分量。

在这里，当图较大时，过程（4）耗时最长。这是因为过程（4）会检查子图中的所有边，以确定是否可以移除某条边。随后会计算图的直径和桥，以确定是否可以移除某条边。如果图中有 N 个节点和 M 条边，则直径的计算复杂度为 $O(MN)$，桥的计算复杂度为 $O(M+N)$。由于每次更新 FEP 图时，图中的边数都会发生变化，因此计算复杂度至少为 $O(M^2N)$。造成这一计算复杂度的最主要因素是需要检查所有边，以确定它们是否可以移除。Lomap 算法按得分从差到优的顺序移除边，但必须保持这一顺序。因此，图生成阶段无法并行计算。

B. 分块检查过程

在本节中，我们提出分块检查过程。由于初始图中的边数与节点数的平方成正比，如果不使用阈值对链接分数进行剪枝，则不满足约束的边所占比例与节点数成反比。换言之，节点数越多，可移除的边就越多。不满足约束的边不均匀地分布在检查过程的后期。图3展示了这一趋势，该趋势出现在边移除过程。这是因为在流程开始时存在大量冗余边，因此通过移除边更容易满足约束条件。然而，随着流程推进，条件变得更加严格，边移除也更具挑战性。因此，应尽可能在一次检查中移除这些空白区域中的边，而在不可移除边密集的区域，则需要逐条移除边。

基于这些观察，我们提出了一种分块检查过程。图 4 概述了分块检查过程。

首先，将块尺度 s 设为与块大小相关的参数，将块级别 k 设为块处理层的深度，并将 C_k 设为块级别 k 时的块。此外，对于边的数量M，$k_{max} = \lfloor \log_s M \rfloor$ 表示最大块级别，即块处理可以递归执行的次数。第一个 $k=0$ 的块大小为$S_0 = s^k_{max}$。例如，当 $M=5100$ 和 $s=10$ 时，$S_0 = s^{k_{max}} = 10^3$。分块检查过程的算法步骤如下：

1) 按链接分数降序排列，对每条 S_0 边检查步骤 (2) 中的约束。2) 约束检查。如果不满足约束（即块中存在不可移除边），则检查失败，恢复所有边并转到下一步 (3)。即使移除了 S_0 条边，只要满足约束，约束检查就会成功。如果所有边的约束检查均成功，则返回步骤 (1)。3) 令 $S_k = \max(\lfloor S_{k-1}/s \rfloor, 1)$，检查块 C_{k-1}中每 S_k 条边的 C_k 约束。4) 如果识别出不可删除边，则检查块 C_{k-1}中其余边的约束。如果满足约束，则完成块 C_{k-1} 的块处理并返回步骤 (2)。如果不满足，则继续执行步骤 (3)，直到块中所有边均通过分块检查。

分块检查过程允许移除不可移除边较少区域中的每个大块。当块尺度较大时，可以快速识别不可移除边众多区域中的不可移除边。

C. 对直径计算的细微修改

原始的 Lomap 实现使用 NetworkX 的 [18] 偏心率方法，但我们已将其改为使用界定直径算法 [19], [20]计算直径。该算法可通过 NetworkX 直径函数中的 usebounds 选项使用。由于需要计算全点对最短路径（ASAP），直径计算通常具有 $O(MN)$ 的计算复杂度。该算法利用节点偏心率与直径之间的下界和上界关系，在与 $O(N)$ 成正比的计算时间内确定精确直径。构造 FEP 图时，也应考虑包括完全图和环图在内的边界情况，

