1 方法

1.1 环境设计

我们将卡宾分子的构建过程建模为一个有限步长的序列决策问题。构建单元（blocks）分为两类：*core* 与 *substructure*。候选集合包含 30 个 core 和 852 个 substructure。初始状态 s_0 为空图。智能体首先必须选择且仅选择一个 core，随后在其反应位点上执行结构扩展。

在任意状态 s_t，动作空间由两类操作组成：*add* 与 *combine*。其中，*add* 表示在选定位点接入一个 substructure；新接入片段可引入额外可反应位点，从而支持后续扩展。*combine* 表示连接两个已有可反应位点（可来自 core 或不同 substructure），用于形成更紧凑的拓扑结构。最终，包含单一 core 且任意数量 substructure 的分子作为有效终止状态。

由于真实评估函数(oracle)的计算代价较高，我们以代理模型 $\hat{f}_\phi$ 近似真实性质映射 f，并将其用于候选分子的快速筛选。对任意分子图 $G = (V, E)$，代理模型的输入由两部分组成：其一是原子/键构成的拓扑与化学特征；其二是图级全局标量 $u = dE\_triplet$。模型首先通过图编码器得到结构表示 $\mathbf{h}_G = \mathrm{Enc}_\phi(G)$，再将 $\mathbf{h}_G$ 与 u 融合并输入回归层，输出三维性质预测向量

$$\hat{\mathbf{y}} = \hat{f}_\phi(G, u) = \big[\hat{y}_1, \hat{y}_2, \hat{y}_3\big] = \big[\widehat{dE\_triplet}, \widehat{vbur\_ratio\_vbur\_vtot}, \widehat{dE\_AuCl}\big].$$

训练阶段采用多目标联合回归，记真实标签为 $\mathbf{y} = [y_1, y_2, y_3]$，则优化目标写为

$$\mathcal{L}_{proxy} = \sum_{k=1}^{3} \lambda_k \, \ell(\hat{y}_k, y_k)\,,$$

其中 $\ell(\cdot,\cdot)$ 为回归损失（实验中使用 MAE），λ_k 为各性质权重。该设计的核心动机是：$dE\_triplet$ 不仅是预测目标，同时也可作为描述电子结构状态的先验信号，与其余目标存在耦合关系；将其显式注入图级表示后，模型可在共享表示空间中学习跨目标相关性，从而提高多目标预测的稳定性与样本效率。推理时，代理模型输出 $\hat{\mathbf{y}}$ 作为后续多目标打分与 Pareto 筛选的依据。

1.2 GFlowNet

传统强化学习（RL）通常以“给定目标下求解单一最优策略”为核心，即倾向于生成奖励最高的一条动作序列。近年来研究表明，在许多实际任务

中，“生成一组具有多样性的高质量候选解”往往比“仅输出一个全局最优解”更有价值，这一点在分子设计与强化学习探索中尤为明显。以分子设计为例，模型不应只给出一个分数最高但难以合成的分子，而应提供一批性能接近最优、但在可合成性等维度上更具可操作性的候选分子，以支持后续实验筛选与决策。

GFlowNet 的目标不是学习单一路径最优，而是学习一个在终止状态集合 $\mathcal{X}$ 上的采样分布 $p_\theta(x)$，使其与奖励函数成正比：

$$p_\theta(x) \propto R(x), \quad x \in \mathcal{X},\ R(x) > 0.$$

在状态转移图中，设 $F_\theta(s \to s')$ 为边流、$F_\theta(s)$ 为状态总流，则对任一非初始且非终止中间状态需满足流守恒：

$$\sum_{s'' \to s} F_\theta(s'' \to s) = \sum_{s \to s'} F_\theta(s \to s').$$

对终止状态 x，其入流等于奖励，即 $F_\theta(x) = R(x)$。在本文场景中，终止状态对应完整分子；我们将代理模型给出的多性质评分聚合为标量奖励并做正值化处理，以保证可用于流匹配训练。

实现上，我们采用前向策略 $P_F(a_t|s_t)$ 逐步执行 *add*/*combine*/*stop* 动作生成分子，并以反向策略 $P_B(s_t|s_{t+1})$ 近似逆过程。给定轨迹 $\tau = (s_0 \to s_1 \to \cdots \to x)$，训练时使用 Trajectory Balance 目标：

$$\mathcal{L}_{TB} = \left(\log Z_\theta + \sum_{t=0}^{T-1} \log P_F(s_{t+1}|s_t) - \log R(x) - \sum_{t=0}^{T-1} \log P_B(s_t|s_{t+1}) \right)^2,$$

其中 Z_θ 为可学习配分函数。该目标直接约束整条生成轨迹的前向概率与终止奖励一致，从而在高奖励区域保持采样强度的同时避免策略塌缩到单一结构，提高候选分子的多样性与可探索性。

