耦合分子结构与深度学习，提升定量结构-性质关系模型预测性能

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/41571.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

耦合分子结构与深度学习，提升定量结构-性质关系模型预测性能。传统定量结构-性质关系（QSPR）模型依赖于分子描述符来量化分子结构并建立与物理性质的相关性，难以实现多品类有机化合物超临界物性的高精度和广适配预测。近期上海交通大学罗正鸿团队发表的研究论文，将完整分子结构优化与深度学习进行融合，构建出兼具高预测精度、强鲁棒性的CNN增强型ANN-QSPR模型，突破了传统QSPR方法的核心局限，为化工领域有机化合物物性预测提供了一定支撑。

研究框架

本研究整体框架如下：收集了1359种有机物的超临界温度Tc与压力pc数据，采用密度泛函理论（DFT）获取分子结构；利用RDKit库计算了400个分子描述符，通过数据驱动的最大信息系数（MIC）方法筛选出20个信息量最大的描述符。基于上述输入构建了三类模型：（1）仅使用描述符的传统QSPR模型；（2）仅使用分子结构的CNN模型；（3）CNN增强的描述符模型。最后与工程常用的JOBACK基团贡献法对比，验证了CNN增强型QSPR模型的优越性能。

文章亮点

精准筛选分子描述符特征

基于Python平台的RDKit库计算得到了400个有效分子描述符，通过MIC法系统量化了非线性相关性，明确与Tc相关性最高的分子描述符为LabuteASA，与pc相关性最高的分子描述符为Chi0n，同时确定Tc预测模型的最优描述符数量为18个，pc预测模型的最优描述符数量为14个，为模型输入优化提供了量化依据。

多模型性能边界清晰

同一标准化数据集完成四类方法的横向对比显示，仅以分子结构为输入的3DResNet模型在千级数据量下存在严重过拟合，不适用于超临界物性预测；传统ANN模型可满足基础工程需求，但精度仍有提升空间；JOBACK方法操作简便但存在预测盲区（无法预测74种分子）且对Tc > 600 K的高温体系偏差显著。

CNN增强型ANN模型性能领先

CNN增强的ANN模型在验证集中实现了超临界物性预测的领先精度，其中Tc预测R2 = 0.888、r = 0.947、MAPE = 5.03%、MSE = 1682；pc预测R2 = 0.919、r = 0.960、MAPE = 6.37%、MSE = 11.7，综合性能显著优于传统ANN模型与JOBACK基团贡献法。

模型鲁棒性与普适性优异

完成模型鲁棒性验证，发现模型在不同数据划分方式下的平均性能与基础验证结果高度一致，无极端预测偏差。数据集覆盖烷烃、烯烃、炔烃、卤代烃、醇、胺等13大类1359种有机化合物，相较于单一品类专用模型普适性更强，可适配绝大多数有机物种的超临界物性工程预测需求。

结论与展望

融合完整分子结构与深度学习的CNN增强型ANN模型，有效弥补了传统QSPR方法的信息丢失缺陷，实现了多品类有机化合物超临界Tc和pc的高精度预测。但受DFT结构优化计算成本较高的限制，模型实际应用门槛仍有待降低。未来研究将考虑采用更快且资源要求更低的分子结构优化方法，或从预先计算的分子数据集中进行迁移学习，以降低使用成本并进一步提升模型的工程适用性。（来源：EngineeringJournals微信公众号）

相关论文信息：https://journal.hep.com.cn/fcse/EN/10.1007/s11705-026-2638-6

特别声明：本文转载仅仅是出于传播信息的需要，并不意味着代表本网站观点或证实其内容的真实性；如其他媒体、网站或个人从本网站转载使用，须保留本网站注明的“来源”，并自负版权等法律责任；作者如果不希望被转载或者联系转载稿费等事宜，请与我们接洽。
作者：罗正鸿等 来源：《工程·化学工程》

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发