



ELSEVIER

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng



Research
AI in Chemical Engineering—Article

基于氢键模式解析的机器学习辅助低共熔溶剂设计

Usman L. Abbas^a, Yuxuan Zhang^b, Joseph Tapia^a, Selim Md^c, Jin Chen^c, Jian Shi^b, Qing Shao^{a,*}

^a Department of Chemical and Materials Engineering, University of Kentucky, Lexington, KY 40506, USA

^b Department of Biosystems and Agricultural Engineering, University of Kentucky, Lexington, KY 40506, USA

^c Institute for Biomedical Informatics, Department of Computer Science, University of Kentucky, Lexington, KY 40506, USA

ARTICLE INFO

Article history:

Received 9 March 2023

Revised 24 August 2023

Accepted 29 October 2023

Available online 9 July 2024

关键词

机器学习

低共熔溶剂

分子动力学模拟

氢键

分子设计

摘要

非离子型低共熔溶剂(DESs)是一种非离子设计溶剂,广泛应用于催化、萃取、碳捕集和制药等领域。然而,由于缺乏准确预测DES形成的有效工具,发现新的DES候选物具有挑战性。对DES的搜索在很大程度上依赖于直觉或试错过程,导致成功率低或错失机会。鉴于氢键(HB)在DES形成中起着核心作用,我们的目标是识别区分DES和非DES系统的HB特征,并利用它们开发机器学习(ML)模型来发现新的DES系统。我们首先使用分子动力学(MD)模拟轨迹分析了38个已知DES和111个已知非DES系统的HB性质。分析表明,与非DES系统相比,DES系统有两个独特的特点:①其组分内两种HB数存在更大的不平衡;②组分间HB数更多且强度更大。基于这些结果,我们使用十种算法和三种HB描述符开发了30个ML模型。首先使用平均和最小受试者工作特征(ROC)-曲线下面积(AUC)值对模型性能进行基准测试。文中还分析了模型中各个特征的重要性,结果与基于模拟的统计分析一致。最后,我们使用34个系统的实验数据验证了这些模型。极端随机树模型在验证中优于其他模型,ROC-AUC为0.88。本文的工作说明了HB在DES形成中的重要性,并展示了ML在发现新的DES中的潜力。

© 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

低共熔溶剂(DESs)是由氢键受体(HBA)和氢键供体(HBD)组成的液体混合物,具有可调特性[1–13]。DESs作为可持续溶剂在许多应用中已经引起了人们的关注,包括碳捕集[2,14–16]、制药[9,14–15,17–23]、材料合成[8,19,24]、电化学[9,14,25–37]、净化[17–18]和萃取[6,8,24,38–41],这得益于其可回收[42]和再利用[43]的特性。与传统溶剂相比,非离子型DES具有几种理想的特性,包括生物可降解性、高导电性、低挥发性和低毒性[2,8,38,43

–44]。非离子型DES通常被归类为V型DES[4,14–16,38],可使用天然化合物制成,并具有低黏度特性,特别适合工业应用,如液-液萃取和碳纳米材料生产[12,38,45]。

DES领域的主要挑战之一是发现大量DES候选物,这将使社区能够有一个庞大的池来探索和搜索具有所需性质的候选物。许多实验和计算研究表明,氢键(HB)在DES的形成和性质调控中起着重要作用[1–2,4,9,14–15,39,42,46]。Farias等[43]进行了一项实验研究,以了解DES的HBD在双水相体系中的作用。他们得出结论,具有高相对亲水性的HBD主要在双相体系中作为佐剂,而具有中

* Corresponding author.

E-mail address: qshao@uky.edu (Q. Shao).

等亲水性的HBD控制着双相体系的形成，具有低亲水性（高疏水性）的HBD形成双水相体系，其中HBA在这种体系中起佐剂作用。Abranches等[1]研究了甜菜碱（一种极性不平衡的分子）作为通用HBA在DES形成中的适用性。他们的研究结合了实验和密度泛函理论计算，证实甜菜碱因其非选择性、低成本和低毒性的特点，是制备天然DES的理想选择。这些基础研究强调了HB在DES形成和性质调控中的重要作用，表明基于HB的描述符可以作为发现新DES的合适输入。

机器学习（ML）模型在预测DES的物理化学特性和热物理特性方面日益受到关注[17,19,46–51]。Hansen等[15]的综述总结了开发定量构效关系模型以预测DES性质的研究[6,15]。Halder等[51]使用化学信息学方法来确定工业应用中准确预测密度所必需的DES的结构属性。他们利用了一种共识建模方法，得出的结论是，HBDs的数量、亲脂性、极化性和范德华表面积等特征可用于高精度预测新型DES的密度。Dietz等[6]使用微扰链统计缔合流体理论（PC-SAFT）模型来预测疏水性DES与水或羟甲基糠醛混合物的液-液平衡和固-液平衡，证明了这种方法在预测疏水性DES混合物的相行为方面的有效性。

其他研究已经使用ML算法来估计DES的密度和黏度。Abdollahzadeh等[19]比较了七种ML算法，结果表明最小二乘支持向量回归在预测149个DES的密度方面具有最高的准确性，比通过经验相关性获得的最佳结果高出74.5%。Zamora等[16]比较了五种基于实验数据训练的ML算法对预测V型DES的密度和黏度的适用性。他们的研究结论表明，支持向量机在预测密度方面表现最佳，高斯过程回归模型在预测黏度方面表现最佳。Xu等[50]使用梯度提升模型预测DES黏度；他们的模型在训练和测试实验和模拟数据时显示出令人满意的结果。总的来说，这些研究证明了将ML和分子模拟结合起来预测DES特性的潜力。

与其他专注于预测DES属性的研究相比，我们的工作旨在使用ML模型来预测DES系统的形成。分子动力学（MD）模拟已经成为一种有价值的技术，可用于确定用作机器学习模型输入的描述符[14,24,46]。我们假设HB特性可以作为DES形成的预测因子。然而，确定相关的HB属性并非易事。我们之前的工作[52]根据组分内和组分间HB数的比值将非离子型DES分为三组。这些观察结果启发我们探索使用基于HB的描述符开发ML模型的可能性。据我们所知，这是首次使用机器学习模型将溶剂分类为DES或非DES的研究。

在机器学习模型训练中，数据是核心要素。为了便于

研究，我们从文献中筛选并构建了一个包含38个已知DES和111个非DES系统的数据库。这个库的构建使我们能够对分子模拟数据进行统计分析，这些数据可用于开发模型的训练和测试数据集。我们整理了一个包含34个系统的独立数据库来验证模型性能。鉴于我们数据库的规模，本文重点介绍传统的机器学习算法。我们使用了十种机器学习算法；然而，我们承认深度学习算法已经成为一种有前景的材料设计技术。使用深度学习算法来预测DES形成的一个障碍是文献中经过实验验证的DES相对稀少。我们开发的模型可以通过产生可能形成DES的溶剂混合物来帮助加速发现新的DES。本文的其余部分结构如下：第2节提供了计算方法的详细信息；第3节给出了结果和讨论；第4节给出了我们的结论。

2. 方法论

2.1. DES和非DES系统的库

附录A中的表S1至表S8提供了本研究中模拟的183个系统的详细信息。如文献中所报道，在这些系统中，38个被鉴定为已知DES，111个为已知非DES系统。这些构成了我们的训练集和测试集。此外，我们还保留了34个经过实验验证的系统（17个DES和17个非DES）用于验证。DES和非DES系统的分类基于van Osch等[53–54]的实验结果，仅考虑其列表中的非离子型DES。本研究排除了所有缺乏三种类型HB（A–A、B–B和A–B）的DES。模拟中使用的化合物用三个字母的缩写表示，如“DEA”代表“癸酸”。我们遵循van Osch等的命名规范，其中系统A-B中的组分A是预期的HBD，化合物B是预期的HBA。这些系统以其化合物的三个字母缩写和相应的摩尔比表示；例如，DEA-MEN11表示癸酸和薄荷醇的1:1混合物。附录A中的表S9至表S11列出了本研究中使用的化合物的缩写。

