arXiv:2303.10179v1 [cs.LG] 17 Mar 2023

# 受QUBO启发的分子指纹：化学性质预测

Koichiro Yawata日立有限公司研究与开发组日本国分寺市
koichiro.yawata.rt@hitachi.com

Yoshihiro Osakabe日立有限公司研究与开发组日本国分寺市
yoshihiro.osakabe.fj@hitachi.com

Takuya Okuyama日立有限公司研究与开发组日本国分寺市
takuya.okuyama.mn@hitachi.com

Akinori Asahara日立有限公司研究与开发组日本国分寺市
akinori.asahara.bq@hitachi.com

**摘要—分子指纹广泛用于预测化学性质，选择合适的指纹至关重要。我们基于这样的假设生成新的指纹：使用更有效指纹进行预测时，预测性能会更好。我们生成由多个基础指纹组合而成的有效交互指纹。由于计算资源有限，难以评估交互指纹的所有组合。为解决这一问题，我们将搜索更有效交互指纹的问题转化为二次无约束二元优化问题。本研究利用QM9数据集找到了有效的交互指纹。**

索引词—**分子指纹，QUBO，退火机，化合物设计，分子设计。**

## I. 引言

利用机器学习预测化学性质是材料信息学中最热门的研究主题之一。分子指纹得到广泛应用，因为它们可以将分子结构表示为表格数据 [1]。目前已有许多种指纹，选择合适的指纹至关重要 [2]。然而，决定化学性质的因素十分复杂。即使关键因素由多个指纹描述，在机器学习中将它们独立处理也并不有效。事实上，由多个特征相乘得到的交互特征被用作一种特征工程技术[3]。

我们假设，基于更有效指纹的预测性能更好，并据此生成新的指纹。我们通过使用每种指纹将训练样本划分为决策树（深度为 1），并根据化学性质的预测误差评估每种指纹的性能。随后，我们定义交互指纹。交互指纹是由多个基础指纹的乘积描述的指纹，我们尝试从中搜索有效指纹。我们使用现有指纹作为基础指纹。图 1 展示了交互指纹的示例及满足交互指纹的分子。在本研究中，

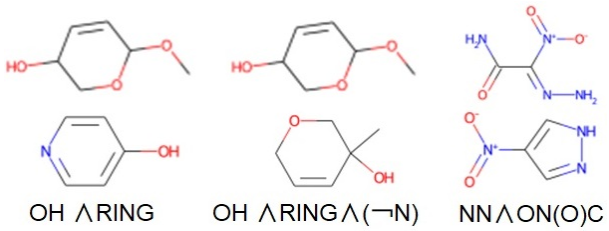


图 1. 使用多个 MACCS键指纹构造交互指纹的示例。OH∧RING∧(¬N)表示交互指纹，即分子含有羟基和环结构，但不含氮原子。

MACCS 键指纹 [2]用作基础指纹，每种基础指纹的表示法均基于 RDKit [4]。

然而，受计算资源限制，难以评估交互指纹的所有组合。为解决这一问题，我们将搜索更有效交互指纹的问题转化为二次无约束二元优化（QUBO）问题，这种方法称为 QUBO决策树 [5]。借助退火机，可以更快地近似求解 QUBO问题 [6]。因此，预期通过使用 QUBO决策树能够获得有效的交互指纹。

## II. 相关工作

**退火机**。关于退火机，无论是在硬件还是软件方面，都已有大量研究。它们在解决组合优化问题方面具有优势，并已应用于现实世界 [7]。量子退火[6] 和绝热量子演化算法 [8] 推动了退火机的发展。D-Wave 在云端提供量子退火机 [9]。此外，也有人推出了采用一种称为动量退火的技术、使用图形处理器的退火机 [10], [11]。

**将机器学习优化为 QUBO 问题**。人们曾尝试将问题表述为 QUBO 问题，以利用退火技术解决机器学习

IEEE 版权声明 版权所有 (c) 2023 IEEE 允许将本材料用于个人用途。已获准发表于
2022 IEEE国际大数据会议（大数据）。数字对象标识符：10.1109/BigData55660.2022.10020236

表 I 交互指纹与指纹向量。

| 交互指纹 | OH | ¬OH | RING | ¬环 | N | ¬N | U |
|---|---|---|---|---|---|---|---|
| OH∧环 | 1 | 0 | 1 | 0 | 0 | 0 | 2 |
| OH∧环∧(¬N) | 1 | 0 | 1 | 0 | 0 | 1 | 3 |

