利用带有物理描述符的可解释机器学习探索高热导率聚合物

Xiang Huang[1, †], Shengluo Ma[1, †], C. Y. Zhao[1], Hong Wang[2], 和 Shenghong Ju[1, 2, *]

[1] 中英低碳学院，上海交通大学，中国上海[2] 材料基因组计划中心，材料科学与工程学院，上海交通大学，中国上海

†共同贡献

摘要

高热导率聚合物的高效、经济开发，对于解决有机器件的散热问题至关重要。目前，由于合成和表征过程中存在多重自由度，功能性高热导率聚合物的实验制备仍然是一个反复试错的过程。聚合物信息学将数据科学、机器学习与聚合物实验/模拟高效结合，从而能够高效设计具有所需性能的聚合物材料。然而，现有的聚合物热导率数据库十分稀少，建立合适的聚合物表征仍颇具挑战。在本研究中，我们提出了一种结合可解释机器学习与物理特征工程、用于高热导率聚合物链的高通量筛选框架。首先，通过分子动力学模拟收集用于训练的聚合物热导率数据集。受类药小分子表征和分子力场的启发，我们计算了 320 个聚合物单体描述符，并通过 分层筛选，提取出 20 个具有物理意义的优化描述符。所有机器学习模型的预测准确率 R^2 均大于 0.80，优于采用传统图描述符的模型。此外，我们确定了横截面积和二面角刚度描述符对热导率具有正向/负向贡献，并获得了 107 种热导率高于 20.00 瓦每米开尔文的潜力聚合物结构。我们还利用符号回归构建了用于预测聚合物热导率的数学公式。高热导率聚合物结构大多为π-共轭结构，其重叠的 p轨道使其能够轻易保持较高的链刚度和较大的群速度。所提出的数据驱动框架有望促进对具有理想性能的聚合物进行理论设计和实验设计。

* 通讯作者：shenghong.ju@sjtu.edu.cn。

1. 引言

聚合物具有化学惰性、机械柔性和质量轻等优点，因此广泛应用于工业和日常生活中[1]。随着有机电子器件日益小型化，而功率密度持续升高，其热管理和散热能力引起了广泛关注[2]。然而，传统聚合物是热绝缘体，其热导率据报道仅为 0.1 至 0.5 瓦每米开尔文，阻碍了有机电子学的发展[3]。有机储能和电子器件亟需高热导率聚合物，以适应有机电子学和光电子学的革命性创新[4]。研究发现，聚合物形态和拓扑结构与热导率密切相关[5]。提高晶粒取向和结晶度能够显著减少声子散射，并增强沿链方向的热导率，这一点已通过实验[6-8]和理论模拟[9-12]得到证实。近期一项研究通过解缠结并排列非晶链，制备出了聚乙烯（PE）薄膜，其类金属热导率达到 62 瓦每米开尔文，比经典非晶聚合物高出两个数量级以上[6]。此外，分子动力学模拟表明，单条结晶态 PE 链的热导率极高，甚至可能发散[11]。这些发现为解决聚合物器件的散热问题带来了机遇。

尽管链取向、结晶度、聚合物纤维甚至单链聚合物结构都会对热学特性产生显著影响[13-15]，但聚合物库规模庞大，化学空间中已知存在多达10^8种有机单体分子[16]。目前关于聚合物热导率的研究仍是爱迪生式过程，依靠直觉或经验进行试错法，既耗时又昂贵[17]。此外，大多数研究都针对 PE[4,6,11]等简单结构开展，因此难以掌握影响聚合物热导率的因素的一般规律，也难以在广阔的化学空间中发现具有高热导率的聚合物分子结构。

聚合物信息学[18]与人工智能和机器学习（ML）方法的发展密切相关，旨在利用数据驱动方法调控有机材料的物理性质或推动器件开发，从而解决传统试错法中结构自由度与效率/成本之间的矛盾。近年来，聚合物信息学研究受到广泛关注并取得了成功[19-21]，其研究内容涉及对有机材料光学[22-24]、电学[25-27]和热学性质[28-32]的预测。尤其是，已有多项研究致力于寻找

或设计具有高热导率的结构，研究对象包括晶态聚合物[30]、非晶聚合物[31,32]和共聚物[33]。这些研究大多采用图描述符[31]或聚合物化学片段统计学[30,32,33]来描述信息学算法中的单体结构，这类描述也称为指纹或表征。生成的图描述符依赖分子/单体图信息，这些信息通过基于领域知识的特征工程[34]构建，或通过尝试形成通用描述符[35]获得。此外，分子访问系统（MACCS）等描述符[36]是通过对不同化学片段进行统计而获得的，并且与分子图密切相关。因此，这些描述符统称为图描述符。指纹必须能够对每个候选对象进行唯一、完整且精简的表征，而构建成功的指纹是一项艰巨任务[37]。此外，聚合物由许多重复单元组成，其结构比有机小分子更复杂，因此需要准确捕捉单体连接位点的信息[26]。图描述长期以来已应用于类药小分子的开发并得到验证[38]，而 RDKit[39]和 Mol2vec[34]等开源工具包的出现提高了图描述的可及性，这也是图描述在聚合物信息学中广受欢迎的原因之一。然而，图描述符以数值向量串的形式呈现。分子结构的完整性决定了各数字之间的耦合关系。因此，分子单体与材料性质之间的关系难以把握。

探索由物理独立描述符构成的集成，以表征分子结构，这对定性构效关系（QSPR）建模十分重要，并有助于为分子结构评估提供更直观的指导[40]。用于收集和精简物理描述符的特征工程，是决定聚合物信息学有效能力的关键步骤。研究人员关注如何开发自动化、通用且高效的工具来计算有机分子的描述符；这类工具可将分子符号表示中编码的化学信息转化为有用的数值或某些标准化的实验结果[41]。目前已有多种开源和商业软件[41-43]可用于计算各类分子描述符，例如碳原子序数、分子量和扩展拓扑化学原子（ETA）[44]，这些描述符已成功应用于有机化学合成[45]、分子抗菌活性预测[46]等领域。此外，实验或模拟中的参数条件会影响分子性质。例如，受力场启发的描述符，如键、角度和二面角的类型，已被证实可用于预测聚合物的比热，即使数据集来自实验[29]。聚合物特征数量的精简

也是一个值得关注的问题，因为某些描述符与目标性质的相关性可能很低，而低维描述符空间更易于构建机器学习模型[47]。特征提取和特征选择是降低特征维度的主要方法。特征提取通过原始数据空间创建子集，例如主成分分析（PCA）；但新特征的具体含义往往难以理解[48]。特征选择则保留单个描述符的物理意义，而基于相关性评估的筛选方法依赖于数学模型，例如皮尔逊系数和斯皮尔曼系数，分别用于考察数据的线性关系和单调关系[49]。此外，过滤法不涉及机器学习模型，因此筛选所得特征可能并不适用。基于包装器的特征选择技术结合机器学习模型来消除冗余特征，包括递归特征消除（RFE）、序列特征选择（SFS）和穷举特征选择（EFS）[50]。针对信息学算法测试不同描述符子集，是包装器方法的关键特征，其核心在于组合不同描述符的策略。典型的RFE通过在迭代构建的机器学习模型中，持续从剩余特征中剔除影响较小的特征，以提升模型性能；这一过程依据随机森林等模型分配的特征权重排序进行[48]。因此，RFE依赖于机器学习模型的特征权重评估机制。

本文聚焦于聚合物单体表示和特征选择方面的挑战，提出了一种将机器学习可解释框架与高通量分子动力学模拟相结合的方法，用于发现高热导率聚合物结构，如图 1 所示。该框架包含四个组成部分：1）聚合物库构建；2）聚合物热导率的分子动力学模拟；3）单体特征表示与分层筛选；4）用于热导率预测的机器学习模型构建。训练数据收集自文献[51,52]，并利用 PolyInfo [53]和 PI1M [54] 数据库中的候选结构进行高热导率结构的虚拟筛选。所有聚合物单体均通过 SMILES（简化分子线性输入规范）字符串表示，并通过重复复制形成一维聚合物链。训练数据集中聚合物的热导率采用第二代通用 AMBER 力场 GAFF2[55]进行分子动力学模拟计算。受药物类分子表示和分子力场的启发，我们通过 Mordred 软件[41]计算和提取力场参数文件，获得了 320 个物理描述符，并通过分层筛选保留了 20 个优化描述符。随后，我们分别训练了随机森林（RF）、极端梯度提升（XGBoost）树模型以及多层感知机（MLP）神经网络模型

以建立优化描述符与这些基准聚合物数据集热导率之间的关系。此外，我们分析了各优化描述符的特征重要性，并通过 SHAP 分析 [56]提取了用于设计高热导率聚合物的化学启发式规则。利用训练好的机器学习模型，我们筛选出 107 种热导率高于 20.00 瓦每米开尔文的有前景的聚合物，并通过符号回归推导数学公式，以表示这些聚合物的热导率。最后，我们讨论了八种典型聚合物的热导率机制。总而言之，所提出的方法有助于开展高热导率聚合物的理论或实验研究。

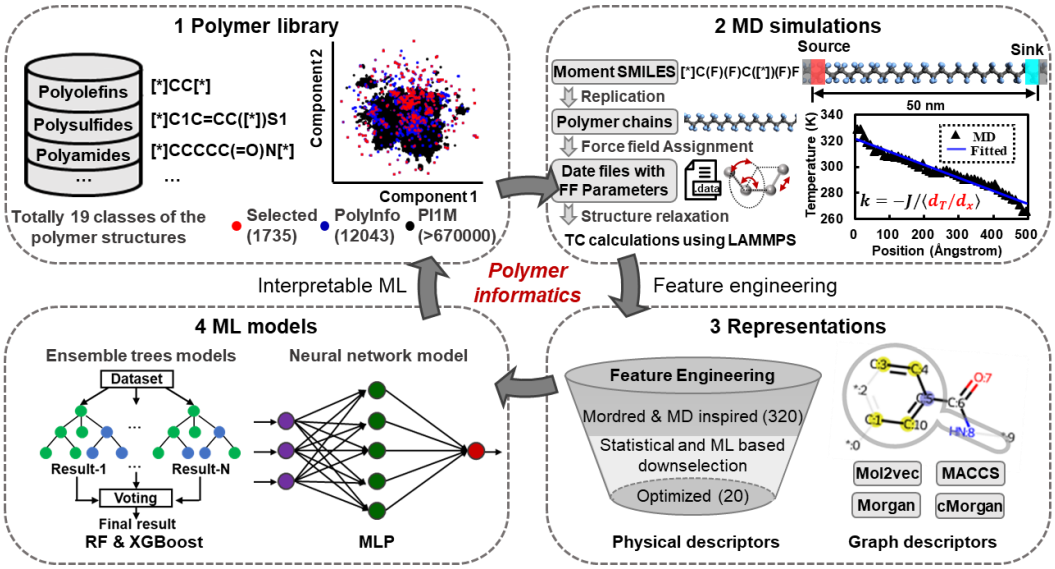


图 1。 通过可解释机器学习对高热导率聚合物进行高通量筛选的示意图。

2. 结果与讨论

2.1 聚合物数据集在化学空间中的分布

文献中的聚合物数据[51,52] 被用作训练机器学习模型的基准数据库；此外，PolyInfo [53] 和 PI1M [54] 数据库用于虚拟筛选具有高热导率的聚合物结构。根据不同的元素和化学官能团，聚合物被分为 19 类，例如聚烯烃、聚醚和聚酰胺 [57]。为验证所选的 1735 条基准数据在另外两个数据集中的分布情况，我们利用均匀流形逼近与投影（UAMP）[58]在二维空间中可视化其化学结构；其中，每种单体的化学结构都被转换为半径为 2 个原子的 1024 维 Morgan指纹[35] 。结果表明，所选基准数据集中的聚合物结构（图 2a）充分覆盖了 PolyInfo 中的化学空间分布（图 2b）

以及 PI1M 数据库（图 2c）中的化学空间分布。需要注意的是，PI1M 数据集由一个利用 PolyInfo 数据训练的循环神经网络生成模型构建，它填补了 PolyInfo 数据集化学空间中的稀疏区域，但整体分布仍保持一致 [54]。因此，利用所选数据训练的机器学习模型能够很好地学习所有候选物的化学特征，可有效用于虚拟筛选具有高热导率的聚合物结构。

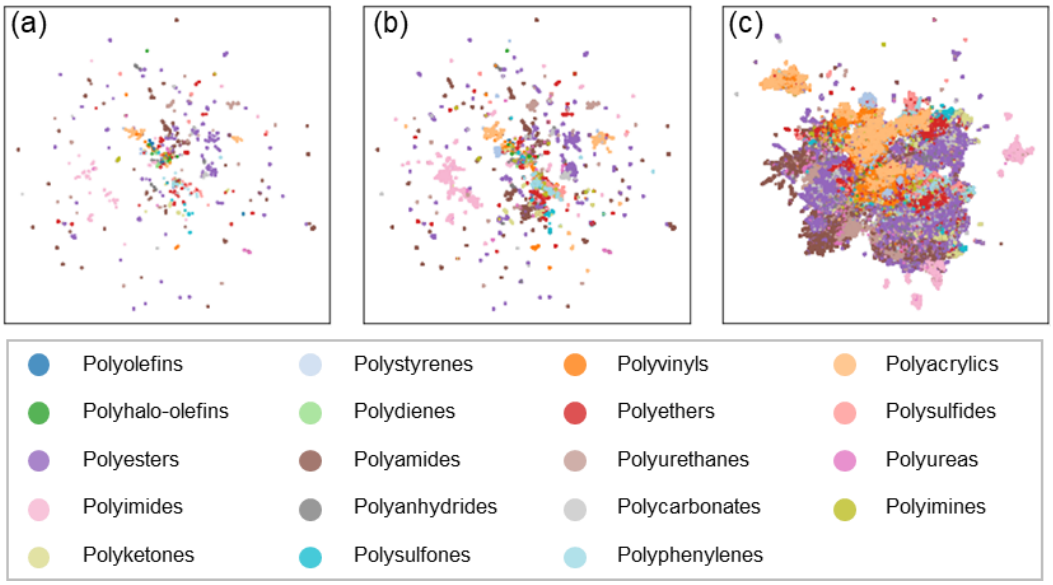


图 2。 使用 UMAP 将聚合物数据分布可视化于二维空间中。(a)、(b) 和 (c) 分别对应筛选后的数据集、PolyInfo 数据集和 PI1M 数据集。

2.2 聚合物描述符分层筛选与机器学习模型训练

聚合物描述符通过三个阶段进行层级式筛选：去除低方差特征、依据不同相关系数进行初步筛选，以及借助机器学习模型进行最终选择（见补充材料第 S1 节）。收集到的初始单体物理描述符由 286 个基于 Mordred 的描述符和 34 个受分子动力学启发的描述符组成。受分子动力学启发的描述符和基于 Mordred 的描述符列于补充材料第 S2 节。去除低方差描述符旨在剔除方差低于特定阈值的描述符，因为这些描述符对所有聚合物数据的目标性质（本研究中为热导率）的贡献被认为几乎一致。将方差阈值设为 0.01 后，保留了 264 个描述符进入下一阶段。鉴于不同数学模型各自关注的方面不同，我们基于不同的相关系数建立了权重分配机制（WAM），以进一步对描述符进行初步筛选。皮尔逊、斯皮尔曼和距离相关系数分别用于评估数据之间的线性、单调和非线性关系，而最大信息系数（MIC）则通过信息熵反映两个变量之间的关联，无论其关系是线性

还是非线性。MIC 的可靠性取决于数据样本量，只有在数据集规模较大时，其值才可靠。我们纳入了皮尔逊、斯皮尔曼、距离相关系数和 MIC 这四种度量指标系数，并为每种系数赋予 0.25 的权重因子；相应阈值依次设为 0.05、0.05、0.153 和 0.132。通过 VAM 保留了累计权重值为 1 的 53 个描述符。随后，我们将随机序贯特征选择（RFSF）与随机森林模型结合，用于确定优化描述符。考虑机器学习模型训练中描述符的所有可能组合既耗时又昂贵，因此传统的 SFS 通常会导致次优解；此时，推荐的优化描述符集并不唯一，且受描述符输入顺序的影响 [59]。在本研究中，我们打乱输入描述符的顺序，将其与 100 轮随机森林模型训练结合，并通过统计方法确定最终的优化描述符。阈值设为 0.34，即保留在 100 轮随机森林模型训练中出现次数超过 34 次的描述符。基于 VAM 和 RSFS 得到的优化描述符结果见图3a，其详细说明列于补充材料第 S3 节。此外，图3e 展示了优化描述符之间的皮尔逊相关矩阵（其他度量指标见图 S1）。结果发现，大多数描述符彼此呈正相关，而与热导率呈负相关。只有三个描述符与热导率呈正相关，其中两个是受分子动力学启发的描述符。例如，描述符 MW_ratio 表示主链分子量与单体分子量之比，取值介于 0 和 1 之间。MW_ratio 为 1 表示聚合物没有侧链，这会减少热通量沿链传递时的损失，从而有可能获得较高的热导率。

图3b展示了使用优化描述符训练的随机森林模型的结果，其训练集和测试集的 R^2 分别为 0.87 和 0.84。为验证优化描述符的可扩展性，我们使用 XGBoost 和多层感知机模型进行训练（见图S2）。XGBoost 的训练集和测试集准确度分别为 0.95 和 0.87，多层感知机的准确度则分别为 0.81 和 0.88，与随机森林模型相当，甚至更优。因此，后续讨论中将采用这三种模型。

图3c展示了机器学习模型在不同特征筛选阶段的预测准确率（训练集和测试集的预测结果见图S3）。我们还进行了保留95%方差的主成分分析，以便与随机序贯特征选择技术进行比较。根据图S4中主成分数量与累积方差之间的关系，至少需要19个主成分才能超过95%方差，这与优化描述符的组数相近。如图3c所示，基于树的