2.2. 分子模拟

2.2.1. 分子模型

本研究中，使用液体模拟的全原子优化势（OPLS-AA/M）力场[55]来描述分子。由于OPLS-AA/M力场已被证明能够准确模拟有机分子的行为，因此该系统中的非键合和键合参数是基于OPLS-AA/M力场确定的。力场参数是通过LigParGen [56]网络服务器生成的。

2.2.2. 模拟细节

模拟系统是通过在立方体中随机插入特定数量的（基

于摩尔比)所选有机分子而创建的。图1显示了使用可视化分子动力学(VMD)生成的THY-MEN11系统的快照[57]。

对于每个系统,模拟过程包括三个步骤:①能量最小化,以消除任何原子重叠;②50 ns等压等温[NPT,其中, N 是粒子数, P 是压力($P = 1 \text{ atm}$, $1 \text{ atm} = 101.325 \text{ kPa}$), T 是温度($T = 295 \text{ K}$)]系综MD模拟,使系统达到热力学平衡;③10 ns正则(NVT,其中, V 是系统的体积, $T = 295 \text{ K}$)系综MD模拟,以10 ps的频率收集数据。在步骤②中,分子动力学模拟使用Berendsen等[58]的方法来控制系统压力,而速度重标度[59]方法用于控制系统温度。

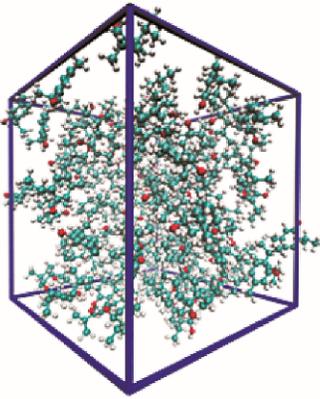


图1. 平衡的THY-MEN11系统的快照。分子以Corey-Pauling-Koltun(CPK)模型显示(原子配色方案:C,青色;O,红色;H,白色)。

分别使用Lennard-Jones 12-6和库仑势(E),通过方程(1)计算OPLS-AA/M力场中的短程和长程非键相互作用。

$$E = \sum_i \sum_{j < i} \left\{ \frac{1}{4\pi\epsilon_0} \frac{q_i q_j e^2}{r_{ij}} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \right\} \quad (1)$$

式中, r_{ij} 为原子*i*和*j*之间的距离; q_i 和 q_j 分别为原子*i*和*j*的部分电荷; ϵ_0 为自由空间电容率; ϵ_{ij} 和 σ_{ij} 分别为能量和几何参数。粒子网格Ewald(PME)[60]求和用于计算长程势能,线性约束求解器(LINCS)算法[61]用于约束涉及氢原子的键。所有能量最小化和分子动力学模拟均使用GROMACS 2021.2[62]进行。

2.3. 氢键分析

我们使用Luzar和Chandler[63]制定的标准对HB进行了表征:①O(供体)和O(受体)之间的距离 $\leq 0.35 \text{ nm}$;②O(受体)-H(供体)-O(供体)的角度 $\leq 30^\circ$ 。我们分两步计算HB寿命:

(1) 计算相关函数 $C(t)$,如方程(2)所示:

$$C(t) = \frac{\langle N_{\text{HB}}(t) \rangle}{\langle N_{\text{HB}}(0) \rangle} \quad (2)$$

式中, $\langle N_{\text{HB}}(0) \rangle$ 是初始状态下HB数的系综平均值; $\langle N_{\text{HB}}(t) \rangle$ 是时间*t*时仍然存在的HB数的系综平均值。根据Rappaport的定义[64],即使HB间歇性断裂,也应计算在内。

(2) 通过数值积分 $C(t)$ 曲线来计算寿命。

2.4. 机器学习模型

第2.1节中介绍的基于文献构建的数据库包含的非DES系统比DES系统多,因此数据不平衡可能导致模型训练中的偏差。为了减弱这种潜在的人为影响,我们在机器学习模型开发过程中为每轮训练整理了一个包含38个DES和38个非DES系统的数据库。从库中最初的111个非DES系统中随机选择了38个非DES系统。我们将这个数据库进一步分为由30个DES和30个非DES系统组成的训练集及由8个DES和8个非DES系统组成的测试集。我们在从DES和非DES数据集中采样时使用了固定种子,以确保所有模型都在相同的数据集切片上进行评估。所有模型均通过实验验证的DES和非DES系统进行了进一步验证,如第3节所述。

我们使用scikit学习[65–66]和XGBoost[67]包提供的算法训练了10种不同的机器学习算法。这些算法是:①逻辑回归;②决策树;③梯度提升;④AdaBoost;⑤随机森林;⑥极端随机树;⑦支持向量机;⑧*k*近邻;⑨XGBoost;⑩XGBoost随机森林。使用scikit学习的网格搜索方法进行超参数优化。使用受试者工作特征(ROC)-曲线下面积(AUC)指标,通过*k*折交叉验证(6折和10次重复)来测量每个模型的性能。在优化过程中具有最高ROC-AUC值的模型被认为是最佳模型。对于每种机器学习算法,后续的训练和测试仅在训练得最好的模型上进行。

HB在DES的形成中起着决定性的作用。为了全面了解HB环境,必须知道系统中有多少分子相互作用形成HB(即HB数),以及这些HB持续多长时间(即HB寿命)。所有ML算法都考虑三种类型的输入特征:①HB数;②HB寿命;③HB数与寿命的组合。通过分子动力学模拟生成的输入特征如表S1至表S8所示。本研究共训练了30个模型;模型的超参数详见附录A中的表S12至表S14。

本研究中使用以下Python软件包进行工作:Python(版本3.10.8)、scikit学习[66](版本1.2.0)、pandas[68](版本1.5.2)、NumPy[69](版本1.22.3)、matplotlib[70]

(版本 3.6.2)、SciPy [71] (版本 1.7.3) 和 XGBoost [67] (版本 1.7.3)。所有机器学习工作均在第八代英特尔酷睿 i7-8750H 处理器上执行。

2.5. 实验

为了验证训练模型的有效性, 我们使用了之前研究 [72] 中的溶剂配方列表, 以确定这些配方是否可以形成 DES。为了制备该系统, 将两种组分以特定的摩尔比混合, 加热并不断搅拌以确保完全混合。更具体地说, 首先根据摩尔比计算每个组分所需的质量, 然后依次用分析天平 (VWR-224AC, VWR International, 美国) 称量到玻璃瓶中。使用玻璃棒预混合化合物, 随后将磁力搅拌子加入瓶中。然后将瓶子密封并置于油浴中加热。温度通常保持在 80 °C, 在磁力搅拌器加热板 (德国 Heidolph Instruments 公司 Hei-Tec 品牌) 上以 $500 \text{ r} \cdot \text{min}^{-1}$ 的速度持续搅拌 1 h。对于在该温度下没有形成均匀液体的组合, 进一步应用更高温度 100 °C 和 120 °C 来检查这些组合是否可以在高温下转化为液态。加热过程后, 将混合物空气冷却至室温, 并在干燥器中保持 24 h。在 24 h 内保持液态无结晶的样品被认为是 DES 候选物。然而, 我们观察到一些系统最初表现出类似 DES 的行为, 但几天后最终形成了固相。这些系统被排除在本研究之外。最终, 我们选择了 17 个 DES 系统和 17 个非 DES 系统。

3. 结果与讨论

3.1. 氢键特征的统计分析

3.1.1. 氢键数特征

我们首先分析了 DES 和非 DES 系统的实际组分内和组分间 HB 数的概率密度分布。图 2 显示了基于平均组分间和组分内 HB 数的 38 个 DES 和 111 个非 DES 系统的分

布。图 2 中的模式分布并没有显示出明显的差异。如图 2 (a) 所示, DES 的组分内 HB 数 (A-A 和 B-B) 呈现左偏特征, 表明我们数据集中的大多数 DES 的平均 HB 数小于 20。与 A-A HB 相比, B-B HB 数集中在分布区间的低端。组分间 HB 数呈现右偏分布, 表明大多数 DES 的组分间 HB 数高于组分内 HB 数。在图 2 (b) 中, 非 DES 的组分内 HB 数也呈现左偏分布。此外, 大多数组分间 HB 数右偏。因此, 根据图 2 的分析, 组分内和组分间 HB 的实际数量可能不是区分 DES 和非 DES 系统的合适 HB 特征。

