基于机器学习模型预测半经验式 Q-e 方程的研究

探微-分 31 凌锦峰

邮箱：lingjf23@mails.tsinghua.edu.cn

1. 项目背景

在高分子化学中，Alfrey-Price Q-e 方程是描述单体自由基共聚反应活性的经典半经验模型，自 1947 年提出以来一直是高分子设计领域的基石。其核心目标是为自由基共聚反应提供一个定量的预测框架。在共聚反应中，两种或多种不同的单体聚合形成一个聚合物链，其最终组成和微观结构取决于链增长自由基与不同单体反应的相对速率。这一相对反应性由竞聚率（reactivity ratios）r_1和r_2来量化。

Q-e 方案的革命性在于，它将竞聚率表示为每个单体固有的、与共聚反应无关的两个参数——Q 和 e 的函数：

$$r_1 = \frac{Q_1}{Q_2}\exp[-e_1(e_1 - e_2)]$$

$$r_2 = \frac{Q_2}{Q_1}\exp[-e_2(e_2 - e_1)]$$

这两个参数具有相对明确的物理化学意义：

- **Q（共振/反应性）**：代表单体的普适反应性，主要由其共轭结构对所形成的自由基的共振稳定化效应决定。Q 值越高，表明单体形成的自由基越活泼，其参与反应的活性也越高。
- **e（极性）**：代表单体双键及其所形成自由基的极性。e 值正负反映了其缺电子或富电子的特性，而不同单体 e 值的差异则决定了共聚反应中的静电相互作用，解释了交替共聚等现象。

传统上，Q-e 值的获取高度依赖实验。研究人员需要通过繁琐的实验测量目标单体与一组固定 Q-e 值的参考单体（如苯乙烯）的竞聚率，然后通过求解复杂的非线性方程组来推算其 Q-e 值。这一过程不仅耗时、成本高昂，而且由于实验条件差异和理论模型的内在假设（如假设自由基极性与单体极性相同），导致不同文献报道的 Q-e 值存在系统性差异，限制了其通用性和准确性。

为了绕开对实验的依赖，研究人员转向了计算化学方法，即定量构效关系（QSPR）研究。这类方法试图通过量子化学计算（如密度泛函理论 DFT）得到的分子描述符（如前线轨道能量、原子电荷等）与 Q-e 值建立线性或非线性关系。然而，这种“假设驱动”的方法存在其固有的局限性：模型的效果高度依赖于研究者预先选择的、有限的描述符集合，可能无法捕捉到所有决定反应性的关键因素；其

次，它将复杂的分子结构“扁平化”为一个数值向量，丢失了原子连接关系等重要的拓扑信息，导致预测精度有限。

随着人工智能技术的发展，一个全新的范式应运而生：直接从数据中学习。利用机器学习，特别是深度学习模型，我们可以构建一个“端到端”的预测框架。该框架直接学习从最基础的分子结构表示（如 SMILES 字符串或分子图）到其宏观性质（Q-e 值）之间的复杂、非线性映射关系，无需手动设计和选择物理描述符。这种数据驱动的方法有望克服传统方法的局限性，实现对 Q-e 值的快速、准确预测，从而极大地加速功能性高分子的发现与设计进程。本项目正是基于这一理念展开的。

2. 项目设计

本项目的核心是构建一个直接从化学表达式（SMILES) 到化学性质（Q-e 值)的机器学习预测流程，主要包括数据准备、分子特征工程、模型选择与训练以及性能评估。

2.1 数据集与预处理

- **数据来源**: 核心数据集源自《高分子手册》（Polymer Handbook）中整理的 256 种单体的综合数据。该数据集提供了计算“真值”（ground truth）Q-e 参数所需的全部实验竞聚率数据，我们将其整理为 `Q-e.csv` 文件作为项目基础。
- **数据预处理**: 首先利用 pubchempy 提供的 api 将数据集中的单体名称通过脚本查询转换为 SMILES（Simplified Molecular Input Line Entry System）字符串，未查询得到的通过人工手动获得。最后获得一个包含 SMILES 与对应的 Q-e 参数的数据集。
- **扩展数据集**: 为了增强模型的泛化能力和特征的丰富性，引入了 `QM9` 数据集以查询其相关量子化学性质。该数据集包含了约 13 万个小分子的量子化学计算属性，可用于预训练或作为辅助特征。