学习算法问题。在 [12],中，退火机已用于聚类。在 [13], 中，线性回归、支持向量机 (SVM) 和平衡 k 均值聚类这三种方法被转化为 QUBO 问题。退火机也已用于深度学习 [14]–[16],图像处理 [17]–[19]以及贝叶斯网络 [20], [21]。

## III. 问题设定

设 $\boldsymbol{X}_{i,\ j}$ 为基础指纹的取值，$j$为样本 $i$的目标值，$y^{(i)}$ 为样本 $i$的预测值，$g^{(i)}$ 为样本 $i$的生成指纹取值。$g^{(i)}$ 可写为:

$$g^{(i)} = \prod_{f_j=1, j\in\mathbb{F}} \boldsymbol{X}_{i,j}, \tag{1}$$

其中，$\mathbb{F}$ 为基础指纹集合，$f$ 为指纹向量。$f_j = 1$表示生成的相互作用指纹使用了指纹 $j$，此类指纹在本研究中称为生成指纹。表 I 第一列中的生成的相互作用指纹为 OH∧环，基础指纹为 OH、¬OH、环、¬环、N 和 ¬N，生成指纹为 OH 和环。设 $N_{\mathbb{F}}$ 为预处理指纹的数量。那么，用于交互指纹的指纹数量 $U$ 可写为:

$$U = \sum_{j\in\mathbb{F}} f_j.$$

由于指纹有 $2^{N_{\mathbb{F}}}$ 种不同的组合，一般而言很难搜索到最优解。因此，本研究采用擅长解决组合优化问题的退火机。此外，设 $\mathbb{S}$ 为所有样本的集合，$\mathbb{S}_1$ 为满足 $g^{(i)} = 1$的样本集合，$\mathbb{S}_0$ 为不满足的样本集合。$\mathbb{S}$ 中的样本数量记为 $N_{\mathbb{S}}$。

**QUBO问题**。在退火机中，要优化的参数由取二进制值 $\theta_l \in \{0, 1\}$的 QUBO变量表示，而目标变量则由 QUBO 形式的哈密顿量表示:

$$H = -\sum_{l<m} Q_{lm}\theta_l\theta_m - \sum_l b_l\theta_l, \tag{2}$$

其中，$Q_{lm}$ 和 $b_l$ 是用于表征 QUBO问题的系数。

通过 QUBO 形式将哈密顿量表示为损失函数 $L$ 和约束函数$C$ 之和，从而构建 QUBO问题，如下所示:

$$H = L\left(\boldsymbol{X}, \boldsymbol{t}|\boldsymbol{\theta}\right) + C(\boldsymbol{\theta})$$

其中，$\boldsymbol{\theta}$ 是表示模型参数的二进制向量。$\boldsymbol{X}$ 是样本 $i$的解释变量，而 $\boldsymbol{t}$ 是表示目标变量的向量。此外，$L$ 是优化目标，$C(\boldsymbol{\theta})$是用于获得有效解的集合，通常要求为 $C(\boldsymbol{\theta}) = 0$。退火机用于寻找使哈密顿量最小化的 $\boldsymbol{\theta}$ 。

## IV. QUBO 决策树

均方误差（MSE）常用于学习决策树。回归决策树中的均方误差可写为:

$$\text{MSE} = \frac{1}{N_{\mathbb{S}}}\Big(\sum_{i\in S_1}\left(\text{pred}_1 - t^{(i)}\right)^2 + \sum_{i\in\mathbb{S}_0}\left(\text{pred}_0 - t^{(i)}\right)^2\Big),$$

其中，当交互指纹满足条件时，$\text{pred}_1$ 是估计值；当不满足条件时，$\text{pred}_0$ 是估计值。回归决策树的学习包括寻找用于分割的交互指纹以及 $\text{pred}_1$,$\text{pred}_0$ ，以降低均方误差。显然，$\text{pred}_1$和 $\text{pred}_0$ 是由交互指纹划分得到的样本的均值。因此，均方误差等于各分割样本组的方差乘以相应样本组比例后求和。据此，均方误差可写为以下公式:

$$\begin{aligned}\text{MSE} &= \sum_{b=0,1} \text{Var}\left(\left\{t^{(i)}\middle| i\in\mathbb{S}_b\right\}\right)\frac{N_{\mathbb{S}_b}}{N_{\mathbb{S}}} \\ &= \sum_{b=0,1}\Big(\frac{1}{N_{\mathbb{S}}}\sum_{i\in\mathbb{S}_b} t^{(i)2} - \frac{1}{N_{\mathbb{S}}N_{\mathbb{S}_b}}\big(\sum_{i\in\mathbb{S}_b} t^{(i)}\big)^2\Big).\end{aligned}$$