随机森林和XGBoost模型即使在描述符维度较高时，仍能保持相对较高的准确度，因为它们具有很强的数据过拟合抑制能力。此外，特征筛选过程通常会伴随信息损失，导致模型准确度下降。然而，特征筛选过程也减少了数据间的冗余，从而抑制过拟合并提高多层感知机模型的准确度。总体而言，采用RFSF筛选出的优化描述符训练的三种模型，其准确度均高于采用主成分分析得到的描述符训练的模型，这证明了我们方法的有效性。

图3d展示了采用不同图描述符的机器学习模型进行比较的结果（训练集和测试集的预测结果见图S5）。Mol2vec [34] 是一种无监督机器学习方法，用于学习分子子结构的向量表征，需要使用基准数据集进行分子结构训练。此处使用的预训练聚合物嵌入模型来自其他研究[54]，该模型利用 PolyInfo 和 PI1M 数据集构建。MACCS [36] 描述符是一种基于结构键的描述符，采用 166位键集。Morgan 和 Morgan 计数（cMorgan） [35] 描述符属于扩展连接指纹，可捕捉与分子活性相关的分子特征。图3d中的结果表明，无论采用随机森林、XGBoost 还是多层感知机，使用优化描述符训练的机器学习模型都表现更优。物理描述符的筛选过程考察单个或组合描述符与热导率的关系，而图描述符则旨在尽可能全面地表征分子信息或单体信息。尽管已有研究指出，分子图中的元素或基团与聚合物链的热导率相关 [30]，但使用相关的物理描述符来预测聚合物链的热导率更直观、有效。不过，这并非绝对，因为热导率也与链刚度等参数有关 [60]。

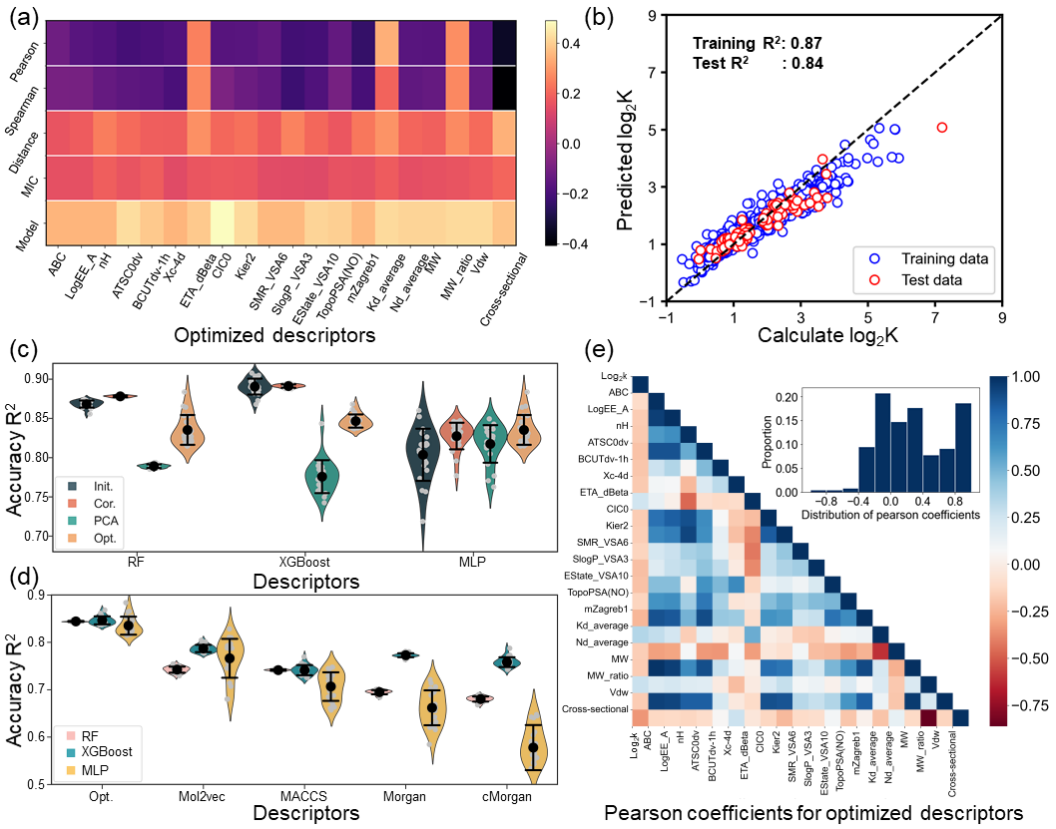


图 3。聚合物描述符筛选与机器学习模型训练。(a) 通过四种系数——皮尔逊、斯皮尔曼、距离相关系数和最大信息系数——以及随机森林模型筛选得到的优化描述符。(b) 基于优化描述符的随机森林模型准确度，其中训练集 R^2 为 0.875，测试集 R^2 为 0.844。(c) 不同筛选流程下机器学习模型的准确度，包括初始（Init.）、数学相关性（Cor.）系数筛选和随机森林模型优化（Opt.）阶段。此外，还采用了主成分分析方法进行比较。(d) 不同聚合物表征方法下机器学习模型的准确度。小提琴图表示各值的分布，灰色表示各个子样本，黑色表示 R^2 的均值和标准差。(e) 皮尔逊相关矩阵显示了优化描述符与热导率之间的相关性。插图为皮尔逊系数分布的统计数据。

2.3 可解释机器学习模型的物理见解

图4总结了使用 SHAP 分析基于优化描述符训练的随机森林模型中特征的影响。SHAP 方法通过计算特征对模型输出的边际贡献，尝试解决机器学习算法难以解释的黑箱难题 [56]。因此，训练数据集中每种聚合物结构的特征均分别被赋予 SHAP值。如图4a所示，优化描述符的重要性排名依据平均 SHAP 值确定。在排名前8的优化描述符中，受分子动力学启发的描述符与基于 Mordred 的描述符数量相同，这反映出随机森林模型的构建由这两类描述符共同贡献。图4b展示了各描述符的 SHAP 值分布，蜂群图中数据点的颜色深浅表示热导率的大小

训练集中的聚合物结构。排名靠前的特征的SHAP值分布相对较宽，并且总体上随特征值呈单调变化（图 S6）。

这里，我们重点介绍两个受分子动力学启发的描述符：横截面和 Kd_平均值。最重要的描述符“横截面”表示聚合物链的有效横截面积，这与热导率的关系直观明了。从图 4c 可以看出，横截面的 SHAP 值随描述符值单调递减。根据傅里叶定律，沿一维聚合物链的热流率可表示为 $Q/d_t = -kA\,d_T/d_x$，其中 Q 表示热流，d_t 表示时间间隔，k表示热导率，A 表示横截面积，d_T 和 d_x 分别表示冷热两端之间的温差和距离 [61]。因此，热导率与横截面积呈负相关，热导率高的聚合物通常具有较小的横截面积（图 S7a）。此外，与无定形结构相比，聚合物链结构不存在无序性，因而能够保持晶体的对称性并减少声子散射。然而，聚合物链可能会因温度及其他因素而旋转并变得无序，导致热导率迅速下降 [62]。二面角能量常数与聚合物链刚度之间的密切关系已得到证实；在分子动力学模拟中，人为提高二面角力常数 Kd，可维持聚乙烯链的刚度并提高热导率 [62,63]。Kd_平均值是聚合物链中 GAFF2 力场所有类型二面角力常数的平均值，与图 4d 中相应的 SHAP 值大致成正比。尤其是，Kd_平均值较大的聚合物结构（>4 kcal/mol）通常具有较大的 SHAP 值和热导率（图 S7b）。值得注意的是，聚合物链的热导率受多个参数影响，因此很难仅凭单个描述符确定其值。例如，研究已证实，晶态聚降冰片烯对链刚度不太敏感，即使在分子动力学模拟中增加二面角力常数项也是如此 [62]。这证实了我们提出的机器学习框架在预测聚合物热导率方面的重要性。

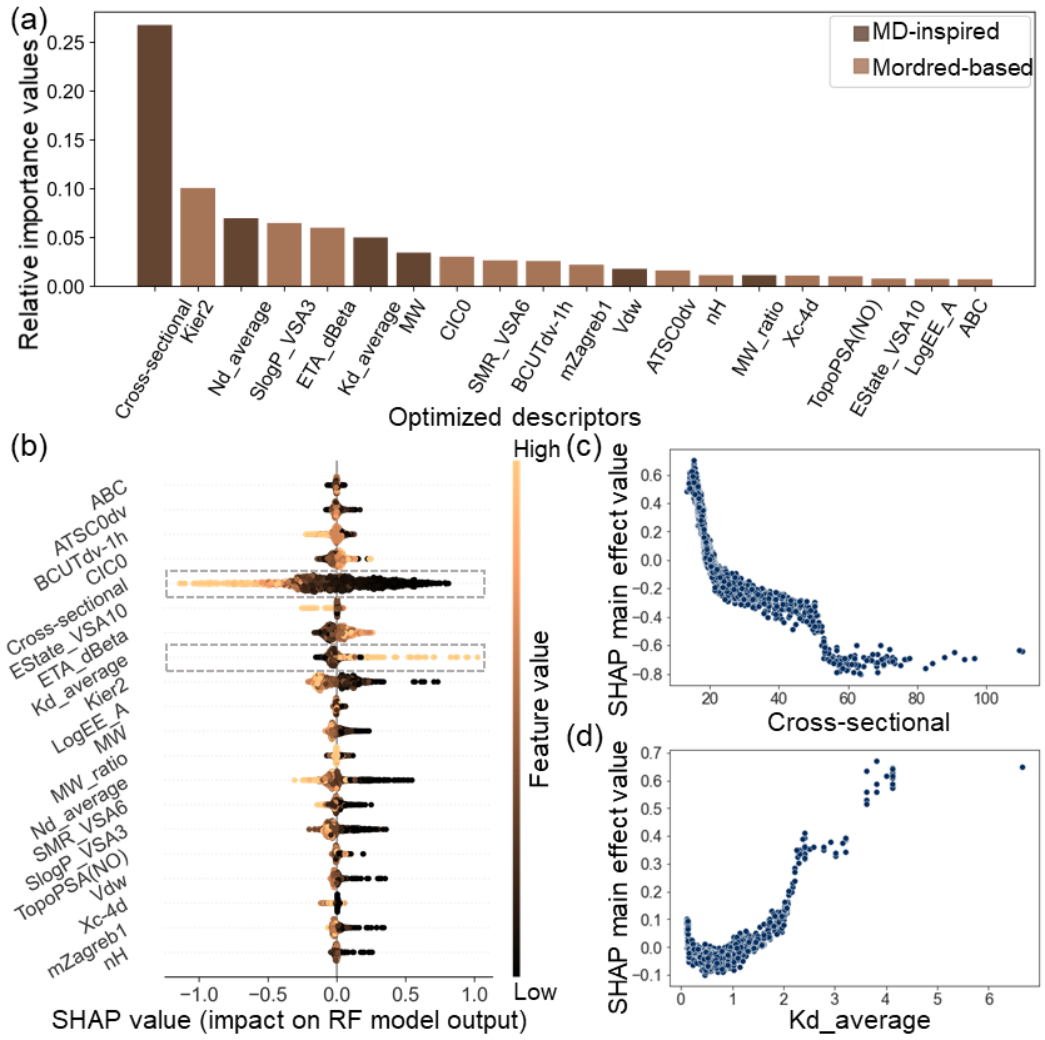


图 4. 分析 使用 SHAP 分析基于优化描述符训练的随机森林模型的特征重要性。（a）20 个优化描述符的平均 SHAP 值。(b) 以蜂群图表示训练数据集中聚合物各描述符的 SHAP 值。(c) 和 (d) 分别展示训练数据集中聚合物的横截面和 Kd_平均值的 SHAP 值随描述符值的变化。横截面是聚合物链的有效横截面积，Kd_平均值是 GAFF2 力场中二面角力常数的平均值。

2.4 高热导率聚合物的发现

通过与不同机器学习模型相结合，我们验证了优化描述符在预测聚合物链热导率方面的可靠性。接下来，我们将这些机器学习模型应用于预测 PolyInfo 和 PI1M 数据库中聚合物结构的热导率，以虚拟筛选具有高热导率的有前景聚合物。图 5a–c 分别展示了优化描述符集成与随机森林、XGBoost 和多层感知机结合后预测的聚合物热导率与横截面积的关系。其中星号标示的是随机森林、XGBoost 和多层感知机预测的聚乙烯，其 $\log_2$TC 分别为 3.91、4.66 和 5.30

而分子动力学模拟计算得到的值为 5.28。这里可以明显看出热导率与横截面积之间的依赖关系，因为几乎所有预测的高热导率聚合物都具有较小的横截面积。此外，PI1M 与 PolyInfo 具有相同的化学分布空间，并填补了其中的数据稀疏区域；它涵盖了 PolyInfo 中的大部分热导率范围，也丰富了高热导率区域的聚合物结构。

比较不同机器学习模型的结果发现，随机森林和XGBoost这两种树模型对聚合物热导率的预测范围比多层感知机更窄。尽管树模型在防止过拟合方面表现出色，但其外推能力通常不足，预测结果仍局限于训练集中聚合物结构的热导率范围。相比之下，多层感知机神经网络模型通常具有更好的外推能力，尽管模型的训练准确率R^2相对较低，但它在发现高热导率聚合物结构这类小样本数据方面更具优势。这一发现与先前使用机器学习预测气体分离膜渗透率的研究相似 [17]。此外，先前研究揭示了聚合物链热导率对链长的依赖性。在一定长度范围内，发散的热导率k 与链长L可用$k \sim L^{\beta}$拟合，其中β 表示相对占主导地位的声子输运机制[13]。在本研究中，我们将有效长度为50 nm、热导率大于20.00瓦每米开尔文的聚合物链视为具有高热导率的优异聚合物。随后，我们制定了平衡策略，综合三种机器学习模型的性能，以推荐适合通过分子动力学模拟计算热导率的潜在聚合物结构。我们利用随机森林、XGBoost和多层感知机，在PolyInfo数据集中筛选出预测$\log_2$热导率最高分别为3.51、3.50和4.33的聚合物结构，并仅选取出现次数不少于2次的结构进行分子动力学模拟。最终发现并验证了24种高热导率聚合物结构。同样，我们采用这一方法，在PI1M数据库中筛选出84种高热导率聚合物结构。去重后，本研究共发现107种高热导率聚合物结构，并计算了合成可及性评分（SA），结果如图5d所示。具体的聚合物结构见补充材料S8节。从图S8可以看出，大多数高热导率聚合物是简单的线性结构，或主链中含有芳香环；它们的重复单元长度较短，且不含侧链。SA评分最初用于根据分子复杂度和片段贡献估算类药分子的合成可及性[64]，随后也被用于聚合物[31,32]。SA评分的取值范围为1至10，值越高，合成难度越大。为考虑单体连接方式的影响，我们对聚合度为6的聚合物分子计算了SA评分。其中

这些结构中，有28种聚合物结构的SA不超过3.00，包括聚乙烯、聚四氟乙烯和聚对苯撑等。尽管目前制造这些结构中的每一种都很困难，但我们相信，未来不久，随着微机械拉伸、静电纺丝和纳米模板制备等先进工艺的发展，将制备出更多类似聚乙烯链的聚合物，以探索聚合物热导率的极限[4,6,11]。

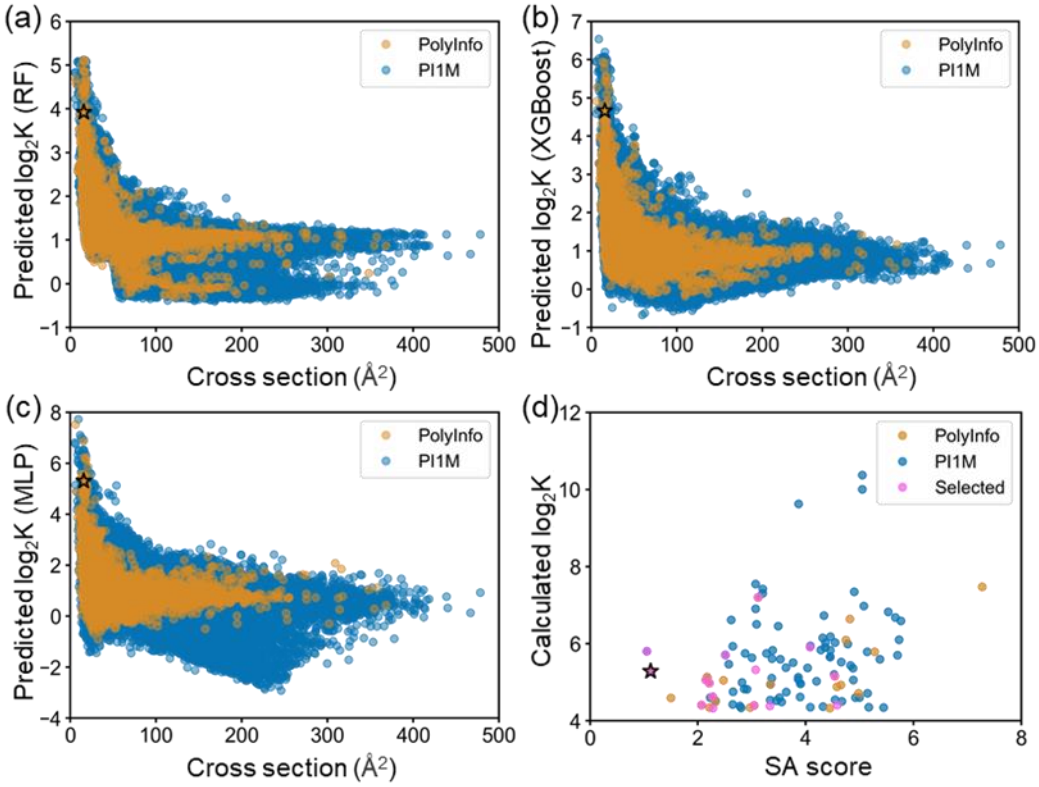


图5. 预测 使用构建的机器学习模型预测PolyInfo和PI1M数据库中的高热导率聚合物 .

