

Research
AI in Chemical Engineering—Article

一种改进的纯组合物性预测机器学习模型

曹欣羽^a, 贡铭^b, Anjan Tula^{a,*}, 陈曦^{a,*}, Rafiqul Gani^{c,d,e}, Venkat Venkatasubramanian^f^a State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China^b Department of Physics, Bard College at Simon's Rock, Great Barrington, MA 01230, USA^c PSE for SPEED Company, Charlottenlund DK-2920, Denmark^d Sustainable Energy and Environment Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 510000, China^e Department of Applied Sustainability, Széchenyi István University, Győr 9026, Hungary^f Complex Resilient Intelligent Systems Laboratory, Department of Chemical Engineering, Columbia University, New York, NJ 10027, USA

ARTICLE INFO

Article history:

Received 17 February 2023

Revised 1 June 2023

Accepted 1 August 2023

Available online 27 April 2024

关键词

基团贡献法

高斯过程

扭曲函数

先验预测检查

摘要

了解物质的物理化学性质是进行工艺设计和产品设计等任务的重要前提。然而,数据的匮乏和高昂的实验成本阻碍了这些性质预测技术的发展。此外,准确性和预测能力也限制了大多数性质预测方法的范围和适用性。本文提出了一种新的基于高斯过程的建模框架,旨在处理由基团贡献法表示分子结构的离散和高维输入空间。扭曲函数被用来将离散输入映射到连续域,以调整不同化合物之间的相关性。在机器学习建模过程中,本文还应用了先验选择技术(包括先验推断和先验预测检查)以提供更多来自之前研究结果的信息。该框架使用不同规模的数据集对20种纯组合物性进行了评估。对于其中18种纯组合物性,新模型相比其他已发表的模型(无论是否使用机器学习)表现出更高的准确性和预测能力。

© 2024 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. 引言

预测化学物质行为的能力主要取决于对其物理化学性质的了解。因此,纯组分的物性数据是有效执行工艺和产品设计任务的重要前提[1–2]。在工艺设计中,需要确定已识别化合物的性质(溶解度、蒸气压等)与设计目标匹配的条件(温度、压力等);在工艺设计中,需要选择具有理想目标性质(沸点、临界压力等)的化合物[3]。然而,缺乏纯组合物性的数据是对所需性质建模的一个关键限制。通过做实验的方法来获得物性数据,存在要求高、费用昂贵等劣势,有些性质甚至不可测量[4]。因此,需

要准确和鲁棒的性质预测模型。近些年来,统计分析计算机软件方面飞速进展,可以弥补不完整测量物性数据的不足[5]。已知一类纯组合物性,即主要性质,与分子的结构有关[6]。本文重点研究有机化学品主要纯组分的物性,利用其分子结构信息建模。

在过去几十年中,纯化合物性质预测研究取得了迅速发展。从简单的多项式函数到非常大的微分代数系统集的数学模型都被用来估计所需的纯化合物性质[3]。其中,传统方法包括基团贡献(GC)方法[7]、定量结构性质关系(QSPR)建模[8]、*ab initio*量子力学方法[9]等。GC方法是最广泛用于主要纯组合物性的物理化学性质预测的方

* Corresponding authors.

E-mail addresses: anjantula@zju.edu.cn (A. Tula), xi_chen@zju.edu.cn (X. Chen).

法。在这种方法中，物性被确定为代表分子的功能基团贡献的函数[10]。其快速预测无需大量计算资源，并且易于嵌入其他模型，加速了GC方法在毒性预测[11]、熔点预测[12]和生物质转化处理[13]等不同领域的应用。另一方面，预测准确性低下限制了GC模型的使用[7]。因此，不少研究提出了不同的方法来提高GC模型的性能。比如，更高级的描述符可以通过二阶功能基团揭示更多结构信息[10,14]，GC⁺模型则是为表示缺失的GC而创建的[1,7,15]。

最近的研究显示，基于机器学习的模型进行性质预测[16]以及基于人工智能(AI)技术来识别具有良好性质的潜在分子结构[17]的趋势正在增加。Venkatasubramanian和Mann [18]使用AI进行反应预测和化学合成。此外，在最近的一篇观点文章中，Mann等[19]强调了在AI时代性质预测在化学产品设计中的应用。基于机器学习的方法，如神经网络和随机森林算法，在使用不同分子描述符进行性质预测中起主要作用。许多方法相比传统建模技术具有显著优势，包括灵活性、准确性和执行速度[20]。它们的可行性在减少与量子力学/分子力学计算相关的计算成本[21]、新型QSPR方法[22]等方面得到了证明。例如，Zhou等[23]将简化的分子输入线条条目系统(SMILES)符号视为句子，并使用自然语言处理技术进行分子信息挖掘和化学性质探索；Zhang等[24]开发了一个准确且可解释的深度神经网络(DNN)模型用于性质预测；Wen等[25]提出了一种系统方法，结合多种机器学习技术解决DNN-based QSPR建模中的适用领域和预测不确定性等关键问题。此外，在功能基团表示领域，许多机器学习模型在估计和预测能力方面表现更好，但代价是消耗更多的计算资源，这扩展了GC-based应用的范围。例如，Papuszyński和Domańska [26]采用了基于GC类型分子描述符的两层前馈人工神经网络，它被证明是当时文献中描述离子液体(ILs)黏度的最佳GC模型；Li等[27]通过MATLAB回归学习模块使用23种机器学习算法开发回归模型预测燃料点火质量，达到了高精度。在现有机器学习模型中，值得注意的是，具有置信区间的高斯过程(GP)[28]是一种广泛使用的性质预测方法。它对可能的目标函数先给定先验，并在训练过程中通过更新基于观察数据的贝叶斯后验逐步优化模型[29]。置信区间促进了GP在许多领域的发展，包括安全关键环境[30-31]、不确定性预测[32-33]和贝叶斯优化[34-35]。GP的另一个优点是，在建模前需要较少的模型结构信息(如架构和学习率)[28]。鉴于这些突出的特点，Alshehri等[36]开发了基于GP的下一代有机化学品25种纯组合物性模型，比简单的GC模型具有更高的物性预测准确性。

虽然研究人员经常使用各种预测方法，但在开发性质模型方面仍存在几个挑战。在大多数情况下，简单的GC模型表现较差，平均误差阈值约为10% [36]。对于机器学习方法，尽管误差较小，但在训练集之外的外推能力仍然有限。其中，部分原因在于原始化合物与分子描述符数学表示之间的信息差距。此外，严格应用模型而不考虑数据特性可能导致次优的预测结果。尽管与传统模型相比，基于GP的纯组合物性模型已被开发出更高的准确性，但可以使用一些技术进一步改进GP模型。功能基团输入空间是离散且高维的，许多成熟的建模方法已经存在，可以使用这些信息。通过考虑输入空间的特征和给定更好的先验信息，可以构建更高准确性和外推能力的机器学习模型。

本文提出了一种基于GP的功能基团表示下的性质预测新框架。其中，高维和离散的输入空间通过扭曲函数进行处理，并利用GC方法给定先验信息。本文的结构如下：第2节提供了一些与GC方法相关的基本概念以及最新的机器学习方法；第3节介绍了完整的模型结构，并应用于第4节中的纯组分回归；第4节依次分析了纯组分回归中单个技术(即扭曲函数、先验推断和先验预测检查)的贡献，并列出了比较了一些其他主流机器学习方法建立的模型；第5节强调了本文的主要贡献，并简要介绍了未来的研究方向。

2. 物性建模基础

2.1. GC建模方法

在基于GC的模型中，化合物的性质是通过代表分子结构的不同基团的贡献来预测的，具有无需大量计算资源即可快速预测的重要优势[10]。图1展示了化合物methoxylor的GC表示。methoxylor的分子结构用一个424维的向量表示，分为三个基团阶层。在这种情况下，值为“2”的第二维对应于methoxylor中甲基(-CH₃)出现两次。其他维度的值可以类似地获得，从而以仅包含整数的向量形式提供信息。需要注意的是，仅使用一级基团就可以准确预测简单和单功能分子的性质，因为一级基团捕捉到了近邻效应(不考虑分子中不同基团的效应)。然而，更高的阶层(第二和第三)提供了多功能和结构基团，这些基团提供了关于更复杂化合物分子结构的更多信息。然而，原子平衡是通过一级基团来确保的。

在传统的GC框架中，性质预测模型采用公式(1)的形式。

$$f(\mathbf{x}) = \sum_{i=1}^{NF} \mathbf{c}_i^F \mathbf{x}_i^F + \sum_{j=1}^{NS} \mathbf{c}_j^S \mathbf{x}_j^S + \sum_{k=1}^{NT} \mathbf{c}_k^T \mathbf{x}_k^T + b \quad (1)$$

式中, \mathbf{c}_i^F 、 \mathbf{c}_j^S 、 \mathbf{c}_k^T 分别代表第一、第二和第三阶层的贡献; $\mathbf{x}([\mathbf{x}_i^F, \mathbf{x}_j^S, \mathbf{x}_k^T])$ 代表功能基团数; b 是截距; $f(\mathbf{x})$ 是性质值的预测; i 、 j 、 k 是第一、第二和第三阶层的贡献度; NF、NS、NT 分别代表三个阶层的基团总数。Hukkerikar 等[37]提出了两种优化贡献系数的方法(同时和顺序), 它们在系数预测的顺序上有所不同。虽然同时方法在一个步骤中考虑了所有参数, 但顺序方法逐步使用第二和第三阶层的项来减少第一阶层的残差。