当绘制 DES 和非 DES 系统的平均组分间和组分内 HB 数的箱线图时, 二者展现出显著差异模式。如图 3 (a) 和 (b) 所示, 与非 DES 系统相比, DES 系统在两类组分内 HB 数 (A-A 与 B-B) 的中位数之间存在很大差异。此外, DES 系统中的组分间 HB 的中位数 (56.07) 比 A-A 和 B-B 的中位数高 6%~83%。对于非 DES 系统, A-B HB 仅呈现中位数值 48.36, 分别比 A-A 和 B-B 高 55%~59%。在 DES 和非 DES 系统中, 即使将两类组分内 HB 数 (A-A 和 B-B) 相加, 组分间 HB 数 (A-B) 仍然更大。这种中位数的差异意味着两类组分内和组分间/组分内 HB 数的比值可以作为区分 DES 和非 DES 系统的重要特征。

图 3 (c) 中的 A-A/B-B 和 A-B/(A-A+B-B) 分布图进一步证实了我们的假设。对于某些 DES, 组分间 HB 数与组分内 HB 数的比值远高于 1.5。平均而言, DES 的组分间 HB 数比组分内 HB 总数大 35%。最后, 我们研究了组分内键的比值, 以更深入了解其差异的大小。DESs 的 A-A 与 B-B HB 的比值平均为 8.01。对于非 DES, A-A 与 B-B HB 的比值下降到 3.44。非 DES 的组分内平均 HB 数 (A-A 和 B-B) 大致相同 (分别为 24.31 和 24.08)。在非 DES 中, A-A 和 B-B 的 HB 数中位数也相似, 分别为 20.01 和 21.41。这一发现表明, 在非 DES 中不存在占主导地位的分内 HB。这在图 3 (c) 中也有显示, 其中大多

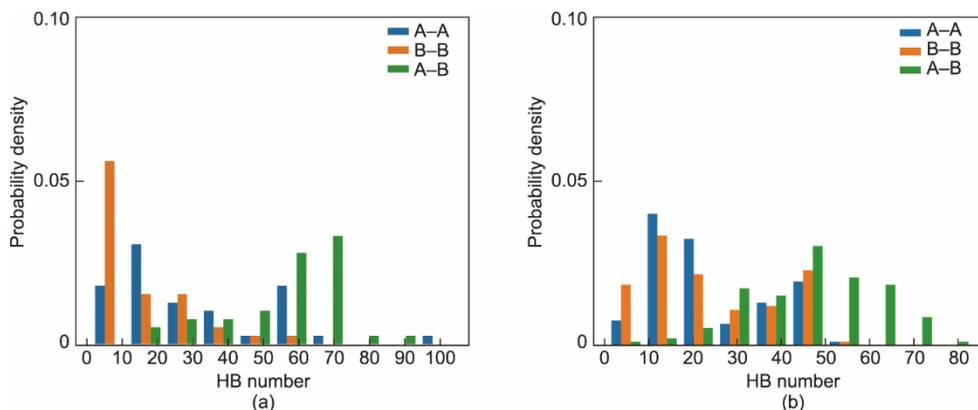


图 2. 平均组分间 (A-B) 和组分内 (A-A 和 B-B) HB 数的概率密度分布。(a) DES 系统; (b) 非 DES 系统。

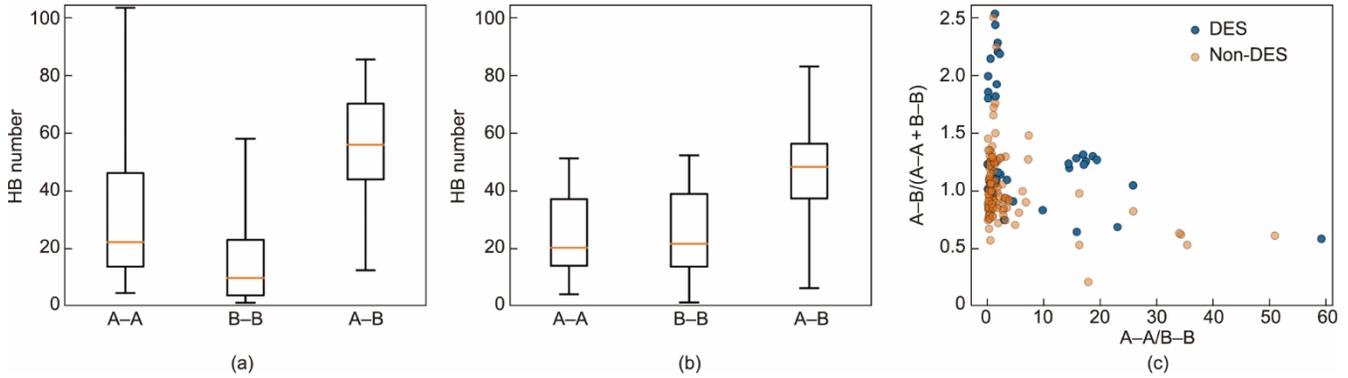


图3. DES和非DES的HB数特征。(a)、(b)平均HB数的分布：(a) DES系统和(b)非DES系统。从底部开始，箱线图显示了最小值，第25、50和75百分位数以及最大值。(c)组间与组内HB数的比值vs.组内HB数的比值。

数非DES的组分内HB数比值都集中在1.0左右，只有少数异常值。对于非DESs，平均组间(A-B)HB数约为平均组分内(A-A和B-B)HB数的两倍(1.93~1.95倍)。相对于DESs，非DESs中的组分内HB的比例较小。例如，在第25、50和75百分位数上，DESs的A-A/B-B值分别为1.13、1.91和15.51，而非DESs的A-A/B-B值分别为0.15、0.49和0.99。附录A中的表S15和表S16提供了更多详细信息。

3.1.2. 氢键寿命

我们还分析了38个DES系统和111个非DES系统的组间以及组分内HB寿命的概率密度分布。在不同分布区间中，我们观察到DESs的两种不同情况：①组间(A-B)HB占主导地位；②组分内HB(A-A或B-B)占主导地位。这一发现与我们之前的工作一致，我们在之前的工作中将几个已知的DESs分类为组间或组分内HB占主导地位的组。在图4(a)中，组分内HB寿命(A-A)集中在2.0~4.0 ns，而DES的B-B寿命则集中在0.25~2.50 ns。组间HB寿命(A-B)似乎右偏，并且持续时间比组分内HB寿命长。

图4(b)显示，非DES的组分内HB寿命在不同的区间中占主导地位，但并没有明显的趋势；例如，B-B在寿命小于1.25 ns时占主导地位，但A-A在寿命大于3.00 ns时占主导地位。在每个区间中，A-B寿命似乎比组分内HB寿命更占主导地位，同时与其他组分内HB寿命相似。缺乏明确的模式意味着，仅凭实际的组分内和组间HB寿命特征，可能不足以区分DES和非DES系统。

当我们以箱线图的方式绘制组分内和组间的HB寿命分布时，出现了一些差异。如图5(a)所示，组间(A-B)与其中一个组分内(A-A)HB寿命的中位数之间存在微小差异；而与其他组分内(B-B)HB寿命中位数之间的差异更大。A-B寿命的中位数为2.67，分别比A-A和B-B寿命的中位数大14%和39%。如图5(b)所示，非DESs的组间HB寿命值和组分内HB寿命值的中位数之间存在较小的差异。A-B HB寿命的中位数为2.72，分别比A-A和B-B HB寿命的中位数大3.6%和14.0%。这些差异表明，与实际寿命相比，组间与组分内HB寿命的比值作为特征可能更实用。