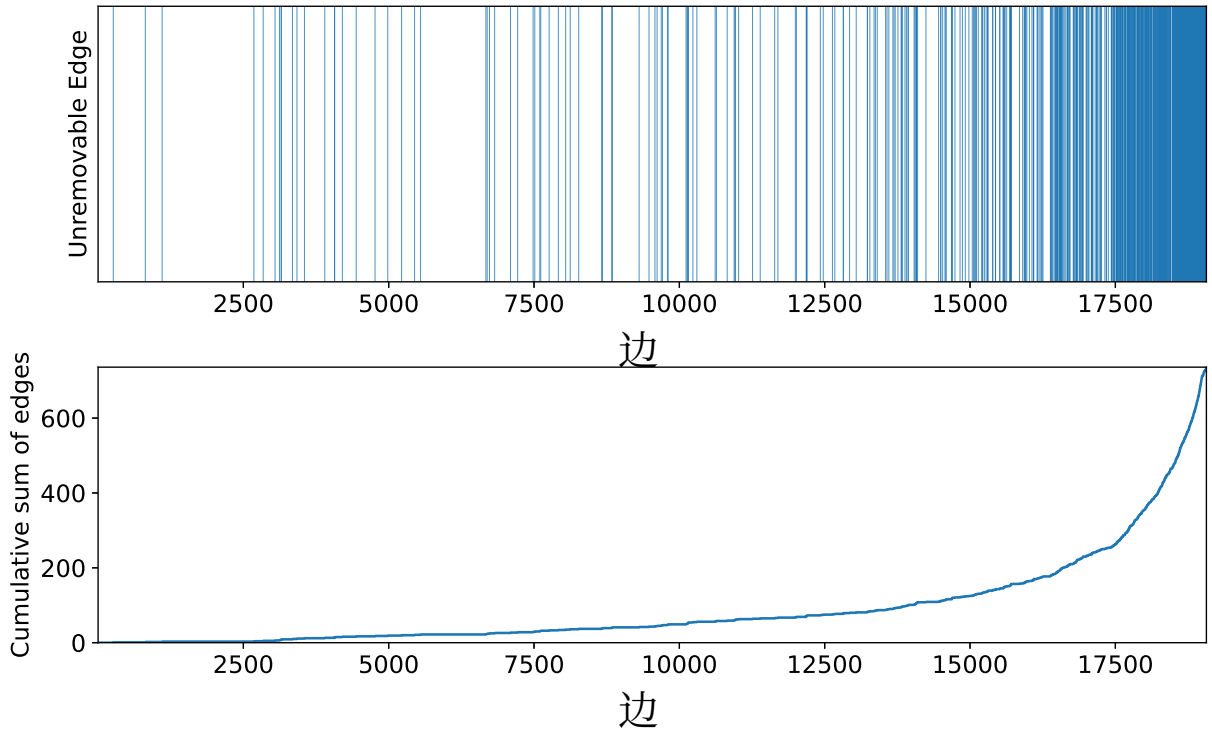


图3. 不可移除边的出现情况及其在约束检查中的累积和（Thompson TYK2 数据集，500 个节点）。在所有可能的边中组合，19,066 条带有链接 s 的边对 0.4 或更高的核心值进行检查，最终得到 740 条边。

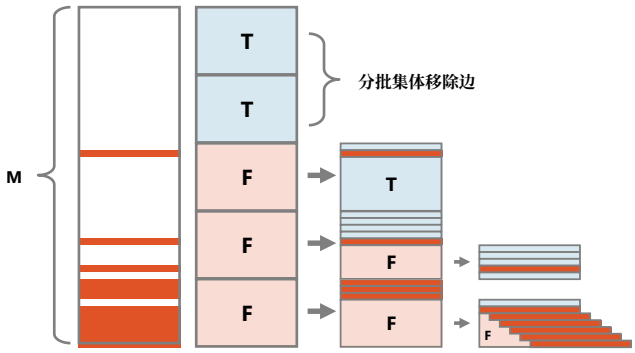


图4. 分块检查处理概览。

但也应在构建 FEP 图时加以考虑。

D. 代码可用性

我们提出的 Lomap 代码改进版本（称为 FastLomap）可在 https://github.com/ohuelab/ FastLomap 获取，其采用与原始 Lomap 相同的 MIT 许可证。