2.2. 支持向量机回归模型

支持向量机回归(SVR)模型[38]是由支持向量机(SVM)生成的监督学习模型。用线性核建立的SVR模型和线性回归之间的最明显区别在于, 前者在误差小于某个正数 ε 时会忽略误差。此功能使模型对聚合数据的敏感性降低, 从而更鲁棒。

本文中, 模型中的GC系数通过具有线性核的SVR进行回归(在本文中称为“SVR模型”), 使用第2.1节中介绍的同时和顺序方法进行。SVR模型的超参数包括作为违反条件的比例系数的惩罚 \mathcal{P} 和最大容差偏差的精度 ε , 如公式(2)所示。

$$\begin{aligned} \min \quad & \frac{1}{2} \|\mathbf{c}\|^2 + \mathcal{P} \sum_{i=1}^N (\delta_i + \delta_i^*) \\ \text{s.t.} \quad & y_i - (\mathbf{c}^T \mathbf{x}_i + b) \leq \varepsilon + \delta_i \quad i = 1, 2, 3 \dots N \\ & (\mathbf{c}^T \mathbf{x}_i + b) - y_i \leq \varepsilon + \delta_i^* \quad i = 1, 2, 3 \dots N \\ & \text{Nf}(\mathbf{x}_i) = \mathbf{c}^T \mathbf{x}_i + b \quad i = 1, 2, 3 \dots N \\ & \delta_i, \delta_i^* \geq 0 \quad i = 1, 2, 3 \dots N \end{aligned} \quad (2)$$

式中, δ_i 和 δ_i^* 是处理超出 ε 精度范围的点的两个松弛变量; \mathbf{x}_i 代表第 i 个分子, y_i 代表其测量结果; $\text{Nf}(\mathbf{x}_i)$ 代表同时方法的预测结果或顺序方法的第一、第二、第三阶层预测项; \mathbf{c} 是贡献向量, 等于公式(1)中的 $[\mathbf{c}_i^F, \mathbf{c}_j^S, \mathbf{c}_k^T]$; \mathbf{c}^T 是 \mathbf{c} 的转置; N 代表数据数量。本文中, 超参数调优通过栅格搜索方法实现。

2.3. GP模型

GP是一种随机过程, 其中每个有限集合的随机变量都具有多变量高斯分布。在最常见的情况下, GP先验满足公式(3)。通过贝叶斯推断, 获得的未观测点的后验也服从正态分布。均值和方差分别如公式(4)和公式(5)所示。

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N} \left(0, \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_\varepsilon^2 \mathbf{I} & \mathbf{K}(\mathbf{X}, \mathbf{X}_*) \\ \mathbf{K}(\mathbf{X}_*, \mathbf{X}) & \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix} \right) \quad (3)$$

$$\boldsymbol{\mu}(\mathbf{f}_*) = \mathbf{K}(\mathbf{X}_*, \mathbf{X}) [\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_\varepsilon^2 \mathbf{I}]^{-1} \mathbf{y} \quad (4)$$

$$\text{var}(\mathbf{f}_*) = \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) - \mathbf{K}(\mathbf{X}_*, \mathbf{X}) [\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_\varepsilon^2 \mathbf{I}]^{-1} \mathbf{K}(\mathbf{X}, \mathbf{X}_*) \quad (5)$$

式中, \mathbf{X} 和 \mathbf{X}_* 分别对应训练集和测试集中的输入变量; \mathbf{y} 是输入 \mathbf{X} 的真实输出值; \mathbf{f}_* 是 \mathbf{X}_* 的预测输出, 其均值为 $\boldsymbol{\mu}(\mathbf{f}_*)$, 方差矩阵为 $\text{var}(\mathbf{f}_*)$; σ_ε 表示测量噪声; \mathbf{I} 表示单位矩阵; \mathcal{N} 代表联合正态分布。 $\mathbf{K}(\cdot, \cdot)$ 是由核函数计算值组成的矩阵[39], 其第 i 行($i = 1, 2, 3, \dots, N$)和第 j 列($j = 1, 2, 3, \dots, N$)的元素等于由第 i 和第 j 个输入获得的核函数的值, 如公式(6)所示。

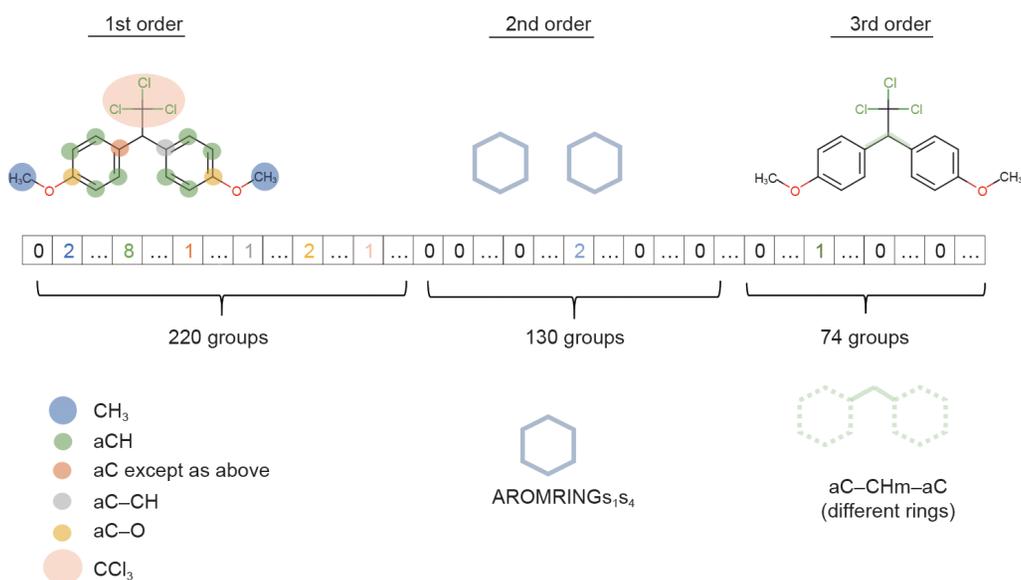


图1. methoxylor化合物的GC表示。图中展示了methoxylor化合物的一级基团(CH_3 , aCH, aC, aC-CH, aC-O, CCl_3)、二级基团($\text{AROMRINGS}_1\text{S}_4$)和三级基团(aC-CHm-aC)。

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \dots & k(\mathbf{x}_1, \mathbf{x}_N) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \dots & k(\mathbf{x}_2, \mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & k(\mathbf{x}_N, \mathbf{x}_2) & \dots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix} \quad (6)$$

核函数 $k(\cdot, \cdot)$ 表示两个变量之间的相关程度，应谨慎选择。Alshehri 等[36]总结了用于 GP 下性质预测的四个指数核（在本文中称为“GP 模型”），其中使用的核函数形式为公式（7）。

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^4 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_i^2}{2l_i^2}\right) \quad (7)$$

式中， \mathbf{x} 和 \mathbf{x}' 是代表 GC 框架下两个特定化合物的 424 维向量， $l_i (i=1, 2, 3, 4)$ 是 GP 模型调优的超参，这些参数对于模型性能也至关重要。常用的超参数优化方法包括最大似然估计和交叉验证。

3. 基于高斯过程的改进机器学习模型

由官能团向量表示的分子是高维的，且向量元素均为整数。尽管 GP 模型相比线性的官能团贡献度模型在整个数据集上具有提升的预测性能[36]，但其预测能力还可以进一步提高。本节提出了一种改进的 GP 模型，旨在处理高维离散输入空间。

3.1. 离散输入空间的扭曲函数

具有有序级别的分类变量称为序数，可以被视为连续空间的离散化[40]。GP \mathbf{Y} 在 $\{1, 2, 3, \dots, L\}$ 上定义为以整数向量为输入变量的序数 GP，GP \mathbf{Z} 是一个普通的连续 GP。GP \mathbf{Y} 可以通过非递减函数[41]（也称为扭曲函数）在连续域中转换为 GP \mathbf{Z} ，如方程（8）所示。因此，为序数处理的核函数可以如方程（9）所示进行变换。

$$\mathbf{Y}(u) = \mathbf{Z}(F(u)) \quad (8)$$

$$t(u, u') = k(F(u), F(u')), u, u' = 1, 2, 3, \dots, L \quad (9)$$

式中， u 和 u' 均是具有 L 个不同级别的序数； $t(\cdot, \cdot)$ 是 GP \mathbf{Y} 中的核函数。核函数是 GP 中两个变量之间的相关程度。许多常用的核函数只取决于两个变量间的距离，而不是变量本身。因此，具有某个官能团的 0 和 1 的分子之间的相关性与具有 19 和 20 的分子之间（因为这两个距离都是“1”）的相关性相同。然而，直观来看，表示数量的两个序数之间的相关性应该与它们自己的数值有关。比如在 GC 模型中，具有特定官能团的 0 和 1 的分子之间的相关性应该小于具有 19 和 20 的分子之间（第一对的距离应该更大）。因此，扭曲函数的形式定义为方程式（10）。

$$F(u) = \log_\alpha(u+1), \alpha > 1 \quad (10)$$

图 2 展示了扭曲函数如何处理一维序数，以 GC 表示和常用的指数核为例。组合 1 包括具有 0 和 1 的某个官能团的分子，而组合 2 包括具有 19 和 20 的某个功能基团的分子。在离散变量直接输入核函数的情况下，两对组合的相关性是相同的。然而，在添加扭曲函数后，组合 1 的相关性较小，满足了模型的需要。当化合物 \mathbf{x} 用 424 维向量表示时，方程（10）可以这样展开，即扭曲向量 $\mathbf{F}(\mathbf{x})$ 的第 i 个元素是用 \mathbf{x} 的第 i 维分量计算的。

