arXiv:2603.02761v1 [physics.chem-ph] 3 Mar 2026

混合机器学习用于增强液体中扩散系数的预测

Jens Wagner、Zeno Romero、Kerstin M¨unnemann、Sebastian Schmitt、Thomas Specht、Hans Hasse和Fabian Jirasek*

工程热力学实验室（*LTD*），*RPTU*凯泽斯劳滕，德国

电子邮箱：fabian.jirasek@rptu.de

电话： +49(0)631 - 205 4685

摘要

扩散系数是用于模拟液体中质量传递的关键热物理性质，但实验数据稀缺，因此可靠的预测方法不可或缺。在本研究中，我们将斯托克斯–爱因斯坦（SE）方程与机器学习（ML）相结合，提出了一种预测分子组分在纯液态溶剂中无限稀释时扩散系数的新方法。与以往的机器学习方法不同，由此得到的混合增强型斯托克斯-爱因斯坦（ESE）模型能够严格遵循物理规律，预测广泛二元混合物体系中扩散系数随温度的变化。ESE模型利用大量文献数据进行训练和验证，这些数据涵盖二元液体系统中无限稀释时的扩散系数。与此前最先进的模型SEGWE相比，ESE的预测准确度显著更高，同时只需将相关组分分子式编码为SMILES 字符串作为额外输入，而这类信息始终可以获取。这种简便性使ESE具有广泛的适用性，例如可用于过程设计与优化。ESE模型及其源代码均已完整公开，并可通过交互式网页界面 https://ml-prop.mv.rptu.de/ 直接访问。

引言

扩散是一种普遍存在的现象，支配着质量传递，并在众多自然和技术过程中发挥核心作用。因此，准确的扩散系数对于定量描述和模拟传递现象至关重要。然而，实验测定扩散系数往往要求高且耗时，而许多相关系统的现有数据仍然稀少。因此，可靠且适用范围广的扩散系数预测模型不可或缺。在此背景下，液体中的扩散尤为重要，因为它不仅难以预测，而且对许多应用至关重要，尤其是化学工程中的反应与分离过程。

一般而言，必须区分两种扩散：*互扩散* 指混合物中不同组分分子群体的运动，可通过菲克方法或麦克斯韦-斯特凡方法进行描述，每种方法均对应一组自身的扩散系数。相比之下，*自扩散* 涉及单个分子的布朗运动，其定义适用于纯组分和混合物。当所研究的溶质在特定溶剂中处于无限稀释状态时，菲克互扩散系数、麦克斯韦-斯特凡互扩散系数与该溶质自扩散系数之间的差异趋于消失。因此，“无限稀释扩散系数”这一术语概括描述了无限稀释状态下的互扩散和自扩散，本文也沿用这一用法。

无限稀释扩散系数数据与对扩散组分高度稀释的混合物进行建模直接相关。此外，无限稀释数据还可用于预测二元混合物和多组分混合物在有限浓度下的麦克斯韦-斯特凡互扩散系数，例如使用Vignes方法。[1,2]然而，即使对于二元混合物，即单一溶质存在于纯溶剂中的体系，无限稀释扩散系数的实验数据仍极为稀缺，[3] 因此，准确预测这些系数至关重要。

扩散系数　溶质在液体溶剂中无限稀释时的扩散系数 D_{ij}^{∞}，在 i 某些假设下 j 可以 r

用斯托克斯-爱因斯坦（SE）方程描述，该方程[4] 是通过考虑硬球（扩散粒子）在连续流体（溶剂）中的运动推导得出的：

$$D_{ij}^{\infty,\mathrm{SE}} = \frac{k_B T}{6\pi\eta_j r_i} \tag{1}$$

其中， k_B 为玻尔兹曼常数， T 为绝对温度， η_j 为溶剂 j 在温度 T下的动力黏度， r_i 为溶质 i的有效半径。后者可根据溶质的摩尔质量 M_i、比密度 ρ_i和阿伏伽德罗常数 N_A 估算，得出：

$$r_i = \sqrt[3]{\frac{3M_i}{4\pi\rho_i N_\mathrm{A}} f} \tag{2}$$

其中， f 为球形体积的经验校正因子（有时称为堆积率），通常取值范围为 $f = 0.6 \ \ldots \ 0.8$。在液体中的预测应用中，通常使用默认值 $f = 0.64$ 。[5]