然而，由于 MSE 涉及除以变量 $N_{\mathbb{S}_b}$，而该变量是 QUBO 变量之和，因此很难将其表述为 QUBO问题。这里，$N_{\mathbb{S}}$ 是样本总数，因而是常数，所以不存在这个问题。因此，我们提出平方加权均方误差（SWMSE）。在 SWMSE 中，计算方差之和时，不使用样本组所占比例作为权重，而是使用该比例的平方:

$$\begin{aligned}\text{SWMSE} &= \sum_{b=0,1} \text{Var}\left(\left\{t^{(i)}\middle| i\in\mathbb{S}_b\right\}\right)\left(\frac{N_{\mathbb{S}_b}}{N_{\mathbb{S}}}\right)^2 \\ &= \sum_{b=0,1}\Big(\frac{N_{\mathbb{S}_b}}{N_{\mathbb{S}}}\sum_{i\in\mathbb{S}_b} t^{(i)2} - \frac{1}{N_{\mathbb{S}}}\big(\sum_{i\in\mathbb{S}_b} t^{(i)}\big)^2\Big).\end{aligned} \tag{3}$$

通过这一转换，MSE 最小化问题转化为 QUBO问题。

### A. 表述为 *QUBO*问题

有两类需要优化的 QUBO 变量 $\boldsymbol{\theta}_F$和 $\boldsymbol{\theta}_X$，其元素分别表示为 $\theta_{F,\ j}$ 和 $\theta_{X,\ i,\ c}$。首先，$\theta_{F,\ j}$ 是表示生成的相互作用指纹的二元变量；最后，根据计算结果，我们使用 $\theta_{F,\ j}$ 作为 $f_j$ ，并生成新的相互作用指纹。$\theta_{X,\ i,\ c}$ 是用于计算 SWMSE 的辅助二元变量，其中 $\theta_{X,\ i,\ c} = 1$ 表示样本 $i$ 未满足 $c$ 个基础指纹。

IEEE版权声明 版权所有 (c) 2023 IEEE。允许将此材料用于个人用途。已接受并将发表于2022年IEEE国际大数据会议（大数据）。数字对象标识符：10.1109/BigData55660.2022.10020236

请注意，$\theta_{X,\,i,\,0} = 1$ 表示样本 $i$ 满足所有生成指纹，即，$\theta_{X,\,i,\,0}$ 表示划分结果。这里，$c$ 是一个取值范围为 0 到 $M$的整数，其中 $M$ 是用于生成交互指纹的生成指纹的最大数量，也是一个学习参数。样本 $i$ 未满足的生成指纹数量记为未满足的指纹$_i$，可由以下公式表示：

$$\text{unsatisfied_fingerprint}_i = \sum_j (1 - \boldsymbol{X}_{i,j})\,\theta_{F,j}.$$

如果存在未满足的指纹$_i = 0$，则样本 $i$ 满足所有生成指纹。

**损失函数**。方程 (3) 中 $t^{(i)}$ 出现的时刻可以使用 QUBO变量表示，如下列方程所示：

$$\sum_{i\in\mathbb{S}_1} t^{(i)n} = \sum_i \theta_{X,i,0} t^{(i)n} \qquad (n = 1, 2, ..),$$
$$\sum_{i\in\mathbb{S}_0} t^{(i)n} = \sum_i (1 - \theta_{X,i,0}) t^{(i)n} \quad (n = 1, 2, ..).$$

类似地，由交互指纹划分的样本组数量 $N_{S_1}$ 和 $N_{S_0}$可写为如下形式：

$$N_{\mathbb{S}_1} = \sum_i \theta_{X,i,0}$$
$$N_{\mathbb{S}_0} = \sum_i (1 - \theta_{X,i,0}).$$