为了更好地处理离散变量的相关变化程度，扭曲函数中的参数 α 和核函数引起的参数被视为超参数，并通过 GP \mathbf{Y} 中的交叉验证或最大对数似然进行调整。

3.2. 物性预测模型的先验推断

给定先验的策略包括向专家小组寻求建议，从样本数据中获得信息[42]。许多学者研究了适用于不同场景的 GP 的先验推断[43–45]。大多数情况下，如果 GP 缺乏某些信息或经验，则 GP 的先验设置为零，而良好的先验推断过程会使模型具有更好的性能。

维度信息应该通过核函数或先验添加到 GP 模型中，因为它在 GC 表示下起着重要作用。平方指数核可以根据超参数进行参数化，如公式（11）所示：

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^\top \mathbf{M}(\mathbf{x} - \mathbf{x}')\right) \quad (11)$$

式中， \mathbf{M} 表示对称矩阵。矩阵 \mathbf{M} 的可能选择包括[28]：

$$\mathbf{M}_1 = l^{-2} \mathbf{I}, \mathbf{M}_2 = \text{diag}(\mathbf{l})^{-2} \quad (12)$$

式中， l 、 \mathbf{l} 和 σ 都是核函数的参数。需要注意的是，虽然 $k(\mathbf{x}, \mathbf{x}')$ 的形式似乎与输出空间无关，但在模型训练过程中，超参数将在很大程度上取决于输出值，从而使 GP 模型在预测多个属性时更加灵活。一方面， \mathbf{M}_2 显著增加了超参数的数量，从而使超参数调整变得困难；另一方面， \mathbf{M}_1 无法提供维度信息。对于具有高维和线性特征或意义的模型，如 GC 模型，先验可以设置为提供维度信息的输入空间的线性组合。在此框架下，GP 模型转化为方程式（13）。后验的均值和方差分别对应于方程（14）和方程（15）[46]。

$$\begin{pmatrix} \mathbf{y}(\mathbf{X}) \\ \mathbf{f}(\mathbf{X}_*) \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \text{SVR}(\mathbf{X}) \\ \text{SVR}(\mathbf{X}_*) \end{pmatrix}, \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_c^2 \mathbf{I} & \mathbf{K}(\mathbf{X}, \mathbf{X}_*) \\ \mathbf{K}(\mathbf{X}_*, \mathbf{X}) & \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix} \right) \quad (13)$$

$$\boldsymbol{\mu}(\mathbf{f}) = \text{SVR}(\mathbf{X}_*) + \mathbf{K}(\mathbf{X}_*, \mathbf{X}) [\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_c^2 \mathbf{I}]^{-1} \times (\mathbf{y} - \text{SVR}(\mathbf{X})) \quad (14)$$

$$\text{var}(\mathbf{f}) = \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) - \mathbf{K}(\mathbf{X}_*, \mathbf{X}) [\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_c^2 \mathbf{I}]^{-1} \mathbf{K}(\mathbf{X}, \mathbf{X}_*) \quad (15)$$

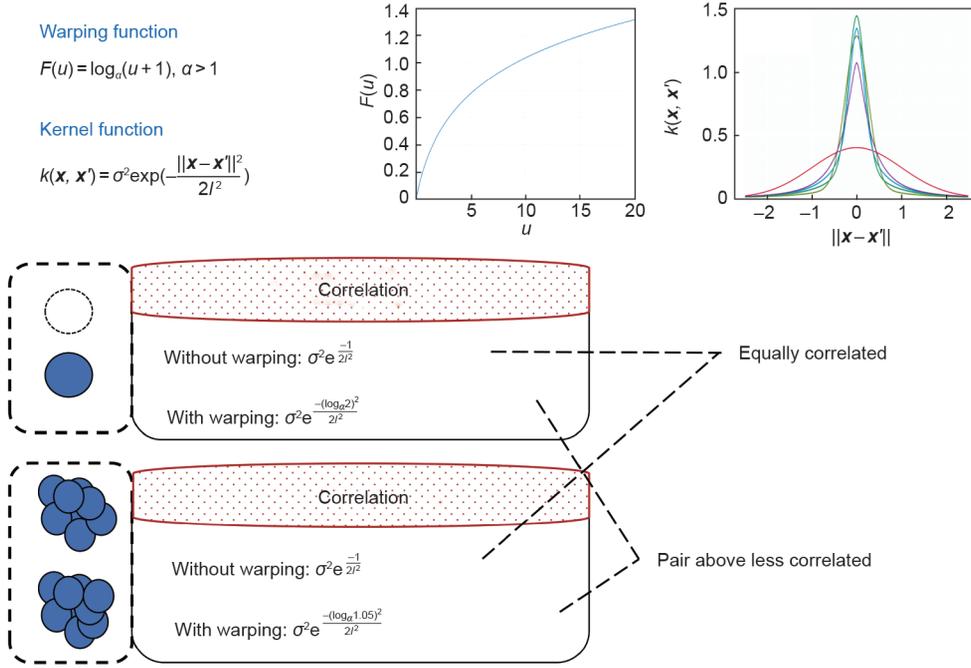


图2. 使用扭曲函数后相关性的变化。 σ 、 l : 核函数的参数。

式中， \mathbf{y} 是匹配 \mathbf{X} 的真实输出值； \mathbf{f} 是匹配 \mathbf{X}_* 的预测输出。**SVR**是通过SVR技术获得的线性模型输出值。

3.3. GP模型的先验预测检查

如果先验推断程序有效，那么基于数据分析或专家实验的先验分布在本质上是利好的。然而，即使在有效的先验推断的情况下，也必须检查先验是否产生了不正确的数据[42]。研究显示，有许多用于先验预测检查的方法，如先验预测 p 值[47]和贝叶斯因子[48]。虽然**SVR**模型被添加为GP的先验，以提供维度信息，但事实上，它不一定优于零先验的GP。因此，在GP建模之前必须进行预先预测检查。由于只能使用训练集并对超参数调优进行交叉验证，因此使用不同折叠上的平均交叉验证损失来比较零和非零先验。考虑到**SVR**模型是在训练集上训练的，这使得具有**SVR**先验的模型的先验预测检查过程具有固有优势，因此具有非零先验的平均交叉验证损失会乘以一个惩罚因子。优先选择标准如方程式(16)所示。

$$\begin{cases} \mathbf{GP}_{\text{prior}} = \text{SVR} & \text{if } CV_{\text{non-zero}} \times (1+p) < CV_{\text{zero}} \\ \mathbf{GP}_{\text{prior}} = \mathbf{0} & \text{else} \end{cases} \quad (16)$$

式中， $\mathbf{GP}_{\text{prior}}$ 为最终GP模型的先验； $CV_{\text{non-zero}}$ 和 CV_{zero} 分别为具有先验**SVR**和零训练集上的平均交叉验证损失； p 为惩罚项。

3.4. 集成的模型结构

基于扭曲函数、先验推断和先验预测检查，为高维离

散输入构建最终的GP模型（本文称为“GP-WP”）。方程(17)、方程(18)给出了模型的数学表示。

$$\begin{pmatrix} \mathbf{y}(\mathbf{X}) \\ \mathbf{GP-WP}(\mathbf{X}_*) \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \mathbf{GP}_{\text{prior}}(\mathbf{X}) \\ \mathbf{GP}_{\text{prior}}(\mathbf{X}_*) \end{pmatrix}, \begin{bmatrix} \mathbf{T}(\mathbf{X}, \mathbf{X}) + \sigma_e^2 \mathbf{I} & \mathbf{T}(\mathbf{X}, \mathbf{X}_*) \\ \mathbf{T}(\mathbf{X}_*, \mathbf{X}) & \mathbf{T}(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix} \right) \quad (17)$$

$$\mathbf{T}(\mathbf{X}, \mathbf{X}) = \begin{bmatrix} t(\mathbf{x}_1, \mathbf{x}_1) & t(\mathbf{x}_1, \mathbf{x}_2) & \cdots & t(\mathbf{x}_1, \mathbf{x}_N) \\ t(\mathbf{x}_2, \mathbf{x}_1) & t(\mathbf{x}_2, \mathbf{x}_2) & \cdots & t(\mathbf{x}_2, \mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ t(\mathbf{x}_N, \mathbf{x}_1) & t(\mathbf{x}_N, \mathbf{x}_2) & \cdots & t(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix} \quad (18)$$