图5(c)中A-A/B-B与A-B/(A-A+B-B)的曲线证

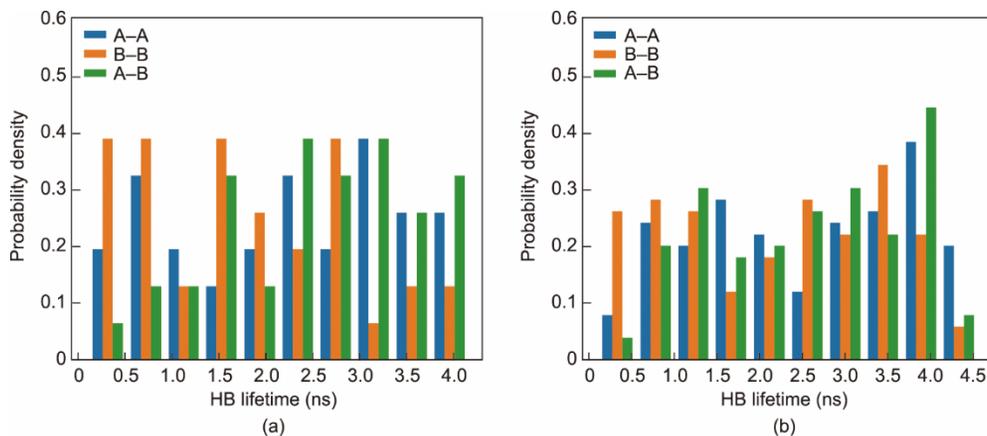


图4. 平均组间(A-B)和组分内(A-A和B-B)HB寿命的概率密度分布。(a) DES系统；(b)非DES系统。

实了这一假设。在DES中，A-A寿命中位数比B-B寿命长约7%，而在非DES中寿命则长约13%。尽管组分间HB比组分内HB多，但组分内HB持续的时间更长。DES中A-B/(A-A+B-B)寿命比值的中位数是0.63，非DES中A-B/(A-A+B-B)寿命比值的中位数是0.53。DES中组分间HB寿命与组分内HB寿命的比值在0.5~2.0之间变化，而大多数非DES中组分间HB寿命与组分内HB寿命的比值集中在0.5左右。与HB数类似，HB寿命的比值作为特征可能比实际的寿命值更有用。

3.2. 模型开发

我们使用十种算法（逻辑回归、决策树、梯度提升、AdaBoost、随机森林、极端随机树、支持向量机、 k 近邻、XGBoost和XGBoost随机森林）和三种类型的输入特征（HB数、HB寿命以及HB数和寿命特征的组合）训练了30个模型，可以预测系统是否是DES。对于每种类型的输入特征，我们使用了第3.1节中提到的五个变量[A-A、B-B、A-B、A-A/B-B和A-B/(A-A+B-B)]。我们对每个模型进行了100轮训练，并计算了这100轮训练中每个模型的平均ROC-AUC值。ROC是概率曲线，而AUC代表可分离性的程度或度量。ROC-AUC显示了模型在区分类别方面的能力。AUC越高，表明模型在区分DES和非DES类别时越好。对于每一轮，我们从DES和非DES数据集中随机抽取38个（30个用于训练，8个用于测试）条目。在每一轮中，使用六折网格搜索交叉验证进行超参数优化，以ROC-AUC作为评估标准。为了确保公平比较，每个模型都使用来自DES和非DES数据集的相同样本进行训练和测试。附录A中的图S1至图S11显示了训练过程中每个模型的ROC-AUC的变化。

我们使用两个标准对模型进行排名：①平均ROC-AUC得分（表1）；②最小ROC-AUC得分（表2）。

当使用HB寿命特征进行训练时，AdaBoost和极端

随机树分类器的平均ROC-AUC得分为0.70，并列为表现最佳模型。当使用HB数特征进行训练时，XGBoost随机森林、随机森林和XGBoost是表现最好的模型，平均ROC-AUC分别为0.82、0.81和0.81。当HB数和寿命特征结合在一起时，表现最好的模型是随机森林和XGBoost随机森林分类器，两者的平均ROC-AUC均为0.79。总体而言，根据平均ROC-AUC值和最小ROC-AUC值，表现最好的模型分别是XGBoost随机森林和极端随机树。

100个训练周期中的最小ROC-AUC得分也可以用于评估模型的性能。表2列出了30个模型的最小ROC-AUC得分。单独使用HB寿命特征训练的最小ROC-AUC得分在0.15~0.30之间，低于使用HB数或HB数和寿命特征训练的最小ROC-AUC得分。这些观察表明，仅凭HB寿命可能不足以开发用于分类DES系统的ML模型。在所有类别中，极端随机树分类器在使用HB数进行训练时，最小ROC-AUC得分最高，为0.70。

无论使用何种模型选择标准，一些算法都是表现最好的。对于使用HB数字训练的模型，表现最好的模型是基

表1 30个模型的平均ROC-AUC值(最佳值以粗体显示)

Algorithm	Lifetime	Number	Number + lifetime
Logistic regression	0.68	0.78	0.77
Decision tree	0.63	0.74	0.68
Gradient boost	0.66	0.78	0.76
AdaBoost	0.70	0.78	0.75
Random forest	0.69	0.81	0.79
Extra trees forest	0.70	0.80	0.78
Support vector machine	0.64	0.77	0.77
k -nearest neighbors	0.63	0.77	0.77
XGBoost	0.67	0.81	0.77
XGBoost-random forest	0.62	0.82	0.79

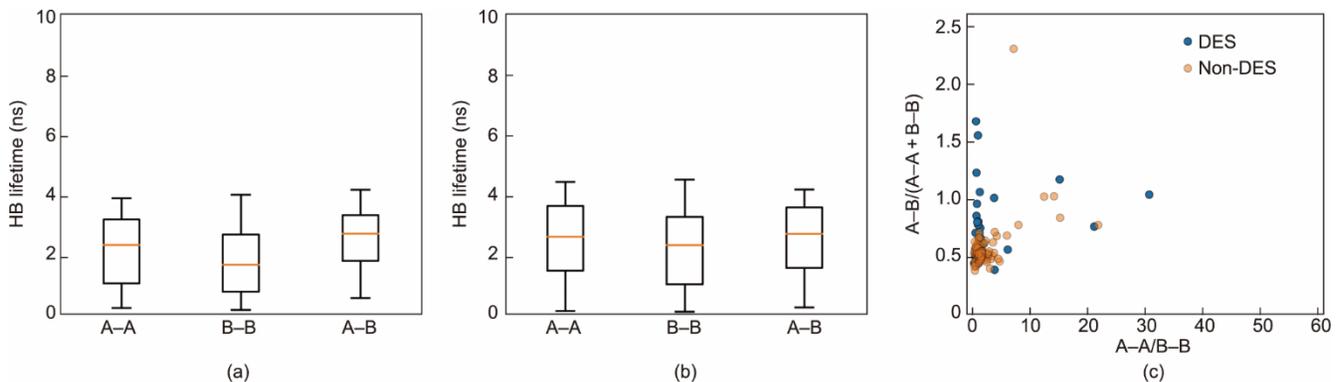


图5. DES和非DES的HB寿命特征。(a)、(b)平均HB寿命的分布：(a) DES系统；(b)非DES系统。从底部开始，箱线图显示了最小值，第25、50和75百分位数以及最大值。(c)组分间HB寿命与组分内HB寿命之比 vs. 组分内HB寿命之比。

表2 30个模型的最小ROC-AUC得分(最佳值用粗体表示)

Algorithm	Lifetime	Number	Number + lifetime
Logistic regression	0.25	0.55	0.50
Decision tree	0.30	0.50	0.45
Gradient boost	0.25	0.50	0.40
AdaBoost	0.30	0.40	0.38
Random forest	0.30	0.45	0.45
Extra trees forest	0.30	0.70	0.55
Support vector machine	0.15	0.10	0.10
<i>k</i> -nearest neighbors	0.30	0.45	0.45
XGBoost	0.20	0.50	0.45
XGBoost-random forest	0.20	0.55	0.45

于最小ROC-AUC的极端随机树，并且当根据平均ROC-AUC得分判断时，该模型仅略微落后于XGBoost随机森林。对于使用HB寿命特征训练的模型，无论是使用平均ROC-AUC得分还是最小ROC-AUC得分，极端随机树和AdaBoost都是表现最好的。在使用组合HB数和寿命特征训练的模型中，极端随机树分类器是使用最高的最小ROC-AUC得分或平均ROC-AUC得分的最佳表现者。然而，应该指出的是，在训练期间观察到的出色表现并不一定意味着在验证阶段表现出色，这将在下一节中讲述。