III. 实验

A. Thompson TYK2 数据集

在本研究中，我们使用 100–1000 个节点进行了大规模 FEP 图构建实验。因此，我们采用了 Thompson等人 [21]提供的、以氨基嘧啶骨架为基础的一系列 TYK2 抑制剂。

(a) Thompson TYK2 数据集

我们从原始数据集中随机选取了 1000 个案例，并构建了一个包含 100 至 1000 个节点的实验集子集。

(b) Thompson $TYK2_{0.7}$ 数据集

从与参考配体（蛋白质数据库：4GIH [22]）的链接分数为0.7或更高的原始数据集中，以相同方式从100到1000个案例构建了实验集。

图 5 展示了 Thompson TYK2 数据集和 Thompson $TYK2_{0.7}$ 数据集的初始图中边数与节点数之间的关系。初始图的边经过剪枝，链接分数阈值设为 0.4。因此，由相似化合物组成的 Thompson $TYK2_{0.7}$ 数据集中的边数最多是 Thompson TYK2 数据集的五倍。图 6 展示了两个数据集的连接分数分布。Thompson $TYK2_{0.7}$ 数据集的链接分数峰值高于 Thompson TYK2 数据集。图 7 展示了生成的 FEP 图中每个节点的边数。两个数据集的 FEP 图中边数相近。请注意，所提方法生成的图与现有的 Lomap 生成的图等效。