式中， $\mathbf{GP-WP}$ 是预测结果； $\mathbf{T}(\cdot, \cdot)$ 是由扭曲函数和核函数计算值组成的矩阵，其第 i 行和第 j 列中的元素等于方程(9)为第 i 和第 j 个输入获得的核函数的值。**SVR**的参数由方程式(2)获得。最后，方程(19)、方程(20)给出了预测值和不确定性。

$$\mu(\mathbf{GP-WP}) = \mathbf{GP}_{\text{prior}}(\mathbf{X}_*) + \mathbf{T}(\mathbf{X}_*, \mathbf{X}) [\mathbf{T}(\mathbf{X}, \mathbf{X}) + \sigma_e^2 \mathbf{I}]^{-1} (\mathbf{y} - \mathbf{GP}_{\text{prior}}(\mathbf{X})) \quad (19)$$

$$\text{var}(\mathbf{GP-WP}) = \mathbf{T}(\mathbf{X}_*, \mathbf{X}_*) - \mathbf{T}(\mathbf{X}_*, \mathbf{X}) [\mathbf{T}(\mathbf{X}, \mathbf{X}) + \sigma_e^2 \mathbf{I}]^{-1} \mathbf{T}(\mathbf{X}, \mathbf{X}_*) \quad (20)$$

完整的建模过程如图3所示。在训练之前，数据集被分为训练集和测试集。在下一阶段，扭曲函数将离散GP \mathbf{Y} 转换为连续GP \mathbf{Z} ，其形式已经给出。接着，在训练集上进行SVR模型。基于SVR模型，进行交叉验证以调整超

参数，并使用方程（16）对具有零和非零先验的模型进行先验预测检查。最后，利用先验信息建立 GP 模型，并用扭曲函数对核进行处理。超参数包括 α 、 l 和 σ_e 。

图4描述了用于预测新分子性质的机器学习模型的结构。在步骤1中，需要新化合物的分子式来获得其基团贡献表示。接下来，在步骤2中，使用扭曲函数将积分向量转换为连续域，并在 GP 的训练过程中确定该函数的参数。在步骤3中，生成两个协方差矩阵：一个捕获训练数

据集中不同分子之间的相关性，另一个表示新分子与其他分子之间的关联。此外，使用通过贝叶斯统计推断的公式计算预测值。值得注意的是，在训练阶段之后，核函数公式中的超参数是固定的。

4. 结果与讨论

本节将利用前几节中提到的模型（即第2.2节中的

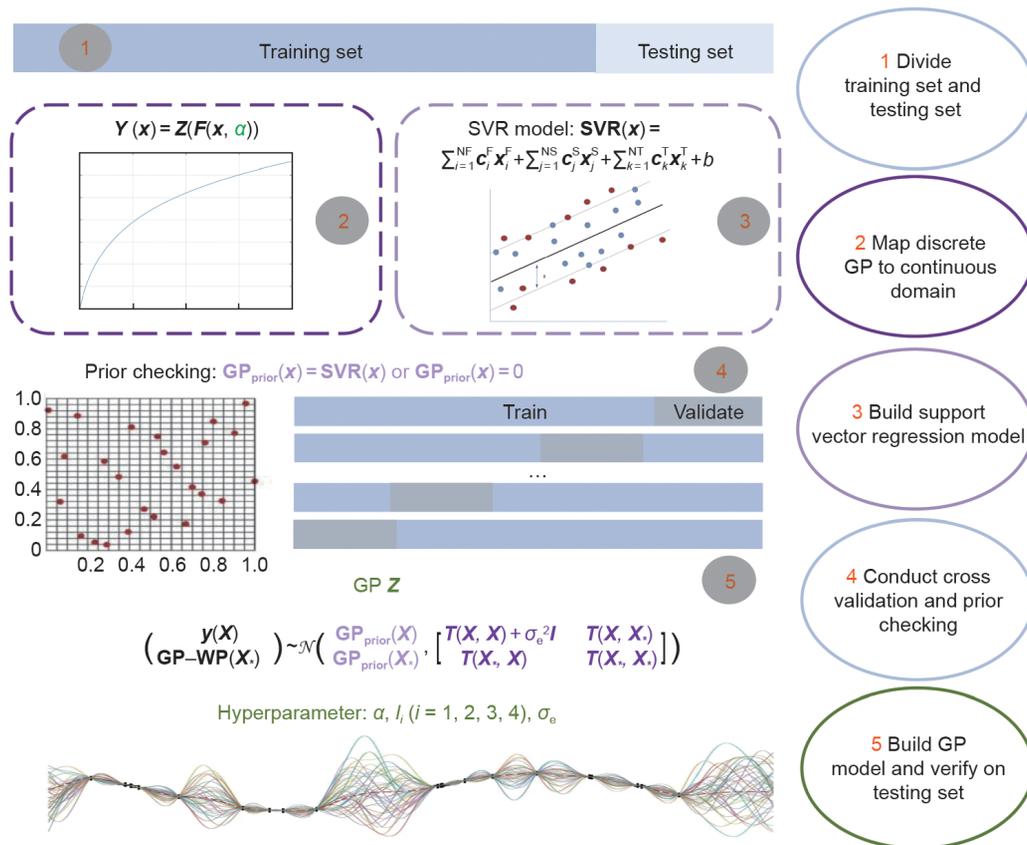


图3. 建立具有高维离散输入的GP-WP模型的过程。

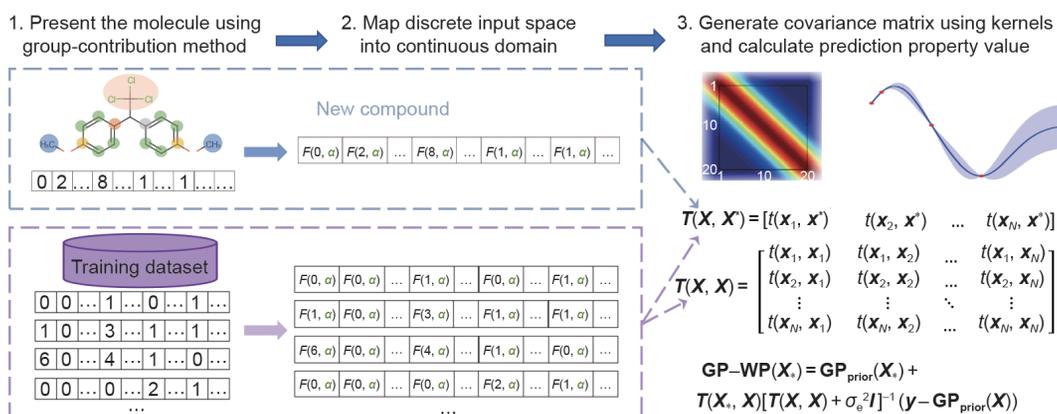


图4. 机器学习模型结构。

SVR、第2.3节中的GP和第3节中的GP-WP)预测20种纯组分的主要物性。表1列出了20种物质的数据库信息,表2则给出了有关三个模型的详细信息。

第4节结构如下:首先,使用误差阈值图和定量误差指数(RMSE和 R^2)显示三个模型的结果;其次,分析了三种技术(扭曲函数、先验推断和先验预测检查)对整个模型框架的贡献;最后,将GP-WP的性能与其他主流机器学习模型(包括神经网络和决策树)的性能进行比较。

4.1. 20种物性的仿真结果

本节比较了SVR(即使用SVR开发的模型)的预测精度、GP(即使用常规GP框架开发的模型)以及GP-WP(即使用GP开发的模型,具有扭曲函数、先验推断和

先验预测检查)模型的预测效果。共测试了20种性质,包括正常沸点(K)、临界体积($\text{mL}\cdot\text{mol}^{-1}$)、临界温度(K)、临界压力(bar, $1\text{ bar}=10^5\text{ Pa}$)、自燃温度(K)、生物富集因子、298 K时的吉布斯自由能($\text{kJ}\cdot\text{mol}^{-1}$)、标准生成焓($\text{kJ}\cdot\text{mol}^{-1}$)、298 K时的熔化焓($\text{kJ}\cdot\text{mol}^{-1}$)、298 K时的Hildebrandt溶解度参数($\text{MPa}^{0.5}$)、298 K时的生成焓($\text{kJ}\cdot\text{mol}^{-1}$)、LC50(胖头鲮)($\text{mol}\cdot\text{L}^{-1}$)、大鼠口服毒性($\text{mol}\cdot\text{kg}^{-1}$)、298 K时的液态摩尔体积($\text{mL}\cdot\text{mol}^{-1}$)、正辛醇-水分配系数、水溶性($\text{mol}\cdot\text{L}^{-1}$)、允许暴露限值($\text{mol}\cdot\text{m}^{-3}$)、光化学氧化电位、酸解离常数和常压熔点(K)。训练集和测试集的划分与原始数据集保持一致;然而,对于输入相同、输出接近的异构体,只保留最接近平均输出值的异构体。

