赛题八：基于 AI 的化学反应过渡态结构预测

一、赛题背景

在化学反应研究中，过渡态结构（Transition State, TS）位于反应路径的能量峰值，是连接反应物与产物的“关键桥梁”，其几何构型直接决定反应的活化能和路径选择，是理解和调控化学反应机理的核心。然而，在传统研究中，过渡态结构的获取严重依赖高昂的量子化学计算资源与研究人员的专业经验，效率低且缺乏通用性，成为制约化学反应自动化建模与智能设计的重要瓶颈。

随着人工智能在科学研究中的广泛应用，基于机器学习的过渡态预测方法正逐步兴起，为破解该难题提供了全新思路。目前已有部分研究尝试通过深度学习、图神经网络、生成模型等手段在一定程度上重建反应路径中的 TS 结构，但普遍存在泛化能力不足、对复杂体系适应性弱、数据依赖度高等问题，尚未形成被学界广泛接受的标准化解决方案。

本赛题聚焦 “过渡态结构预测”这一前沿挑战问题，结合最新的开源反应数据集和反应三维结构表示需求，面向算法、化学、材料等多学科交叉背景的参赛者开放，旨在推动高效、通用、准确的过渡态结构预测算法的发展。该问题不仅是计算化学自动化、绿色催化剂设计、药物筛选和新材料发现等领域的基础核心任务，同时也具备广阔的产业转化前景，特别是在新药研发、能源转化、环境化工等方向具有重要现实意义。

通过设置具有代表性与挑战性的反应数据集，本赛题将引导参赛者探索具有创新性和推广性的算法，进一步推动人工智能驱动的科学发现（AI4Science）的发展，助力跨学科人才的培养与行业技术进步，为智能化分子设计与复杂反应系统建模提供关键支撑。

二、赛题应用场景

在新药研发、绿色催化、能源、材料设计等高价值产业中，化学反应路径的理解和调控是核心环节。其中，过渡态结构的准确预测决定了研究人员能否高效评估反应活化能、筛选反应路径以及设计高选择性的催化剂。然而，当前主流的过渡态搜索方法高度依赖量子化学计算和专家经验，不仅计算资源消耗巨大，而且在应对大规模反应筛选任务时效率极低，难以满足高通量化学设计的实际需求。

以药物分子的合成路径设计为例，制药企业在先导化合物的筛选和优化过程中，往往需要探索成百上千条可能的反应路径。若能快速获得每条路径的过渡态结构并计算其能垒，将大大提升反应可行性筛查的效率和准确度，缩短药物开发周期。又如，在电催化材料开发中，不同材料表面对关键中间态的稳定性决定了其反应性能，精确

预测这些关键中间态的几何结构成为提升催化剂活性的关键。

本赛题所提出的过渡态结构预测任务，正是解决上述现实问题的关键一环。通过引导参赛者利用 AI 算法挖掘反应物与产物之间的隐含化学信息，快速生成高可信度的 TS 结构，本赛题有望推动 “AI 辅助反应路径搜索”技术在工业研发和基础科学研究中的实际落地。该任务具有明确的工程需求、现实可行的技术路线和广泛的行业适配性，是 AI 技术向分子科学高端应用延伸的典型代表。

三、赛题任务

本赛题的核心任务是：开发一个能够根据反应物与产物的三维结构预测其过渡态结构的机器学习模型。参赛者需在提供的开源训练数据集上进行模型训练，并在测试集上对未知反应的过渡态结构进行准确预测。

参赛者具体需完成以下工作：

1. 数据处理：从组委会提供的开源反应数据集中提取有效结构信息，包括反应物、产物及过渡态三维坐标，完成数据清洗、标准化与特征工程；

2. 模型构建与训练：基于合适的机器学习算法（如图神经网络、生成模型、机器学习势能面等），构建能够从反应物和产物结构中推断过渡态结构的模型，并完成训练；

3. 过渡态结构预测：针对测试集中给出的反应物与产物结构，使用训练好的模型生成对应的过渡态结构，并以标准的分子坐标文件格式（如 XYZ）输出；

4. 预测评估：模型输出将通过与真实过渡态结构的几何误差（RMSE）进行比较，低误差表示更高预测精度；

5. 结果提交：提交完整的预测结果、模型代码和说明文档，便于组委会进行自动评分与专家评审。

任务范围明确限于“由反应物和产物预测三维过渡态结构”，不涉及反应类型分类、路径搜索优化或电子结构分析等工作，确保参赛者聚焦于结构预测算法本身的构建与优化。

四、数据集及数据说明

本赛题所使用的训练数据来自公开发布的化学反应结构数据集（Transition1x），测试数据为自行采集数据，所有数据来源合法合规，广泛用于前沿机器学习势能面研究与反应路径建模工作中。