3.3. 模型验证与实验结果

我们使用34个实验结果（17个DES和17个非DES系统）验证了30个训练模型。结果如表3所示。

对于使用HB寿命特征训练的模型，XGBoost随机森林、逻辑回归和极端随机树是表现最好的，在验证过程中ROC-AUC值分别为0.68、0.65和0.65。当模型用HB数特征训练时，支持向量机、极端随机树和梯度提升是表现最好的，ROC-AUC值分别为0.80、0.79和0.77。在使用HB数和寿命训练的模型中，极端随机树、逻辑回归和梯度提升是表现最好的模型，ROC-AUC值分别为0.88、0.84和0.81。总的来说，集成算法（集成和提升）表现良好。诸如随机森林、极端随机树和决策树等集成算法会构建和训练独立的估计器，然后对这些估计器的独立预测结果进行平均，以做出最终预测。这有助于减少预测方差并提高准确性。另一方面，提升算法（XGBoost、XGBoost随机森林、AdaBoost和梯度提升）会顺序训练多个估计器。每个估计器都专注于减少前一个估计器的误差，这通常会减少偏差。

图6显示了验证期间三种输入特征类别中表现最佳的模型的混淆矩阵。混淆矩阵显示了每个模型预测的真阳性、真阴性、假阳性和假阴性结果。在本研究中，DES为

阳性，而非DES为阴性。敏感性衡量有多少DES被模型正确地预测为DES，而特异性衡量有多少非DES被模型正确地预测为非DES。一些模型在预测DESs（高敏感性）方面表现更好，而一些模型在预测非DESs（高特异性）方面表现更好。

表3 使用验证数据测试时训练模型的ROC-AUC值(每种特征类型下表现最佳模型的ROC-AUC值以粗体显示)

Algorithm	Lifetime	Number	Number + lifetime
Logistic regression	0.65	0.66	0.84
Decision tree	0.52	0.69	0.65
Gradient boost	0.57	0.77	0.81
AdaBoost	0.61	0.74	0.66
Random forest	0.54	0.76	0.79
Extra trees forest	0.65	0.79	0.88
Support vector machine	0.56	0.80	0.80
<i>k</i> -nearest neighbors	0.47	0.53	0.57
XGBoost	0.61	0.65	0.74
XGBoost-random forest	0.68	0.69	0.79

在使用HB寿命特征训练的模型中，XGBoost随机森林是表现最佳的算法。它在预测哪些系统是DES方面表现最佳，如0.82的高敏感性所示[图6(a)]，但它不擅长预测哪些系统是非DES[特异性低，为0.47，图6(a)]。在用HB数特征训练的模型中，支持向量机是表现最佳的算法。它的特异性为0.88[图6(c)]，这意味着它在预测哪些系统是非DESs方面表现最佳。其敏感性低至0.35[图6(c)]，这意味着它不擅长预测DES。当模型以组合HB寿命和HB数特征作为输入进行训练时，极端随机树模型表现最佳。它的敏感度为0.76[图6(e)]，表明它在预测哪些系统是DES方面表现最佳。它的特异性为0.94[图6(e)]，表明它在预测哪些系统是非DES方面表现最佳。相对于其他输入特征类别中表现最好的模型，极端随机树算法在预测DES和非DES方面总体上表现最佳。所有模型的混淆矩阵如附录A中的图S18至图S20所示。

3.4. 预测概率

预测概率是衡量每个模型分离DES与非DES能力的有用指标。具有良好分离能力的模型，其所有非DES预测的DES概率小于0.5并尽可能接近0，而其DES预测的DES概率大于0.5并尽可能接近1。图7显示了验证期间最佳模型的预测概率分布。从图7(a)中的概率分布可以看出，XGBoost随机森林的预测值主要集中在0.49~0.51

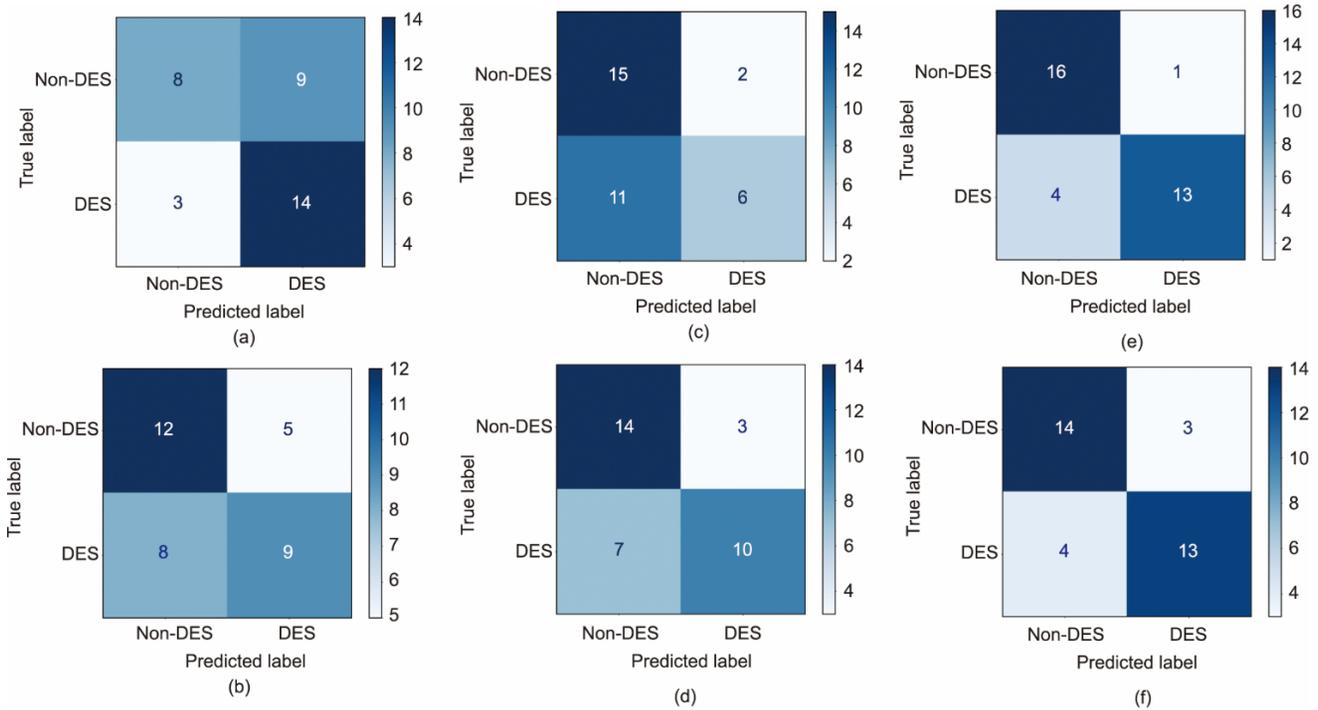


图6. 验证期间表现最佳的模型的混淆矩阵。(a)、(b) HB寿命特征:(a) XGBoost随机森林;(b) 逻辑回归。(c)、(d) HB数特征:(c) 支持向量机;(d) 极端随机树。(e)、(f) 组合HB数和寿命特征:(e) 极端随机树;(f) 逻辑回归。右侧的颜色条表示模型的性能,从白色到蓝色表示性能从差到好。