(a)、(b)和(c)分别基于随机森林、XGBoost和多层感知机模型。(d) 筛选出的高热导率聚合物的合成可及性（SA）评分与计算得到的$\log_2 K$之间的关系（TC > 20.00 瓦每米开尔文）。星号表示聚乙烯，本研究中的TC为38.98 瓦每米开尔文。

2.5 使用符号回归预测有前景的聚合物

由于聚合物链的热导率受复杂的多参数影响，因此很难仅凭单一描述符预测不同聚合物的热导率变化趋势。符号回归（SR）尝试通过数学公式关联现有描述符，构建新的组合特征，从而加速发现具有优异性能的材料 [65]。只要具有较高的一致性和准确度，SR 就不需要海量 [66,67]。采用具有优化描述符的 107 种有前景的聚合物结构（热导率 > 20.00 瓦每米开尔文）进行符号回归，训练集与测试集的比例为 3:1。基于遗传编程的符号回归（GPSR）通过 gplearn 代码实现，并采用高效的逐步策略获得和筛选数学公式

[68]。超参数设置和公式确定过程的详细信息见补充材料S9节。首先使用皮尔逊系数筛选优化描述符并构建子描述符，得到一个由 22 个描述符组成的新集。图 6a 展示了优化描述符在 158 个数学公式（PC 值 >=0.85，复杂度<=10）中的出现频率，最终保留了前 8 个描述符。值得强调的是，受分子动力学启发的横截面积（横截面）和二面角力常数（Kd_平均值）描述符出现在每个公式中。在图 6b 中，我们计算了这组新描述符与热导率之间的皮尔逊系数，结果表明这些描述符与热导率密切相关。随后，我们重新设置了 gplearn 中的网格搜索超参数，并使用 R^2 作为度量指标。只有 R^2 较高且复杂度（公式长度）较低的公式才适用于预测聚合物结构的热导率 [69]。因此，我们通过图 6c 中的密度图，以复杂度和准确度 R^2 为特征，展示了 9073 个复杂度不超过 30 且 R^2 高于 0.6 的数学公式。采用拉丁超立方抽样方法 [70,71]确定了帕累托前沿上的 c、d、e 和 f 四个点，其对应的公式列于表 S8。四个公式的复杂度介于 20 至 30 之间，拟合准确度均高于 0.70。此外，训练准确率总体上与复杂度呈正相关。例如，复杂度为 20 的 c 点对应公式在这四个点中准确度 R^2 相对较低，但其拟合结果与分子动力学标注的 $\log_2 K$ 一致，如图 6d 所示。同时，所有四个确定的公式均包含横截面、Kd_平均值和 Nd_平均值描述符，这验证了聚合物链的热导率与横截面积、二面角刚度等参数密切相关。这些公式对于高热导率聚合物链结构的初步快速筛选具有重要意义。

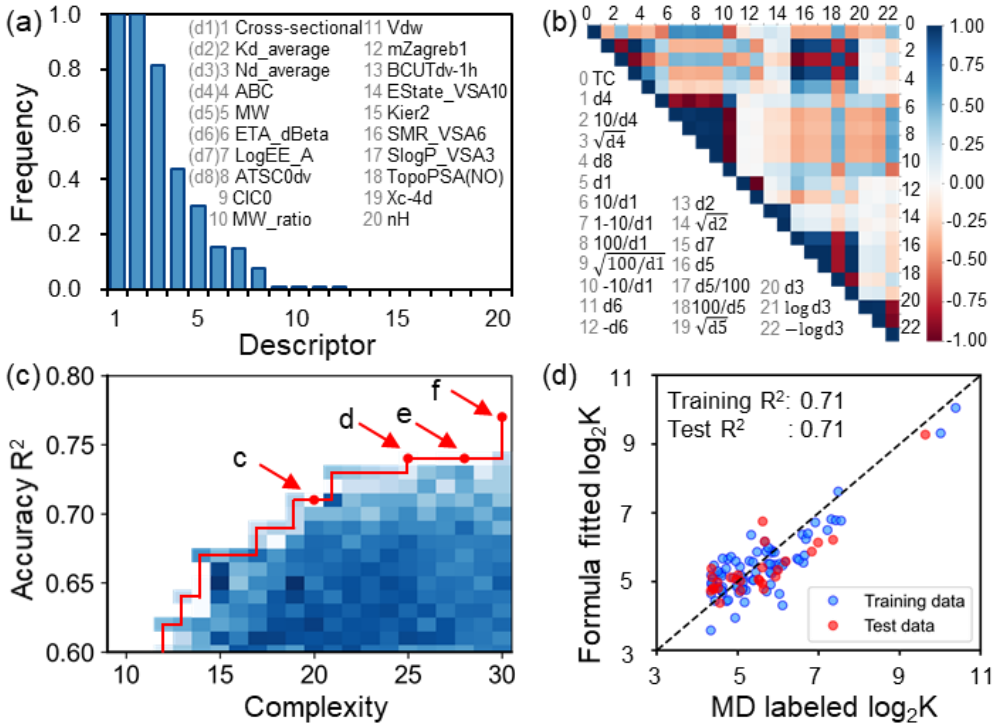


图 6. GPSR 用于预测有前景聚合物的热导率。(a) 158 个数学公式（PC 值 >=0.85，复杂度为 <=10）中优化描述符的出现频率。(b) 显示 22 个描述符与热导率之间相关性的皮尔逊相关矩阵，其中描述符 d1~d8 对应图 6a 中的描述符 1 至 8。(c) 通过密度图展示的 9073 个数学公式的准确度 R^2 与复杂度之间的帕累托前沿。(d) 分子动力学标注值与公式（点 c）的拟合结果对比，该公式的复杂度为 20，训练准确率 R^2 为 0.71。*2.6* 有前景晶态聚合物的热传输机制

综合考虑热导率和合成可及性评分等因素，选取了八种聚合物结构（见图7a）进行声子色散关系分析。目前，诸如 [*]C=C[*]和 [*]N=N[*] 这类聚合物结构尚难以通过实验合成，但有助于我们理解聚合物热导率的热传导机制。除聚乙烯和聚四氟乙烯（PTFE）这两种简单的线性结构外，所有这些聚合物分子均为 π-共轭结构。在 π-共轭聚合物分子中，p轨道的重叠增强了对链旋转的抑制作用，并形成刚性主链 [12]。图7b展示了通过声子谱能量密度（Phonon-SED）分析[72]得到的声子色散关系。关于 Phonon-SED 方法的详细说明见方法部分。由于声学模主导聚合物晶体中热载流子的热输运，因此图中展示了频率低于25 THz的声子模。此外，声子群速度近似为

所有声学支斜率的平均值[12,60]。对于一维聚合物链系统，热导率分析为 $k = v_g C_v l$，其中 v_g 为声子群速度，C_v 为体积热容，l 为声子平均自由程。受经典分子动力学模拟的限制，体积热容可表示为 $C_v = 3k_b N/V$，其中 k_b 为玻尔兹曼常数，N为原子数，V为体积。因此，声子平均自由程可通过热导率除以声子群速度与体积热容的乘积来计算。上述计算采用了近似，因此结果可能较为粗略，但通过比较相关参数的相对趋势，确实有助于我们理解这些有前景的聚合物结构背后的热传导机制，相关参数列于表1。

八种聚合物结构的体积热容介于 3.28 至 6.13 之间，这对聚合物链的高热导率并非关键因素。至于声子群速度，六种π共轭聚合物的声子群速度较大（超过 5900 m/s），这是由于 p 轨道重叠和离域电子的作用。此外，较小的原子质量也使声子群速度较大。与聚乙烯相比，聚四氟乙烯的声子群速度较小，这是因为氟原子的质量相对大于氢原子。声子平均自由程有助于深入了解聚合物链中的声子输运。总体而言，简单的线性聚合物链容易具有较长的声子平均自由程，尤其是[*]C=C[*] 和[*]N=N[*]这两种线性π共轭聚合物。这些结构的链刚度较大，除主链外所含原子较少，因此声子-无序散射较弱。

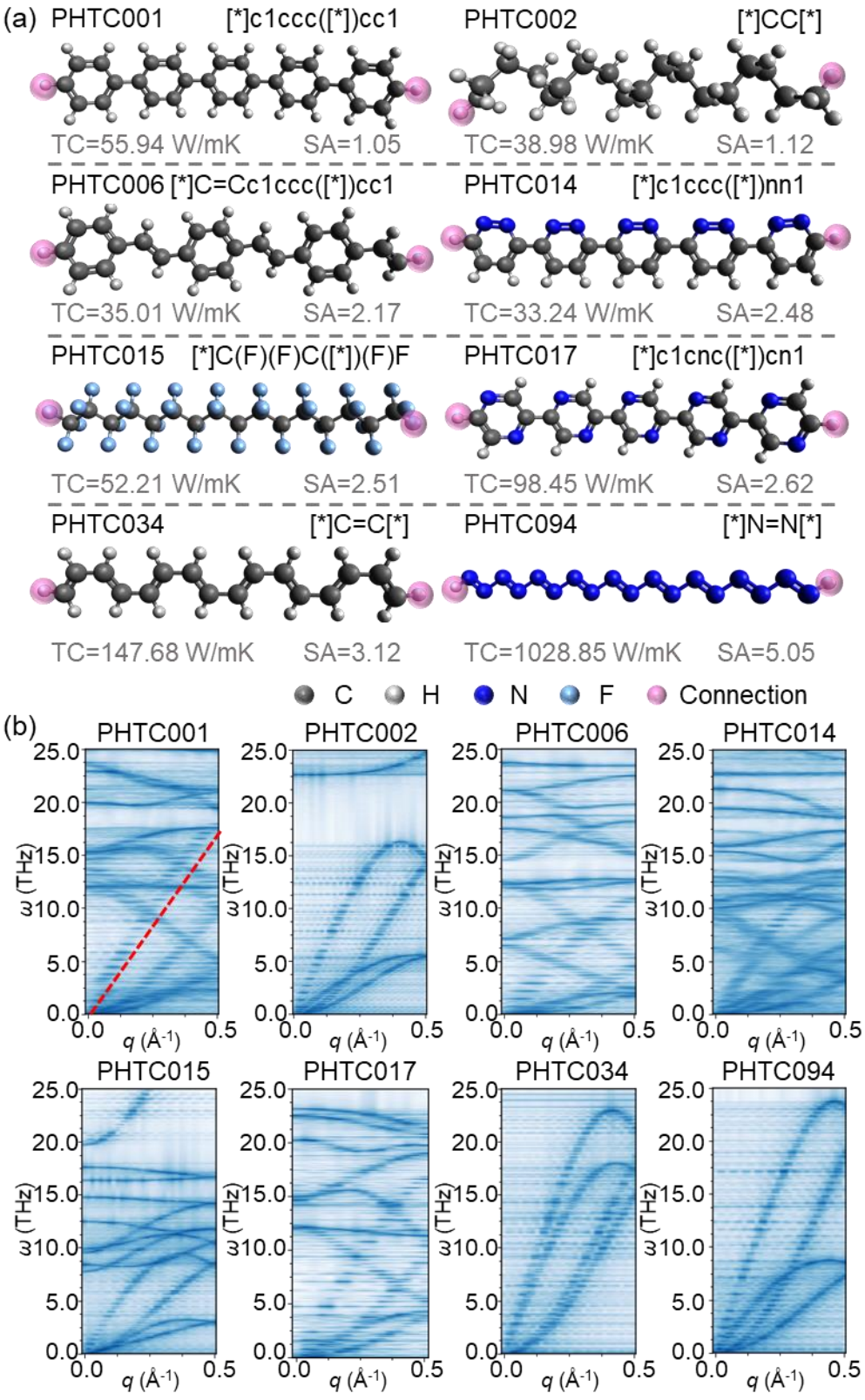


图 7. 八种有前景的聚合物的结构和声子色散关系。(a) 聚合物链结构。(b) 声子色散关系。q表示波矢，ω 表示声子频率；如 PHTC001 结构中的红色虚线所示，一支声子分支的平均声子群速度估算为从原点到最高频率点的斜率。

表1。八种有前景的聚合物的体积热容、声子群速度、声子平均自由程和声子热导率

《聚合物》ID	SMILES	SA	C_v (J/cm³K)	v_g (m/s)	l (nm)	k (瓦每米开尔文)
PHTC001	[*]c1ccc([*])cc1	1.05	5.27	6822.21	1.55	55.94
PHTC002	[*]CC[*]	1.12	6.13	5240.91	1.21	38.98
PHTC006	[*]C=Cc1ccc([*])cc1	2.17	5.25	6295.54	1.06	35.01
PHTC014	[*]c1ccc([*])nn1	2.48	5.03	5927.05	1.12	33.24
PHTC015	[*]C(F)(F)C([*])(F)F	2.51	3.84	2952.11	4.61	52.21
PHTC017	[*]c1cnc([*])cn1	2.62	5.03	7439.50	2.63	98.45
PHTC034	[*]C=C[*]	3.12	4.37	8380.09	4.03	147.68
PHTC094	[*]N=N[*]	5.05	3.28	6378.73	49.22	1028.85

3. 结论

我们开发了一种可解释机器学习框架，通过高通量分子动力学模拟探索高热导率聚合物链。受类药小分子表征和分子力场的启发，我们通过分层筛选，将最初计算/收集的320个物理描述符缩减为20个优化描述符。构建的机器学习模型能够有效反映优化描述符与性质之间的关系，并在热导率预测中表现出较高的准确度。随机森林、XGBoost和多层感知机模型的R^2 均超过0.80，优于基于传统图描述符构建的模型。此外，随机森林模型通过SHAP分析捕捉到了横截面积和二面角刚度等优化描述符对热导率的促进或抑制作用。

利用训练好的机器学习模型，我们发现了107种有前景的聚合物，其热导率高于20.00瓦每米开尔文，其中29种的合成可及性评分不超过3.00。这些聚合物结构已通过高保真分子动力学模拟验证。此外，我们使用带有优化描述符的符号回归拟合有前景聚合物的热导率，得出的数学公式无需依赖机器学习模型，即可初步快速筛选高热导率聚合物，有利于实验研究。最后，我们通过声子谱能量密度分析计算了八种典型聚合物结构的声子色散关系，以揭示其潜在的热导机制。值得注意的是，这些结构大多是π-共轭聚合物，其重叠的p轨道有助于维持较强的链刚度和较大的群速度。我们的方法或可助力研究不限于热导率的高

性能聚合物。

4. 方法

4.1 聚合物建模与横截面积计算

聚合物建模是一个从单体构建聚合物链的过程，通过 STK 工具实现，输入参数为单体 SMILES 和聚合度 [73]。聚合物链的长度统一设为 50 nm，聚合度通过用链长除以单体长度并向上取整为整数获得。以聚合物 SMILES 为起点，使用 RDKit 生成聚合度为 2 的分子链，并用 MMFF 力场进行优化[74]。随后，通过测量热传输方向上两个重复单元中对应原子之间的距离，确定单体长度。建模完成后，利用 PYSIMM 的 Python 流程完成 GAFF2 力场参数的赋值，并生成分子动力学模拟输入结构数据文件 [75]。

横截面积是热导率分析的重要参数之一。在分子动力学模拟中，对于未占据整个模拟盒的体系，横截面积难以计算。横截面积通过范德华体积与单体长度的比率进行估算 [9]。单体的范德华体积通过对原子和键贡献求和计算得出，并已在以往的药物化合物研究中成功测试和应用[76]。

4.2 通过分子动力学模拟计算热导率

聚合物链的热导率是通过在大规模原子/分子大规模并行模拟器（LAMMPS）中进行非平衡分子动力学模拟（NEMD）获得的[75]。在非平衡分子动力学方法中，通过改进版热交换算法实现热能交换；该算法会对储热区内粒子的速度进行缩放和平移，以施加恒定的热通量[77]。聚合物链置于一个尺寸为540×60×60 (x×y×z) Å 的盒子中，其中y和z方向的尺寸设为 60 Å，以避免与相邻聚合物链发生相互作用。在计算热导率之前，对聚合物链结构进行弛豫，使其达到稳定构象。随后，沿x方向将聚合物链划分为 50 个薄层，并将链两端的固定区域设为隔热壁。在非平衡分子动力学模拟中，系统依次在 NVT（原子数、体积和温度恒定）和 NVE 、原子数、

体积和能量恒定）系综下运行 1 ns，温度为 300 K，以释放链应力[30,78]。随后，以恒定速率向聚合物链末端的热源/热汇区域（每个区域20 Å）添加/提取热量，以产生恒定的热通量。施加的热量因聚合物链结构而异，范围为 0.01 eV/ps 至 0.08 eV/ps。最后，对最后2~3 ns 的温度分布取平均值，并依据傅里叶定律计算热导率 $k = -J(d_T/d_x)$，其中 J 表示热通量，d_T/d_x 表示温度梯度。

4.3 描述符计算与机器学习模型构建

理想的聚合物描述符应尽可能精简，同时完整表征聚合物信息；这是决定机器学习算法预测准确率的关键因素之一。本研究所用的物理描述符来自 Mordred 软件计算和 GAFF2 力场参数提取。Mordred 软件最初是为化学信息学中的小分子特征而开发的，可计算 1800 多种描述符 [41]。然而，考虑到聚合物单体存在两种连接方式，我们仅获得了 286 个有效描述符。因此，作为补充，我们还从每个聚合物力场文件中提取参数作为描述符。对于图描述符，我们使用 RDKit 软件包计算了 MACCS、Morgan 和 cMorgan 指纹 [39]。Mol2vec 指纹则通过 Mol2vec 进行嵌入 [34]。我们参考了使用 PoLyInfo 和 PI1M 数据库训练的聚合物表征模型 [54]。

随机森林、XGBoost 和多层感知机机器学习模型均使用 Scikit-learn 实现 [79]。我们使用贝叶斯优化软件包 [80] 对随机森林、XGBoost 和多层感知机进行超参数优化。该软件包是一种全局优化工具，旨在获得良好的预测准确率 R^2。我们选择高斯回归过程和采集函数，并使用 10 组随机参数对进行初始训练；经过 100 次优化迭代后，确定了每种机器学习模型的理想参数 [46]。