（一）数据类型与数据规模

1. 数据类型：数值型三维结构数据，包含反应物（Reactant）、产物（Product）与过渡态（Transition State）的原子坐标。

2. 数据结构：每个反应文件夹包括一个反应三元组（反应物、过渡态、产物），

每个结构以标准 XYZ 文件格式表示，包含原子种类与对应的三维坐标信息。

3. 数据规模：训练集共 10073 条反应，提供反应物、产物和过渡态结构，示例数据可从 https://pan.baidu.com/s/1GSGUE4rnTZBnJVB5FQelSg?pwd=rdfy 中获取；测试集共 1000 条反应（初赛和复赛各 500 条），仅提供反应物与产物结构，需参赛者预测过渡态结构。

4. 数据维度：涵盖多种元素（如 C、H、O、N）与多种类型的反应。

（二）数据分布与覆盖范围

1. 数据广泛覆盖有机化学中的典型反应类型；

2. 包括低能垒、高能垒反应，确保模型泛化能力评估；

3. 原子排序一致，便于结构对齐和几何误差计算。

（三）数据预处理说明

为方便参赛者建模，数据集已完成初步清洗与标准化，具体包括：

1. 原子顺序对齐：确保反应物、产物与过渡态三者之间的原子编号一一对应；

2. 单位规范：所有坐标均以 Å（埃）为单位；

3. 格式统一：结构文件统一采用标准 XYZ 格式，便于通用读取；

4. 异常数据剔除：去除收敛失败或结构不完整的数据项，提升数据质量；

5. 分组说明文件：提供每个反应的数据索引与结构文件路径说明，便于快速定位与批处理。

参赛者可在此基础上进一步进行特征工程，如分子图构建、距离矩阵提取、描述符生成等。

备注：

1. 所有数据将在赛题发布时一并提供，包括训练集结构文件、测试集输入结构文件、基准评估脚本等；

2. 如参赛者需引入额外公开数据增强模型训练，须明确标注数据来源，并确保其合规性。

五、算法设计要求

本赛题推荐参赛者采用监督学习（Supervised Learning）方法，基于提供的反应物、产物与对应过渡态结构三元组数据进行模型训练。算法应能够从已知样本中学习反应物与产物之间的结构映射关系，并在测试集中对未知反应的过渡态结构进行有效预测。