尽管底层模型较为简单，SE 方程仍能捕捉扩散系数的许多关键特征。然而，在定量预测实际液体混合物中的扩散系数时，该方程存在明显不足。为弥补这些不足，文献中提出了许多 SE 模型的经验扩展。[5–9] 在近期对这些扩展的预测性能进行的一项比较中，Evans 等人[5,10] 提出的 Stokes-Einstein Gierer-Wirtz Estimation（SEGWE）被认为是文献中的半经验模型中最准确的。[3]

在 SEGWE 模型中，对原始 SE模型的扩展通过引入溶剂的摩尔质量和一个经验参数——所谓的有效密度 $\rho_\mathrm{eff} = 627$ kg m^{-3}——来考虑溶剂相对于溶质的尺寸。该参数是根据多种二元混合物的实验数据拟合得到的。尽管这些改进显著提升了 SEGWE模型相较于 SE 模型在许多体系中的预测性能，但仅依赖一个全局经验参数不足以

捕捉混合物中可能存在的各种相互作用，这一点 SEGWE 的开发者早已指出。[10] 这些结果与我们近期研究的发现一致，[11–16]研究表明，SEGWE模型存在若干缺陷，包括对极性相互作用的处理不足。

作为现有半经验模型的替代方案，人们提出了数据驱动的定量构效关系（QSPR）和机器学习（ML）方法，用于预测二元混合物中的无限稀释扩散系数[17–21] 和有限浓度下的扩散系数[18,22] 。尽管这些研究成功识别出影响扩散的溶质和溶剂性质，并将其用作输入，但它们往往只有将分析范围限定于特定溶剂类别时才能取得这样的成果。例如，Aniceto 等人[20] 针对极性和非极性溶剂分别开发了模型，尽管将组分归入这些类别往往存在歧义。另一些作者则只研究以水[17,21] 或烃类[18] 为溶剂的情况。此外，这些完全由数据驱动的方法缺少内置的物理约束，例如确保扩散系数随温度升高而增加；因此，尤其是在模型应用条件超出训练所用实验数据的适用范围时，可能会产生不符合物理规律的结果。

矩阵补全方法（MCM）是一类特殊的机器学习方法，可用于预测二元混合物的性质，且不依赖分子描述符。矩阵补全方法利用了这样一个事实：二元混合物的性质可以方便地存储在矩阵中，其中行和列分别对应混合物的各个组分，从而将未研究混合物性质的预测表述为矩阵补全问题。纯数据驱动的矩阵补全方法仅从现有混合物实验数据中学习；混合型矩阵补全方法则融入物理知识和物理模型。这两类方法均已用于预测各种混合物性质，[23–35] 包括二元混合物中的无限稀释扩散系数 D_{ij}^{∞}。[3,36]

将数据驱动的矩阵补全方法与 SEGWE模型相结合的混合模型，相较于所有现有半经验模型（包括单独使用 SEGWE模型的情况），提高了 D_{ij}^{∞} 的预测准确度，[3] 包括

SEGWE模型。然而，矩阵补全方法（包括基于 SEGWE模型的混合模型）的缺点是，它们仅适用于至少有一些实验 D_{ij}^{∞}数据可用的溶质和溶剂混合物，无论这些数据是与其他溶剂还是其他溶质组合获得的。此外，基本形式的矩阵补全方法仅适用于单一温度。

第二个局限可通过将矩阵补全方法扩展为张量补全方法（TCM）来解决，[37] 后者将温度视为额外的张量维度，从而捕捉温度依赖性。然而，这些 TCM 仍仅适用于至少有一些实验数据的溶质和溶剂。因此，对于包含缺乏相应实验训练数据的溶质或溶剂的体系，它们无法用于预测 D_{ij}^{∞} 。