因此，SWMSE可通过以下方程表示为一个 QUBO 形式问题：

$$\begin{aligned}\text{SWMSE} = &(\sum_i \theta_{X,i,0} t^{(i)2})(\sum_i \theta_{X,i,0}) \\ &- (\sum_i \theta_{X,i,0} t^{(i)})^2 \\ &+ (\sum_i (1 - \theta_{X,i,0}) t^{(i)2})(\sum_i (1 - \theta_{X,i,0})) \\ &- (\sum_i (1 - \theta_{X,i,0} t^{(i)}))^2.\end{aligned}$$

**约束函数**。QUBO 变量有三类约束，如下方公式 (4)–(6) 所示。这里，公式 (4) 是对每个样本中 $\theta_{F,j}$ 与 $\theta_{X,i,c}$ 之间关系的约束。满足此约束意味着 $\theta_{X,i,c}$ 可以表示用于交互指纹的已满足基础指纹数量。公式 (5) 是对每个样本中 $\theta_{X,i,c}$ 有效性的约束。公式 (6) 是用于缩小搜索空间的可选约束。尽管 $M$ 最多可以取值至 $N_B$，但出于实际考虑，我们可以对其加以限制。

$$\forall i \sum_j (1 - \boldsymbol{X}_{i,j})\theta_{F,j} - \sum_c c\theta_{X,i,c} = 0 \tag{4}$$
$$\forall i \sum_c \theta_{X,i,c} = 1 \tag{5}$$
$$1 \le \sum_j \theta_{F,j} \le M \tag{6}$$

表II 生成有效指纹的试验次数。

| | M=2 | M=3 | M=4 | M=5 |
|---|---|---|---|---|
| $N_{\mathbb{S}}$=50 | 0 | 2 | 2 | 2 |
| $N_{\mathbb{S}}$=100 | 0 | 0 | 0 | 3 |
| $N_{\mathbb{S}}$=200 | 0 | 3 | 0 | 1 |

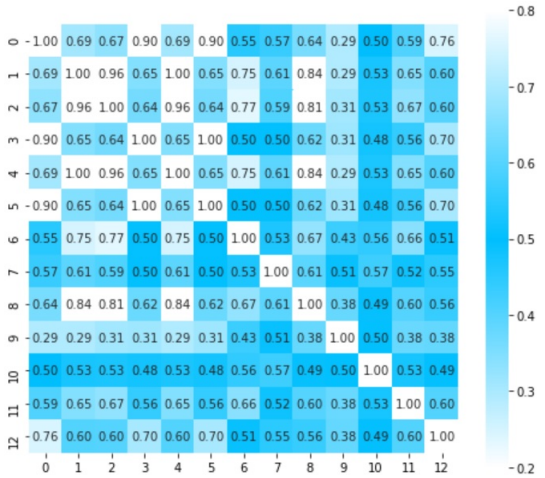


图 2. 展示了生成的交互指纹之间匹配样本的百分比。数值为 0 或 1 表示两个指纹基本相同。

当所有这些约束都得到满足时，实际的 SWMSE 等于退火机中计算出的 SWMSE。

**哈密顿量**。通过最小化下文公式 (7) 中给出的哈密顿量来探索分裂方法。在该公式中，$C_1$、$C_2$ 和 $C_3$ 分别是公式 (4)–(6) 中约束对应的哈密顿量。请注意，这些约束不一定都必须满足。换言之，即使存在违反的约束，也可以使用获得的 $\theta_{F,\,j}$创建估计器。

$$\begin{aligned}H = &\frac{1}{N_S}\text{SWMSE}(\boldsymbol{X}, \boldsymbol{t}|\boldsymbol{\theta}_F, \boldsymbol{\theta}_X) + \frac{1}{N_S}C_1(\boldsymbol{\theta}_F, \boldsymbol{\theta}_X) \\ &+ \frac{1}{N_S}C_2(\boldsymbol{\theta}_X) + C_3(\boldsymbol{\theta}_X).\end{aligned} \tag{7}$$

## V. 实验与讨论

我们使用多个指纹评估了 QUBO决策树是否能够找到有效指纹。有效指纹是指与仅使用一种指纹相比，至少能降低预测误差的指纹。我们使用了 QM9数据集 [22]，预测值为偶极矩。输入值根据 MACCS键生成。实验在图形处理器上使用动量退火进行。针对各种 $N_{\mathbb{S}}$ 和 $M$进行了实验。每种参数均进行了十次试验。为了考察 $M$的影响，将使用相同 $N_{\mathbb{S}}$ 的样本合并。