（一）推荐的算法类型包括但不限于

1. 图神经网络（GNN）：适用于分子结构的图表示建模；

2. 结构生成模型（如 GAN、Diffusion Models）：可用于直接输出三维原子坐标；

3. 势能面建模方法（如 SchNet、DimeNet、PhysNet）：建模势能函数以推导过渡态；

4. 基于反应路径搜索的增强方法（如 NEB+ML、GEO-Predictor）：辅助提高结构合理性。

（二）算法优化要求，为保证模型的实用性与可部署性，参赛者在设计模型时需兼顾以下性能指标

1. 预测精度：核心评估指标为模型生成的过渡态结构与真实结构之间的几何误差（RMSE），误差越小越优；

2. 计算效率：鼓励使用轻量化模型结构，减少训练与推理时的计算开销，适应大规模反应筛选需求；

3. 内存与资源占用：建议对模型参数规模、图神经网络深度、三维结构表示维度等进行合理控制，提升运行效率；

4. 泛化能力：模型应具有较强的跨反应类型泛化能力，避免对特定反应结构的过拟合。

（三）算法开发建议

1. 可在结构预测中引入能量或力作为辅助损失函数；

2. 鼓励融合物理知识（如键长约束、反应路径平滑性）以提升预测合理性；

3. 可使用自定义距离损失、坐标对齐算法提升结构精度；

4. 推理时支持批量预测和快速结构构建，以适应大规模评估。

参赛者最终需提交模型代码、说明文档和预测结果，模型应可在标准计算环境中运行并复现测试集预测过程。

六、性能指标要求

（一）平均几何误差（Root-Mean-Square Error, RMSE）

1. 含义：衡量预测过渡态结构与真实结构的整体几何偏差，需通过刚性对齐消除平移和旋转误差。

2. 计算方法：使用 rmsd 工具（https://github.com/charnley/rmsd），通过 Kabsch 算法进行刚性对齐。

目标值：参赛模型的平均 RMSE 需超过基线模型结果，越低越好。

（二）预测成功率（Success Rate）

1. 含义：衡量预测过渡态结构与真实结构在几何误差可接受范围内的反应数量占测试集总数的比例，以反映模型的可靠性和实用性。

2. 计算方法：对测试集中每个反应，若其预测结构的 RMSE≤0.5Å，则判定为“成功预测”，成功率定义为：成功率= 成功预测的反应数/测试集总反应数（500

个）×100%

3. 目标值：参赛模型的预测成功率需超过基线模型结果，越高越好。

（三）推理时间（Inference Time）

1. 含义：衡量模型预测单个反应过渡态结构的效率，反映实际应用中的可行性。

2. 计算方法：在标准计算环境上，计算模型从输入反应物-产物结构到输出过渡态结构的平均时间（单位：秒/反应）。

3. 目标值：单个反应的推理时间越短越好，鼓励参赛者通过模型优化（如模型压缩、并行计算）提升效率。

七、功能要求

参赛者提交的解决方案需实现以下核心功能，覆盖赛题定义的系统功能需求：

（一）数据预处理与特征工程

1. 支持多格式输入：读取反应物/产物结构（XYZ 格式），提取原子类型、坐标、化学键等基础信息。

2. 特征生成：计算几何特征（键长、键角、质心坐标）、图结构特征（原子-键图表示）或物理化学描述符（如 SOAP、分子指纹、库伦矩阵等）。

（二）机器学习模型构建与训练

1. 模型实现：基于机器学习方法（如自回归模型、扩散模型、机器学习势能面）构建过渡态预测模型，支持从反应物-产物特征到过渡态坐标的映射。

2. 训练流程：包含数据加载、模型训练、验证集评估（如按反应类型分层验证），支持保存训练好的模型权重。

3. 可配置参数：允许调整模型超参数（如学习率、批次大小）、训练轮次、特征选择策略等。

4. 评判标准：训练日志显示损失值逐步下降，验证集 RMSE≤基线模型性能。

（三）过渡态结构预测

1. 输入：反应物与产物的结构文件（XYZ）。

2. 输出：预测的过渡态结构文件（XYZ 格式），包含原子类型与三维坐标，文件名需符合规范（如 ts_pred.xyz）。

3. 评判标准：输出 XYZ 文件格式正确（原子数、坐标精度符合规范），平均 RMSE 和成功率均优于基线模型。单个反应的预测时间需在指定硬件下满足推理时间指标，批量处理耗时与反应数呈线性关系（无异常延迟）。

（四）可复现性与文档

1. 代码完整性：提供完整可运行的代码（含数据预处理、模型训练、预测脚本），支持依赖环境一键部署（如 requirements.txt）。

2. 模型说明文档：描述数据预处理流程、特征工程细节、模型架构图及关键参数。提供示例运行命令与输出文件说明。

3. 评判标准：代码无语法错误，示例输出与预期一致，依赖库版本冲突可通过文档指引解决。

八、开发环境

（一）软件环境

1. 操作系统：Linux
2. 编程语言：Python 3.x （需兼容科学计算库与机器学习框架）
3. 开发工具：不限
4. 相关库和框架：PyTorch、TensorFlow、RDKit、OpenBabel、Scikit-learn 等

（二）硬件环境

1. CPU 型号：不限
2. 内存：32 GB RAM 以上
3. GPU 型号：支持 CUDA 即可

九、成绩评价

（一）指标权重

指标类别	具体指标	权重	说明
核心指标	平均几何误差（RMSE）	40%	反映结构预测精度
	预测成功率	30%	衡量可接受误差的反应比例
	推理时间	10%	体现模型实际应用效率
附加因素	代码规范性与可复现性	10%	考察工程实现质量
	报告质量与创新性	10%	评估方法描述清晰度与技术突破