表1 20种物性的数据库信息

Number	Property	Notation	Number of active groups	Number of data points
1	Normal boiling point (K)	Tb	369	4658
2	Critical volume ($\text{mL}\cdot\text{mol}^{-1}$)	Vc	235	723
3	Critical temperature (K)	Tc	231	717
4	Critical pressure (bar)	Pc	235	724
5	Auto ignition temperature (K)	Ait	183	542
6	Bioconcentration factor	bcf	222	554
7	Gibbs energy of formation at 298 K ($\text{kJ}\cdot\text{mol}^{-1}$)	Gf	230	706
8	Standard enthalpy of formulation ($\text{kJ}\cdot\text{mol}^{-1}$)	Hf	245	877
9	Enthalpy of fusion at 298 K ($\text{kJ}\cdot\text{mol}^{-1}$)	Hfus	246	706
10	Hildebrandt solubility parameter at 298 K ($\text{MPa}^{1/2}$)	Hsolp	291	1278
11	Enthalpy of formation at 298 K ($\text{kJ}\cdot\text{mol}^{-1}$)	Hv	154	395
12	LC50 (Fathead Minnow) ($\text{mol}\cdot\text{L}^{-1}$)	Lc50_fm	250	680
13	Toxicity (oral rat) ($\text{mol}\cdot\text{kg}^{-1}$)	Ld50	373	4728
14	Liquid molar volume at 298 K ($\text{mL}\cdot\text{mol}^{-1}$)	Lmv	255	987
15	Octanol-water partition coefficient	logP	375	11 236
16	Aqueous solubility ($\text{mol}\cdot\text{L}^{-1}$)	logWs	340	2306
17	Permissible exposure limit ($\text{mol}\cdot\text{m}^{-3}$)	osha_twa	180	399
18	Photochemical oxidation potential	pco	149	546
19	Acid dissociation constant	pKa	282	1502
20	Normal melting point (K)	Tm	397	8408

表2 模型信息

Model	SVR	GP	GP-WP
Parameter	$c(\text{vector}), b$	—	—
Hyperparameter	\mathcal{R}, ε	$\sigma_c, l_1, l_2, l_3, l_4$	$\sigma_c, l_1, l_2, l_3, l_4, \alpha$
Training step	Tune hyperparameter via grid search and cross validation; optimize parameter with Eq. (2)	Tune hyperparameter via grid search and cross validation	Tune hyperparameter (α is used for mapping) via grid search and cross validation for both zero-prior or SVR prior; use the cross-validation result to do prior predictive checking, as shown in Eq. (16)
Prediction	Calculate property value using Eq. (1)	Calculate property value using Eq. (4); estimate model uncertainty using Eq. (5)	Calculate property value using Eq. (19); estimate model uncertainty using Eq. (20)

SVR 模型用方程式 (2) 训练, 其超参数 \mathcal{S} 和 ε 通过 python 包 `skopt.sampler` (scikit 优化库) 的网格搜索进行调整。同时使用顺序和同时方法, 选择 RMSE 较低的方法作为 SVR 模型。SVR 系数优化是通过 python 包 `sklearn.svm` (scikit-learn 库) 完成的, 之后获得每个功能组的贡献。GP 的五个超参数使用方程 (7) 中所示的核函数进行调整, 通过贝叶斯推理获得每个属性的预测。然后, 按照图 3 中的结构实现 GP-WP 模型。对于有或没有先验 (先验为零) 的模型, 分别设置六个超参数 (其中额外的一个来自扭曲函数)。在训练过程中, 还可以计算平均验证损失。使用方程式 (16) 确定最终的 GP-WP 模型。在这里, 它的惩罚因子被设置为 0.05。

首先, 通过不同的误差阈值率 (1%、5% 和 10%) 测量不同模型在整个数据集上的性能, 如图 5 所示。前三列对应 SVR, 中间三列对应常规 GP, 最后三列对应此 GP-WP 模型。不同的行表示不同属性的百分比, 而最后一行为计算所得上述所有 20 个属性的平均值。在图 5 中, 红色表示高百分比, 蓝色表示低百分比。

图 5 中 SVR 模型和其他两个模型之间的边界非常清晰, 因为前三列中大面积被蓝色覆盖。当观察三个 1% 的柱时, 这种现象变得尤为突出。尽管不同属性的预测精度不同, 但 SVR 预测的百分比均不高于 90%。然而, GP 框架下的对应模型都高于 90%。因此, 当需要替代原始数据

集以实现高精度预测时, GP 将是一个不错的选择。同时, 从热图中可以看出, GP-WP 在 1%、5% 和 10% 阈值下的平均分数 (分别是 94.57%、96.52% 和 97.64%) 均高于常规 GP 的相应分数 (分别是 94.51%、95.78% 和 96.94%)。

图 6 仅显示了测试集的结果。尽管 GP 模型与 SVR 相比失去了突出的优势, 但根据热图上的分数, GP-WP 模型在大多数属性上仍优于 SVR 模型, 三个阈值的平均百分比在三个模型中最高。必须承认, 对于某些属性 (即 Hfus 和 bcf), 没有一个模型具有非常准确的预测能力。虽然这两个 GP 模型可以提供不确定性范围, 以告知建模者模型的性能不佳, 但 SVR 模型只给出了不准确的预测结果。

为了更好地量化 SVR、正则 GP 和 GP-WP 的误差, 使用方程式 (21) 和方程式 (22) 计算每个属性的 RMSE 和 R^2 。

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (21)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (22)$$

式中, N 表示数据编号; y_i 是真实属性值; \hat{y}_i 对应于预测值; \bar{y} 是所有样本真实值的平均值。20 个属性预测的结果如表 3 (整个数据集) 和表 4 (测试集) 所示。