Wang 数据集

Wang 数据集 [2]被用作评估 RBFE 预测性能的基准，其中八个靶点各包含数十种配体。我们提出的算法

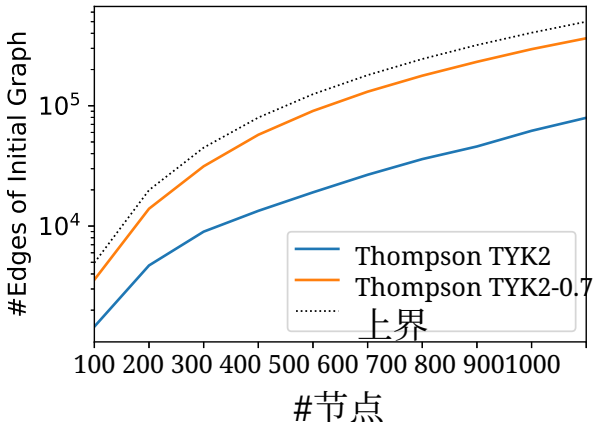


图 5. Thompson TYK2 数据集中，初始图的边数与节点数之比。

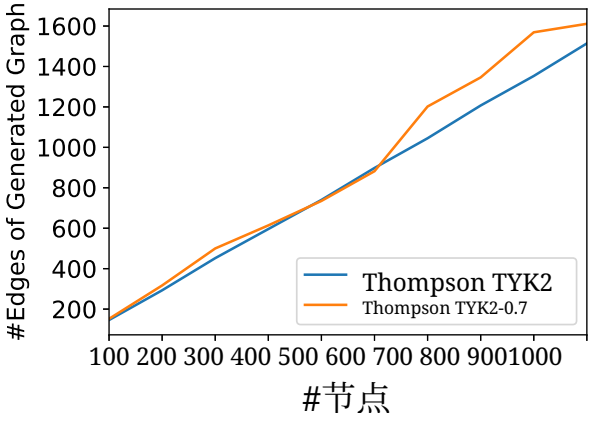


图7。Thompson TYK2 数据集中生成图的边数与节点数之比。

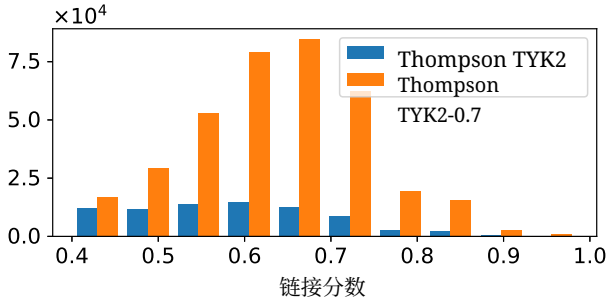


图 6。Thompson TYK2 数据集的连接分数直方图。

假设化合物数量超过 100 种，对于此类数据集，其优势并不明显。不过，我们使用 Wang 数据集来证明所提方法可应用于一般用例。

C. 图生成设置

首先，由于未提供Thompson TYK2 数据集的复合物结构，我们生成了一个复合物结构。复合物结构是使用 CressetFlare [10] 的比对工具（非常宽松选项），通过叠合参考配体（蛋白质数据库：4GIH [22]）生成的。

所有连接分数均使用 Lomap 的默认 3D 模式计算。请注意，对 1000 个化合物的所有配对进行链接分数计算，在 28 个 CPU 上并行处理需耗时 26.4 分钟。

我们将原始 Lomap 计算作为基线进行实验，并在所提方法中采用块尺度 $s = 2$，5，10，50。为生成 FEP 图，我们采用 Lomap 的默认设置，即图直径为 6、不允许使用树结构，且链接分数阈值为 0.4。

IV. 结果与讨论

A. Thompson TYK2 数据集的结果

图 8 和图 9 显示了 Thompson TYK2 和 ThompsonTYK $2_{0.7}$ 数据集中，不同节点数量对应的执行时间和约束检查次数。基线方法

在 Thompson TYK $2_{0.7}$ 数据集中，无法在 24 小时内完成 $N = 500$ 个或更多节点的计算。首先，我们比较了基线方法和所提方法。对于 Thompson TYK 数据集，所提方法最快约为基线方法的 23.7 倍$N = 1000$。在 Thompson TYK $2_{0.7}$ 数据集中，FEP 图的构建速度提高了数百倍 $N = 400$，而基线方法无法为 $N = 500$ 个或更多节点构建 FEP 图。由于剪枝效果不佳，Thompson TYK $2_{0.7}$ 数据集的初始图中包含许多边。因此，显然，约束检查次数与边数相同的基线方法并不适用。因此，所提方法适用于边数较多的情况，因为其约束检查次数取决于节点数。

接下来，对块尺度评分进行了评估。对于两个数据集，当 $s = 5$,10 时，块尺度在执行时间和约束检查次数方面均小于 $s = 2$,50。这比 $s = 2$ 更好，因为更大的块尺度降低了最大块级别 k_{max} 并迅速识别出不可移除边。相反，如果块尺度较大，例如 $s = 50$，由于单个块检查失败会导致拆分为更细的块，因此在不可移除边较少的区域，块处理的效果会变差。此外，需要注意的是，约束检查次数和执行时间并不直接对应，因为删除和恢复边会消耗时间。图 9中 $s = 50$ 的约束检查次数少于 $s = 2$，但由于移除和恢复边所需的时间，执行时间相当。

B. Wang 数据集的结果

图10展示了Wang数据集中每个节点数量对应的执行时间和约束检查次数。尽管执行时间会随节点数量的增加而增加，但即使对于 $s = 5$,10，执行时间仍略短于基线。即使在基线情况下，该过程也耗时