1.3 多目标 GFlowNet

单目标 GFlowNet 仅对应一个标量奖励 $R(x)$，而在本任务中每个分子同时对应多维性质向量 $\mathbf{r}(x) = [r_1(x), \ldots, r_m(x)]$。为在一次训练中覆盖不同目标权衡，我们引入偏好向量 $\boldsymbol{\omega} \in \Delta^{m-1}$，并学习条件化策略

$$\pi_\theta(\cdot|s, \boldsymbol{\omega}), \quad P_F(\cdot|s, \boldsymbol{\omega}), \quad P_B(\cdot|s', \boldsymbol{\omega}).$$

其中，训练时偏好并非固定，而是从分布 $p(\boldsymbol{\omega})$ 采样；$p(\boldsymbol{\omega})$ 将直接影响模型覆盖的 Pareto 前沿区域。本文采用

$$\boldsymbol{\omega} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}),$$

并在输入策略网络时对 $\boldsymbol{\omega}$ 使用 thermometer encoding（离散分桶的单调累计编码）以增强偏好条件信号表达。给定 $\boldsymbol{\omega}$ 后，将向量奖励标量化为

$$R(x|\boldsymbol{\omega}) = g(\boldsymbol{\omega}, \mathbf{r}(x)),$$

并引入奖励指数 $\beta > 0$，使目标分布满足

$$\pi(x|\boldsymbol{\omega}) \propto R(x|\boldsymbol{\omega})^{\beta}.$$

该设计会强化策略对 $R(x|\boldsymbol{\omega})$ 模式区域（高奖励峰值区域）的关注，从而更易生成高质量且保持多样性的候选分子。结合 TB 训练目标，可写为

$$\mathcal{L}_{MTB} = \left(\log Z_\theta(\boldsymbol{\omega}) + \sum_{t=0}^{T-1} \log P_F(s_{t+1}|s_t, \boldsymbol{\omega}) - \log R(x|\boldsymbol{\omega}) - \sum_{t=0}^{T-1} \log P_B(s_t|s_{t+1}, \boldsymbol{\omega}) \right)^2.$$

推理阶段通过采样 $\boldsymbol{\omega}$，可获得覆盖 Pareto 前沿不同区域的一组候选分子；不同 $p(\boldsymbol{\omega})$ 的影响在实验部分进行对比分析。

1.4 主动学习

我们采用“生成–评估–回流训练”的主动学习闭环。首先，使用初始标注数据集 D_0 训练代理模型与生成策略。随后在第 k 轮迭代中，模型先生成一批候选分子并由代理模型进行快速打分；再结合预测不确定性与非支配排序（Pareto ranking）选择高信息样本，优先保留位于或接近 Pareto front 的候选；最后对入选样本进行真实评估，并将新获得的标注数据并入训练集 $D_{k+1} = D_k \cup \mathcal{B}_k$，用于下一轮模型更新。该流程在控制真实评估成本的同时，持续提升模型在 Pareto 前沿附近的采样效率与候选质量。

1.5 离线 + 在线，模型迭代最可靠的方式

随着主动学习迭代推进，会累积大量离线数据（包含代理预测样本与真实评估样本）。在大规模、稀疏高分区域的分子搜索任务中，这些数据不仅能提升样本效率，也能显著稳定训练过程。基于此，我们采用“离线 + 在

线”联合迭代范式：离线阶段利用历史数据学习稳定分布，在线阶段依托代理奖励持续探索新区域，实现“利用–探索”的动态平衡。

具体而言，离线部分采用 COFlowNet 框架，通过区分受支持与不受支持转移，缓解纯离线训练中目标分布偏移问题；在线部分采用 GFlowNet-proxy，在离线覆盖不足区域提供更强的外推能力。我们进一步提出不确定性驱动的混合前向策略。状态 s 的不确定性由深度集成（Deep Ensemble）代理模型估计为 $u(s)$，并将其映射为混合系数

$$\alpha_1(s) = \exp\left(-\frac{u(s)}{\tau}\right), \qquad \alpha_2(s) = 1 - \alpha_1(s),$$

其中 $\tau > 0$ 为温度参数。最终前向策略写为

$$P_F(\cdot|s) = \alpha_1(s)\, P_F^{\mathrm{COF}}(\cdot|s) + \alpha_2(s)\, P_F^{\mathrm{proxy}}(\cdot|s).$$

当 $u(s)$ 较小（离线覆盖充分）时，策略更依赖 COFlowNet；当 $u(s)$ 较大（离线覆盖稀疏）时，策略自动提高对在线策略的权重，从而在保证稳定性的同时增强对新颖高价值分子的发现能力。除上述策略外，我们还设计了其他几种混合策略，包括：