表 3 和表 4 中的结果与误差阈值图中的结果一致。以 2 号属性 (Vc) 为例, GP-WP 模型将 RMSE 从 28.280 降低

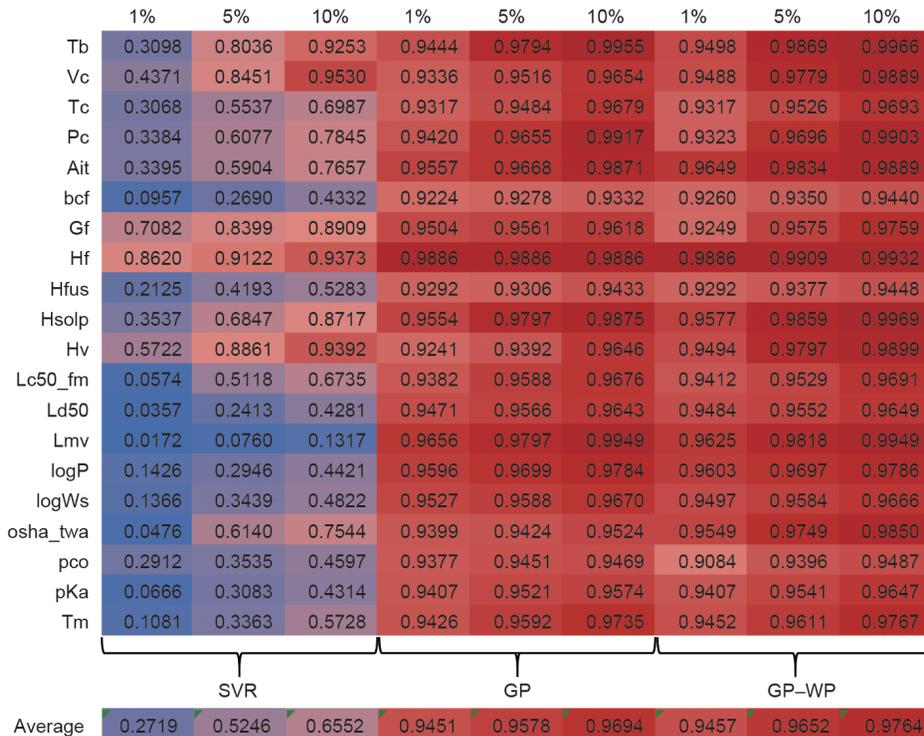


图 5. 热图显示整个数据集在不同误差阈值率下的分数。红色表示高百分比, 蓝色表示低百分比。

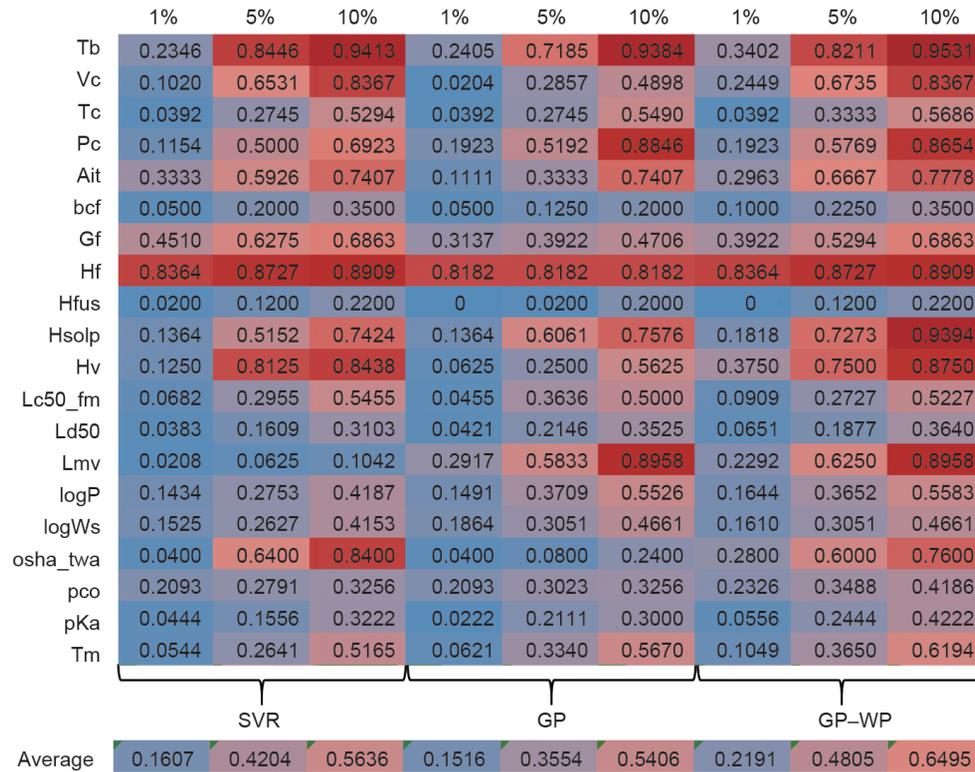


图6. 热图显示测试集在不同误差阈值率下的分数。红色表示高百分比，蓝色表示低百分比。

表3 整个数据集的误差分析

Number	Property	Sample number	RMSE			R^2		
			SVR	GP	GP-WP	SVR	GP	GP-WP
1	Tb	4658	43.956	7.654	6.888	0.775	0.993	0.994
2	Vc	723	22.434	28.280	8.704	0.989	0.983	0.998
3	Tc	717	78.716	24.307	23.152	0.457	0.948	0.953
4	Pc	724	4.782	0.770	0.743	0.833	0.996	0.996
5	Ait	542	73.612	25.093	21.124	0.679	0.963	0.974
6	bcf	554	0.697	0.233	0.262	0.728	0.970	0.962
7	Gf	706	36.542	24.034	11.942	0.972	0.988	0.997
8	Hf	877	66.485	22.995	9.394	0.930	0.992	0.999
9	Hfus	706	5.802	1.380	1.091	0.801	0.989	0.993
10	Hsolp	1278	1.806	0.616	0.526	0.793	0.976	0.982
11	Hv	395	3.310	2.698	1.630	0.942	0.961	0.986
12	Lc50_fm	680	0.685	0.239	0.196	0.788	0.974	0.983
13	Ld50	4728	0.437	0.104	0.103	0.660	0.981	0.981
14	Lmv	987	0.054	0.002	0.002	0.545	0.999	1.000
15	logP	11 236	0.633	0.101	0.087	0.869	0.997	0.998
16	logWs	2306	0.852	0.186	0.179	0.830	0.992	0.993
17	osha_twa	399	0.820	0.373	0.325	0.716	0.941	0.955
18	pco	546	0.245	0.049	0.045	0.813	0.992	0.994
19	pKa	1502	2.529	0.481	0.437	0.491	0.982	0.976
20	Tm	8408	58.184	12.832	12.008	0.681	0.984	0.986

Bold numbers indicate the best results.

表4 测试集的误差分析

Number	Property	Sample number	RMSE			R^2		
			SVR	GP	GP-WP	SVR	GP	GP-WP
1	Tb	341	28.730	28.290	25.397	0.860	0.865	0.891
2	Vc	49	34.957	108.628	33.432	0.984	0.848	0.986
3	Tc	51	93.535	91.140	86.807	0.345	0.378	0.435
4	Pc	52	3.773	2.873	2.770	0.790	0.878	0.887
5	Ait	27	95.601	112.429	94.642	0.441	0.226	0.452
6	bef	40	0.993	0.867	0.973	0.437	0.570	0.458
7	Gf	51	72.592	89.423	44.373	0.816	0.720	0.931
8	Hf	55	48.184	91.825	37.426	0.958	0.849	0.975
9	Hfus	50	4.320	5.187	4.099	0.741	0.626	0.766
10	Hsolp	66	3.715	2.710	2.314	0.424	0.694	0.777
11	Hv	32	7.021	9.480	5.726	0.644	0.351	0.763
12	Lc50_fm	44	0.827	0.939	0.769	0.779	0.715	0.808
13	Ld50	261	0.479	0.442	0.438	0.626	0.681	0.687
14	Lmv	48	0.051	0.009	0.007	0.598	0.988	0.992
15	logP	523	0.602	0.468	0.403	0.878	0.926	0.945
16	logWs	118	0.985	0.824	0.790	0.779	0.845	0.858
17	osha_twa	25	1.305	1.489	1.300	0.426	0.252	0.430
18	pco	43	0.193	0.176	0.160	0.727	0.774	0.814
19	pKa	90	2.586	1.963	1.784	0.390	0.648	0.710
20	Tm	515	59.801	51.847	48.520	0.664	0.748	0.779

Bold numbers indicate the best results.

到 8.704，降低了 71.45%。此外，它将“1% 误差阈值以下的分数”从 93.36% 提高到 94.88%，预测性能提高了 1.63%。由于 RMSE 和“1% 误差阈值以下的分数”都代表了预测的准确性，因此 GP-WP 模型在误差阈值比和定量相对误差方面都提高了大多数属性的准确性。在整个数据集上，两种基于 GP 的模型与 SVR 模型相比具有显著优势。此外，对于大多数物性，GP 模型的表现不如 GP-WP 模型。这在 Vc、Gf 和 Hf 的情况下很明显，在采用扭曲函数和先验推断技术后，RMSE 降低了 50% 以上（分别从 28.280 降低到 8.704、24.034 降低到 11.942、22.995 降低到 9.394）。在测试集上，GP-WP 模型始终优于 SVR 模型，显示出较小的误差和较高的 R^2 值。可以看出，将 SVR 模型用作 GP-WP 模型的先验是非常实用的。

4.2. GP-WP 模型中各项技术分析

前一小节展示了 GP-WP 相对于传统 SVR 和 GP 方法的优势。本节进一步说明了扭曲函数、先验推断和先验预测检查技术如何分别对整个模型框架做出贡献。通过引入每种技术可以提高大部分物性估计的性能。

首先，使用扭曲函数来更好地处理两个离散向量之间的相关性。由于扭曲函数的参数是 GP 的超参数，在超参数

调整过程中会自动确定相关变化的程度。与常规 GP 相比，合并扭曲函数后的物性预测模型的性能略有但稳定地提高。换句话说，对于 20 个物性中的大多数物性，仅添加扭曲函数可以略微提高预测精度。补充信息中给出了添加所有物性的扭曲函数后的 RMSE 结果，表 5 中列出了一些代表性示例。很明显，Tb、Gf 和 Tc 等性质的 RMSE 变小（分别从 7.65、24.034 和 24.31 变为 6.89、22.908 和 23.68）。

表5 加入扭曲功能后的均方根误差比较

Property	RMSE		
	SVR	GP	GP _{warping}
Tb	43.960	7.650	6.890
Gf	36.542	24.034	22.908
Tc	78.720	24.310	23.680

GP_{warping}: GP with warping function only. Complete results for all properties are provided in the Appendix A.

其次，当 SVR 模型表现良好时，传统方法的先验推断尤其有效。从方程 (14) 可以确认，SVR 先验使 GP 模型拟合 SVR 模型的残差。补充信息中给出了添加所有物性的事先启动程序后的 RMSE 结果，表 6 中列出了一些代表性示例。对于 Vc 等物性，其中 SVR 模型的预测精度已经很高，先验推断过程大大提高了 GP 的模型性能。

表6 加入先验推断程序后的均方根误差比较

Property	RMSE		
	SVR	GP	GP _{prior}
Vc	22.430	28.280	8.720
Hf	66.485	22.995	10.516
Hfus	5.802	1.380	1.108

GP_{prior}: GP with an SVR prior only. Complete results are provided in the Appendix A.