为解释优化描述符与 TC 之间的关联，我们使用 SHAP 工具包结合随机森林模型评估特征重要性 [56]。SHAP 分析基于一种博弈论方法，将最优贡献分配与模型的局部解释相关联；该方法通过忽略每个特征来考察模型性能，并给出每个描述符影响的方向 [46]。*4.4* 通过符号回归（*SR*）拟合 *TC* 的数学公式

数学公式通过一种高效的逐步策略获取和筛选，该策略采用基于遗传编程的符号回归（GPSR），并由 gplearn 实现 [68]。这107种聚合物结构

热导率高于20.00 瓦每米开尔文的样本被随机按3:1的比例分别划分为训练集和测试集。首先，使用皮尔逊系数作为训练适应度的评价指标，以筛选优化描述符并生成子描述符，进而构建了一个包含22个描述符的新数据集。随后，采用表2所列的超参数和度量指标 R^2，通过网格搜索策略确定数学公式。最终，我们讨论了通过拉丁超立方抽样方法识别出的帕累托前沿上的四个公式 [70,71]。有关符号回归的更多信息，请参阅补充材料S9节。

表2。 gplearn 中 GPSR 的超参数设置

参数	**值**
代数	300
每代种群规模	5000
交叉概率（pc）	[0.30,0.90], 步长=0.05
子树变异 (ps)	[(1-pc)/3,(1-pc)/2] (step= 0.01)
提升变异 (ph)	[(1-pc)/3,(1-pc)/2] (步长 = 0.01)
点突变 (pp)	1-pc-ps-ph
函数集	$\{+, -, \times, \div, \sqrt{x}, \ln\ x, \lvert x\rvert, -x, 1/x\}$
简约系数	0.001、0.003、0.005
度量指标	R^2
停止准则	0.900
随机_状态	0, 1, 2, 3, 4
初始化_深度	[2, 6], [4, 8], [6, 10], [2, 10]

4.5 基于声子谱能量密度（声子谱能量密度）的声子色散关系分析

为理解聚合物的热导机制，采用分子动力学模拟结合声子谱能量密度方法，计算聚合物的色散关系。将长度为 100 Å 的聚合物链构建为 SMILES 输入，并置于横截面为60×60 Å 的盒子中。能量最小化后，系统先在 2 K 下的 NVT（原子数、体积和温度恒定）系综中运行 0.25 ns，以释放链应力。随后，系统在 NVE（原子数、体积和能量恒定）系综中运行 200 万步，时间步长为 0.25 fs。在此期间，每隔 20 步记录一次聚合物主链中每个原子的速度和位置。声子谱能量密度方法通过二维傅里叶变换，将原子速度和位置的时域信息转换为波矢与角频率的关系，表示为 [72]

$$\Phi(q,\omega)=\frac{1}{4\pi\tau_0 N_T}\sum_{\alpha}^{\{x,y,z\}}\sum_{b}^{B} m_b\left|\int_0^{\tau_0}\sum_{n}^{N_T}\dot{u}_\alpha(n,b;t)\times e^{iq\cdot r(n,0;t)-i\omega t}dt\right|^2 \quad (1)$$

其中， q 为波矢， ω 为频率， τ_0 为模拟时间， m_b 为b原子的质量， α 为笛卡尔方向， N_T 为聚合物链中晶胞的数量， $\dot{u}_\alpha(n,b;\ t)$表示 α 方向上、时间t时晶胞n中b原子的速度， $r(n,0;t)$表示晶胞n的平衡位置。

致谢

本研究得到上海浦江计划（编号：20PJ1407500）、国家自然科学基金委员会（编号：52006134）、上海市基础研究重点项目（编号：21JC1403300）和上海交通大学-华威大学联合种子基金的资助。本文中的计算均在由上海交通大学高性能计算中心支持的 π 2.0 集群上完成。

作者贡献

概念化：S. J.

方法学：X. H.、S. M. 和 S. J.

研究：X. H.、S. M. 和 S. J.

监督：S. J.、C. Y. Z. 和 H. W.

撰写——初稿：X. H.、S. M. 和 S. J.

撰写——审阅与编辑：X. H.、S. M.、S. J.、C. Y. Z. 和 H. W.

竞争性利益

作者声明，他们没有竞争性利益。

补充信息

补充信息可免费获取。

参考文献

1 Zhang, B. 等 聚合物与界面中的热输运调控：理论、模拟与实验。*ES Energy & Environment* **5**, 37-55, DOI:10.30919/esee8c306 (2019)。2 Ghaffari-Mosanenzadeh, S. 等 高热导聚合物复合材料综述。

《聚合物复合材料》43, 692-711, DOI:https://doi.org/10.1002/pc.26420 (2022)。3 Qian, X., Zhou, J. & Chen, G. 声子工程化极高热导率材料。《自然·材料》20, 1188-1202, DOI:10.1038/s41563-021-00918-3 (2021)。4 Guo, Y., Zhou, Y. & Xu, Y. 制备具有类金属热导率的工程化聚合物——现状与未来展望。《聚合物》233, 124168, DOI:https://doi.org/10.1016/j.polymer.2021.124168 (2021)。5 Ma, H. & Tian, Z. 聚合物拓扑结构和形貌对热输运的影响：瓶刷状聚合物的分子动力学研究。Applied Physics Letters 110, 091903, DOI:10.1063/1.4976946 (2017)。6 Xu, Y. 等. 具有类金属热导率的纳米结构聚合物薄膜。Nature Communications 10, 1771, DOI:10.1038/s41467-019-09697-7 (2019)。7 Ma, J. 等. 静电纺聚乙烯纳米纤维的热导率。《纳米尺度》7, 16899- 16908, DOI:10.1039/C5NR04995D (2015)。8 Singh, V. 等. 链取向非晶聚噻吩的高热导率。《自然·纳米技术》9, 384-390, DOI:10.1038/nnano.2014.44 (2014)。9 Liu, X., Lin, C. & Rao, Z. 直链聚四氟乙烯的热导率：分子动力学研究。《国际热科学杂志》159, 106646, DOI:https://doi.org/10.1016/j.ijthermalsci.2020.106646 (2021)。10 Crnjar, A., Melis, C. & Colombo, L. 评估单根一维PEDOT链中的反常超扩散热输运。《物理评论·材料》2, 015603, DOI:10.1103/PhysRevMaterials.2.015603 (2018)。11 Henry, A. & Chen, G. 利用分子动力学模拟研究单根聚乙烯链的高热导率。《物理评论快报》101, 235502, DOI:10.1103/PhysRevLett.101.235502 (2008)。12 Zhang, T., Wu, X. & Luo, T. 具有卓越热导率和热稳定性的聚合物纳米纤维：分子特征与宏观热学性质之间的基本联系。《物理化学杂志 C》118, 21148-21159, DOI:10.1021/jp5051639 (2014)。13 Liu, J. & Yang, R. 单根伸展聚合物链的长度依赖性热导率。《物理评论 B》86, 104307, DOI:10.1103/PhysRevB.86.104307 (2012)。14 Zhang, T. & Luo, T. 形貌影响的聚乙烯单链热导率

和晶体纤维。《应用物理学杂志》112，094304，DOI:10.1063/1.4759293（2012）。15 Liu, J., Ju, S., Ding, Y. & Yang, R. 超薄聚苯乙烯薄膜热导率的尺寸效应。Applied Physics Letters 104，153110，DOI:10.1063/1.4871737（2014年）。16 Kim, S. 等。PubChem 物质和化合物数据库。《核酸研究》44，D1202-D1213，DOI:10.1093/nar/gkv951（2015）。17 Yang, J., Tao, L., He, J., McCutcheon, J. R. & Li, Y. 机器学习助力发现用于气体分离膜的创新聚合物，并实现可解释发现。《科学进展》8，eabn9545，DOI:doi:10.1126/sciadv.abn9545（2022）。18 Chen, L. 等。聚合物信息学：现状与关键后续步骤。《材料科学与工程：R：综述》144，100595，DOI:https://doi.org/10.1016/j.mser.2020.100595（2021年）。19 Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. 材料信息学中的机器学习：近期应用与前景。npj《计算材料》3，54，DOI:10.1038/s41524-017-0056-5（2017年）。20 Wu, S., Yamada, H., Hayashi, Y., Zamengo, M. & Yoshida, R. 聚合物信息学的潜力与挑战：利用机器学习进行聚合物设计。arXiv 预印本 arXiv:2010.07683（2020）。21 Audus, D. J. & de Pablo, J. J. 聚合物信息学：机遇与挑战。《ACS大分子快报》6，1078-1082，DOI:10.1021/acsmacrolett.7b00228（2017年）。22 Huang, Y. 等。采用数据科学方法研究有机光伏聚合物材料的结构–性能相关性。《物理化学杂志 C》124，12871-12882，DOI:10.1021/acs.jpcc.0c00517（2020）。23 Nagasawa, S., Al-Naamani, E. & Saeki, A. 计算机辅助筛选用于有机太阳能电池的共轭聚合物：采用随机森林进行分类。《物理化学快报》9，2639-2646，DOI:10.1021/acs.jpclett.8b00635（2018年）。24 Afzal, M. A. F., Haghighatlari, M., Ganesh, S. P., Cheng, C. & Hachmann, J. 通过第一性原理分子建模、虚拟高通量筛选和数据挖掘，加速发现高折射率聚酰亚胺。《物理化学杂志 C》123，14610-14618，DOI:10.1021/acs.jpcc.9b01147（2019年）。25 Wu, C. 等。具有大带隙的柔性温度不变型聚合物电介质。

26 Wu, K. 等. 使用无限链描述符（ICD）和机器学习预测聚合物性质：迈向优化的介电聚合物材料。《聚合物科学杂志B辑：聚合物物理》54, 2082-2091，DOI：https://doi.org/10.1002/polb.24117（2016）。27 Sahu, H. 等. 一种用于设计导电聚合物的信息学方法。《ACS应用材料与界面》13, 53314-53322，DOI：10.1021/acsami.1c04017（2021）。28 Tao, L., Varshney, V. & Li, Y. 聚合物信息学机器学习模型的基准测试：以玻璃化转变温度为例。《化学信息与建模杂志》61, 5395-5413，DOI：10.1021/acs.jcim.1c01031（2021）。29 Bhowmik, R., Sihn, S., Pachter, R. & Vernon, J. P. 基于实验数据和机器学习方法预测聚合物的比热。《聚合物》220, 123558，DOI：https://doi.org/10.1016/j.polymer.2021.123558（2021）。30 Zhu, M.-X., Song, H.-G., Yu, Q.-C., Chen, J.-M. & Zhang, H.-Y. 机器学习驱动的高热导率聚合物分子结构发现。《国际传热传质杂志》162, 120381，DOI：https://doi.org/10.1016/j.ijheatmasstransfer.2020.120381（2020）。31 Ma, R., Zhang, H. & Luo, T. 利用强化学习探索高热导率非晶聚合物。《ACS应用材料与界面》14, 15587-15598，DOI：10.1021/acsami.1c23610（2022）。32 Wu, S. 等. 利用分子设计算法辅助机器学习发现高热导率聚合物。npj《计算材料》5, 66，DOI：10.1038/s41524-019-0203-2（2019）。33 Zhou, T., Wu, Z., Chilukoti, H. K. & Müller-Plathe, F. 使用基于分子动力学的遗传算法，通过序列工程设计高热导率聚乙烯–聚丙烯共聚物。《化学理论与计算杂志》17, 3772-3782，DOI：10.1021/acs.jctc.1c00134（2021）。34 Jaeger, S., Fulle, S. & Turk, S. Mol2vec：兼具化学直觉的无监督机器学习方法。《化学信息与建模杂志》58, 27-35，DOI：10.1021/acs.jcim.7b00616（2018）。35 Morgan, H. L. 为化学结构生成独特机器描述的方法——化学文摘社开发的一项技术。《化学文献杂志》5, 107-113（1965）。

36 Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. 用于药物发现的MDL键的再优化。《化学信息与计算机科学杂志》42, 1273- 1280, DOI:10.1021/ci010132r (2002年). 37 Capecchi, A., Probst, D. & Reymond, J.-L. 一种通用的分子指纹：药物、生物分子和代谢组。Journal of Cheminformatics 12, 43, DOI:10.1186/s13321- 020-00445-4 (2020). 38 D' Souza, S., Prema, K. V. & Balaji, S. 用于药物–靶点相互作用的机器学习模型：当前认知与未来方向。《今日药物发现》25, 748-756, DOI:https://doi.org/10.1016/j.drudis.2020.03.003 (2020). 39 Landrum, G. RDKit：用于化学信息学、计算化学和预测建模的软件套件。Greg Landrum (2013). 40 Karelson, M., Lobanov, V. S. & Katritzky, A. R. QSAR/QSPR 研究中的量子化学描述符。《化学评论》96, 1027-1044, DOI:10.1021/cr950202r (1996). 41 Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred：分子描述符计算器。Journal of Cheminformatics 10, 4, DOI:10.1186/s13321-018-0258-y (2018年). 42 Haghighatlari, M. 等. ChemML：用于分析、挖掘和建模化学与材料数据的机器学习与信息学程序包。《WIREs 计算分子科学》10, e1458, DOI:https://doi.org/10.1002/wcms.1458 (2020). 43 Mauri, A., Consonni, V., Pavan, M. & Todeschini, R. Dragon 软件：一种简便的分子描述符计算方法。Match 56, 237-248 (2006). 44 Roy, K. & Das, R. N. 基于扩展拓扑化学原子（ETA）指数的定量构效关系（QSTR）。16. 开发用于预测离子液体对大型溞毒性的分类与回归模型。《危险材料杂志》254-255, 166-178, DOI:https://doi.org/10.1016/j.jhazmat.2013.03.023 (2013). 45 Shields, B. J. 等. 贝叶斯反应优化作为化学合成工具。《自然》590, 89-96, DOI:10.1038/s41586-021-03213-y (2021年). 46 Tiihonen, A. 等. 通过机器学习预测共轭低聚电解质分子的抗微生物活性。《美国化学会志》143, 18917-18931, DOI:10.1021/jacs.1c05055 (2021年). 47 Domingos, P. 关于机器学习，有几件值得了解的事情。《通讯

ACM **55**，78-87（2012年）。

48 Marchetti, F.、Moroni, E.、Pandini, A. 和 Colombo, G. 基于分子动力学的变构药物活性机器学习预测。《物理化学快报》12，3724-3732，DOI:10.1021/acs.jpclett.1c00045（2021年）。49 Schober, P.、Boer, C. 和 Schwarte, L. A. 相关系数：适当的使用与解读。《麻醉与镇痛》126，1763-1768（2018年）。50 Chandrashekar, G. 和 Sahin, F. 特征选择方法综述。《计算机与电气工程》40，16-28，DOI:https://doi.org/10.1016/j.compeleceng.2013.11.024（2014年）。51 Kuenneth, C. 等。基于多任务学习的聚合物信息学。《Patterns》2，100238，DOI:https://doi.org/10.1016/j.patter.2021.100238（2021年）。52 Kamal, D. 等。使用计算和数据驱动技术的新型高压聚合物绝缘体。《化学物理学杂志》154，174906，DOI:10.1063/5.0044306（2021年）。53 Otsuka, S.、Kuwajima, I.、Hosoya, J.、Xu, Y. 和 Yamazaki, M. 载于《2011年新兴智能数据与网络技术国际会议》。22-29（IEEE）。54 Ma, R. 和 Luo, T. PI1M：用于聚合物信息学的基准数据库。《化学信息与建模杂志》，DOI:10.1021/acs.jcim.0c00726（2020）。55 Wang, J.、Wolf, R. M.、Caldwell, J. W.、Kollman, P. A. 和 Case, D. A. 通用 AMBER 力场的开发与测试。《计算化学杂志》25，1157-1174，DOI:https://doi.org/10.1002/jcc.20035（2004年）。56 Lundberg, S. M. 和 Lee, S.-I. 解释模型预测的统一方法。《神经信息处理系统进展》30（2017年）。57 Hayashi, Y.、Shiomi, J.、Morikawa, J. 和 Yoshida, R. RadonPy：利用全原子经典分子动力学模拟进行聚合物信息学物理性质自动计算。arXiv预印本 arXiv:2203.14090（2022年）。58 McInnes, L.、Healy, J. 和 Melville, J. UMAP：用于降维的均匀流形逼近与投影。arXiv预印本 arXiv:1802.03426（2018年）。59 Kavzoglu, T. 和 Mather, P. M. 特征选择在人工神经网络应用中的作用。《国际遥感杂志》23，2919-2937，DOI:10.1080/01431160110107743（2002年）。60 Ma, H. 和 Tian, Z. 链旋转显著降低单链聚合物的热导率。《材料研究杂志》34，126-133，DOI:10.1557/jmr.2018.362（2019年）。61 Hua, C.、Lindsay, L.、Chen, X. 和 Minnich, A. J. 非扩散型的广义傅里叶定律