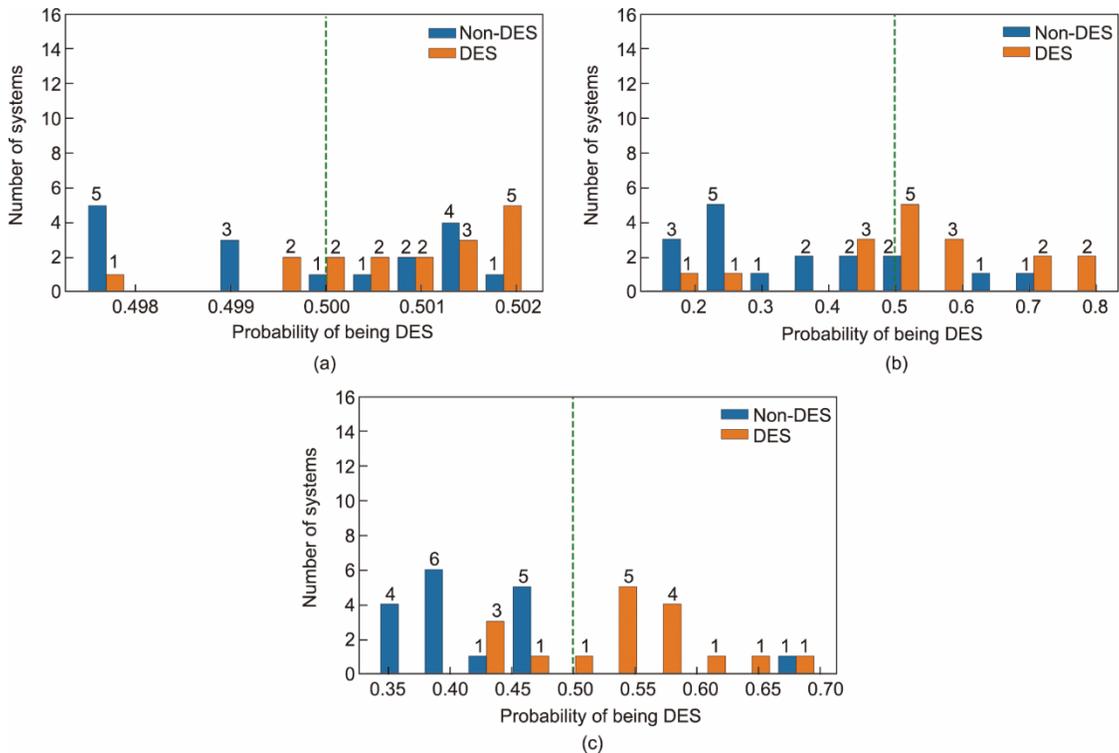


图7. 验证期间表现最佳的模型的预测概率分布。(a) 用HB寿命特征验证的XGBoost随机森林模型;(b) 使用HB数特征验证的支持向量机模型;(c) 使用组合的HB数和寿命特征验证的极端随机树模型。每个箱内的系统数量在条形图上显示。垂直虚线表示分类阈值。理想模型的预测结果将呈现所有非DES位于垂直虚线的左侧,而DES位于右侧。

之间,这表明使用HB寿命特征训练的模型之间没有明显的差异。值得注意的是,XGBoost随机森林的所有14个

DES预测的置信度都大于0.5,因此都是正确的。图7(b)中的分离情况有所改善,概率分布在0.46~0.54之间,这

表明HB数特征帮助模型检测非DES的能力相对优于单独使用HB寿命。这一结论得到了模型的验证，即支持向量机模型做出的所有15个DES概率小于0.5的非DES预测都是正确的。图7(c)中概率分布在0.30~0.70之间，表明当HB数和寿命特征作为输入组合时，极端随机树模型的预测具有更好的置信度。极端随机树模型在分类中表现出更好的分离能力，并且对其非DES预测特异性达0.94 [图6(e)]。它只有一个非DES预测错误。所有其他模型的预测概率如附录A中的图S21至图S23所示。

当ML模型进行预测时，了解哪些输入特征具有最大的权重是有用的。图8显示了模型是如何对输入特征的重要性进行排序的。绝大多数仅使用HB寿命特征训练的模型将组分间与组分内的HB寿命比列为预测的最重要特征，其次是组分间的HB寿命。当仅使用HB数特征进行训练时，模型将组分间HB数列为最重要的特征；但是，应该注意的是，组分间与组分内HB数的比值紧随其后，排在第二位。当将HB数和寿命组合起来时，训练模型将组分间HB数列为最重要的特征，紧随其后的是组分间与组分内HB寿命的比值。

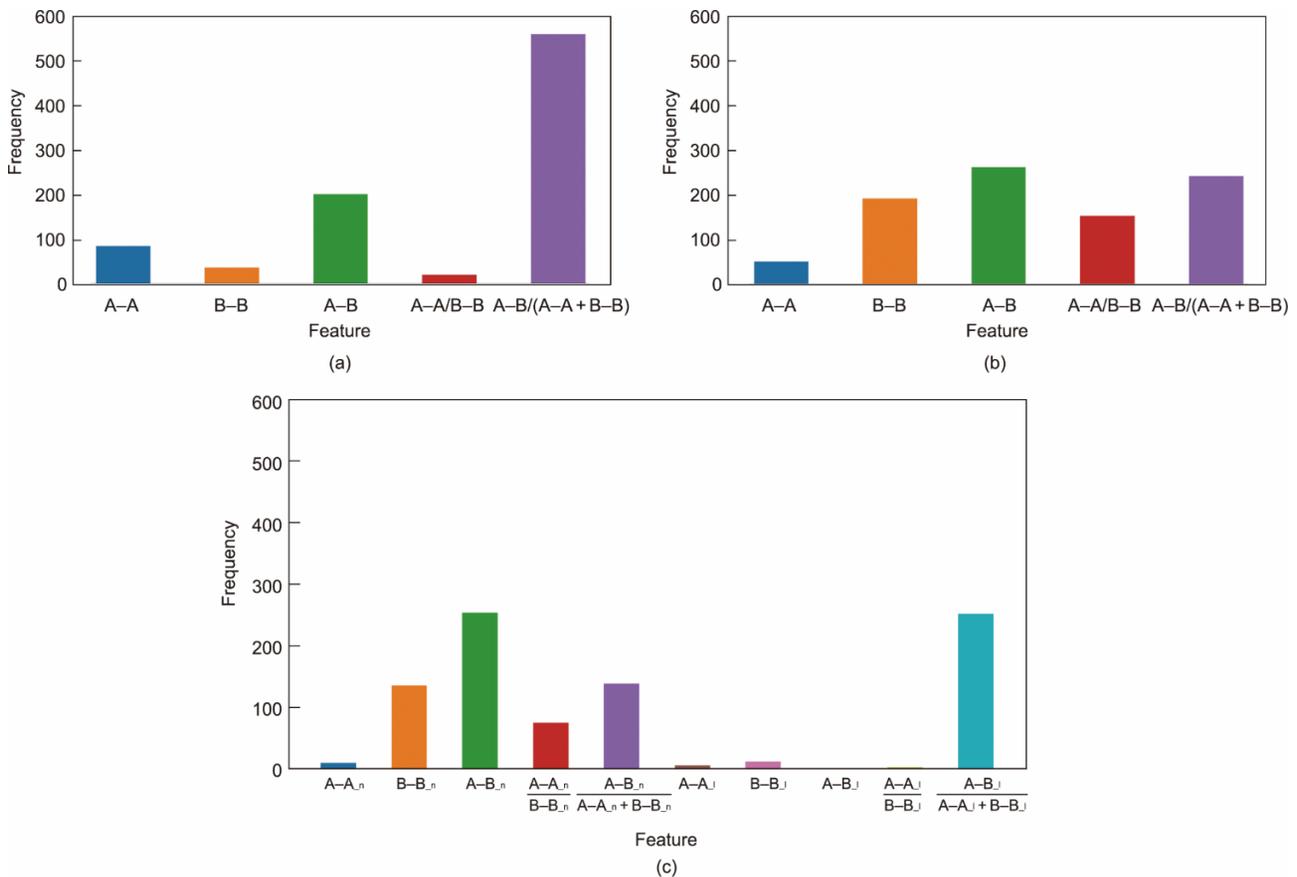


图8. 所有模型在训练迭代过程中的重要特征。(a) 使用HB寿命训练的模型；(b) 使用HB数训练的模型；(c) 使用HB数和寿命训练的模型。(c)中，下标“_n”和“_l”分别表示HB数和寿命特征。

4. 结论

我们使用MD模拟轨迹分析了38个已知DES和111个已知非DES系统的HB特征。对组分内HB数和寿命的统计分析揭示了与非DES系统相比，DES系统具有两个独特特征：DESs在两个组分内HB数之间表现出不平衡；组分间HB更多且更强。然后，我们通过在三种类型的输入特征上训练十种算法，开发了30个ML模型，并使用经过实验验证的17个DES和17个非DES系统对模型进行了验证。使用最高平均值和最高最小ROC-AUC得分这两个标准，我们发现当使用HB寿命、HB数以及寿命和HB数组合特征进行训练时，逻辑回归、梯度提升、支持向量机和极端随机树模型是表现最好的模型。经实验验证集测试，极端随机树分类器是总体上表现最佳的模型，其ROC-AUC为0.88，HB数和寿命组合作为输入。直观地看，当输入有关HB数以及这些HB数持续时间的信息时，模型的表现更好具有合理性。所有模型都将组分间HB数及其与组分内HB数和寿命的比值列为DES分类的最重要特征。