- **阈值切换策略（Hard Switch）**：设不确定性阈值 δ，

$$P_F(\cdot|s) = \begin{cases} P_F^{\mathrm{COF}}(\cdot|s), & u(s) \le \delta, \\ P_F^{\mathrm{proxy}}(\cdot|s), & u(s) > \delta. \end{cases}$$

  该策略具有明确的决策边界，但可能在阈值附近产生不连续切换。

- **竞争式 Softmax 融合（Score-based Softmax）**：构造两路置信分数 $q_1(s), q_2(s)$，并令

$$\alpha_i(s) = \frac{\exp(q_i(s)/\tau)}{\sum_{j=1}^{2} \exp(q_j(s)/\tau)}, \quad P_F(\cdot|s) = \sum_{i=1}^{2} \alpha_i(s)\, P_F^{(i)}(\cdot|s),$$

  其中 $P_F^{(1)} = P_F^{\mathrm{COF}}$，$P_F^{(2)} = P_F^{\mathrm{proxy}}$。该策略平滑且可学习。

- **阶段性退火融合（Iteration Annealing）**：在第 k 轮主动学习时设置

$$\alpha_1^{(k)} = \max\left(\alpha_{\min}, 1 - \frac{k}{K}\right), \qquad \alpha_2^{(k)} = 1 - \alpha_1^{(k)},$$

$$P_F^{(k)}(\cdot|s) = \alpha_1^{(k)} P_F^{\mathrm{COF}}(\cdot|s) + \alpha_2^{(k)} P_F^{\mathrm{proxy}}(\cdot|s).$$

  该策略前期偏向离线稳定训练，后期逐步增强在线探索能力。

由此,我们提出了 Offline with Online Multi-Objective GFlowNet(OWOM-GFN)。其核心思想是统一 COFlowNet 的离线稳定性与 GFlowNet-proxy 的在线探索能力：先利用历史数据进行离线流匹配，在受支持转移上学习稳健策略；再在每轮主动学习中基于代理模型对新候选进行快速评估，并将高价值样本回流至训练集持续迭代。具体而言，我们首先将 COFlowNet 扩展到多目标场景。由于离线数据可提供准确的多目标奖励，训练时可通过采样不同偏好向量 $\boldsymbol{\omega}$（或等价地调节 Dirichlet 参数 $\boldsymbol{\alpha}$）学习对 Pareto 前沿不同区域的覆盖。随后，我们引入不确定性驱动的混合前向策略，在“已知高置信区域”偏向离线策略，在“未知稀疏区域”自动提升在线策略权重，从而在样本效率、稳定性与新颖性之间取得更优平衡。考虑到本任务中 oracle 评估耗时显著高于 GFlowNet 训练与采样开销，我们在每轮主动学习迭代中均基于最新数据重新训练离线策略、在线策略与代理模型。

具体的训练与采样流程如算法 1 所示。

Code Listing 1: OWOMGFN training and sampling pipeline

```
# Input:
# D0: initial labeled dataset
# K : active learning rounds
# tau: uncertainty temperature
# alpha: Dirichlet parameter for preference sampling
#
# Models:
# PF_COF, PB_COF       : offline forward/backward policies (COFlowNet)
# PF_proxy, PB_proxy   : online forward/backward policies (GFlowNet-proxy)
# f_hat_ens            : deep-ensemble proxy for reward + uncertainty

D = D0

for k in range(1, K + 1):
    # Stage 1: retrain all models on the latest dataset
    Train(f_hat_ens, on=D, objective="multi-objective regression")
    Train(PF_COF, PB_COF, on=D,
          objective="offline multi-objective TB/flow matching",
          pref_sampler=Dirichlet(alpha))
```

```
    Train(PF_proxy, PB_proxy, on=D, reward=f_hat_ens,
          objective="online multi-objective TB",
          pref_sampler=Dirichlet(alpha))

    # Stage 2: hybrid policy rollout
    C = []  # generated candidates
    for episode in range(N_rollout):
        omega = SampleDirichlet(alpha)
        s = s0
        while not terminal(s):
            u = Uncertainty(f_hat_ens, s)
            alpha1 = exp(-u / tau)
            alpha2 = 1 - alpha1
            PF_mix = alpha1 * PF_COF(. | s, omega) + alpha2 * PF_proxy(. | s
            a = SampleAction(PF_mix)
            s = Transition(s, a)
        C.append(s)  # terminal molecule

    # Stage 3: proxy evaluation + Pareto selection
    Y_hat = PredictObjectives(f_hat_ens, C)  # multi-objective prediction
    U_hat = PredictUncertainty(f_hat_ens, C)
    B = SelectByParetoAndUncertainty(C, Y_hat, U_hat, budget=B_k)

    # Stage 4: expensive oracle labeling and dataset update
    Y_true = OracleEvaluate(B)
    D = D union {(x, y) for x, y in zip(B, Y_true)}

return Trained policies and Pareto-diverse molecule set
```

参考文献