热输运：理论与实验。Physical Review B 100, 085203, DOI:10.1103/PhysRevB.100.085203 (2019). 62 Robbins, A. B. & Minnich, A. J. 利用分子动力学研究具有极低热导率的晶态聚合物。Applied Physics Letters 107, 201908, DOI:10.1063/1.4936195 (2015). 63 Zhang, T. & Luo, T. 链形态和刚度对非晶聚合物热导率的影响。The Journal of Physical Chemistry B 120, 803-812, DOI:10.1021/acs.jpcb.5b09955 (2016). 64 Ertl, P. & Schuffenhauer, A. 基于分子复杂度和片段贡献估算类药分子的合成可及性评分。Journal of Cheminformatics 1, 8, DOI:10.1186/1758-2946-1-8 (2009). 65 Weng, B. et al. 由符号回归得出的简单描述符加速新型钙钛矿催化剂的发现。Nature Communications 11, 3513, DOI:10.1038/s41467-020-17263-9 (2020). 66 Wang, Y., Wagner, N. & Rondinelli, J. M. 材料科学中的符号回归。MRS Communications 9, 793-805, DOI:10.1557/mrc.2019.85 (2019). 67 Vladislavleva, E. Y. 通过帕累托遗传编程进行符号回归，以模型为基础解决问题。(CentER, Tilburg University Tilburg, 2008). 68 Stephens, T. Python中的遗传编程，采用受 scikit-learn 启发的 API：gplearn, 2022). 69 Zhou, Y., Rao, Y., Zhang, L., Ju, S. & Wang, H. 机器学习预测二元置换型金属固溶体的 Vegard定律因子和体积尺寸因子。Acta Materialia 237, 118166, DOI:https://doi.org/10.1016/j.actamat.2022.118166 (2022). 70 Paulson, N. H., Libera, J. A. & Stan, M. 通过统计学和机器学习优化火焰喷雾热解。Materials & Design 196, 108972, DOI:https://doi.org/10.1016/j.matdes.2020.108972 (2020). 71 Agarwal, G., Doan, H. A., Robertson, L. A., Zhang, L. & Assary, R. S. 利用量子化学引导的多目标贝叶斯优化发现储能分子材料。Chemistry of Materials 33, 8133-8144, DOI:10.1021/acs.chemmater.1c02040 (2021). 72 Thomas, J. A., Turney, J. E., Iutzi, R. M., Amon, C. H. & McGaughey, A. J. H. 根据谱能量密度预测声子色散关系和寿命。Physical Review B

81, 081411, DOI:10.1103/PhysRevB.81.081411（2010）。

73 Turcani, L., Berardo, E. & Jelfs, K. E. stk：用于超分子组装的 Python 工具包。Journal of Computational Chemistry 39, 1931-1942, DOI:https://doi.org/10.1002/jcc.25377 (2018). 74 Tosco, P., Stiefl, N. & Landrum, G. 将 MMFF 力场引入 RDKit：实现与验证。Journal of Cheminformatics 6, 37, DOI:10.1186/s13321-014- 0037-3 (2014). 75 Fortunato, M. E. & Colina, C. M. pysimm：用于分子系统模拟的 Python 软件包。SoftwareX 6, 7-12, DOI:https://doi.org/10.1016/j.softx.2016.12.002 (2017). 76 Zhao, Y. H., Abraham, M. H. & Zissimos, A. M. 通过对原子和键贡献求和快速计算范德华体积及其在药物化合物中的应用。The Journal of Organic Chemistry 68, 7368-7373, DOI:10.1021/jo034808o (2003). 77 Wirnsberger, P., Frenkel, D. & Dellago, C. 热交换算法的增强版本，具有出色的能量守恒特性。The Journal of Chemical Physics 143, 124104, DOI:10.1063/1.4931597 (2015). 78 Ju, S., Palpant, B. & Chalopin, Y. 聚合物涂层对固液界面热输运的不利影响。The Journal of Physical Chemistry C 121, 13474-13480, DOI:10.1021/acs.jpcc.7b02123 (2017). 79 Pedregosa, F. et al. Scikit-learn：Python中的机器学习。《机器学习研究杂志》12, 2825-2830 (2011). 80 Nogueira, F. 贝叶斯优化：用于 Python 的开源约束全局优化工具。网址 https://github. co m/fmfn/BayesianOptimization (2014).

补充材料，题为

利用物理描述符和可解释机器学习探索高热导率聚合物

Xiang Huang[1]、†、Shengluo Ma[1]、†、C. Y. Zhao[1], Hong Wang[2], 和 Shenghong Ju[1, 2, *]

[1] 中国上海上海交通大学中英低碳学院[2] 中国上海上海交通大学材料科学与工程学院材料基因组计划中心

†同等贡献

[*] 通讯作者：shenghong.ju@sjtu.edu.cn。

S1 节。聚合物描述符筛选

聚合物描述符分三个阶段进行筛选：移除方差（Var.）较低的特征、根据不同的相关系数（Cor.）进行初步筛选，以及借助机器学习模型（Opt.）进行最终选择，如图 S1 所示。

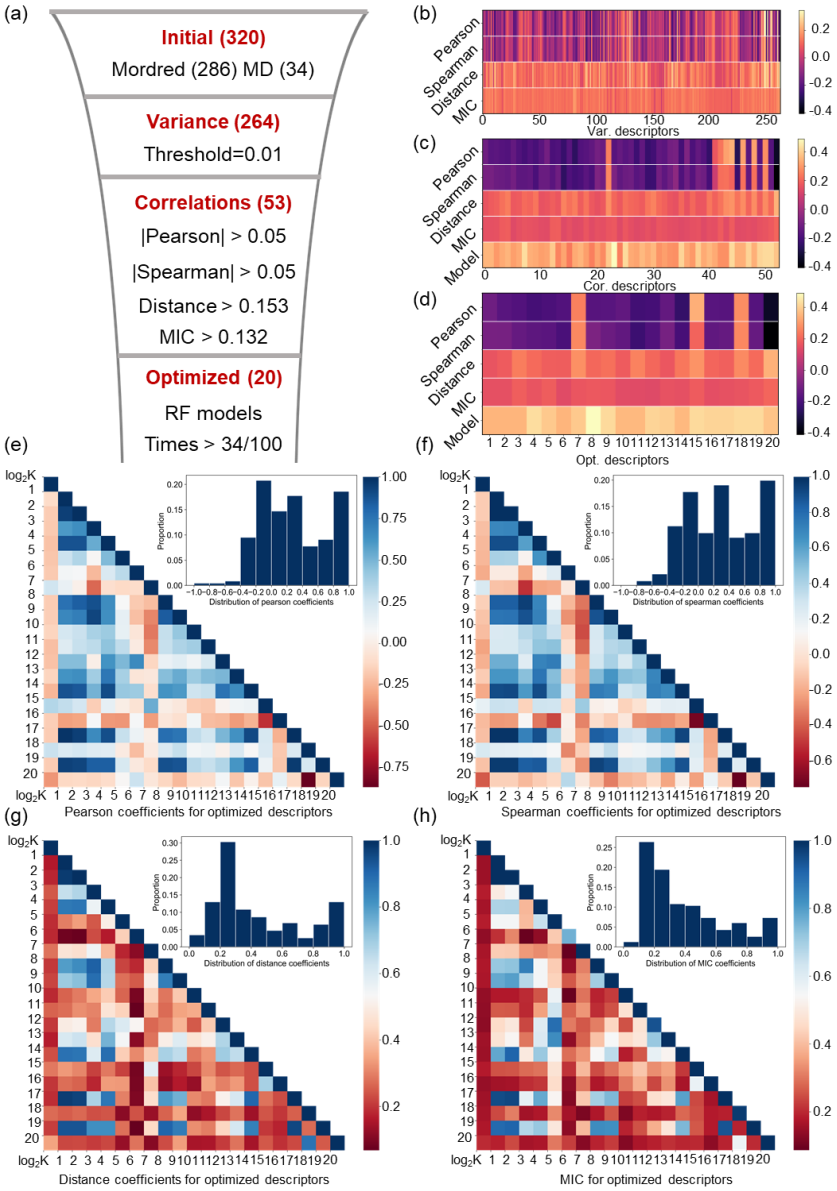


图 S1。描述符筛选与分析。（a）描述符筛选流程。（b）、（c）以及（d）基于各种度量指标或描述符在 100 次随机森林（RF）模型筛选中的出现频率，展示不同阶段（Var.、Cor. 和 Opt.）的描述符与热导率之间的关系。（e）、（f）、（g）和（h）分别为皮尔逊、斯皮尔曼、距离相关系数和最大信息系数（MIC）相关矩阵，展示优化描述符与热导率之间的相关性。插图分别显示相应相关系数值的分布统计。20 个优化描述符的说明见 S3 节。

S2 节：受分子动力学启发的描述符和基于 Mordred 的指纹列表

受分子动力学启发的描述符包括与力场相关的描述符（FF-related）和与热导率相关的描述符（TC-related），列于表 S1 中。

表 S1. 所有受分子动力学启发的描述符的说明

Name	说明
Epsilon_max	能量势深度的最大值（Lennard–Jones 参数）
Epsilon_min	能量势深度的最小值（Lennard–Jones 参数）
Epsilon_平均值	能量势深度的平均值（Lennard–Jones 参数）
Sigma_最大值	平衡距离的最大值（Lennard–Jones 参数）
Sigma_最小值	平衡距离（Lennard–Jones 参数）的最小值
Sigma_平均值	平衡距离（Lennard–Jones 参数）的平均值
Kb_max	键的力常数最大值
Kb_min	键的力常数最小值
Kb_平均值	键的力常数平均值
R0_max	键的平衡结构参数最大值
R0_min	键的平衡结构参数最小值
R0_平均值	键的平衡结构参数平均值
Ka_max	键角的力常数最大值
Ka_min	键角力常数最小值
Ka_平均值	键角力常数平均值
Theta0 _max	键角平衡结构参数最大值
Theta0 _min	键角平衡结构参数的最小值
Theta0_平均值	键角平衡结构参数的平均值
Kd_max	二面角力常数的最大值
Kd_min	二面角力常数最小值
Kd_平均值	二面角力常数平均值
Nd_max	扭转角参数多重性的最大值
Nd_min	扭转角参数多重性的最小值
Nd_平均值	扭转角参数多重性的平均值
德尔塔_max	扭转角参数的相位角最大值
Delta _min	扭转角参数的相位角最小值
Delta_平均值	扭转角参数的相位角平均值
Ki_max	反常角力常数的最大值
Ki_min	反常角力常数最小值
Ki_平均值	反常角力常数平均值
MW	单体分子量
MW_比率	主链分子量与单体分子量的比率
Vdw	单体的范德华体积
横截面	聚合物链的等效横截面积

基于 Mordred 的描述符由 Mordred 软件 [1] 计算得出。本研究中的 286 个基于 Mordred 的描述符列于表 S2，详细含义可参见 https://mordred-descriptor.github.io/documentation/master/descriptors.html（访问日期: 2022年8月10日）。

表 S2。286 个基于 Mordred 的描述符的说明

ABC	酸性基团数	碱性基团数	SpMax_A	SpMAD_A
LogEE_A	VE1 A_	VR1 A_	VR3 A_	芳香原子数
原子数	桥头原子数	杂原子数	nH	nN
nO	nS	nF	nCl	nBr
ATS0dv	ATS0Z	AATS0dv	AATS0d	AATS2d
AATS0Z	AATS1Z	AATS2Z	AATS3Z	ATSC0dv
ATSC1dv	ATSC2dv	ATSC3dv	ATSC4dv	ATSC5dv
ATSC6dv	ATSC7dv	ATSC8dv	ATSC1d	ATSC2d
ATSC3d	ATSC4d	ATSC5d	ATSC6d	ATSC7d
ATSC8d	ATSC0Z	ATSC1Z	ATSC2Z	ATSC3Z
ATSC4Z	ATSC5Z	ATSC6Z	ATSC7Z	ATSC8Z
AATSC0dv	AATSC1dv	AATSC2dv	AATSC3dv	AATSC0d
AATSC1d	AATSC2d	AATSC3d	AATSC0Z	AATSC1Z
AATSC2Z	AATSC3Z	MATS1dv	MATS2dv	MATS3dv
MATS1d	MATS2d	MATS3d	MATS1Z	MATS2Z
MATS3Z	GATS1dv	GATS2dv	GATS3dv	GATS1d
GATS2d	GATS3d	GATS1Z	GATS2Z	GATS3Z
BCUTdv-1h	BCUTdv-1l	BCUTd-1h	BCUTd-1l	BCUTZ-1h
BCUTZ-1l	BalabanJ	BertzCT	nBondsD	nBondsT
C1SP1	C2SP1	C1SP2	C3SP2	C1SP3
C2SP3	C3SP3	C4SP3	HybRatio	Xch-4d
Xch-5d	Xch-6d	Xch-4dv	Xc-3d	Xc-4d
Xc-5d	Xc-6d	Xpc-4d	AXp-0d	AXp-1d
AXp-2d	AXp-3d	NsCH3	NdCH2	NssCH2
NdsCH	NsssCH	NddC	NdssC	NaaaC
NssssC	NsNH2	NssNH	NaaNH	NdsN
NaaN	NsssN	NaasN	NsOH	NdO
NssO	NaaO	NdS	NssS	NaaS
NddssS	SsssCH	SdssC	SaasC	SaaaC
SsssN	SaasN	AETA_beta	AETA_beta_s	ETA_beta_ns_d
AETA_beta_ns_d	AETA_eta_RL	AETA_eta_BR	ETA_epsilon_3	ETA_dBeta
片段复杂度	fMF	氢键受体数	IC0	IC1
IC2	IC3	SIC0	SIC1	SIC2
SIC3	SIC4	BIC5	CIC0	MIC0
MIC1	MIC2	Kier2	Kier3	FilterItLogS
PEOE_VSA1	PEOE_VSA2	PEOE_VSA3	PEOE_VSA4	PEOE_VSA6
PEOE_VSA7	PEOE_VSA8	PEOE_VSA9	PEOE_VSA10	PEOE_范德华表面积11
PEOE_VSA12	PEOE_VSA13	SMR_VSA1	SMR_VSA3	SMR_VSA4
SMR_VSA6	SMR_VSA9	SlogP_VSA1	SlogP_VSA2	SlogP_VSA3

SlogP_VSA4	SlogP_VSA5	SlogP_VSA7	SlogP_VSA8	SlogP_VSA10
SlogP_范德华表面积11	EState VSA1_	EState VSA2_	EState VSA3_	EState VSA4_
EState _VSA5	EState _VSA6	EState _VSA7	EState _VSA8	EState _VSA9
EState _VSA10	VSA_EState1	VSA_EState2	VSA_EState3	VSA_EState4
VSA_EState5	VSA_EState7	VSA_EState8	VSA_EState9	AMID_h
AMID_N	AMID_O	AMID_X	环数	五元环数
六元环数	七元环数	氢原子环数	四元杂环数量	五元杂环数量
六元杂环数量	五元芳香环数量	芳香杂环数量	六元芳香杂环数量	芳香环数量
五元芳香环数量	六元芳香环数量	芳香杂环数量	五元芳香杂环数量	六元芳香杂环数量
氟原子数Ring	n7氟原子数Ring	n8氟原子数Ring	n9氟原子数Ring	n10氟原子数Ring
n11氟原子数Ring	n12FRing	nG12FRing	nFHRing	n7FHRing
n8FHRing	n10FHRing	nG12FHRing	nFaRing	n8FaRing
n9FaRing	n10FaRing	nG12FaRing	nFaHRing	nFARing
n10FARing	n12FARing	nG12FARing	nFAHRing	nG12FAHRing
旋转比	SLogP	TopoPSA(NO) 拓扑极性表面积		GGI4
GGI5	GGI6	GGI7	JGI1	JGI2
JGI3	JGI4	JGI5	JGI6	JGI7
JGI8	JGI9	JGI10	JGT10	拓扑形状指数
mZagreb1				

S3 节. 聚合物热导率（TC）预测的优化描述符说明

预测

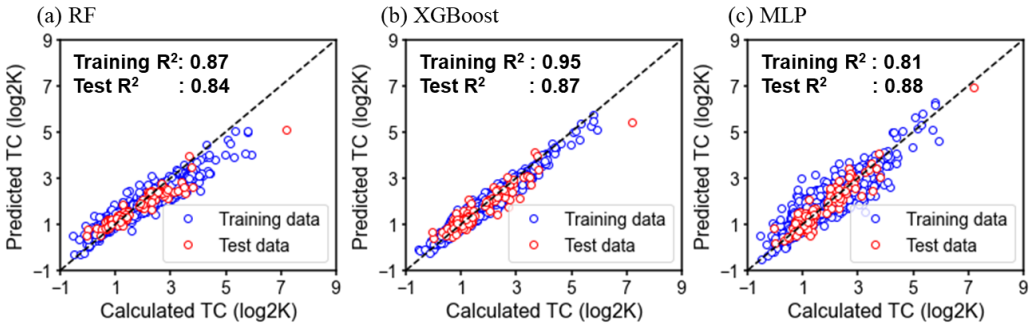
聚合物优化描述符通过筛选获得，列于表 S3。

表 S3. 所有受分子动力学启发的描述符说明

编号	名称	说明	区块
1	ABC	原子键连接指数	原子键连接指数描述符
2	LogEE_A	邻接矩阵的 LogEE	邻接矩阵描述符
3	Nh	氢原子数	原子计数描述符
4	ATSC0dv	中心化 Moreau–Broto 自相关滞后 0 时按价电子加权	中心化自相关拓扑结构描述符
5	BCUTdv-1h	Burden 矩阵的第一大特征值由价电子加权	BCUT 描述符
6	Xc-4d	由……加权的 4 阶 Chi 簇	Chi 描述符
7	ETA_dBeta	ETA δβ	ETA δβ 描述符
8	CIC0	零阶互补信息含量	互补信息含量内容描述符
9	Kier2	Kappa形状指数2	Kappa形状指数2描述符
10	SMR_VSA6	MOE MR VSA描述符6（2.75 <= x < 3.05）	使用MOE类型描述符Wildman-Crippen MR和表面积贡献
11	SLogP_VSA3	MOE logP VSA描述符3（-0.20 <= x < 0.00）	使用MOE类型描述符Wildman-Crippen LogP和表面积贡献
12	EState _VSA10	EState VSA 描述符 10（9.17 <= x < 15.00）	使用 MOE类型描述符EState 指数和表面积贡献
13	TopoPSA(NO)	拓扑极性表面积（仅使用氮和氧）	拓扑极性表面积描述符
14	mZagreb1	修正扎格雷布指数（版本1）	扎格雷布指数描述符
15	Kd_平均值	力常数平均值	FF相关描述符
16	Nd_平均值	二面角的多重性平均值扭转角参数	FF相关描述符
17	MW	单体分子量	与热导率相关的描述符
18	MW_比率	的分子量比率主链与单体的分子量	与热导率相关的描述符
19	Vdw	单体的范德华体积	热导率相关描述符
20	横截面	等效横截面积聚合物链	热导率相关描述符