总而言之，目前尚缺乏能够在不同温度下、针对广泛的溶质和溶剂（包括完全未经研究的对象），对无限稀释时的液相扩散系数进行物理一致且准确预测的模型， D_{ij}^{∞}。

在本研究中，我们提出了一种此类方法，将物理性的斯托克斯-爱因斯坦方程与神经网络（NN）相结合，构建出混合型增强型斯托克斯-爱因斯坦（ESE）模型。神经网络利用实验 D_{ij}^{∞} 数据进行训练，通过溶质和溶剂的简单分子描述符生成特定于混合物的缩放因子；这些描述符可直接从各组分的 SMILES[38] 字符串中自动提取。关键的是，神经网络受到约束，必须保留 SE 模型中蕴含的物理原理，从而确保混合型 ESE 模型在所有温度下都能给出符合物理规律的预测，同时发挥机器学习算法的灵活性和预测能力。我们将 ESE 与 SEGWE、MCM、[36] 以及 TCM[37] 模型进行了基准比较，结果表明，ESE 在广泛的混合物和温度范围内具有更优的预测性能。

方法

模型架构

图 1 示意概述了本研究开发的混合型增强型斯托克斯-爱因斯坦（ESE）模型。该模型将斯托克斯-爱因斯坦方程（SE）与神经网络相结合，以改进二元混合物中无限稀释扩散系数的预测 D_{ij}^{∞}。在此架构中，首先利用 SE 方程（1），根据溶质摩尔质量 M_i、温度 T以及该温度下的溶剂黏度 $\eta_j\,T$，得到初步预测$D^{\infty}{}_{ij}^{\mathrm{SE}}$。为避免需要额外的热物性输入，所有溶质的溶质密度均固定为 $\rho_i = 1050\ \mathrm{kg\,m^{-3}}$，经验校正因子则固定为$f = 0.64$ [5]。初步测试表明，只要这些数值处于物理上合理的范围内，其选择并不关键，对训练后模型结果的影响也很小。因此，我们不将其视为可调超参数。

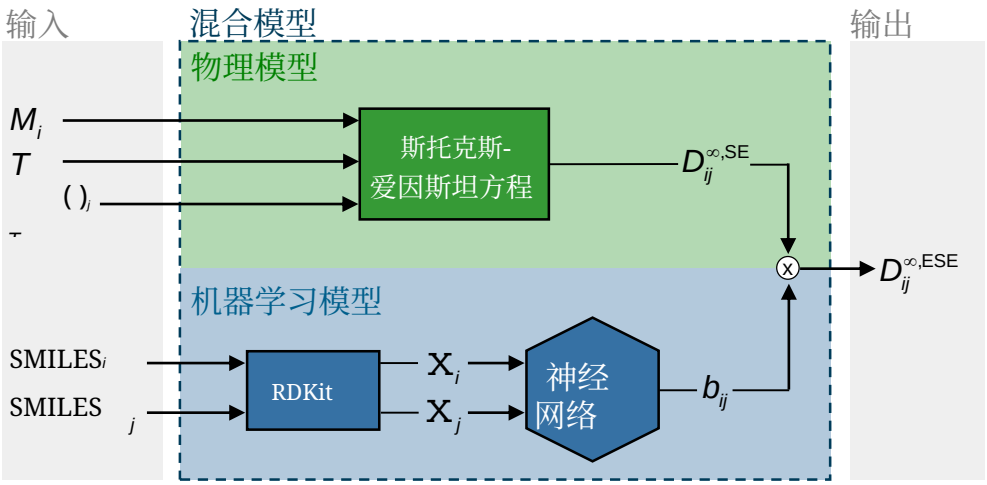


图 1：用于预测二元混合物中无限稀释液相扩散系数的混合型增强型斯托克斯-爱因斯坦（ESE）模型示意图 D_{ij}^{∞}。斯托克斯-爱因斯坦（SE）方程的预测结果 $D^{\infty}{}_{ij}^{\mathrm{SE}}$，通过学习得到的混合物特异性缩放因子 b_{ij} 进行校正；该因子由正值约束神经网络根据分子描述符向量 $\mathbf{X}_i$ 和 $\mathbf{X}_j$ 计算得出（参见表 1）。这些向量是使用 RDKit 根据溶质 i和溶剂 j 的 SMILES 字符串生成的。[39]