（二）评分细则

1. 平均几何误差（RMSE，40 分）

RMSE≥0.5　0 分

0.2<RMSE<0.5　40 - （（RMSE - 0.2） / 0.3） * 40 分

RMSE≤0.2　40 分

2. 预测成功率（30 分）

分数=成功率×30 分

3. 推理时间（10 分）

由于不同机器学习方法的时间尺度不一致，这里需结合实际使用的方法与基线模

型进行对比，由专家进行评判。

4. 代码规范性与可复现性（10 分）

评分项	评分标准
代码结构	模块化设计合理，目录清晰（2 分）
注释与文档	关键算法有注释，参数说明完整（2 分）
依赖管理	提供完整 requirements.txt 且环境可复现（2 分）
可扩展性	支持命令行参数配置，易于替换模型组件（2 分）
结果复现	按文档步骤可复现测试结果（2 分）

5. 报告质量与创新性（10 分）

评分项	评分标准
内容完整性	包含模型架构图、特征工程细节、对齐方法说明（4 分）
分析深度	对比不同方法性能，讨论误差来源（2 分）
创新性	提出新模型架构或优化现有方法（4 分）

备注：在初赛和复赛阶段，总分数仅由平均几何误差和预测成功率决定，满分 70 分。

十、解题思路

（一）主要考核知识点

1. 分子结构表示与特征工程

（1） 几何特征：原子坐标、键长/键角、质心对齐等空间信息的编码。

（2） 图结构建模：将分子表示为原子-键图，利用图神经网络（GNN）提取局部/全局特征。

（3） 物理化学描述符：SOAP、ACSF、分子指纹等特征的生成与应用。

2. 机器学习与深度学习模型

（1） 生成模型：VAE、扩散模型在连续坐标生成中的应用。

（2） 图神经网络：GAT、GCN、SchNet 等模型的实践应用。

（3） 物理启发模型：机器学习势能面（ML-PES）、强化学习（RL）与路径搜索方法、分子动力学（MD）的结合。

3. 泛化能力与算法优化

（1） 数据增强：坐标微扰、反应类型混合等提升模型鲁棒性的方法。

（2） 模型压缩：知识蒸馏、轻量化网络设计以平衡精度与推理效率。

（二）基本解题思路

1. 数据处理：从结构到“可计算的信息”

（1）核心目标：把反应物/产物的三维结构（XYZ 文件）转化为模型能理解的“特征”。

（2）关键思路：提取分子的几何特征（如原子间距离、键角）或图结构特征（分子作为 “图”，原子是节点，键是边）。对比反应物和产物的结构差异（如哪些键断裂/形成，原子位置如何变化），这些差异是预测过渡态的关键线索。

2. 预测与优化：从模型输出到可用结果

（1）核心目标：让模型输出的结构符合化学合理性，并以标准格式（如 XYZ）输出。

（2）关键步骤：模型可能直接预测原子坐标，或预测“从反应物到过渡态的结构变化”（如原子位移）。用简单的几何优化或能量计算（如基于物理的优化方法）修正模型预测结果，确保结构合理。

3. 泛化能力：让模型适应未知反应

（1）核心挑战：测试集中的反应可能和训练数据类型不同，模型需要“举一反三”。

（2）关键策略：用多样化的训练数据（涵盖不同反应类型），或对数据进行“扰动”（如轻微改变结构），增强模型抗干扰能力。利用迁移学习，先在类似任务（如分子生成）上预训练模型，再针对过渡态预测微调。

十一、参考资源

NeuralNEB—neural networks can find reaction paths fast - IOPscience

Optimal transport for generating transition states in chemical reactions | Nature Machine Intelligence

Machine learning transition state geometries and applications in reaction property prediction | Theoretical and Computational Chemistry | ChemRxiv | Cambridge Open Engage

Transition1x - a dataset for building generalizable reactive machine learning potentials | Scientific Data

Comprehensive exploration of graphically defined reaction spaces | Scientific Data