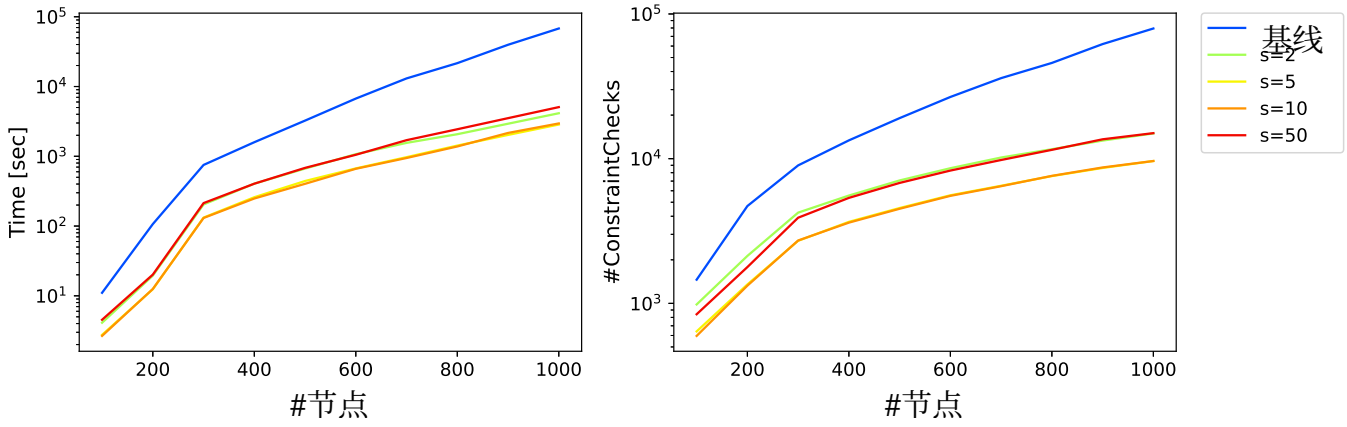


图 8. Thompson TYK2 数据集中的执行时间和约束检查次数。

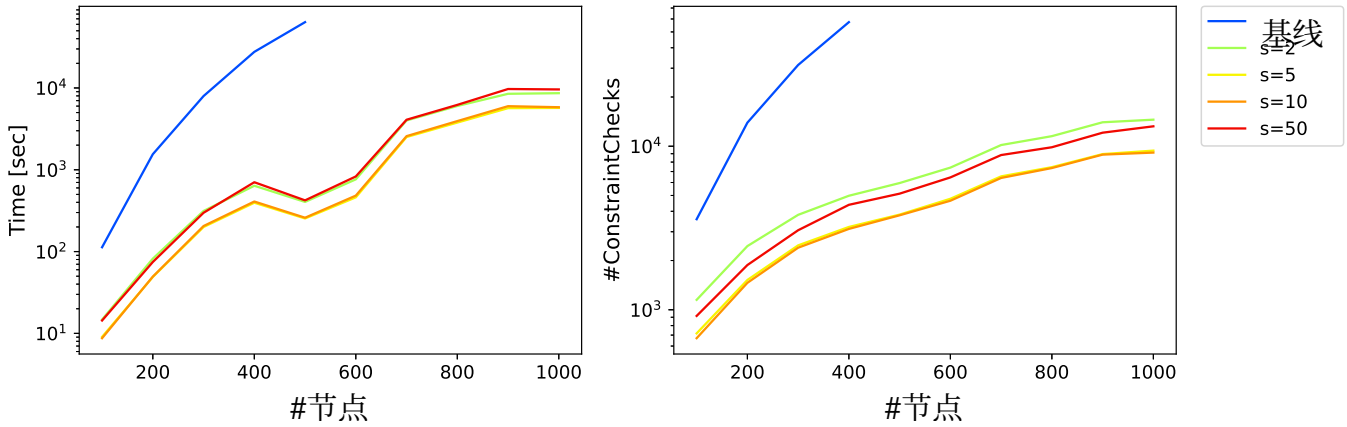


图 9. Thompson $TYK2_{0.7}$ 数据集中的执行时间和约束检查次数。

最多小于 3 秒。因此，所提方法不会对一般用例产生负面影响。

V. 结论

在本研究中，我们提出了一种分块检查流程，以改进 Lomap 算法，使其能够适应不断增加的化合物数量。所提方法在数十个案例中的通用效果可能有限；然而，在数百个案例中应用时，该方法展现出显著的效率，速度提升数十倍。由于所提方法生成的输出图与原始 Lomap 算法生成的图等效，因此可以立即用于替代现有方法。