与此同时，溶质 i 和溶剂 j 的 SMILES 字符串通过开源工具包 RDkit[39] 处理，以自动生成各自的分子描述符向量，$\mathbf{X}_i$ 和 $\mathbf{X}_j$，具体如下文所述（参见表 1）。这些分子描述符向量是神经网络的唯一输入，神经网络会输出混合物特异性缩放因子 b_{ij}，该因子与温度无关。结合神经网络仅被允许为 b_{ij}生成正值这一限制，这确保了混合模型保留 SE 方程的物理背景，尤其是关于 D_{ij}^{∞}温度依赖性的物理背景。最终预测值 $D^{\infty,\mathrm{ESE}}_{ij}$ 通过将物理预测值 $D^{\infty,\mathrm{SE}}_{ij}$ 乘以相应的缩放因子 b_{ij}得到：

$$D_{ij}^{\infty,\mathrm{ESE}} = b_{ij} \cdot D_{ij}^{\infty,\mathrm{SE}} \tag{3}$$

根据超参数优化（见下文）确定的神经网络架构包含两个全连接层，分别具有 32 个和 16 个节点，采用线性整流单元（ReLU）激活函数，并在输出层使用 Softplus 激活函数，以确保 b_{ij}严格为正值。

分子描述符

表 1 概述了 **X**向量中包含的分子描述符，这些描述符编码了分子的结构和极性相关特征，同时还可使用 RDKit 从其 SMILES 字符串中明确且自动地推导出来。溶质和溶剂采用了相同的一组描述符。

我们有意保持描述符集精简，仅包含数量有限的属性，以便利用有限的可用实验数据对 D_{ij}^{∞}进行有效训练，并提高模型及其结果的可解释性。这些描述符的选择基于物理原理，并源自我们在先前研究中发现对扩散具有显著影响的分子性质。[13,15] 获取这些描述符所用的具体

用于获取这些描述符的 RDKit 函数列于补充信息的表 S.1 中。

表 1：用于分别描述溶质 i 和溶剂 j，并作为 ESE模型输入的向量$\mathbf{X}_i$ 和$\mathbf{X}j$ 中包含的分子描述符。

标签	分子描述符
M	摩尔质量，单位为 kg mol^{-1}
R	表示是否存在分子环结构的布尔变量
r_{Het}	杂原子数与非氢原子数之比
r_{Hal}	卤素原子数与非氢原子数之比
r_{Acc}	氢键受体数与非氢原子数之比
r_{Don}	氢键供体数与非氢原子数之比