S4节：使用优化描述符构建不同的机器学习模型

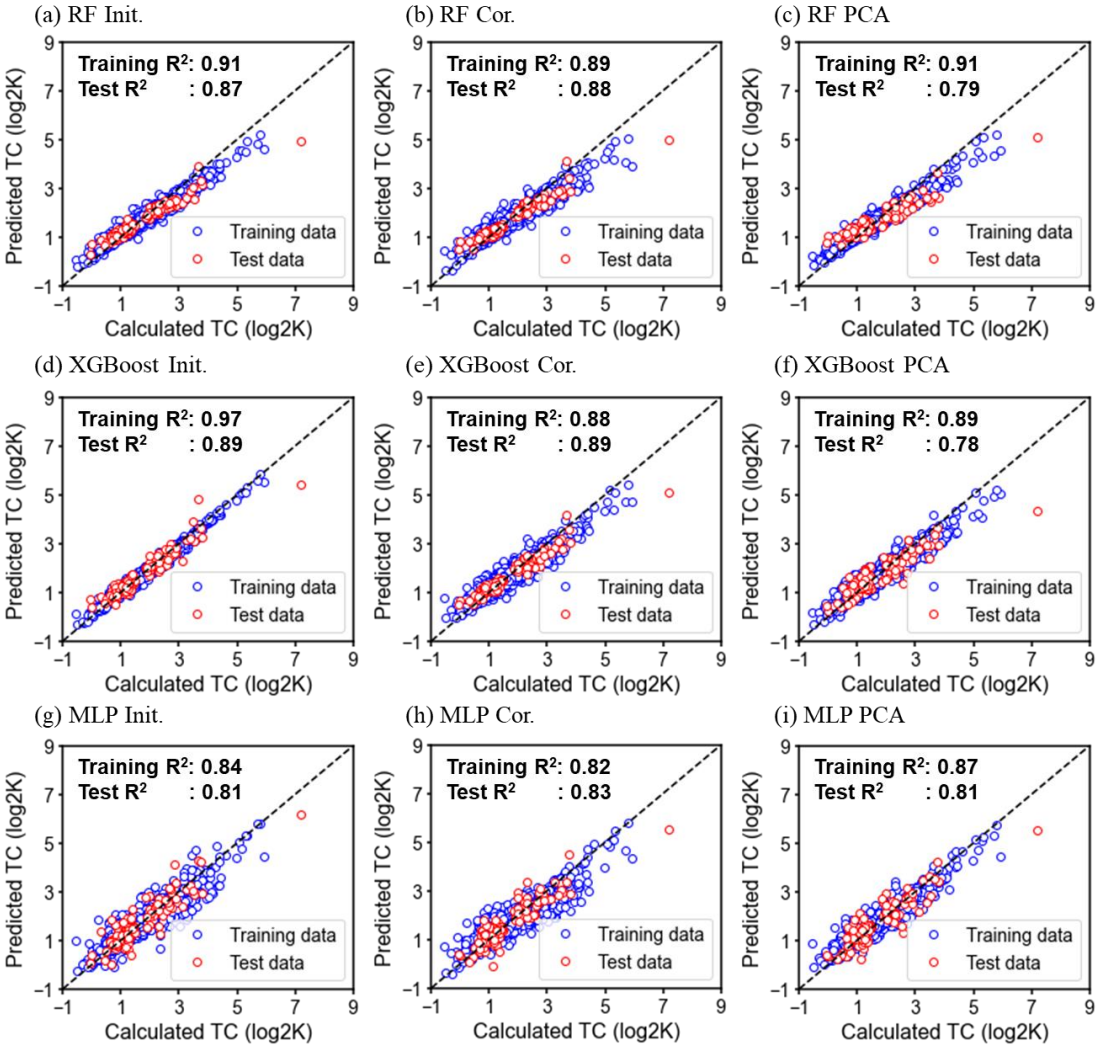
图 S2 展示了使用优化描述符训练的随机森林、XGBoost 和多层感知机模型。测试集由从基准数据集中随机选取的 100 个聚合物结构组成。



图S2。 使用优化描述符构建不同的机器学习模型

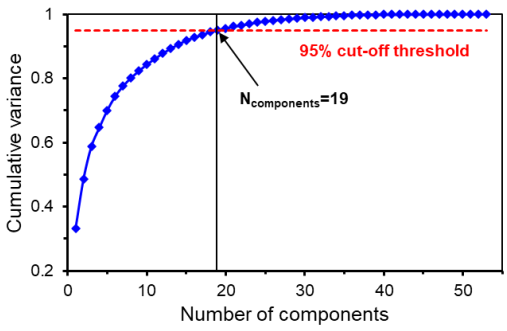
S5节. 不同描述符降选阶段机器学习模型的准确度

图S3展示了不同描述符降选阶段机器学习模型的准确度。为进行比较，采用了具有95%方差的额外主成分分析（主成分分析）[2]技术。



图S3. 不同描述符降选阶段机器学习模型的准确度

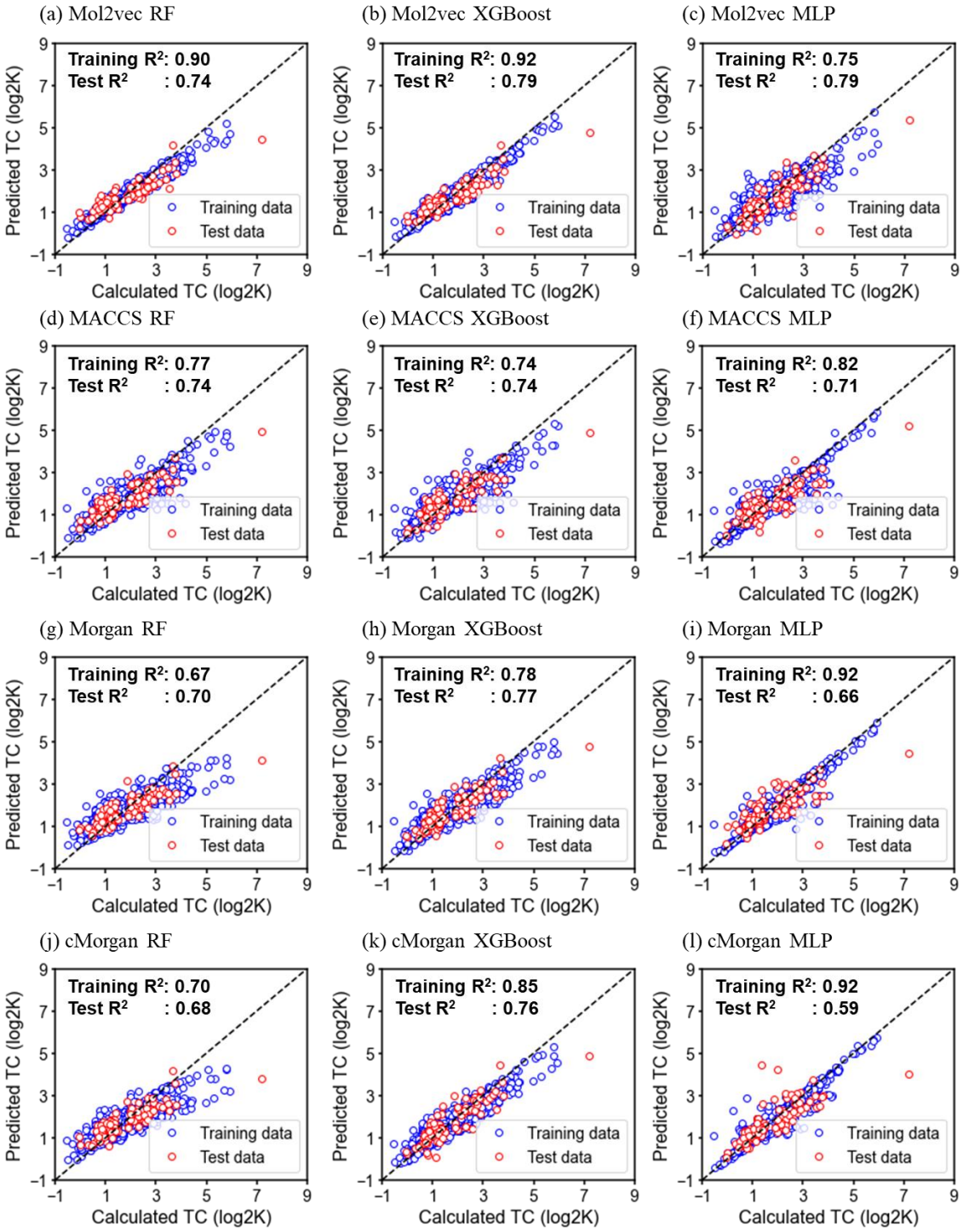
图S4展示了主成分数量与累积方差之间的关系。此处采用了包含19个主成分、方差为95%的主成分分析。



图S4. 主成分数量与累积方差之间的关系 e

S6节。采用各种图描述符的机器学习模型准确度

图 S5 展示了采用 Mol2vec [3]、MACCS [4]、Morgan 和 Morgan 计数（cMorgan）[5] 图 描述符的机器学习模型的训练准确度和测试准确率。



图S5。 采用各种图描述符的机器学习模型准确度

第 S7 节 特征重要性分析

图 S6 展示了训练数据集中聚合物的每个优化描述符对应的 SHAP 值，[6] 并将其绘制为对应描述符值的函数。各子图按平均 SHAP 值排序。总体而言，训练数据集中聚合物对应的排名靠前特征的 SHAP 值与特征值呈单调关系，而排名靠后的特征则难以捕捉这一规律。其中，两个重要的受分子动力学启发的描述符与热导率之间的关系如图 S7 所示。

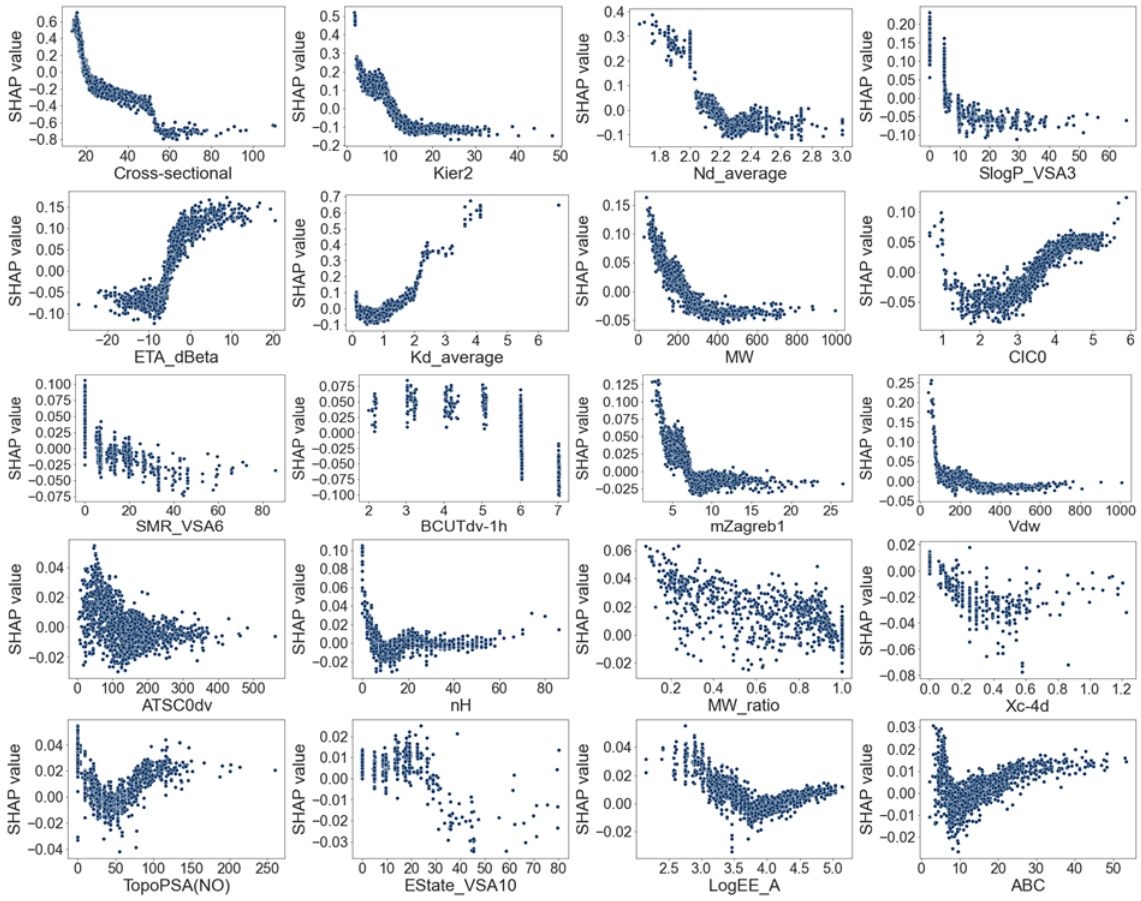


图 S6。各优化描述符的 SHAP 值分布

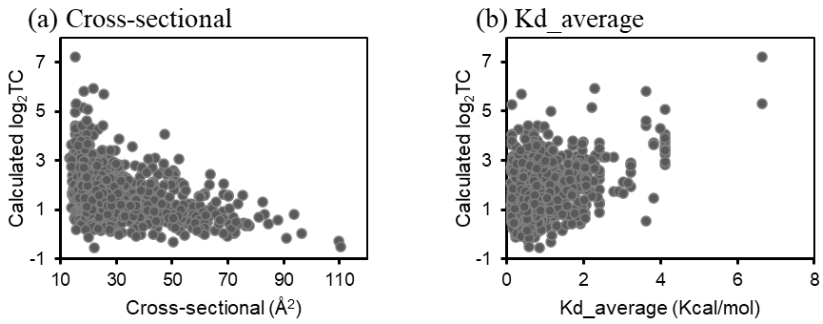


图 S7。两个重要的受分子动力学启发的描述符与 1735 个基准样本的热导率之间的关系

聚合物。

S8节：利用所提出的机器学习方法发现高热导率聚合物（PHTC）

工作流程

我们分别基于训练好的随机森林、XGBoost 和多层感知机机器学习模型，对 PolyInfo 和 PI1M 中的聚合物数据进行了虚拟筛选。本研究共筛选出 107 种热导率大于 20 瓦每米开尔文的聚合物结构，并通过分子动力学模拟进行了验证，结果列于表 S3。本研究共筛选出 107 种热导率大于 20 瓦每米开尔文的聚合物结构，并通过分子动力学模拟进行了验证，结果列于表 4；相关单体结构可根据聚合物编号（PID）在图S8中查看。随机森林、XGBoost 和多层感知机的模型预测及分子动力学模拟所得热导率均为对数$_2$单位为瓦每米开尔文的热导率。聚合物的有效横截面积以Å^{2}表示。

表 S4. 经分子动力学验证的热导率大于 20 瓦每米开尔文的聚合物结构

PID	SMILES	**SA** 评分	交叉-分段	随机森林	XGBoost	多层感知机	分子动力学
PHTC001	[*]c1ccc([*])cc1	1.05	18.32	5.00	5.77	6.28	5.81
PHTC002	[*]CC[*]	1.12	15.69	3.91	4.66	5.30	5.28
PHTC003	[*]c1cccc([*])c1	1.50	20.74	4.02	5.21	5.86	4.59
PHTC004	[*]CC(=O)N[*]	2.07	15.13	4.00	4.34	4.05	4.41
PHTC005	[*]c1cccc([*])n1	2.15	19.72	4.65	4.98	5.65	5.06
PHTC006	[*]C=Cc1ccc([*])cc1	2.17	16.76	4.29	4.36	5.48	5.13
PHTC007	[*]c1ccc2cc([*])ccc2c1	2.17	18.28	3.55	4.37	5.01	5.12
PHTC008	[*]CC(=O)O[*]	2.21	15.09	4.01	4.45	4.36	4.98
PHTC009	[*]c1cccc2c([*])cccc12	2.22	23.86	3.17	3.81	4.36	4.35
PHTC010	[*]c1ccc([*])c(Cl)c1	2.23	21.89	3.73	4.06	5.34	4.61
PHTC011	[*]c1ccc2ccc([*])cc2c1	2.28	19.21	3.81	4.49	5.05	4.62
PHTC012	[*]OCC([*])=O	2.28	20.86	3.31	4.17	3.70	4.33
PHTC013	[*]c1ccc([*])nc1	2.33	16.98	4.85	5.08	5.45	4.50
PHTC014	[*]c1ccc([*])nn1	2.48	15.68	4.87	5.91	4.97	5.05
PHTC015	[*]C(F)(F)C([*])(F)F	2.51	25.57	4.05	5.27	5.07	5.71
PHTC016	[*]NC(=O)N[*]	2.57	14.34	4.53	4.91	3.76	5.47
PHTC017	[*]c1cnc([*])cn1	2.62	15.75	5.11	6.08	5.85	6.62
PHTC018	[*]c1ccc2nc([*])ccc2n1	2.64	16.79	3.96	4.00	5.62	5.93
PHTC019	[*]Oc1ccc([*])nc1	2.65	14.79	4.16	4.46	4.17	4.43
PHTC020	[*]Oc1ccc([*])cn1	2.65	14.74	4.16	4.52	4.17	4.76
PHTC021	[*]c1ccc2nc([*])ccc2c1	2.78	17.62	3.96	4.19	5.59	4.38
PHTC022	[*]c1cncc([*])n1	2.80	18.06	5.08	5.84	5.53	4.34
PHTC023	[*]c1ccnc([*])n1	2.82	18.06	5.08	5.82	5.52	5.00
PHTC024	[*]氮原子数C([*])=O	2.86	14.28	4.19	4.56	3.32	4.83