十二、提交要求

（一）初赛提交内容及要求

1. 核心代码文件

（1）格式：Python 语言脚本，需提供依赖环境配置文件（如 requirements.txt）。

（2）要求：包含数据预处理、模型训练核心逻辑，代码需添加关键注释，可复现基础模型搭建与训练流程。

2. 模型文件

（1）格式：通用机器学习框架保存格式（如 PyTorch 的.pth 格式）。

（2）要求：提交训练后的基础模型，支持读取测试集数据并输出过渡态结构预测结果。

3. 技术方案报告

（1）格式：PDF，格式参考模版。

（2）要求：简述数据处理思路、模型选择依据、训练策略及初步性能指标（如训练集预测误差），突出创新性与可行性。

（二）复赛提交内容及要求

1. 核心代码文件

（1）格式：Python 语言脚本，需提供依赖环境配置文件（如 requirements.txt），README 说明文件，明确各模块功能与运行指令。

（2）要求：在初赛基础上增加模型优化（如调参、算法改进）、结果后处理逻辑等思路。

2. 模型文件

（1）格式：同初赛。

（2）要求：提交在复赛数据集上达到指定性能阈值的模型，支持读取测试集数据并输出过渡态结构预测结果。

3. 技术方案报告

（1）格式：PDF，格式参考模版。

（2）要求：详述数据增强策略、模型训练细节（如超参数搜索方法）、误差分析及泛化性验证结果，附可视化对比图（如预测结构与真实结构的叠合图）。

（三）总决赛提交内容及要求

1. 全流程可复现工程

（1）格式：完整项目压缩包（含代码、数据预处理脚本、模型文件、运行脚本）。

（2）要求：在指定环境下可一键复现从数据处理到结果输出的完整流程，提供详细环境配置说明。

2. 技术方案报告

（1） 格式：PDF，格式参考模版。

（2） 要求：系统性阐述技术创新点（如算法改进的理论依据）、与前沿方法的对比分析，结合实际案例验证模型泛化能力。

3. 答辩材料

（1） 格式：PPT。

（2） 要求：PPT 提炼核心技术与成果，需清晰展示研究思路、技术突破及应用价值。

十三、联系方式

赛项交流 QQ 群：956966549

邮　　箱：libowen990807@163.com

报名官网：www.aicomp.cn

附录、比赛流程及奖项设置

（一）报名阶段

由参赛者在比赛官方网站上完成报名注册，提交个人或团队信息，获取初赛数据下载链接。

（二）初赛阶段

参赛者利用赛事方提供的训练数据集进行算法模型设计，利用提供的初赛测试集进行相应方法的验证与调试。初赛阶段参赛者每天提交结果的次数不限，但是初赛排行榜每隔 1 小时刷新一次。

（三）复赛（省赛）阶段

初赛结束后进入复赛阶段，开放复赛数据下载链接。仅有初赛阶段提交有效结果的参赛团队可以进入复赛。复赛期间，参赛者利用赛事方提供的复赛阶段数据进行算法模型调试，提交对复赛测试数据的推理结果。复赛阶段持续 3 天，每个参赛队伍每天仅能提交 2 次。复赛排行榜每隔 1 小时刷新一次。

（四）复赛（省赛）成绩公布

在比赛官方网站上公布复赛成绩。以进入复赛参赛团队数量作为计奖基数，按照不超过大赛省赛设奖比例，评选出复赛一、二、三等奖（颁发省赛获奖证书）。评选复赛奖过程中，参赛者提交的算法性能低于赛事方提供的基线参考分数的判定为无效成绩，不予授奖。复赛一、二等奖晋级参加国赛总决赛。

（五）决赛（国赛）阶段

1. 决赛线上评选。晋级决赛的参赛团队，依据复赛排行榜结果，以进入决赛参赛团队数量作为获奖基数，按照不超过大赛国赛设奖比例，评选出国赛一等奖候选名单及国赛二、三等奖获奖名单（颁发国赛二、三等奖证书）。

2. 决赛作品提交。国赛一等奖候选团队在规定时间内提交技术文档、算法代码和模型文件、演示视频、补充材料等。提交截止后，不再接受任何形式的修改和补充。

3. 决赛审核阶段。由专业评审团队对国赛一等奖候选团队的参赛作品进行结果复现与审核。评审过程中如有疑问，可要求参赛者进行解释说明。

4.决赛线下答辩。国赛一等奖候选团队在规定时间内提交完善后的技术文档、算法代码和模型文件、演示视频、补充材料，参加国赛线下总决赛复核答辩，最终依据算法性能得分和线下答辩得分确定国赛一等奖获奖名单及其排名（未参加线下复核答辩视同放弃奖项）。国赛一等奖颁发荣誉证书。