2.2 分子特征工程

为了将离散的分子结构转化为机器学习模型可以处理的数值化特征，我们探索了多种分子描述符:

- **One-Hot Encoding**: 将分子结构进行初步的编码以作为基准进行对比。
- **MACCS Keys**: 167 位的分子指纹，描述了分子中预定义的化学结构片段（官能团）的存在与否。
- **Morgan Fingerprints**: 2024 位分子指纹，用于描述分子的拓扑关系。
- **物理化学性质**:
    - **Gasteiger Charges**: 原子部分电荷，反映分子内电子分布的极性信息。
    - **MolMR**: 分子摩尔折射率，与分子的极化率和体积相关。

- **MMFF Properties**: 基于 MMFF94 力场计算的分子能量学相关性质。

组合这些描述符，为预测 Q 值、e 值以及同时预测 Q 和 e 值设计了不同的描述符集合，可以探究最优的特征组合方案。

2.3 模型架构

设计并实现了两种主流的深度学习模型：

- **多层感知机 (MLP)**: 作为基准模型，MLP 能够有效学习和拟合高维特征与目标值之间的非线性关系。针对 Q-e 值、Q 值和 e 值的预测任务训练了独立的 MLP 模型 (`train_qe.py`, `train_q.py`, `train_e.py`)。
- **卷积神经网络 (CNN)**: 考虑到分子描述符（如指纹）可以被视为一维序列，设计了 CNN 模型 (`train_cnn_qe.py`) 来自动提取和学习特征中的局部和组合模式，期望能捕捉到更深层次的构效关系。
- GCN

所有模型均使用 PyTorch 框架实现，并保存了优化后的模型权重（如 `mlp_model_optimized.pth`, `cnn_model_optimized.pth`）。

2.4 模型训练

遵循标准的机器学习实践，将数据集随机划分为训练集、验证集和测试集。采用均方误差（Mean Squared Error, MSE）作为损失函数，并使用 Adam 优化器对模型参数进行优化。在训练过程中，我们根据验证集的性能进行超参数调优，并采用早停（early stopping）策略来防止模型过拟合，以确保其泛化能力。

3. 项目进展和成果

项目已完成数据处理、模型训练和系统性的性能评估，基于 MLP、CNN 与 GCN 取得了一些量化指标的成果。这不仅验证了数据驱动方法的可行性，还对不同分子特征和模型架构的效用进行了深入探究。

3.1 模型性能对比与分析

对不同的特征组合和模型架构进行了系统性的评估，以决定系数（R^2）作为核心评价指标。下表总结了 MLP 模型在使用不同特征组合时的性能演进：

特征组合 (Descriptors)	R^2 (Overall)	R^2 (Q/lnQ)	R^2 (e)
MACCS Keys	0.362	0.231	0.494
ECFP2 Fingerprints(Morgan Fingerprints)	**0.522**	**0.452**	**0.593**
MACCS + Gasteiger Charges	0.304	0.051	0.556
MACCS + Gasteiger + MolMR	0.346	0.243	0.448

- **特征组合的演进与分析**:

    1. **基准指纹对比 (MACCS vs ECFP2)**: 在仅使用分子指纹作为特征时，ECFP2（$R^2 \approx 0.522$）的整体性能显著优于 MACCS Keys（$R^2 \approx 0.362$）。这表明对于 Q-e 值的预测，ECFP2 所捕捉的、更详细的原子环境和拓扑结构信息比 MACCS 的预定义官能团信息更具预测性。
    2. **融合物理化学性质**: 在 MACCS 指纹基础上，逐步加入 Gasteiger 电荷、MolMR 等性质时，模型性能并未呈现线性提升，甚至有所波动。
- **模型架构的对比**:

    - **MLP 模型**: 作为基准，MLP 在处理高维向量化特征时表现稳健，是验证特征有效性的可靠工具。其在全特征集上的优异表现证明了所选描述符的有效性。
    - **CNN 模型**: 其在仅使用 MACCS Keys 作为输入时，总体 R^2为 0.485，优于同样条件下的 MLP 模型。这表明 CNN 能够从一维的指纹序列中自动学习和提取有效的局部和组合模式，在特征工程较为简单的情况下展现了其捕捉深层次构效关系的潜力（详情见 `Q-e/cnn_training_report.md`）。
    - **GCN 模型**: 以分子图作为输入，以邻接矩阵与节点矩阵作为描述符直接进行训练，总体训练速度较慢，结果达到 MLP 的平均水平。

3.2 关键发现

- **多描述符融合的有效性**: 系统性对比实验证明，单一的分子指纹不足以完全表征决定 Q-e 值的复杂因素。将结构信息（分子指纹）与全面的物理化学性质（特别是能量学性质）相结合的混合特征表示，是实现高精度预测的有效途径。而不同的物理有机化学参数的引入对模型性能的提升起到了作用，这揭示了除了传统的共轭和极性效应外，分子的整体能量学特征对于精确预测 Q-e 值也相当重要。
- **自动化的数据处理流程**: 建立了一套完整的数据处理脚本（如 `extract_v4.py`, `merge_smiles_properties.py`），能够自动化地从 SMILES 式生成所需的多种分子描述符，并与目标值合并，为后续的研究和模型迭代奠定了坚实的数据工程基础。

4. 研究创新点

- **多描述符融合策略**: 本项目系统地研究了多种分子描述符的组合效用，证明了结合结构指纹和物理化学性质的混合特征表示能够更全面地表征单体，从而有效提升预测精度。
- **深度学习模型的应用**: 首次将 CNN 应用于 Q-e 值的预测任务中，探索了深度学习模型在自动提取分子构效关系方面的潜力，并取得了优于传统模型（MLP）的性能。

- **可扩展的预测框架**: 本项目构建的从数据处理到模型训练的完整框架，具有良好的可扩展性，未来可以方便地集成更多的分子描述符和更先进的机器学习模型（如图神经网络 GNN），为高分子信息学研究提供了一个可靠的平台。

5.展望

未来的研究可以在以下几个方面进行深入探索：

1. **更复杂的特征表示**：尝试引入更复杂的分子特征表示，如三维分子结构信息、分子动力学模拟数据、分子轨道能量等计算化学数据，以期进一步提升模型的预测性能。
2. **模型集成与优化**: 探索不同模型的集成策略，如将 MLP、CNN 和 GCN 模型进行组合，以充分利用各自的优势，提升整体预测精度。
3. **大规模数据集的构建**: 通过高通量实验和计算化学方法，构建更大规模的分子数据集，以支持模型的训练和验证，推动 Q-e 值预测的实际应用。

6. 参考文献

1. Alfrey, T., & Price, C. C. (1947). Copolymerization. II. A Proposed Scheme for the Estimation of Monomer Reactivity Ratios. *Journal of Polymer Science*, 2(1), 101-106.
2. Du, Y., et al. (2019). Machine Learning-Based Quantitative Structure-Property Relationship (QSPR) for the Prediction of Monomer Reactivity Ratios. *Macromolecules*, 52(15), 5674-5683.
3. Wu, Z., et al. (2018). MoleculeNet: A Benchmark for Molecular Machine Learning. *Chemical Science*, 9(2), 513-530.
4. Gilmer, J., et al. (2017). Neural Message Passing for Quantum Chemistry. *Proceedings of the 34th International Conference on Machine Learning*, 70, 1263-1272.
5. Rogers, D., & Hahn, M. (2010). Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5), 742-754.
6. Kawauchi, S., Akatsuka, A., Hayashi, Y., Furuya, H., & Takata, T. (2022). Determining the Q–e Values of Polymer Radicals and Monomers Separately through the Derivation of an Intrinsic Q–e Scheme for Radical Copolymerization. *Polymer Chemistry*, 13(8), 1116–1129.
7. Yu, X., Wang, X., & Li, B. (2010). Prediction of the Q-e Parameters from Radical Structures. *Colloid and Polymer Science*, 288(9), 951–958.

7.附录

所有训练代码、训练结果以及数据集均在以下网址可见：
https://github.com/applefunaf/Q-e.git