图2给出了我们数据集中三个示例分子的分子结构以及相应的分子描述符向量$\mathbf{X}$ 。

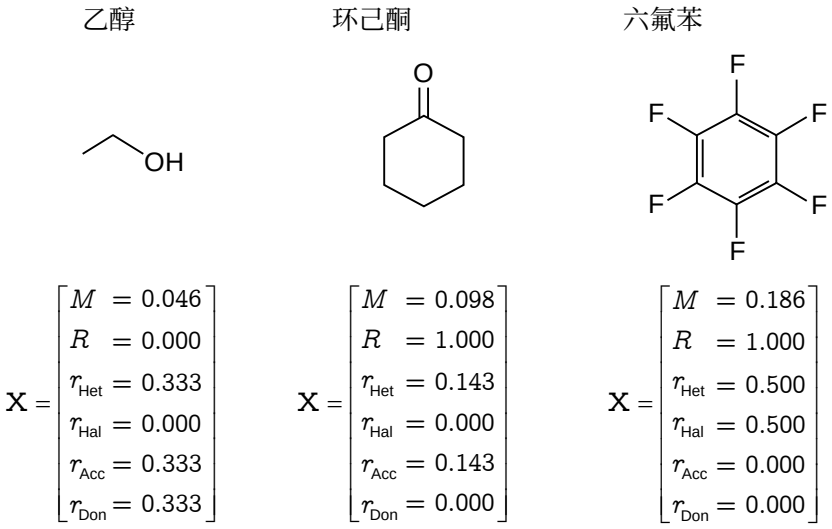


图2：以乙醇、环己酮和六氟苯为例展示的分子结构及相应的分子描述符向量$\mathbf{X}$ （参见表1）。

实验数据库

在本研究中，我们仅考虑非离子有机分子（包括水）的二元液体混合物，其摩尔质量不超过1000 g mol^{-1}，且所含原子均不重于

氯。

$D^{\infty,\mathrm{exp}}_{ij}$ 的实验数据主要来自Großmann等人汇编的数据库，[3] 该数据库随后由Romero等人扩展至其他温度。[36]此外，我们还用文献中的近期结果、[15,37,40,41]多特蒙德数据库（DDB）中新获得的可用数据，[42] 以及本实验室通过脉冲场梯度（PFG）核磁共振波谱[43] 测得的一些此前未发表的数据对数据库进行了补充。实验步骤及所测数据的详细信息见补充信息。

本工作所用的全部数据库包含 $N = 1011$ 个实验数据点，温度范围为273.2 K至363.0 K，涵盖538种二元混合物，其中包括209种不同溶质和42种不同溶剂。所有所考虑模型的应用均需要各溶剂在相应温度下的实验动力黏度，这些数据取自DDB。

训练与评估

采用基于按溶质划分数据的 K折交叉验证（CV）[44] 对ESE进行了训练和评估，其中 K 对应于数据库中不同溶质的数量。为此，在每一折中，与某一特定溶质相关的全部数据被留出并用作测试数据，其余数据则按数据点随机划分为训练数据（80%）和验证数据（20%）。该过程对所有溶质重复进行，直至获得所有数据点的预测结果。

在 ESE 的训练过程中，通过最小化均方相对误差 (MSRE) 来优化神经网络的权重，以预测 b_{ij} （参见方程 3）。MSRE 用作损失函数，其定义为所有纳入考虑的数据点中预测值与实验值之间平方相对误差 (SRE) 的均值：

$$\mathrm{SRE}_{ij} = \left(\frac{D_{ij}^{\infty,\mathrm{pred}} - D_{ij}^{\infty,\mathrm{exp}}}{D_{ij}^{\infty,\mathrm{exp}}} \right)^2 \tag{4}$$

权重更新采用 AdamW[45] 优化器。验证损失连续 25 轮未改善后，学习率将降低至原来的 0.1 倍。若验证损失连续 50 轮未改善，则触发早停，并选取验证损失最低的模型，最终在测试数据上进行评估。

超参数优化以验证损失为依据，采用网格搜索方法进行。优化后的超参数、敏感性分析以及相应的验证损失结果见补充信息。

除用于模型优化的平方相对误差和MSRE外，还使用绝对相对误差（ARE）评估了预测性能：

$$\mathrm{ARE}_{ij} = \left| \frac{D_{ij}^{\infty,\mathrm{pred}} - D_{ij}^{\infty,\mathrm{exp}}}{D_{ij}^{\infty,\mathrm{exp}}} \right| \tag{5}$$

以及其在所有数据点上的平均值，即测试数据上的平均绝对相对误差（MARE）。

虽然上述数据划分流程用于生成下一节所示结果，以评估所开发混合模型的预测能力，但我们也提供了ESE的“最终”版本。对于该最终模型，我们训练了一个由十个模型组成的集成[46]：将数据集按数据点随机划分十次，每次分为95%训练数据和5%验证数据，并利用每次划分从不同初始化出发训练十个单独模型，随后通过对各单独模型的预测进行简单平均，将其合并为一个集成。最终ESE模型可在Zenodo（https://doi.org/10.5281/zenodo.18787099）获取，也可通过我们的交互式网站MLPROP[47]（https://ml-prop.mv.rptu.de/）访问。