DESs 是一种有前景的溶剂，具有巨大的潜力。由于候选化合物的规模庞大，因此拥有能够准确预测混合时哪些化合物会形成或不会形成 DES 的模型非常重要。本工作中开发的 ML 模型的目的是基于 MD 模拟数据确定一个二元体系是否可以形成 DES。这些 ML 模型可以通过加速发现新的 DES 候选物来协助 DES 研究。我们的工作揭示了哪些化合物可能形成 DES，但没有指出它们可能具有的物理化学性质。未来，需要做更多的工作来预测哪些化合物将形成具有特定应用性质的 DES。

Acknowledgements

This work was supported by Ignite Research Collaborations (IRC), Startup funds, and the UK Artificial Intelligence (AI) in Medicine Research Alliance Pilot (NCATS UL1TR001998 and NCI P30 CA177558), University of Kentucky Center for Computational Sciences and Information Technology Services Research Computing for the use of the Lipscomb Compute Cluster of the University of Kentucky.

Compliance with ethics guidelines

Usman L. Abbas, Yuxuan Zhang, Joseph Tapia, Selim Md, Jin Chen, Jian Shi, and Qing Shao 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.2023.10.020>.

References

- [1] Abranches DO, Silva LP, Martins MAR, Pinho SP, Coutinho JAP. Understanding the formation of deep eutectic solvents: betaine as a universal hydrogen bond acceptor. *ChemSusChem* 2020;13(18):4916–21.
- [2] Stephens NM, Smith EA. Structure of deep eutectic solvents (DESs): what we know, what we want to know, and why we need to know it. *Langmuir* 2022; 38(46): 14017–24.
- [3] Celebi AT, Dawass N, Moulton OA, Vlught TJH. How sensitive are physical properties of choline chloride–urea mixtures to composition changes: molecular dynamics simulations and Kirkwood–Buff theory. *J Chem Phys* 2021;154(18): 184502.
- [4] Abranches DO, Coutinho JAP. Type V deep eutectic solvents: design and applications. *Curr Opin Green Sustain Chem* 2022;35:100612.
- [5] Alcalde R, Gutiérrez A, Atilhan M, Aparicio S. An experimental and theoretical investigation of the physicochemical properties of choline chloride–lactic acid based natural deep eutectic solvent (NADES). *J Mol Liq* 2019;290:110916.
- [6] Dietz CHJT, Erve A, Kroon MC, van Sint AM, Gallucci F, Held C. Thermodynamic properties of hydrophobic deep eutectic solvents and solubility of water and HMF in them: measurements and PC-SAFT modeling. *Fluid Phase Equilib* 2019;489:75–82.
- [7] Florindo C, Branco LC, Marrucho IM. Development of hydrophobic deep eutectic solvents for extraction of pesticides from aqueous environments. *Fluid Phase Equilib* 2017;448:135–42.
- [8] Kivela H, Salomäki M, Vainikka P, Mäkilä E, Poletti F, Ruggeri S, et al. Effect of water on a hydrophobic deep eutectic solvent. *J Phys Chem B* 2022;126(2): 513–27.
- [9] Kovács A, Neyts EC, Cornet I, Wijnants M, Billen P. Modeling the physicochemical properties of natural deep eutectic solvents. *ChemSusChem* 2020;13(15):3789–804.
- [10] Křížek T, Bursová M, Horsley R, Kuchař M, Tůma P, Čabala R, et al. Menthol based hydrophobic deep eutectic solvents: towards greener and efficient extraction of phytocannabinoids. *J Clean Prod* 2018;193:391–6.
- [11] Li K, Jin Y, Jung D, Park K, Kim H, Lee J. *In situ* formation of thymol-based hydrophobic deep eutectic solvents: application to antibiotics analysis in surface water based on liquid–liquid microextraction followed by liquid chromatography. *J Chromatogr A* 2020;1614:460730.
- [12] Lukaczynska-Anderson M, Mamme MH, Ceglia A, Van denBergh K, De Strycker J, De Proft F, et al. The role of hydrogen bond donor and water content on the electrochemical reduction of Ni²⁺ from solvents—an experimental and modelling study. *Phys Chem Chem Phys* 2020;22(28):16125–35.
- [13] Martins MAR, Silva LP, Schaeffer N, Abranches DO, Maximo GJ, Pinho SP, et al. Greener terpene–terpene eutectic mixtures as hydrophobic solvents. *ACS Sustain Chem Eng* 2019;7(20):17414–23.
- [14] Tolmachev D, Lukasheva N, Ramazanov R, Nazarychev V, Borzdun N, Volgin I, et al. Computer simulations of deep eutectic solvents: challenges, solutions, and perspectives. *Int J Mol Sci* 2022;23(2):645.
- [15] Hansen BB, Spittle S, Chen B, Poe D, Zhang Y, Klein JM, et al. Deep eutectic solvents: a review of fundamentals and applications. *Chem Rev* 2021;121(3): 1232–85.
- [16] Zamora L, Benito C, Gutiérrez A, Alcalde R, Alomari N, Bodour AA, et al. Nanostructuring and macroscopic behavior of type V deep eutectic solvents based on monoterpenoids. *Phys Chem Chem Phys* 2021;24(1):512–31.
- [17] Bergua F, Castro M, Lafuente C, Artal M. Thymol + *L*-menthol eutectic mixtures: thermophysical properties and possible applications as decontaminants. *J Mol Liq* 2022;368(Pt B):120789.
- [18] Bergua F, Castro M, Muñoz-Embid J, Lafuente C, Artal M. *L*-Menthol-based eutectic solvents: characterization and application in the removal of drugs from water. *J Mol Liq* 2022;352:118754.
- [19] Abdollahzadeh M, Khosravi M, Hajipour Khire Masjidi B, Samimi Behbahan A, Bagherzadeh A, Shahkar A, et al. Estimating the density of deep eutectic solvents applying supervised machine learning techniques. *Sci Rep* 2022;12(1):4954.
- [20] Dai Y, Witkamp GJ, Verpoorte R, Choi YH. Tailoring properties of natural deep eutectic solvents with water to facilitate their applications. *Food Chem* 2015; 187:14–9.
- [21] Gutiérrez A, Aparicio S, Atilhan M. Design of arginine-based therapeutic deep eutectic solvents as drug solubilization vehicles for active pharmaceutical ingredients. *Phys Chem Chem Phys* 2019;21(20):10621–34.
- [22] Gutiérrez A, Atilhan M, Aparicio S. A theoretical study on lidocaine solubility in deep eutectic solvents. *Phys Chem Chem Phys* 2018;20(43):27464–73.
- [23] Zainal-Abidin MH, Hayyan M, Ngoh GC, Wong WF, Looi CY. Emerging frontiers of deep eutectic solvents in drug discovery and drug delivery systems. *J Control Release* 2019;316:168–95.
- [24] Zhong X, Velez C, Acevedo O. Partial charges optimized by genetic algorithms for deep eutectic solvent simulations. *J Chem Theory Comput* 2021;17(5):3078–87.
- [25] Chaabene N, Ngo K, Turmine M, Vivier V. New hydrophobic deep eutectic solvent for electrochemical applications. *J Mol Liq* 2020;319:114198.
- [26] Hanada T, Goto M. Synergistic deep eutectic solvents for lithium extraction. *ACS Sustain Chem Eng* 2021;9(5):2152–60.
- [27] Yurramendi L, Hidalgo J, Siriwardana A. A sustainable process for the recovery of valuable metals from spent lithium ion batteries by deep eutectic solvents leaching. *Mater Proc* 2021;5(1):100.
- [28] Du K, Ang EH, Wu X, Liu Y. Progresses in sustainable recycling technology of spent lithium-ion batteries. *Energy Environ Mater* 2022;5(4):1012–36.
- [29] Neumann J, Petranikova M, Meeus M, Gamarra JD, Younesi R, Winter M, et al. Recycling of lithium-ion batteries—current state of the art, circular economy, and next generation recycling. *Adv Energy Mater* 2022;12(17):2102917.
- [30] Tang S, Zhang M, Guo M. A novel deep-eutectic solvent with strong coordination ability and low viscosity for efficient extraction of valuable metals