表II列出了生成有效指纹的试验次数。在 $M = 2$ 的情况下，没有生成有效的交互指纹

IEEE版权声明 版权所有 (c) 2023 IEEE 允许个人使用本材料。已获准发表于2022年IEEE国际大数据会议（大数据）。
数字对象标识符：10.1109/BigData55660.2022.10020236

表 III 获得的有效交互指纹。

| ID | $N_S$ | M | U | I | 生成的相互作用指纹 |
|---|---|---|---|---|---|
| 1 | 50 | 3 | 3 | 1756 | ¬ N-O∧¬ C$=C($A)$A∧¬QCH3 |
| 2 | 50 | 3 | 3 | 821 | ¬ NC(O)O∧¬ACH2N∧¬ 片段 |
| 3 | 50 | 4 | 3 | 1689 | ¬ NC(O)O∧¬QHAQH∧¬ACH2N |
| 4 | 50 | 4 | 2 | 1027 | RING∧¬ QCH3 |
| 5 | 50 | 5 | 2 | 151 | ¬NC(O)O∧¬ ACH2N |
| 6 | 50 | 5 | 2 | 116 | RING∧¬ QCH3 |
| 7 | 100 | 5 | 5 | 2793 | ¬ 7M RING∧¬ NAAN∧¬ ACH2N∧ ¬ N=A∧¬ X (卤素) |
| 8 | 100 | 5 | 4 | 2660 | ¬ N-O∧¬QCH2Q∧¬NC(O)N∧¬ACH2AACH2A |
| 9 | 100 | 5 | 3 | 1900 | ¬ C=CN∧¬QNQ∧¬ACH2QH |
| 10 | 200 | 3 | 3 | 1478 | NCO∧¬ NC(O)N∧¬ 芳香环 |
| 11 | 200 | 3 | 3 | 3399 | ¬ NC(O)N∧¬ CH3CH2A∧¬ 6M 环 |
| 12 | 200 | 3 | 3 | 2844 | ¬ 杂环原子>1(&...)∧ ¬ CH2QCH2∧¬ O=A>1 |
| 13 | 200 | 5 | 6 | 3452 | ¬ N-O∧¬ NC(O)N∧¬ C=C(C)C∧¬ A$A!A$A∧ ¬ A!O!A∧¬ Anot%A%Anot%A |

。表 III 列出了获得的有效交互指纹。 $I$ 是使用所有样本训练时每个指纹的特征重要性。粗略地说， $N_{\mathbb{S}}$ 和 $U$ 越大， $I$ 就越高。如果 $U$ 较低，由于它与基础指纹高度相关， $I$ 往往也较低。如果 $N_{\mathbb{S}}$ 较低，由于样本量较小时存在不稳定性， $I$ 往往也较低。图2显示了生成的交互指纹之间匹配样本的百分比。可以看出，生成了一些相似的指纹，但总体而言，它们各自生成的指纹都不相同。

我们讨论了与朴素穷举搜索方法相比的计算成本。需要搜索的组合数量为$_{N_{\mathbb{F}}}C_M$。当 $M = 3$ 且 $N_{\mathbb{S}} = 50$时，在我们的环境中，朴素穷举搜索模型的计算时间超过三小时，这表明当 $M$ 超过4时，很难进行穷举搜索。使用所提出的方法进行的实验每次均在两小时内完成，即使在 $M = 5$ 时也是如此，这表明我们的方法在计算时间方面具有优势。

## VI. 结论

本研究尝试使用 QUBO决策树寻找用于确定化学性质的有效指纹。结果发现，利用多种指纹的交互指纹能够降低预测误差。如何利用生成的交互指纹以及如何提升性能，仍是未来需要解决的问题。

