Result

Background & Issues

背景 (Background)

二芳基乙烯(**DAE**)是光开关分子的明星材料，但其**热稳定性（半衰期）** 对分子结构极其敏感，预测难度极大。微小的取代基变化可能导致半衰期从“几秒”变为“几万年”。传统的试错法效率低下，本项目通过引入**机器学习**，结合 **深度学习分子表征(ChemBERTa)** 与物理化学描述符（芳香性、电荷）**，旨在构建一个能够精准预测 DAE 半衰期的智能模型，从而指导特定功能光开关分子的理性设计。

问题（Issues）

本项目的最大瓶颈在于**高质量实验数据的极度稀缺。**这种“小样本”特性限制了 ChemBERTa 等深度学习模型的微调效果，导致高维特征容易引入噪声，限制了模型的**鲁棒性**——即在面对结构差异较大的新分子时，预测误差可能显著增加。

曾试图利用 **DFT 计算**来补充数据，但因时间和算力限制难以规模化实现。同时，尝试引入的**虚拟设计分子**往往表现为严重的离群值，由于会引入偏差，因此这部分数据最终**未予采用。**

方法总结（Methods）

本研究采用**随机森林（Random Forest）** 算法构建预测模型。在特征工程阶段，我们采用了一种融合**物理化学描述符（Physicochemical Descriptors）** 与**深度学习表征（Deep Learning Representations）** 的混合策略。

具体而言，首先利用 **MMFF（Merck Molecular Force Field）** 对分子进行构象搜索与几何优化，基于优化后的构象提取芳香性、部分电荷分布及环张力能作为显式的物理特征。同时，利用主成分分析（PCA）对 **ChemBERTa** 预训练模型生成的分子嵌入（Embeddings）进行降维处理，以提取关键的潜在语义特征。最终，将上述两组特征向量拼接后输入随机森林模型进行训练与预测。

改进尝试 (Attempted Improvements)

曾评估过引入更先进的 **3D 机器学习模型（如 Uni-Mol)。**虽然这类方法能利用量子化学计算的先验知识，但它们通常需要大量数据进行再训练才能适应特定任务。在测试中发现，其预训练提取的特征主要针对通用的物理化学属性，与 DAE 的**半衰期预测**相关性较弱。鉴于数据量不足以支撑重新训练，最终决定暂不采用此方案。

Version 2.0

> **Updates:**
>
> 1. **Feature Encoder Upgrade**: Utilized **ChemBERTa** on **Open-form SMILES** to

capture the initial chemical semantic state before the photoreaction.
2. **PCA Denoising**: Applied **Principal Component Analysis (PCA)** with a hard limit of 10 components to strictly control model complexity.
3. **Hybrid Integration**: Integrated **dHOMA (Sum of Core & Fused Rings)** and **dQ (Gasteiger Charge Difference)** as physical constraints, forming a "Transformer + Physics" hybrid feature set.

Features

1. ChemBERTa (Open-Form State)

   Encoder: `Seyonec/ChemBERTa-zinc-base-v1` (Transformer-based)

   Input: The SMILES string of the Open-ring isomer is used as input. This captures the "starting point" semantics of the molecule, which contains the potential conjugated system information required for the cyclization reaction.
2. PCA Denoising & Dimensionality Reduction

   The raw output from ChemBERTa is a high-dimensional tensor ($768D$). Given our extremely small sample size, the raw embedding space contains significant redundant information that leads to overfitting. We applied PCA as a crucial bottleneck:
   - **Denoising**: By projecting the data onto orthogonal principal components, we isolate the most variant structural features.
   - **Complexity Control**: We enforce a reduction of $768D \rightarrow 10D$. This low-dimensional dense vector forces the Random Forest to learn from the most dominant semantic signals rather than fitting to high-dimensional noise.
3. Physical Features (Physics Tower)

   To compensate for the "black-box" nature of Deep Learning, we explicitly calculate geometric and electronic descriptors based on 3D conformational analysis:
   - **dHOMA (Sum)**: $\Delta\text{HOMA} = \sum(\text{HOMA}_{\text{Closed}}) - \sum(\text{HOMA}_{\text{Open}})$. This aggregates the aromaticity changes of the **Central Ring** and all **Fused/Side Rings**, representing the total aromaticity recovery energy.
   - **dQ (Charge Diff)**: $\Delta Q = Q_{\min}(\text{Closed}) - Q_{\min}(\text{Open})$. This captures the shift in electron density at the most electronegative site of the reactive core.

Result:

R^2	**MAE**
0.5820	1.7209

Code

see: `https://github.com/thovet55/OrganicChemistry_AIhomework/tree/main/Final/version 2.0`

Version 1.0

Features

1. **MorganPrint**
   - **Generator**: `rdFingerprintGenerator` (RDKit Implementation, equivalent to ECFP4).
   - **Configuration**: `Radius=2`, `nBits=512`.
   - **Representation**: We utilized Morgan fingerprints to encode the local topological environment of the **Closed-form** isomer.
   - **Dimensionality Control**: Unlike the standard 2048-bit length, we restricted the vector size to **512 bits**. Given the small dataset size ($N \approx 50$), this reduction prevents the "Curse of Dimensionality" and reduces the risk of overfitting in the Random Forest model while preserving essential substructure information.
2. **Molecule matching (Topology Alignment)**
   - **Algorithm**: **MCS (Maximum Common Substructure)** with Skeletonization.
   - **The Challenge**: DAE photochromism involves a rearrangement of bond orders (alternating double/single bonds) between Open and Closed states. A standard exact match would fail because the bond types differ.
   - **Skeletal Alignment Strategy**:
     - **All-Single-Bond Conversion**: Before matching, we force-convert all bonds in both isomers to single bonds and strip aromatic flags. This extracts the pure "molecular skeleton."
     - **Atom Mapping**: We run MCS on these skeletons to generate a precise index mapping dictionary ($\text{Idx}_{\text{closed}} \rightarrow \text{Idx}_{\text{open}}$). This allows the algorithm to track specific rings and atoms across the chemical reaction.
3. **dHOMA & dQ calculation (Physics-Informed Features)**
   - **3D Conformer Generation**: Structures are embedded using **ETKDG** and optimized via the **MMFF force field** to obtain realistic bond lengths.
   - **Reaction Center Locking**: The script automatically identifies the "Break Bond" (the bond present in the Closed form but absent in the Open form) to locate the reactive carbon atoms (C_a, C_b).
   - **Physical Descriptors**:
     - **dHOMA (Geometric)**: We calculate the Harmonic Oscillator Model of Aromaticity (HOMA) for the central and adjacent rings. The feature is defined as $\Delta\text{HOMA} = \text{HOMA}_{\text{closed}} - \text{HOMA}_{\text{open}}$. This quantifies the **loss of aromaticity** during cyclization, which is the primary thermodynamic driving force for thermal stability.

- **dQ (Electronic)**: Using **Gasteiger Partial Charges**, we compute the absolute charge difference at the reactive center: $\Delta Q = |q_{C_a} - q_{C_b}|$. This metric captures the electronic asymmetry/polarity at the photocyclization site, offering complementary information to the geometric HOMA index.

Result

R^2	**MAE**
0.2020	2.9685

Code

see `https://github.com/thovet55/OrganicChemistry_AIhomework/tree/main/Final/version 1.0`