PID	SMILES	SA 评分	交叉-截面	随机森林	XGBoost	多层感知机	分子动力学
PHTC025	[*]c1csc([*])n1	2.87	15.82	5.11	6.20	5.76	5.75
PHTC026	[*]c1cnc2cc([*])ccc2c1	2.91	17.29	3.99	4.24	5.63	4.82
PHTC027	[*]c1ccc2oc([*])nc2c1	2.97	15.17	3.91	3.55	3.59	4.34
PHTC028	[*]c1ccc(-c2ccc([*])s2)s1	3.00	15.98	4.06	4.26	4.22	5.64
PHTC029	[*]C=Cc1ccc([*])c(F)c1	3.04	17.60	3.89	4.34	4.80	4.41
PHTC030	[*]C=C/C=C\[*]	3.07	15.68	5.06	5.35	5.79	5.33
PHTC031	[*]碳碳键/碳=碳键/[*]	3.07	14.06	5.05	5.10	5.97	7.55
PHTC032	[*]/C=C/C=C/[*]	3.07	14.26	5.05	5.10	5.95	6.91
PHTC033	[*]c1nc([*])nc(Cl)n1	3.09	13.85	4.53	5.64	5.63	6.51
PHTC034	[*]C=C[*]	3.12	15.35	5.08	5.44	6.92	7.21
PHTC035	[*]C=C/C=C/C=C/[*]	3.20	13.73	4.65	4.34	4.86	7.42
PHTC036	[*]/C=C/C=C/C=C/[*]	3.20	13.60	4.65	4.34	4.87	7.31
PHTC037	[*]c1nnc([*])s1	3.27	8.76	5.09	6.04	6.14	5.75
PHTC038	[*]CCOC(=O)NCCNC(=O)O[*]	3.33	15.55	3.20	4.10	3.65	4.38
PHTC039	[*]C(=O)c1cnc([*])cn1	3.34	15.80	4.11	4.34	4.35	4.54
PHTC040	[*]c1cnc([*])s1	3.35	15.34	4.86	5.77	5.40	4.95
PHTC041	[*]C1=Cc2ccc([*])cc21	3.37	18.78	4.37	4.64	5.50	5.11
PHTC042	[*]OC(=O)O[*]	3.38	11.93	3.97	4.05	3.96	5.61
PHTC043	[*]c1ccc2nc([*])ncc2n1	3.39	15.82	3.98	3.95	5.54	5.83
PHTC044	[*]c1cc(Cl)c([*])nn1	3.45	19.29	4.19	5.08	4.97	4.52
PHTC045	[*]c1cnc([*])nn1	3.49	14.33	5.06	5.97	5.57	6.46
PHTC046	[*]c1nnc([*])c(Cl)c1Cl	3.51	15.39	4.11	5.12	5.41	4.60
PHTC047	[*]c1cnnc([*])n1	3.53	16.38	5.04	5.92	5.33	5.60
PHTC048	[*]c1cnc([*])n1C	3.55	19.00	4.25	4.46	5.63	4.49
PHTC049	[*]c1ccc(-c2nnc([*])s2)s1	3.63	14.53	4.15	4.39	4.83	5.39
PHTC050	[*]c1nnc([*])c(F)c1F	3.64	12.41	4.40	5.32	4.51	4.75
PHTC051	[*]c1ccc(-c2cnc([*])s2)s1	3.76	15.25	4.17	4.44	4.85	5.51
PHTC052	[*]C(Cl)=C([*])Cl	3.86	20.41	4.17	4.38	4.81	5.13
PHTC053	[*]/C(Cl)=C(/[*])Cl	3.86	20.56	4.13	4.42	4.80	5.08
PHTC054	[*]C#C[*]	3.87	5.85	4.23	4.58	5.15	9.63
PHTC055	[*]C1=CC=C1[*]	3.90	22.30	4.36	5.72	5.65	4.62
PHTC056	[*]c1nnc([*])c(O)n1	3.90	16.45	4.50	5.18	4.20	4.96
PHTC057	[*]c1noc([*])c1Br	3.90	11.67	4.61	5.55	4.77	4.97
PHTC058	[*]C(=O)N1C(=O)N([*])C1=O	4.04	16.85	3.66	4.36	5.49	5.37
PHTC059	[*]NC(=O)C(=S)C([*])=O	4.08	21.85	4.01	5.07	4.60	5.91
PHTC060	[*]C(=O)c1nnc([*])nn1	4.09	11.89	4.36	5.35	4.15	5.95
PHTC061	[*]c1cc2nc([*])sc2nn1	4.09	14.51	4.38	4.81	5.11	4.35
PHTC062	[*]C(=O)C([*])=S	4.20	16.81	3.78	4.92	4.41	5.03
PHTC063	[*]C1=C(F)C(Cl)=C1[*]	4.28	17.13	4.69	4.59	5.00	5.84

PID	SMILES	SA 评分	交叉-截面	随机森林	XGBoost	多层感知机	分子动力学
PHTC064	[*]C1=CC=C2C([*])=CC=C12	4.32	22.02	3.94	3.73	4.63	6.16
PHTC065	[*]NC(=O)C([*])=N	4.32	18.67	3.86	5.17	3.12	4.37
PHTC066	[*]SC([*])=O	4.34	8.22	4.70	5.29	4.52	6.73
PHTC067	[*]/C=C/C=C([*])Cl	4.36	17.41	4.65	3.72	4.63	5.90
PHTC068	[*]/C(F)=C(/[*])C#N	4.36	20.54	3.77	5.11	3.47	6.02
PHTC069	[*]C=Cc1csc(-c2cc([*])cs2)c1	4.45	15.88	3.69	3.61	3.08	4.33
PHTC070	[*]/C=C/C=C/C=C([*])Cl	4.45	15.96	4.28	4.78	3.99	5.66
PHTC071	[*]/C=C/C=C/C=C(/[*])Cl	4.45	15.96	4.28	4.78	3.99	5.66
PHTC072	[*]/C(F)=C(/[*])Cl	4.46	18.39	4.44	5.10	4.90	6.18
PHTC073	[*]C(F)=C([*])Cl	4.46	17.50	4.69	5.25	4.94	5.83
PHTC074	[*]OC([*])=S	4.53	14.49	4.63	5.07	3.76	5.20
PHTC075	[*]C#CC=C[*]	4.54	18.03	4.51	5.02	5.18	5.16
PHTC076	[*]OC([*])=C	4.54	19.19	4.07	4.36	3.57	4.55
PHTC077	[*]/C=C(/[*])Cl	4.55	21.62	4.28	4.80	5.14	6.03
PHTC078	[*]C=C([*])C#N	4.58	24.93	3.53	4.00	2.70	4.88
PHTC079	[*]C(=S)C(=O)C(F)(F)C([*])(F)F	4.59	25.30	3.19	4.28	3.92	4.41
PHTC080	[*]SS[*]	4.66	13.04	3.77	3.91	0.95	4.93
PHTC081	[*]/C=C/C=C/C=C([*])N	4.66	15.44	4.20	4.41	4.69	5.53
PHTC082	[*]NNC([*])=N	4.67	15.08	4.45	4.61	4.41	4.51
PHTC083	[*]C1=CC2=NC([*])=CC2=N1	4.75	17.77	4.55	4.07	4.51	6.10
PHTC084	[*]C#CC#CC=C[*]	4.78	12.75	4.11	3.79	4.61	5.99
PHTC085	[*]C=C([*])F	4.82	17.93	4.71	5.35	4.81	6.64
PHTC086	[*]/C=C/N[*]	4.83	15.68	5.02	5.28	6.52	5.44
PHTC087	[*]C#C/C(C#N)=C(/[*])C#N	4.88	15.14	3.46	5.00	4.71	5.37
PHTC088	[*]/N=C/C(=S)N[*]	4.88	18.77	4.37	4.29	5.00	5.06
PHTC089	[*]/N=N/C([*])=O	4.90	8.82	4.55	5.76	6.04	7.35
PHTC090	[*]/C=C/O[*]	4.92	15.71	3.94	3.83	3.43	4.60
PHTC091	[*]=CC=氮原子数=[*]	4.98	5.95	4.69	4.92	7.53	4.71
PHTC092	[*]NN[*]	5.00	11.88	4.79	5.37	6.33	4.60
PHTC093	[*]/N=N/[*]	5.05	5.21	4.83	5.95	6.81	10.38
PHTC094	[*]氮=氮[*]	5.05	5.93	4.83	5.95	6.76	10.01
PHTC095	[*]/N=C(/[*])Cl	5.08	9.62	5.06	5.71	6.50	6.98
PHTC096	[*]C#C/C(C#N)=C(/[*])Cl	5.14	14.06	3.63	5.06	4.58	5.79
PHTC097	[*]/C=N/C=C(/[*])C	5.14	17.30	4.02	3.64	4.36	4.72
PHTC098	[*]/C=C/C=CS[*]	5.17	14.14	4.14	3.28	4.13	4.36
PHTC099	[*]C1=CC2=CC([*])=NC2=N1	5.28	17.50	4.62	4.10	4.55	5.79
PHTC100	[*]/C=C/N=C([*])Cl	5.38	19.65	4.16	4.37	4.61	5.60
PHTC101	[*]/C=C/S[*]	5.44	14.40	4.57	4.20	5.74	4.35
PHTC102	[*]/C=N/[*]	5.53	14.02	5.10	6.07	6.07	6.82

PID	SMILES	SA 评分	交叉-分区	随机森林	XGBoost	多层感知机	分子动力学
PHTC103	[*]/N=N/N=C(/[*])C#N	5.66	11.24	3.69	5.07	4.33	6.68
PHTC104	[*]/C=N/C=C/C=C(/[*])Cl	5.70	15.50	4.33	4.46	4.07	5.70
PHTC105	[*]/C=C(/[*])[O-]	5.73	18.54	4.07	4.66	2.87	6.10
PHTC106	[*]/C=C/C=C/N=N/[*]	5.76	12.71	4.74	4.11	7.01	6.59
PHTC107	[*]C([2H])=C([*])[2H]	7.27	11.30	4.32	4.33	1.07	7.48

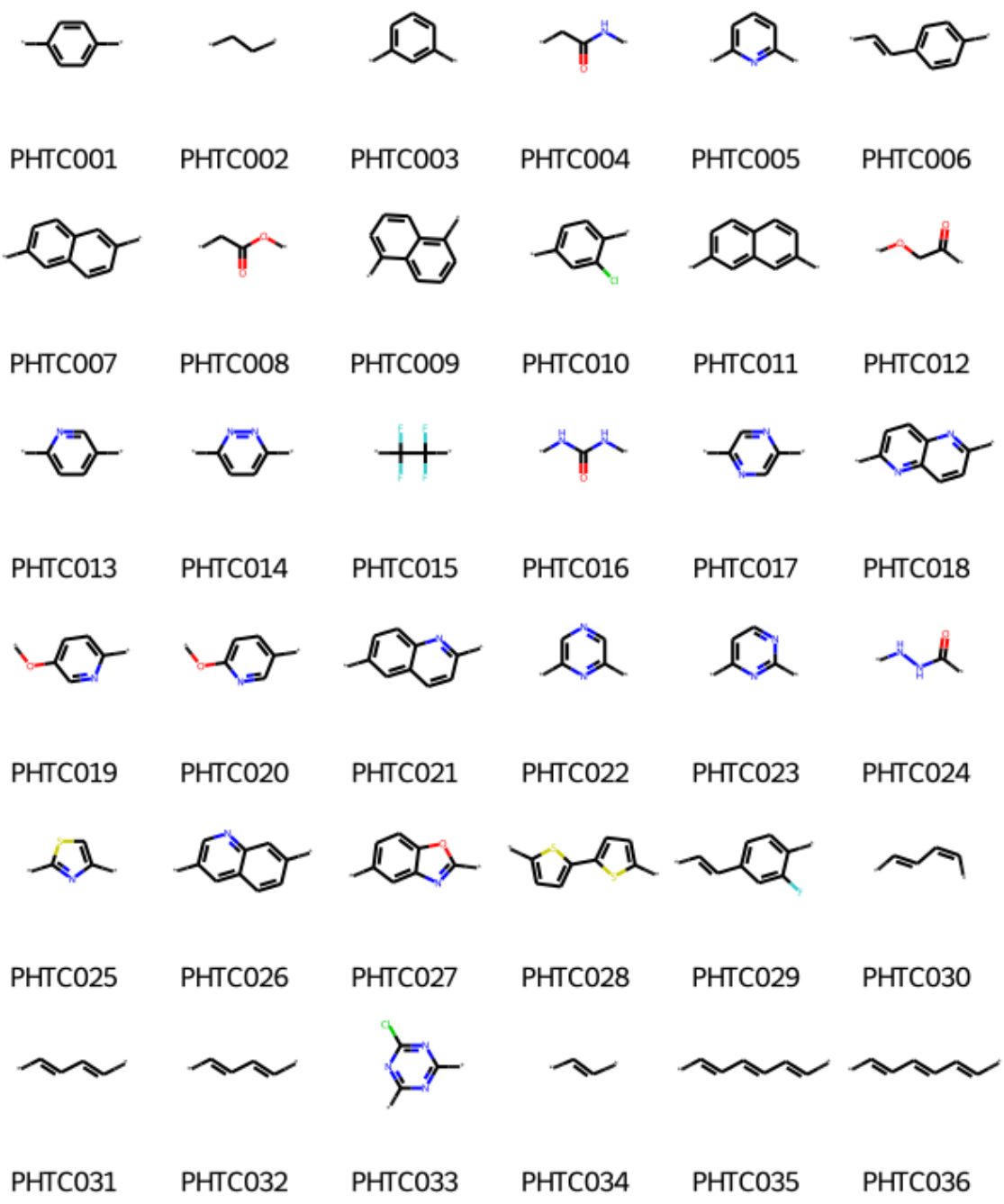


图 S8。续

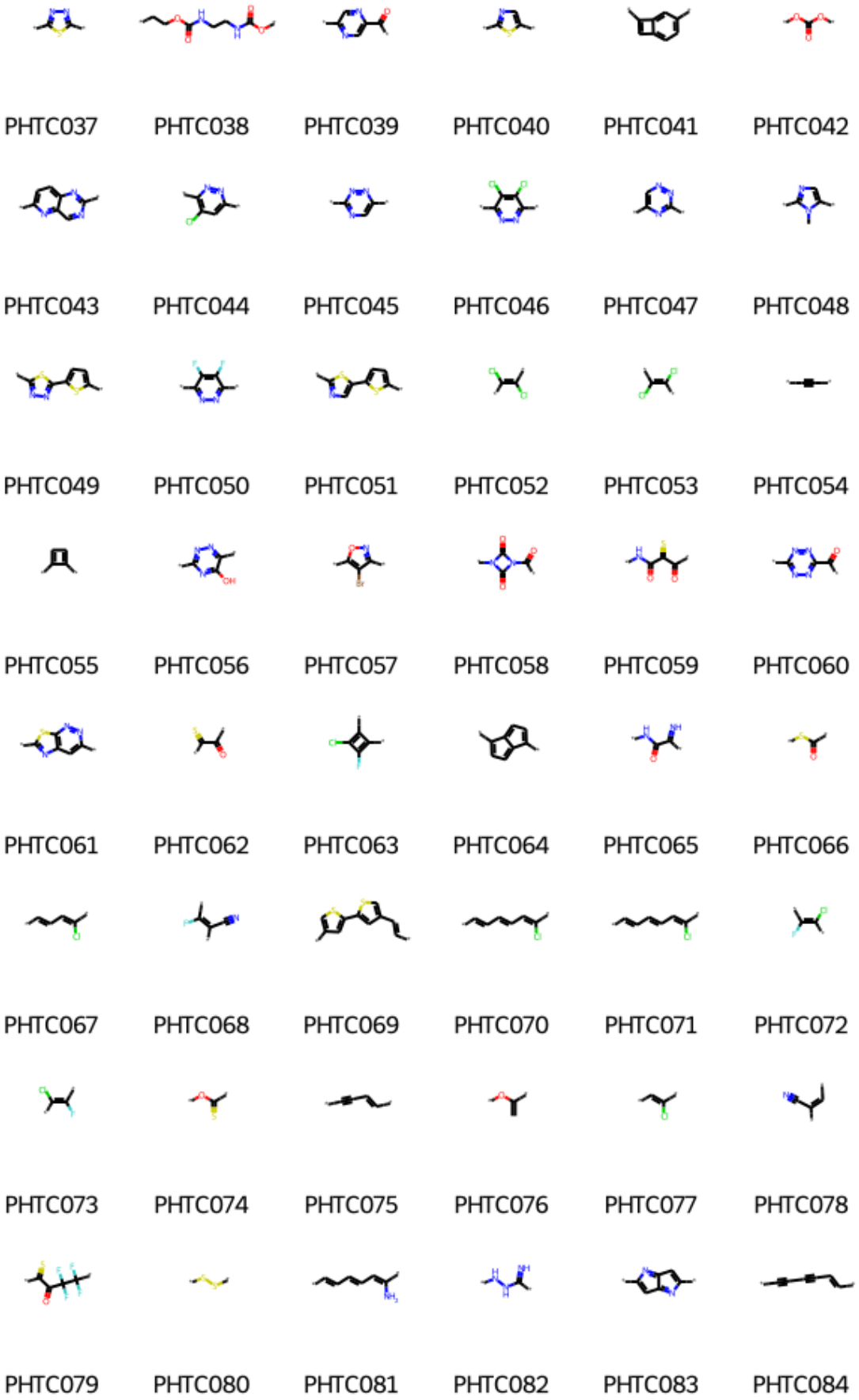

PHTC037
PHTC038
PHTC039
PHTC040
PHTC041
PHTC042
PHTC043
PHTC044
PHTC045
PHTC046
PHTC047
PHTC048
PHTC049
PHTC050
PHTC051
PHTC052
PHTC053
PHTC054
PHTC055
PHTC056
PHTC057
PHTC058
PHTC059
PHTC060
PHTC061
PHTC062
PHTC063
PHTC064
PHTC065
PHTC066
PHTC067
PHTC068
PHTC069
PHTC070
PHTC071
PHTC072
PHTC073
PHTC074
PHTC075
PHTC076
PHTC077
PHTC078
PHTC079
PHTC080
PHTC081
PHTC082
PHTC083
PHTC084