尽管本研究未涉及这一点，但加速链接分数计算是未来研究需要解决的问题。由于需要计算所有化合物对的链接分数，计算复杂度会呈平方级增长。最近提出的 HiMap [9]使用 DBSCAN方法对相似化合物进行无监督聚类。使用化合物指纹进行聚类，并基于每个簇的最大公共子结构计算分数，可能会降低总体计算成本。此外，还需要进一步讨论 Lomap 的图是否适用于大型FEP设计。分析环闭合冗余对大型FEP图中扰动图准确性的影响至关重要。

对于边较少的案例，我们所提方法生成图的速度最高可比原始 Lomap 快 20 倍；对于边较多的案例，速度最高可快数百倍。所提方法不仅适用于大规模 FEP 实验的规划，也有望用作相对结合自由能计算的主动学习方法。

致谢

本研究部分得到了 JST FOR-EST (JPMJFR216J)、JSTACT-X (JPMJAX20A3)、JSPS KAKENHI(20H04280, 22K12258)和 AMEDBINDS (JP22ama121026) 的支持。

参考文献

[1] K. H. Bleicher, H.-J. B¨ohm, K. M¨uller 等，“命中化合物与先导化合物的生成：超越高通量筛选”，《自然评论：药物发现》，第 2 卷，第 5 期，第 369–378 页，2003 年。[2] L. Wang, Y. Wu, Y. Deng 等，“通过现代自由能计算方案和力场，在前瞻性药物发现中准确可靠地预测配体相对结合效力”，《美国化学学会志》，第 137 卷，第 7 期，第 2695–2703 页，2015年。