最后, 先验预测检验的功能在于确定最终模型。结果如表7所示, 其中 CV_{zero} 和 $CV_{non-zero}$ 分别对应于具有零和非零先验的GP模型的平均交叉验证损失。根据 CV_{zero} 和 $CV_{non-zero} \times (1+p)$, 确定最终GP-WP模型的先验形式, 如“GP-WP”列所示。测试集上的RMSE也列在表中以供参考; 在实际过程中是未知的。需要明确的是, 在事先检查过程中, 有时会出现次优的事先选择。例如, bcf和ld50的GP_{zero}模型在测试集上的表现优于GP_{non-zero}模型, 但GP_{non-zero}被先验预测检查技术选为最终模型。然而, 在大多数情况下, 判断是恒定的, 测试集上的误差也是恒定的。

4.3. 与其他主流机器学习模型的对比

为了充分证明GP-WP模型的效果, 我们进一步将其与基于GP的模型以外的其他主流机器学习技术进行了比

较, 包括神经网络和决策树。神经网络是标准的主流建模方法之一, 已被广泛应用于物性预测[49]。由于能够学习数据中的复杂模式和关系, 神经网络经常被用作与其他模型进行比较的基准。相比之下, 决策树擅长处理离散值输入, 使其特别适合输入为离散值GC的问题。在不失一般性的情况下, 这里选取了三个具有代表性的物性, 样本数量从11 236 (logP) 到4658 (Tb), 最终到717 (Tc)。

神经网络是一组被称为神经元的连接节点, 它们形成不同的层。数据通过输入层、隐藏层, 最后通过输出层。在这项工作中, 使用了两种类型的神经网络: 一种具有完全连接的密集隐藏层, 其宽度和深度可以作为超参数进行调整, 以优化拟合[本文称为“BP-*i*layer”, 具有*i*个隐藏层 ($i = 1, 2$)]; 另一种具有在密集隐藏层之前添加的卷积隐藏层 (此处称为“CNN”模型)。与神经网络的神经元和层结构不同, 决策树使用流程图对物性进行分类并分割不同的情况[50], 节点代表一个物性, 分支对应于一个单独的类别, 最后留下叶子来指示结果。同样, 可以调整树的深度、叶子的数量和其他参数来优化拟合。这里, 使用光梯度增强机 (LightGBM) [51]回归器来优化和实现梯度增强决策树 (GBDT)。logP、Tb和Tc的性质预测结果分别如表8至表10所示。

表7 20个物性的先验预测检查结果和验证

Number	Property	Sample number	RMSE		CV_{zero}	$CV_{non-zero}$	$CV_{non-zero} \times (1+p)$	GP-WP
			GP _{zero}	GP _{non-zero}				
1	Tb	341	25.397	26.620	40.590	39.300	41.265	GP _{zero}
2	Vc	49	36.779	33.432	92.636	19.539	20.516	GP _{non-zero}
3	Tc	51	88.795	86.807	92.885	81.427	85.498	GP _{non-zero}
4	Pc	52	2.930	2.770	4.483	2.746	2.883	GP _{non-zero}
5	Ait	27	112.237	94.642	96.667	71.463	75.037	GP _{non-zero}
6	bcf	40	0.809	0.973	0.873	0.646	0.678	GP _{non-zero}
7	Gf	51	85.228	44.373	97.640	41.666	43.750	GP _{non-zero}
8	Hf	55	85.378	37.426	116.904	57.303	60.168	GP _{non-zero}
9	Hfus	50	5.167	4.099	6.740	6.378	6.697	GP _{non-zero}
10	Hsolp	66	2.732	2.314	2.355	1.738	1.825	GP _{non-zero}
11	Hv	32	6.898	5.726	6.530	3.105	3.260	GP _{non-zero}
12	Lc50_fm	44	0.919	0.769	0.916	0.705	0.740	GP _{non-zero}
13	Ld50	261	0.426	0.438	0.461	0.425	0.447	GP _{non-zero}
14	Lmv	48	0.008	0.007	0.023	0.014	0.015	GP _{non-zero}
15	logP	523	0.455	0.403	0.518	0.442	0.464	GP _{non-zero}
16	logWs	118	0.817	0.790	0.892	0.818	0.859	GP _{non-zero}
17	osha_twa	25	1.461	1.300	1.093	0.756	0.793	GP _{non-zero}
18	pco	43	0.172	0.160	0.385	0.286	0.300	GP _{non-zero}
19	pKa	90	1.784	2.072	2.450	2.407	2.527	GP _{zero}
20	Tm	515	48.520	49.871	49.843	49.456	51.929	GP _{zero}

RMSE: not known during training, for verification only. Bold numbers indicate better results.

表8 物性logP的主流机器学习模型(11 236个样本)

Model	Testing set		Whole dataset				
	RMSE	R ²	RMSE	R ²	1% error	5% error	10% error
SVR	0.602	0.878	0.633	0.869	14.26%	29.46%	44.21%
GP	0.468	0.926	0.101	0.997	95.96%	96.99%	97.84%
LightGBM	0.425	0.939	0.311	0.968	17.99%	43.89%	63.71%
BP-1layer	0.424	0.939	0.280	0.974	26.28%	63.54%	77.21%
BP-2layer	0.449	0.932	0.277	0.975	31.05%	66.18%	79.02%
CNN	0.400	0.946	0.107	0.996	29.83%	78.38%	88.85%
GP-WP	0.403	0.945	0.087	0.998	96.03%	96.97%	97.86%

Bold numbers indicate the best results.

表9 物性Tb的主流机器学习模型(4658个样本)

Model	Testing set		Whole dataset				
	RMSE	R ²	RMSE	R ²	1% error	5% error	10% error
SVR	28.730	0.860	43.956	0.775	30.98%	80.36%	92.53%
GP	28.290	0.865	7.654	0.993	94.44%	97.94%	99.55%
LightGBM	39.826	0.732	38.323	0.829	20.55%	68.61%	86.41%
BP-1layer	35.571	0.786	32.353	0.878	34.16%	86.88%	96.31%
BP-2layer	38.017	0.755	19.024	0.958	52.86%	94.27%	97.72%
CNN	38.471	0.750	38.624	0.826	18.38%	71.75%	91.41%
GP-WP	25.397	0.891	6.888	0.994	94.98%	98.69%	99.66%

Bold numbers indicate the best results.

表10 物性Tc的主流机器学习模型(717个样本)

Model	Testing set		Whole dataset				
	RMSE	R ²	RMSE	R ²	1% error	5% error	10% error
SVR	93.535	0.345	78.716	0.457	30.68%	55.37%	69.87%
GP	91.140	0.378	24.307	0.948	93.17%	94.84%	96.79%
LightGBM	96.178	0.307	82.872	0.398	7.11%	34.31%	63.46%
BP-1layer	95.222	0.321	76.930	0.481	11.16%	44.21%	67.78%
BP-2layer	96.354	0.304	74.632	0.512	12.69%	46.72%	70.71%
CNN	91.837	0.368	68.267	0.591	7.95%	47.84%	74.20%
GP-WP	86.807	0.435	23.152	0.953	93.17%	95.26%	96.93%

Bold numbers indicate the best results.

在表8至表10中, 1%误差、5%误差和10%误差分别表示在1%、5%和10%的误差阈值率下预测的样本百分比。

首先, 很明显, 对于所有三个不同大小的数据集, GP-WP模型在测试集和整个数据集上优于几乎所有其他机器学习模型, 这进一步提高了其外推能力和拟合能力。当只关注测试集时, 与大样本集上的GP相比, 神经网络和决策树表现出类似的预测误差。然而, 随着样本量的减少, 它们的性能往往会下降。此外, GP作为整个数据集的替代模型, 比其他主流机器学习模型具有显著优势。换句话说, GP-WP模型在不同的误差阈值率下, 对于预测

误差和分数总是能获得最理想的结果。

5. 总结

本文基于官能团贡献度开发纯组分物性机器学习模型。模型开发方法适用于对各种物性进行建模, 与基于机器学习的模型开发的其他选项相比, 无论数据集大小如何, 都表现出卓越的模型性能。与其他基于GC的物性模型类似, 本文开发的模型是预测模型, 在使用基团表示分子方面存在局限性。它们不适合非常小的分子, 如气体, 但外推到较大的分子时, 已被发现是可靠的。虽然模型可以处理一些异构体, 但部分异构体可能是没有办法通过输入来区分的。Alshehri等[36]也承认了这些局限性。感兴趣的读者可以使用Github上提供的方法访问训练数据集、预测结果和GP-WP模型。

与现有方法相比, 本文所提出的方法具有以下优点: ①基于GP, 其构建过程不需要使用大量数据, 也不需要假设模型的结构; ②增加的维度不涉及超参数数量的增加, 但有效的维度信息包含在模型中; ③在大多数情况下, 它比其他模型表现更好, 对新样本的预测也更准确。该方法的改进主要归因于两个因素: ①使用扭曲函数将离散变量打包, 将其映射到连续域, 通过可调超参数在不同程度上改变变量的相关程度; ②在确定模型之前, 会提取并仔细检查先验信息, 使后验更接近真实值。

根据本文所提出的GC模型, 未来的工作可能涉及同时关注多个物性预测的多个输出, 因为不同物性之间的相关性对模型建立非常有用。通过提取这些特征, 可以进一步获得精确的预测。此外, 与表示空间信息(如原子之间的角度和距离)的其他表示不同, 异构体无法通过当前的官能团组进行区分。在机器学习的指导下, 还可以进一步研究和拓展其他分子表示方法的建模。

致谢

衷心感谢国家自然科学基金项目(22150410338, 61973268)的资助。

Compliance with ethics guidelines

Xinyu Cao, Ming Gong, Anjan Tula, Xi Chen, Rafiqul Gani, and Venkat Venkatasubramanian declare that

they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.eng.2023.08.024>.