图 S8. 续

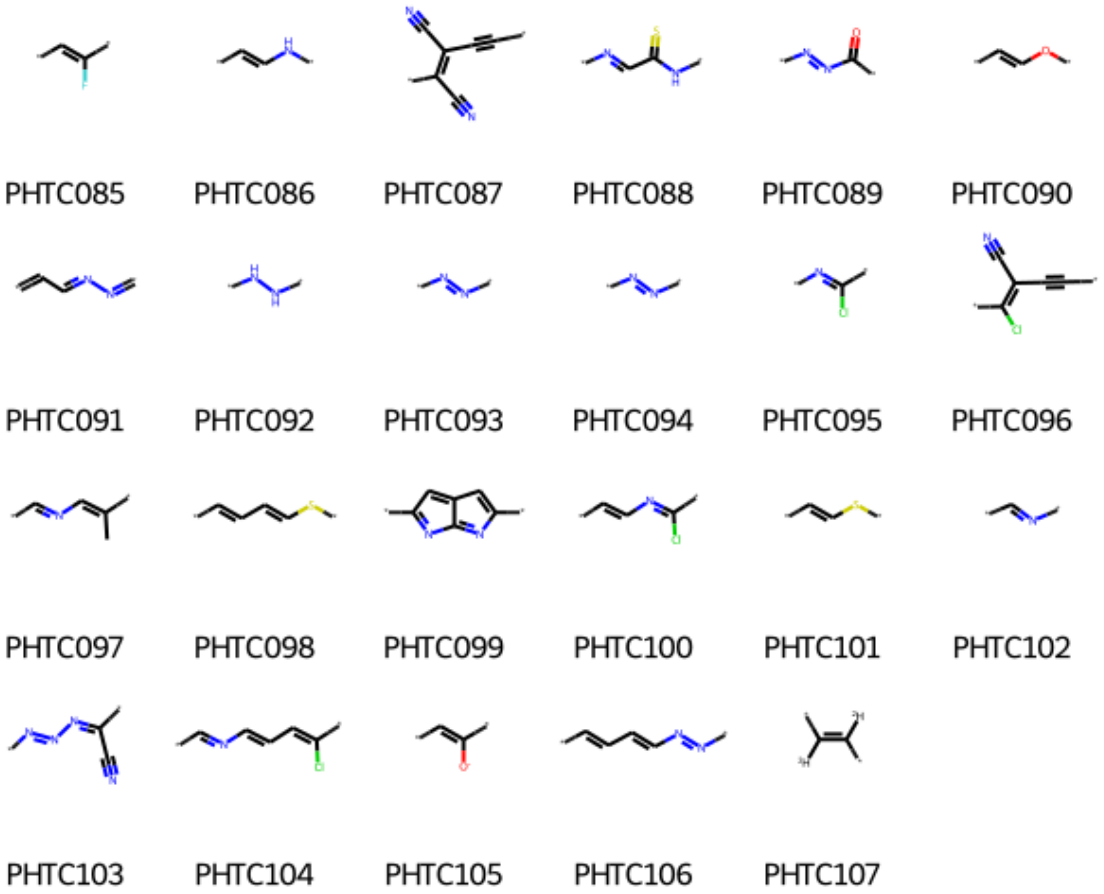


图 S8。 通过分子动力学模拟验证的热导率大于 20 瓦每米开尔文的聚合物单体结构。使用 PID 可在表 S4 中找到聚合物性质。

S9节. 用符号回归表征有前景聚合物的热导率

将 107 种具有优化描述符的有前景聚合物结构（热导率 > 20 瓦每米开尔文）用于符号回归（符号回归），并随机选取其中 20% 的结构作为测试集。采用基于遗传编程（GPSR）的高效逐步策略进行符号回归，并通过 gplearn 代码获取和筛选数学公式 [7]。首先，使用皮尔逊系数作为训练适应度的评价指标，以筛选优化描述符并生成子描述符；同时，在 GPSR 中应用网格搜索策略，所用超参数列于表S5。对应于 22380 个公式的超参数组合，其复杂度和皮尔逊系数（PC）训练适应度如图S9a所示。在 gplearn 中，复杂度等同于公式长度。在这些公式中，有 4365 个公式的 PC 值不低于 0.85，其复杂度统计结果见图S9b。总体而言，复杂度较高的公式具有更高的训练适应度，但这类公式通常也相当冗长，不利于计算和分析。因此，只有适应度高且复杂度低的公式才合适；我们选取了 158 个 PC 值不低于 0.85 且复杂度不超过 10 的公式进行后续分析。通过统计 158 个公式中 20 个优化描述符的出现频率，最终保留了前 8 个描述符，如图S9c所示。值得强调的是，受分子动力学启发的横截面积（横截面）和二面角力常数（Kd_平均）描述符出现在每个公式中。此外，我们还统计了通过取倒数（$-x$）、对数（ln x）等方式生成的子描述符的出现频率，如图S9d所示。因此，表 S6 列出了一个包含 22 个与热导率密切相关描述符的新集成，用于在 gplearn 中重新训练。

此外，我们重置了表 S7 中列出的网格搜索超参数，并采用准确度 R^2作为评估指标。图 S10a 展示了拟合准确度大于 0.6 的 19580 个公式与复杂度之间的关系。有效准确度是训练准确率和测试准确率的平均值。尽管部分公式的有效准确度超过 0.8，且复杂度大于 50。图 S10b 通过密度图展示了复杂度不超过 30 的 9073 个数学公式的准确度 R^2 与复杂度之间的关系，其中 c、d、e 和 f 点通过拉丁超立方抽样方法 [8,9]确定。这四个点对应的公式列于表 S8，其拟合结果与通过分子动力学标记的 $\log_2$热导率进行比较，结果见图10c-f。变量x_0 至 x_{21} 的定义见表 S6。

表S5。用于筛选和生成新子描述符的 gplearn 超参数设置

参数	值	组合
代数	300	1
每代种群规模	1000,2000	2
交叉概率 (pc)	[0.30,0.90], 步进=0.05	746
子树变异 (ps)	[(1-pc)/3,(1-pc)/2] (step= 0.01)	
提升变异 (ph)	[(1-pc)/3,(1-pc)/2] (步长 = 0.01)	
点突变 (pp)	1-pc-ps-ph	
函数集	$\{+, -, \times, \div, \sqrt{x}, \ln\ x, \lvert x\rvert, -x, 1/x\}$	1
简约系数	auto	1
度量指标	皮尔逊系数	1
停止准则	0.900	1
随机_状态	0, 1, 2, 3, 4	5
初始_深度	[2, 6], [4, 8], [6, 10]	3

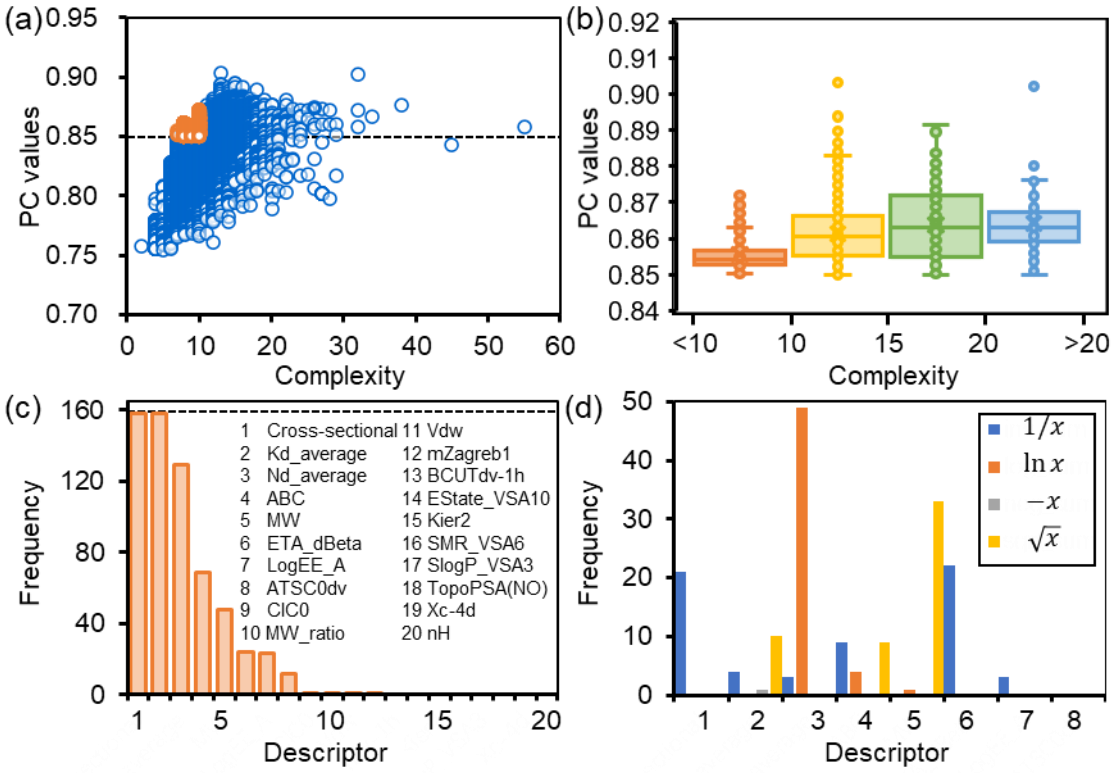


图 S9. 通过皮尔逊系数（PC）值筛选优化描述符，并在 gplearn 中创建子描述符。(a) 数学公式复杂度与 PC 值的关系。(b) 按复杂度统计皮尔逊系数不低于 0.85 的公式。(c) 和 (d) 优化描述符和子描述符在 158 个数学公式（PC 值 >=0.85，复杂度 <=10）中的出现频率。

表 S6. 用于 GPSR 的新描述符集成

变量	格式	PC 值	变量	格式	PC 值
x_0	ABC	-0.37	x_{11}	-ETA dBeta_	0.11
x_1	10/ABC	0.44	x_{12}	Kd_平均值	0.20
x_2	$\sqrt{\text{ABC}}$	-0.39	x_{13}	$\sqrt{\text{Kd_平均值}}$	0.12
x_3	ATSC0dv	-0.32	x_{14}	LogEE_A	-0.40
x_4	横截面	-0.52	x_{15}	MW	-0.39
x_5	10/横截面	0.63	x_{16}	分子量/100	-0.39
x_6	1-10/横截面	0.59	x_{17}	100/分子量	0.52
x_7	100/横截面	0.63	x_{18}	$\sqrt{\text{MW}}$	-0.43
x_8	$\sqrt{100/\text{横截} - \text{面}}$	0.62	x_{19}	Nd_平均值	-0.31
x_9	-10/横截面	-0.63	x_{20}	log Nd_平均值	-0.35
x_{10}	ETA dBeta_	-0.11	x_{21}	− log Nd_平均值	0.44

表 S7。 GPSR 中 gplearn 超参数的重置

参数	值	组合
代数	300	1
每代种群规模	5000	1
交叉概率 (pc)	[0.30,0.90], 步长=0.05	746
子树变异 (ps)	[(1-pc)/3,(1-pc)/2] (step= 0.01)	
提升变异 (ph)	[(1-pc)/3,(1-pc)/2] (步长 = 0.01)	
点突变 (pp)	1-pc-ps-ph	
函数集	$\{+, -, \times, \div, \sqrt{x}, \ln\ x, \lvert x\rvert, -x, 1/x\}$	1
简约系数	0.001, 0.003, 0.005	3
度量指标	R^2	1
停止准则	0.900	1
随机_状态	0, 1, 2, 3, 4	5
初始_深度	[2, 6], [4, 8], [6, 10], [2, 10]	4

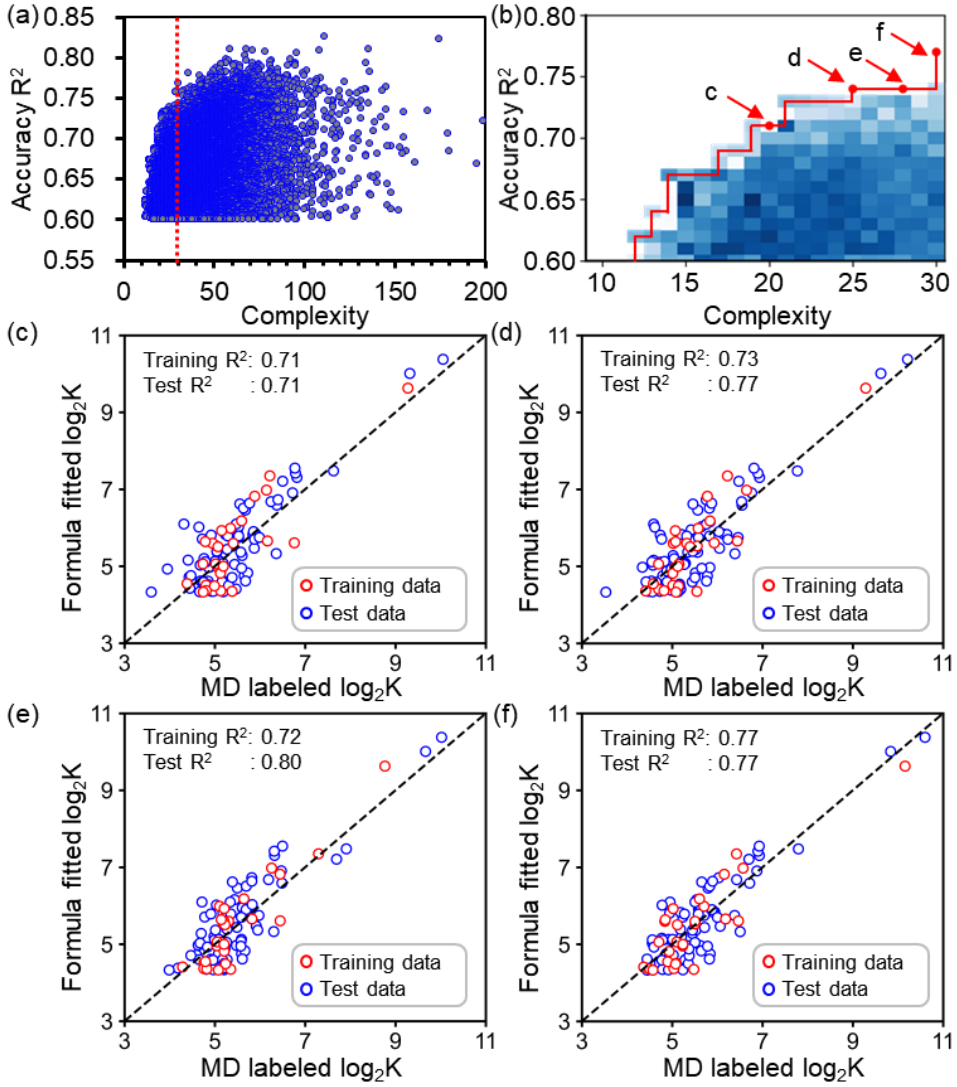


图 S10. 利用 GPSR 对有前景的聚合物进行热导率预测。(a) 19,580 个拟合准确度大于 0.6 的公式与复杂度的关系。(b) 通过密度图展示的 9,073 个数学公式的准确度决定系数与复杂度之间的帕累托前沿。c、d、e 和 f 四个点代表帕累托前沿上的四个公式，其拟合结果与分子动力学标注的热导率的对比图分别显示在 (c)、(d)、(e) 和 (f) 中。

表 S8。 图 S10b 中帕累托前沿上的四个数学公式

点	公式	R^2	复杂度
c	$log_2 K = \sqrt{(x_{12} - x_8 - x_{16} + 0.49) \times x_9 \times (x_6 + x_{20})} + \sqrt{ln\ (x_{12})}$	0.71	20
d	$log_2 K = \sqrt{\dfrac{+ x_{21}}{x_{12} - ln\left[(x_8 - x_{11}) \times \dfrac{x_8}{ln\ (\sqrt{x_{19}})}\right] \times (x_6 + x_{20}) \times x_7}}$	0.74	25
e	$log_2 K = ln\left[\dfrac{+ x_{21} + \sqrt{ln\, x_{12}}}{x_{10} \times \dfrac{1}{(x_1 - x_2)}} - x_{12}\right] \times (x_5 - 0.824)$ $+ \sqrt{\dfrac{x_{17}}{\dfrac{x_3}{x_{12} - x_{11}() \times x_5}}} + x_{21} + x_{13}$	0.74	28
f	$log_2 K = \sqrt{x_8 + \left[\dfrac{x_{10}}{\dfrac{ln\ ln\, x_{17} - \chi_{12}() + x_8}{}} - x_{12}\right] \times (x_6 + x_{20}) \times \dfrac{x_8}{x_{20}}}$ $+ x_{21} + \sqrt{x_{19} - x_{16}}$	0.77	30

参考文献

1 Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred：分子描述符计算器。*Journal of Cheminformatics* **10**，4，doi:10.1186/s13321-018-0258-y（2018年）。2 Abdi, H. & Williams, L. J. 主成分分析。*Wiley interdisciplinary reviews:computational statistics* **2**，433-459（2010）。3 Jaeger, S., Fulle, S. & Turk, S. Mol2vec：融合化学直觉的无监督机器学习方法。《化学信息与建模杂志》 **58**，27-35，doi:10.1021/acs.jcim.7b00616（2018年）。4 Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. 用于药物发现的 MDL 键再优化。《化学信息与计算机科学杂志》 **42**，1273-1280，doi:10.1021/ci010132r（2002年）。5 Morgan, H. L. 化学结构的唯一机器描述的生成——一项由化学文摘社开发的技术。《化学文献杂志》 **5**，107-113（1965年）。6 Lundberg, S. M. & Lee, S.-I. 解释模型预测的统一方法。《神经信息处理系统进展》 **30**（2017年）。7 Stephens, T. Python中的遗传编程，采用受 scikit-learn 启发的 API：gplearn，2022年）。8 Paulson, N. H., Libera, J. A. & Stan, M. 通过统计学和机器学习优化火焰喷雾热解。*Materials & Design* **196**，108972，doi:https://doi.org/10.1016/j.matdes.2020.108972（2020）。9 Agarwal, G., Doan, H. A., Robertson, L. A., Zhang, L. & Assary, R. S. 利用量子化学引导的多目标贝叶斯优化发现储能分子材料。*Chemistry of Materials* **33**，8133-8144，doi:10.1021/acs.chemmater.1c02040（2021年）。