## 参考文献

[1] A. Mauri，《alvaDesc：一种用于计算和分析分子描述符与指纹的工具》，载于《生态毒理学 QSAR》，《药理学与毒理学方法》，纽约：Humana，2020。[2] J. L. Durant、B. A. Leland、D. R. Henry 和 J. G. Nourse，《重新优化用于药物发现的 MDL 键》，J. Chem. Inf. Comput. Sci.，第 42 卷，第 6 期，第 1273–1280 页，2002 年 11 月。[3] A. Zheng 和 A. Casari，《机器学习特征工程：数据科学家的原理与技术》，美国马萨诸塞州 Newton：O'Reilly Media，2018。[4] "RDKit" 2022， [在线]。获取地址：https://github.com/rdkit/rdkit-orig/blob/master/rdkit/Chem/MACCSkeys.py。[访问日期：06.08。2022][5] K. Yawata、Y. Osakabe、T. Okuyama 和 A. Asahara，《QUBO 决策树：退火机扩展决策树分裂》，即将发表。[6] T. Kadowaki 和 H. Nishimori，《横场伊辛模型中的量子退火》，Phys. Rev. E，第 58 卷，第 5355–5363 页，1998 年 11 月。

[7] T. Stollenwerk 等，《将量子退火应用于空中交通管理中最优轨迹的冲突消解》，IEEE Trans. Intell. Transp. Syst.，第 21 卷，第 1 期，第 285–297 页，2020 年 1 月。[8] E. Farhi、J. Goldstone、S. Gutmann、J. Lapan、A. Lundgren 和 D. Preda，《将量子绝热演化算法应用于随机NP完全问题实例》，《科学》，第 292 卷，2001 年，第 5516 期，第 472–475 页。[9] S. W. Shin、G. Smith、J. A. Smolin 和 U. Vazirani，《D-Wave机器有多"量子"？》，arXiv 预印本 arXiv:1401.7087，2014。[10] T. Okuyama、T. Sonobe、K.-I. Kawarabayashi 和 M. Yamaoka，《通过动量退火进行二元优化》，Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.，第 100 卷，第 1 期，2019 年 7 月。[11] M. Tao、K. Nakano、Y. Ito、R. Yasudo、M. Tatekawa、R. Katsuki 等，《适用于 QUBO 和伊辛模型的工作时间最优并行穷举搜索算法及其图形处理器实现》，《国际并行与分布式处理研讨会论文集》，第 557–566 页，2020。[12] K. Kurihara、S. Tanaka 和 S. Miyashita，《用于聚类的量子退火》，载于《人工智能中的不确定性》（UAI），2009。[13] P. Date、D. Arthur 和 P. Lauren，《用于训练机器学习模型的 QUBO表述》，《科学报告》，第 11 卷，2021 年，第 1 期，第 1–10 页。[14]M. Sasdelli 和 T. -J. Chin，《二值神经网络的量子退火表述》，2021 年数字图像计算：技术与应用（DICTA），2021 年，第 1–10 页，数字对象标识符：10.1109/DICTA52665.2021.9647321。[15] L. M. R. Rere、M. I. Fanany 和 A. M. Arymurthy，《用于深度学习的模拟退火算法》，《计算机科学过程》，第 72 卷，第 137–144 页，2015。[16] G. Sato、M. Konoshima、T. Ohwa、H. Tamura 和 J. Ohkubo，《用于整流线性单元类型函数的二次无约束二元优化表述》，Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.，第 99 卷，第 4 期，2019 年 4 月。[17] T. Birdal、V. Golyanik、C. Theobalt 和 L. Guibas，《量子置换同步》，IEEE计算机视觉与模式识别会议（CVPR），2021。[18] J. Li 和 S. Ghosh，《用于精确目标检测的量子软QUBO抑制》，载于欧洲计算机视觉会议（ECCV），2020。[19] W. Cruz-Santos、S. E. Venegas-Andraca 和 M. Lanzagorta，《面向D-Wave量子退火机的立体匹配问题 QUBO 表述》，《熵》，第 2 卷，2018。[20] B. O'Gorman、R. Babbush、A. Perdomo-Ortiz、A. Aspuru-Guzik 和 V. Smelyanskiy，《使用量子退火进行贝叶斯网络结构学习》，《欧洲物理期刊专题》，第 224 卷，2015 年，第 188 期，第 163–1 页。[21]Y. Shikuri，《将贝叶斯网络学习高效转换为二次无约束二元优化》，arXiv 预印本 arXiv:2006.06926，2020。[22] R. Ramakrishnan、P. O. Dral、M. Rupp 和 O. A. von Lilienfeld，《134 千个分子的量子化学结构与性质》，《科学数据》，第 1 卷，第 1 期，2014 年 8 月。