模型训练在 A40 GPU 上进行。所有模型以及训练和评估脚本均使用 Python 3.12.7 和 PyTorch 2.2.1 实现。[48] 每折的典型训练时间为 60 至 120 秒。

结果与讨论

下文将讨论 ESE模型在预测无限稀释条件下扩散系数方面的性能：我们在测试数据上对其进行评估，即针对未见过的溶质进行预测，并将结果与使用 SEGWE 得到的预测结果进行基准比较。[10] 此外，比较还包括使用固定值 $\rho_i = 1050$ kg m^{-3}和 $f = 0.64$的 SE模型预测结果；该模型对应于 ESE框架中使用的变体。关于 ESE 与 MCM[36] 及 TCM 在各自受限应用领域内预测 D_{ij}^{∞}的进一步比较，详见补充信息。

图3通过箱线图展示了将预测结果D_{ij}^{∞}与实验测试数据进行比较后计算得出的 ARE（上）和平方相对误差（下）。结果表明，ESE在两项误差分数上均显著优于SEGWE模型。具体而言，与SEGWE相比，ESE的 ARE 均值和中位数减半；与基准模型相比，ESE的平方相对误差均值和中位数降低至三分之一。这尤其表明，使用ESE时，预测结果非常不准确的测试数据点数量显著减少。

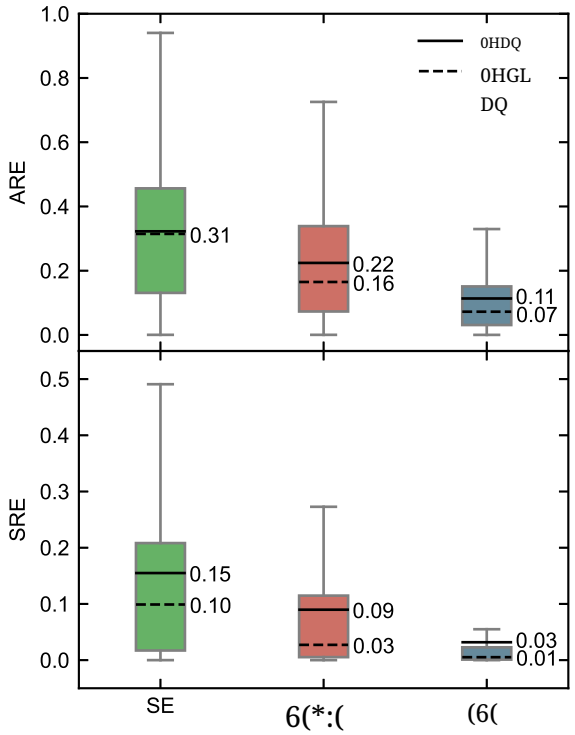


图3：SE、SEGWE 和 ESE 预测的无限稀释扩散系数的绝对相对误差（ARE，上图）和平方相对误差（SRE，下图）的箱线图 D_{ij}^{∞}。箱体宽度表示四分位距，须线长度为四分位距的 1.5 倍。为便于观察，图中未显示离群值。

图4通过直方图比较了ESE、SEGWE和SE对 D_{ij}^{∞} 的预测准确度，展示了能够以特定ARE（上）和平方相对误差（下）预测的测试数据点数量。此外，图中还显示了累积分数，即误差分数小于指定值的测试数据点所占比例。