References

- [1] Hukkerikar AS, Sarup B, Ten Kate A, Abildskov J, Sin G, Gani R. Group contribution⁺ (GC⁺) based estimation of properties of pure components: improved property estimation and uncertainty analysis. *Fluid Phase Equilib* 2012;321:25–43.
- [2] Mackay D, Boethling RS. *Handbook of property estimation methods for chemicals: environmental health sciences*. Boca Raton: CRC Press; 2000.
- [3] Hukkerikar AS. *Development of pure component property models for chemical product-process design and analysis [dissertation]*. Denmark: Technical University of Denmark; 2013.
- [4] Zhou T, Gani R, Sundmacher K. Hybrid data-driven and mechanistic modeling approaches for multiscale material and process design. *Engineering* 2021;7(9): 1231–8.
- [5] Joback KG. Knowledge bases for computerized physical property estimation. *Fluid Phase Equilib* 2001;185(1–2):45–52.
- [6] Joback KG, Reid RC. Estimation of pure-component properties from group contributions. *Chem Eng Commun* 1987;57(1–6):233–43.
- [7] Gani R. Group contribution-based property estimation methods: advances and perspectives. *Curr Opin Chem Eng* 2019;23:184–96.
- [8] Le T, Epa VC, Burden FR, Winkler DA. Quantitative structure-property relationship modeling of diverse materials properties. *Chem Rev* 2012;112(5): 2889–919.
- [9] Wen S, Nanda K, Huang Y, Beran GJO. Practical quantum mechanics-based fragment methods for predicting molecular crystal properties. *Phys Chem Chem Phys* 2012;14(21):7578–90.
- [10] Constantinou L, Gani R. New group contribution method for estimating properties of pure compounds. *AIChE J* 1994;40(10):1697–710.
- [11] Gao C, Govind R, Tabak HH. Application of the group contribution method for predicting the toxicity of organic chemicals. *Environ Toxicol Chem* 1992;11(5): 631–6.
- [12] Aguirre CL, Cisternas LA, Valderrama JO. Melting-point estimation of ionic liquids by a group contribution method. *Int J Thermophys* 2012;33(1):34–46.
- [13] Terrell E. Estimation of Hansen solubility parameters with regularized regression for biomass conversion products: an application of adaptable group contribution. *Chem Eng Sci* 2022;248:117184.
- [14] Marrero J, Gani R. Group-contribution based estimation of pure component properties. *Fluid Phase Equilib* 2001;183–184:183–208.
- [15] Gani R, Harper PM, Hostrup M. Automatic creation of missing groups through connectivity index for pure-component property prediction. *Ind Eng Chem Res* 2005;44(18):7262–9.
- [16] Jirasek F, Hasse H. Perspective: machine learning of thermophysical properties. *Fluid Phase Equilib* 2021;549:113206.
- [17] Venkatasubramanian V. The promise of artificial intelligence in chemical engineering: is it here, finally? *AIChE J* 2019;65(2):466–78.
- [18] Venkatasubramanian V, Mann V. Artificial intelligence in reaction prediction and chemical synthesis. *Curr Opin Chem Eng* 2022;36:100749.
- [19] Mann V, Gani R, Venkatasubramanian V. Group contribution-based property modeling for chemical product design: a perspective in the AI era. *Fluid Phase Equilib* 2023;568:113734.
- [20] Dobbelaere MR, Plehiers PP, Van de Vijver R, Stevens CV, Van Geem KM. Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. *Engineering* 2021;7(9):1201–11.
- [21] Nagai R, Akashi R, Sugino O. Completing density functional theory by machine learning hidden messages from molecules. *npj Comput Mater* 2020; 6(1):43.
- [22] Goh GB, Siegel C, Vishnu A, Hodas NO, Baker N. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. 2017. arXiv:1706.06689.
- [23] Zhou Z, Eden M, Shen W. Treat molecular linear notations as sentences: accurate quantitative structure–property relationship modeling via a natural language processing approach. *Ind Eng Chem Res* 2023;62(12):5336–46.
- [24] Zhang J, Wang Q, Su Y, Jin S, Ren J, Eden M, et al. An accurate and interpretable deep learning model for environmental properties prediction using hybrid molecular representations. *AIChE J* 2022;68(6):e17634.
- [25] Wen H, Su Y, Wang Z, Jin S, Ren J, Shen W, et al. A systematic modeling methodology of deep neural network-based structure–property relationship for rapid and reliable prediction on flashpoints. *AIChE J* 2022;68(1):e17402.
- [26] Padaszyński K, Domańska U. Viscosity of ionic liquids: an extensive database and a new group contribution model based on a feed-forward artificial neural network. *J Chem Inf Model* 2014;54(5):1311–24.
- [27] Li R, Herreros JM, Tsolakis A, Yang W. Machine learning regression based group contribution method for cetane and octane numbers prediction of pure fuel compounds and mixtures. *Fuel* 2020;280:118589.
- [28] Rasmussen CE. Gaussian processes in machine learning. In: Bousquet O, Von Luxburg U, Rätsch G, editors. *Advanced lectures on machine learning*. Berlin: Springer; 2003. p. 63–71.
- [29] Lu X, Jordan KE, Wheeler MF, Pyzer-Knapp EO, Benatan M. Bayesian optimization for field-scale geological carbon storage. *Engineering* 2022; 18: 96–104.
- [30] Capone A, Lederer A, Hirche S. Gaussian process uniform error bounds with unknown hyperparameters for safety-critical applications. In: *Proceedings of the 39th International Conference on Machine Learning*; 2022 Jul 17–23. Baltimore, MD, USA. New York: PMLR; 2022. p. 2609–24.
- [31] Akazaki T. Falsification of conditional safety properties for cyber–physical systems with Gaussian process regression. In: FalconeY, SánchezC, editors. *Proceedings of the 16th International Conference on Runtime Verification*; 2016 Sep 23–30; Madrid, Spain. Cham: Springer; 2016. p. 439–46.
- [32] Mori H, Kurata E. Application of Gaussian process to wind speed forecasting for wind power generation. In: *Proceedings of the 2008 IEEE International Conference on Sustainable Energy Technologies*; 2008 Nov 24–27; Singapore. Piscataway: IEEE; 2008. p. 956–9.
- [33] Sun AY, Wang D, Xu X. Monthly streamflow forecasting using Gaussian process regression. *J Hydrol* 2014;511:72–81.
- [34] Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: a review of Bayesian optimization. *Proc IEEE* 2016; 104(1): 148–75.
- [35] Gelbart MA, Snoek J, Adams RP. Bayesian optimization with unknown constraints. 2014. arXiv:1403.5607.
- [36] Alshehri AS, Tula AK, You F, Gani R. Next generation pure component property estimation models: with and without machine learning techniques. *AIChE J* 2022;68(6):e17469.
- [37] Hukkerikar AS, Kalakul S, Sarup B, Young DM, Sin G, Gani R. Estimation of environment-related properties of chemicals for design of sustainable processes: development of group-contribution⁺ (GC⁺) property models and uncertainty analysis. *J Chem Inf Model* 2012;52(11):2823–39.
- [38] Smola AJ, Schölkopf B. A tutorial on support vector regression. *Stat Comput* 2004;14(3):199–222.
- [39] Hofmann T, Schölkopf B, Smola AJ. Kernel methods in machine learning. *Ann Stat* 2008;36(3):1171–220.
- [40] Roustant O, Padonou E, Deville Y, Clément A, Perrin G, Giorla J, et al. Group kernels for Gaussian process metamodels with categorical inputs. *SIAM/ASA J Uncertain Quantif* 2020;8(2):775–806.
- [41] Qian PZG, Wu H, Wu CFJ. Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics* 2008;50(3):383–96.
- [42] Van de Schoot R, Depaoli S, King R, Kramer B, Märten K, Tadesse MG, et al. Bayesian statistics and modelling. *Nat Rev Methods Primers* 2021;1(1):1.
- [43] Ghosal S, Roy A. Posterior consistency of Gaussian process prior for nonparametric binary regression. *Ann Stat* 2006;34(5):2413–29.
- [44] Casale FP, Dalca AV, Saglietti L, Listgarten J, Fusi N. Gaussian process prior variational autoencoders. In: BengioS, WallachHM, LarochelleH, GraumanK, Cesa-BianchiN, editors. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*; 2018 Dec 3–8; Montréal, QC, Canada. Red Hook: Curran Associates Inc.; 2018. p. 10390–401.
- [45] Kaufman CG, Sain SR. Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Anal* 2010;5(1):123–49.
- [46] Astudillo R, Frazier PI. Thinking inside the box: a tutorial on grey-box Bayesian optimization. In: *Proceedings of the 2021 Winter Simulation Conference*; 2021 Dec 15–17; Phoenix, AZ, USA. Piscataway: IEEE; 2021. p.

- 1–15.
- [47] Nott DJ, Drovandi CC, Mengersen K, Evans M. Approximation of Bayesian predictive p-values with regression ABC. *Bayesian Anal* 2018;13(1):59–83.
- [48] Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc* 1995;90(430):773–95.
- [49] Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. *J Chem Inf Model* 2020;60(8):3770–80.
- [50] Fang J, Gong B, Caers J. Data-driven model falsification and uncertainty quantification for fractured reservoirs. *Engineering* 2022;18:116–28.
- [51] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017 Dec 4–9; LongBeach, CA, USA. Red Hook: Curran Associates Inc.; 2017. p. 3149–57.