- from spent lithium-ion batteries. *ACS Sustain Chem Eng* 2022;10(2):975–85.
- [31] Zhang J, Wenzel M, Steup J, Schaper G, Hennersdorf F, Du H, et al. 4-Phosphoryl pyrazolones for highly selective lithium separation from alkali metal ions. *Chemistry* 2022;28(1):e202103640.
- [32] Chen Y, Wang Y, Bai Y, Duan Y, Zhang B, Liu C, et al. Significant improvement in dissolving lithium-ion battery cathodes using novel deep eutectic solvents at low temperature. *ACS Sustain Chem Eng* 2021;9(38):12940–8.
- [33] Wang K, Hu T, Shi P, Min Y, Wu J, Xu Q. Efficient recovery of value metals from spent lithium-ion batteries by combining deep eutectic solvents and coextraction. *ACS Sustain Chem Eng* 2022;10(3):1149–59.
- [34] Zante G, Boltoeva M. Review on hydrometallurgical recovery of metals with deep eutectic solvents. *Sustain Chem* 2020;1(3):238–55.
- [35] Chen L, Chao Y, Li X, Zhou G, Lu Q, Hua M, et al. Engineering a tandem leaching system for the highly selective recycling of valuable metals from spent Li-ion batteries. *Green Chem* 2021;23(5):2177–84.
- [36] Tran MK, Rodrigues MTF, Kato K, Babu G, Ajayan PM. Deep eutectic solvents for cathode recycling of Li-ion batteries. *Nat Energy* 2019;4(4):339–45.
- [37] Wang S, Zhang Z, Lu Z, Xu Z. A novel method for screening deep eutectic solvent to recycle the cathode of Li-ion batteries. *Green Chem* 2020;22(14):4473–82.
- [38] Aguilar N, Barros R, Antonio Tamayo-Ramos J, Martel S, Bol A, Atilhan M, et al. Carbon nanomaterials with thymol + menthol type V natural deep eutectic solvent: from surface properties to nano-Venturi effect through nanopores. *J Mol Liq* 2022;368:120637.
- [39] Tiecco M, Cappellini F, Nicoletti F, Del Giacco T, Germani R, Di Profio P. Role of the hydrogen bond donor component for a proper development of novel hydrophobic deep eutectic solvents. *J Mol Liq* 2019;281:423–30.
- [40] Zainal-Abidin MH, Hayyan M, Wong WF. Hydrophobic deep eutectic solvents: current progress and future directions. *J Ind Eng Chem* 2021;97:142–62.
- [41] Paul R, Mitra A, Paul S. Phase separation property of a hydrophobic deep eutectic solvent–water binary mixture: a molecular dynamics simulation study. *J Chem Phys* 2021;154(24):244504.
- [42] Makoš P, Šlupček E, Gębicki J. Extractive detoxification of feedstocks for the production of biofuels using new hydrophobic deep eutectic solvents—experimental and theoretical studies. *J Mol Liq* 2020;308:113101.
- [43] Farias FO, Pereira JFB, Coutinho JAP, Igarashi-Mafra L, Mafra MR. Understanding the role of the hydrogen bond donor of the deep eutectic solvents in the formation of the aqueous biphasic systems. *Fluid Phase Equilib* 2020;503:112319.
- [44] Vainikka P, Thallmair S, Souza PCT, Marrink SJ. Martini 3 coarse-grained model for type III deep eutectic solvents: thermodynamic, structural, and extraction properties. *ACS Sustain Chem Eng* 2021;9(51):17338–50.
- [45] Atilhan M, Aparicio S. Molecular dynamics simulations of mixed deep eutectic solvents and their interaction with nanomaterials. *J Mol Liq* 2019;283:147–54.
- [46] Alkhatib III, Bahamon D, Llovel F, Abu-Zahra MRM, Vega LF. Perspectives and guidelines on thermodynamic modelling of deep eutectic solvents. *J Mol Liq* 2020;298:112183.
- [47] Adeyemi I, Abu-Zahra MRM, AlNashef IM. Physicochemical properties of alkanolamine-choline chloride deep eutectic solvents: measurements, group contribution and artificial intelligence prediction techniques. *J Mol Liq* 2018;256:581–90.
- [48] Shahbaz K, Bagh FSG, Mjalli FS, AlNashef IM, Hashim MA. Prediction of refractive index and density of deep eutectic solvents using atomic contributions. *Fluid Phase Equilib* 2013;354:304–11.
- [49] Bagh FSG, Shahbaz K, Mjalli FS, AlNashef IM, Hashim MA. Electrical conductivity of ammonium and phosphonium based deep eutectic solvents: measurements and artificial intelligence-based prediction. *Fluid Phase Equilib* 2013;356:30–7.
- [50] Xu X, Range J, Gygli G, Pleiss J. Analysis of thermophysical properties of deep eutectic solvents by data integration. *J Chem Eng Data* 2020;65(3):1172–9.
- [51] Halder AK, Haghbakhsh R, Voroshylva IV, Duarte ARC, Cordeiro MNDS. Density of deep eutectic solvents: the path forward cheminformatics-driven reliable predictions for mixtures. *Molecules* 2021;26(19):5779.
- [52] Abbas UL, Qiao Q, Nguyen MT, Shi J, Shao Q. Molecular dynamics simulations of heterogeneous hydrogen bond environment in hydrophobic deep eutectic solvents. *AIChE J* 2022;68:e17382.
- [53] van Osch DJGP, Dietz CHJT, Warrag SEE, Kroon MC. The curious case of hydrophobic deep eutectic solvents: a story on the discovery, design, and applications. *ACS Sustain Chem Eng* 2020;8(29):10591–612.
- [54] van Osch DJGP, Dietz CHJT, van Spronsen J, Kroon MC, Gallucci F, van Sint AM, et al. A search for natural hydrophobic deep eutectic solvents based on natural components. *ACS Sustain Chem Eng* 2019;7(3):2933–42.
- [55] Robertson MJ, Tirado-Rives J, Jorgensen WL. Improved peptide and protein torsional energetics with the OPLS-AA force field. *J Chem Theory Comput* 2015;11(7):3499–509.
- [56] Dodda LS, Cabeza de Vaca I, Tirado-Rives J, Jorgensen WL. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res* 2017;45(W1):W331–6.
- [57] Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph* 1996;14(1):33–8.
- [58] Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *J Chem Phys* 1984;81(8):3684–90.
- [59] Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. *J Chem Phys* 2007;126(1):014101.
- [60] Darden T, York D, Pedersen L. Particle mesh Ewald: an $N \cdot \log(N)$ method for Ewald sums in large systems. *J Chem Phys* 1993;98(12):10089–92.
- [61] Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: a linear constraint solver for molecular simulations. *J Comput Chem* 1997;18(12):1463–72.
- [62] Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 2015;1–2:19–25.
- [63] Luzar A, Chandler D. Hydrogen-bond kinetics in liquid water. *Nature* 1996;379(6560):55–7.
- [64] Luzar A. Resolving the hydrogen bond dynamics conundrum. *J Chem Phys* 2000;113(23):10663–75.
- [65] Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project [presentation]. In: *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*; 2013 Sep 23–27; Prague, Czech Republic; 2013.
- [66] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;12:2825–30.
- [67] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016 Aug 13–17; FranciscoSan, CA, USA. New York City: Association for Computing Machinery; 2016. p. 785–94.
- [68] McKinney W. Data structures for statistical computing in Python. In: van der Walt S, Millman J, editors. *Proceedings of the 9th Python in Science Conference*; 2010 Jun 28–Jul 3; Austin, TX, USA; 2010. p. 56–61.
- [69] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature* 2020;585(7825):357–62.
- [70] Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng* 2007;9(3):90–5.
- [71] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020;17(3):261–72.
- [72] Zhang Y, Qiao Q, Abbas UL, Liu J, Zheng Y, Jones C, et al. Lignin derived hydrophobic deep eutectic solvents as sustainable extractants. *J Clean Prod* 2023;388:135808.