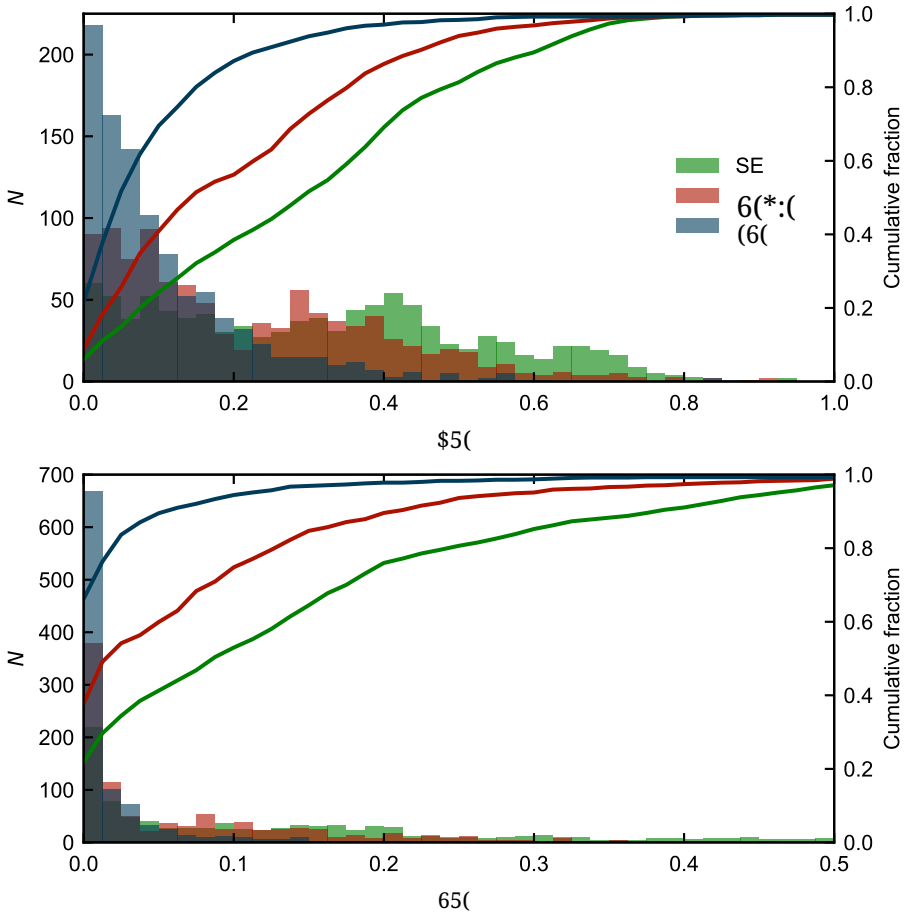


图4：直方图（柱形）和累积分数（曲线）显示了使用 SE、SEGWE 和 ESE 时，绝对相对误差（ARE，上图）或平方相对误差（SRE，下图）处于特定水平的 D_{ij}^{∞}预测结果数量。所有模型的 ARE 显示范围涵盖了 > 99 % 的预测结果。SRE 的显示范围涵盖了 SE 预测结果的 96.53%、SEGWE 预测结果的 98.51% 和 ESE 预测结果的 99.31%。

这些结果证实，与 SEGWE 和 SE 相比，ESE 的预测准确度有所提高。例如，使用 SEGWE 只能预测约 18% 的数据点，使其 ARE < 0.05 处于针对 D_{ij}^{∞} 报告的典型实验不确定度范围内；而使用 ESE 时，约 38% 的数据点可达到这一准确度。

为了更详细地分析 ESE 和基准模型在不同类型混合物中的预测性能，根据溶质和溶剂的氢键受体位点与供体位点之比，分别将其归类为非极性、极性非质子或极性质子，r_{Acc} 和 $r_{\mathrm{Don}} = 0$。因此，具有 $r_{\mathrm{Acc}} = 0$ 和 $r_{\mathrm{Don}} = 0$ 的组分被

归类为非极性，具有 $r_{\mathrm{Acc}} \neq 0$ 和 $r_{\mathrm{Don}} = 0$ 的组分被归类为极性非质子，而具有 $r_{\mathrm{Acc}} \neq 0$ 和 $r_{\mathrm{Don}} \neq 0$ 的组分被归类为极性质子。图 5 展示了由非极性、极性非质子或极性质子溶质 i 与溶剂 j组合形成的九种混合物类别中各项预测的 MARE 和 MSRE，以及我们数据集 $i-j$ 中各类别对应的数据点数量 N。