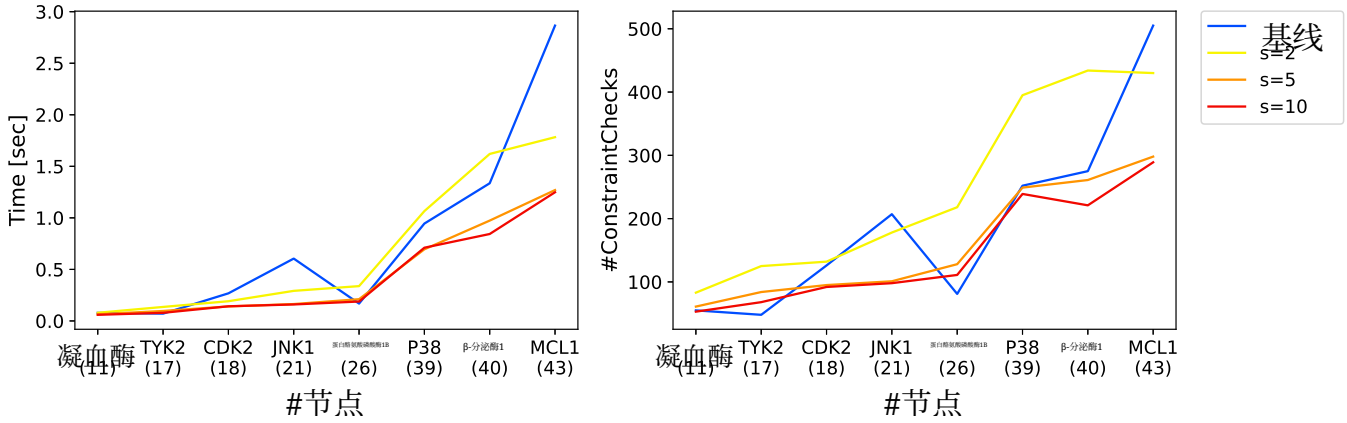


图10。Wang 数据集中的执行时间和约束检查次数。

[3] M. Kuhn、S. Firth-Clark、P. Tosco 等，“通过炼金术自由能计算评估结合亲和力”，《化学信息与建模杂志》，第60卷，第6期，第3120–3130页，2020年。[4] V. Gapsys、L. P´erez-Benito、M. Aldeghi 等，“利用非平衡炼金术大规模计算蛋白质-配体结合亲和力”，《化学科学》，第11卷，第4期，第1140–1152页，2020年。[5] I. Muegge 和 Y. Hu，“用于药物发现的炼金术结合自由能计算近期进展”，《ACS药物化学快报》，第14卷，第3期，第244–250页，2023年。[6] C. E. M. Schindler、H. Baumann、A. Blum 等，“活性药物发现项目中结合自由能计算的大规模评估”，《化学信息与建模杂志》，第60卷，第11期，第5457–5474页，2020年。[7] S. Liu、Y. Wu、T. Lin 等，“先导化合物优化映射器：自动化先导化合物优化的自由能计算”，《计算机辅助分子设计杂志》，第27卷，第9期，第755–770页，2013年。[8] H. Xu，“成对差异的最优测量网络”，《化学信息与建模杂志》，第59卷，第11期，第4720–4728页，2019年。[9] M. Pitman、D. F. Hahn、G. Tresadern 等，“要设计可扩展的自由能微扰网络，仅有最优性还不够”，《化学信息与建模杂志》，第63卷，第6期，第1776–1793页，2023年。[10] M. R. Bauer 和 M. D. Mackey，“静电互补性：优化蛋白质-配体复合物结合与选择性的快速有效工具”，《药物化学杂志》，第62卷，第6期，第3036–3050页，2019年。[11] L. Carvalho Martins、E. A. Cino 和 R. S. Ferreira，“PyAutoFEP：集成增强采样方法的 GROMACS 自动化自由能微扰工作流程”，《化学理论与计算杂志》，第17卷，第7期，第4262–4273页，2021年。[12] B. Settles，“主动学习文献综述”，2009年。[13] D. Reker 和 G. Schneider，“计算机辅助药物发现中的主动学习策略”，《今日药物发现》，第20卷，第4期，第458–465页，2015年。[14] D. Reker、P. Schneider、G. Schneider 等，“计算化学基因组学中的主动学习”，《未来药物化学》，第9卷，第4期，第381–402页，2017年。[15] K. D. Konze、P. H. Bos、M. K. Dahlgren 等，“基于反应的枚举、主动学习和自由能计算：快速探索可合成化学空间并优化细胞周期蛋白依赖性激酶2抑制剂的效力”，《化学信息与建模杂志》，第59卷，第9期，第3782–3793页，2019年。[16] F. Gusev、E. Gutkin、M. G. Kurnikova 等，“基于相对结合自由能建模的主动学习引导药物设计与先导化合物优化”，《化学信息与建模杂志》，第63卷，第2期，第583–594页，2023年。[17] Y. Khalak、G. Tresadern、D. F. Hahn 等，“结合主动学习与炼金术自由能的化学空间探索”，《化学理论与计算杂志》，第18卷，第10期，第6259–6270页，2022年。[18] A. A. Hagberg、D. A. Schult 和 P. J. Swart，“使用 NetworkX 探索网络结构、动力学和功能”，载于《第7届 Python科学会议论文集》，2008年，第11–15页。[19] F. W. Takes 和 W. A. Kosters，“确定小世界网络的直径”，载于《第20届 ACM国际信息与知识管理会议论文集》，2011年，第1191–1196页。[20] M. Borassi、P. Crescenzi、M. Habib 等，“基于广度优先搜索的快速直径和半径计算：应用于（弱连通）真实世界图中的六度分隔游戏”，理论计算机科学，第586卷，第59–80页，2015年。[21] J. Thompson、W. P. Walters、J. A. Feng 等，“优化用于自由能计算的主动学习”，人工智能生命科学，第2卷，100050页，2022年。[22] J. Liang、V. Tsui、A. Van Abbema 等，“新型选择性 TYK2 抑制剂的先导化合物鉴定”，欧洲药物化学杂志，第67卷，第175–187页，2013年。