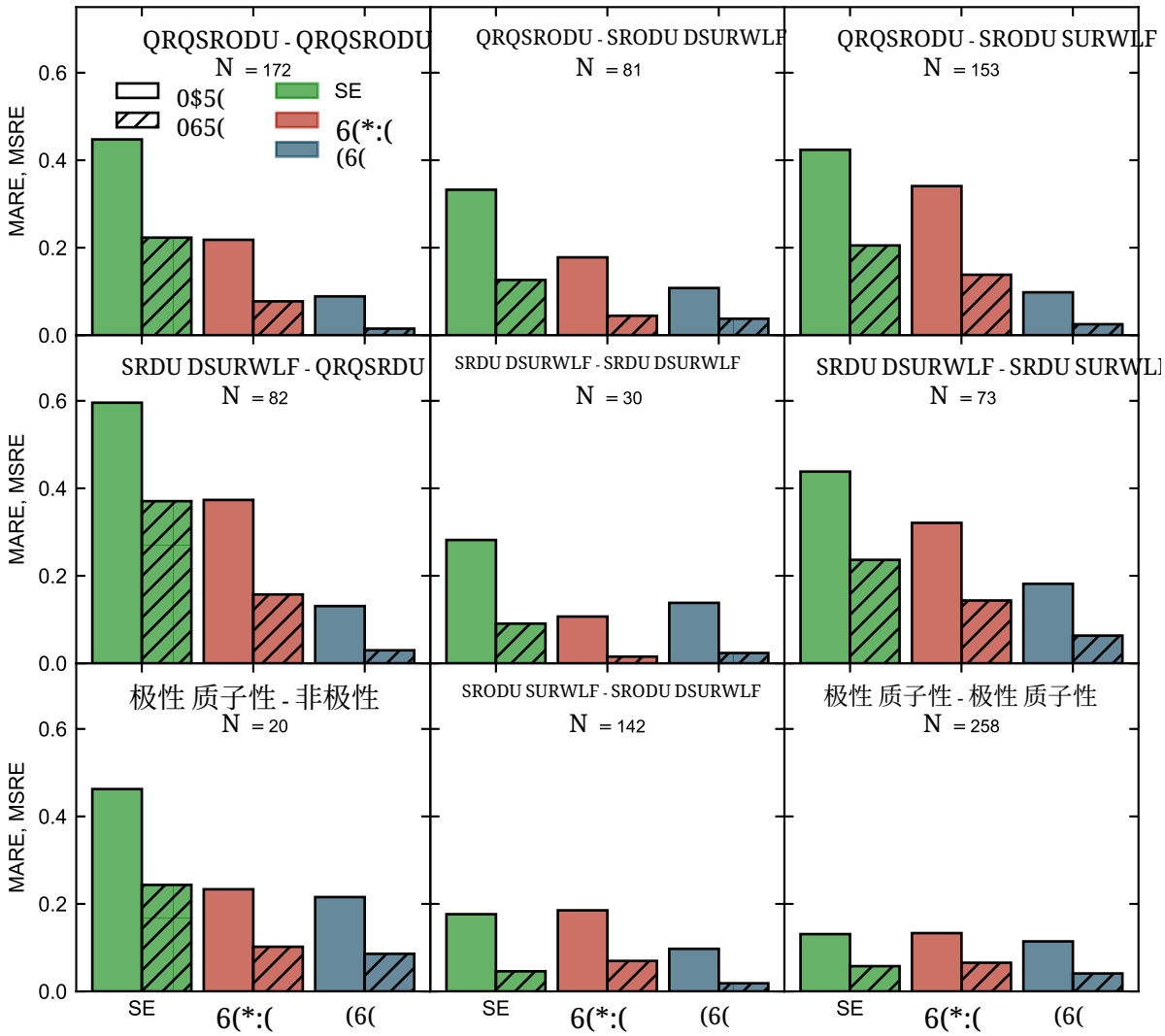


图 5：对于九种不同的溶质-溶剂类别，SE、SEGWE 和 ESE 对预测的 D_{ij}^{∞} 的平均绝对相对误差（MARE）和均方相对误差（MSRE），参见正文。 N 表示数据库中每种类别的测试数据点数量。

ESE 在几乎所有混合物类别中均表现最佳，即误差分数最低；与参考模型相比，其在

涉及非极性组分的类别中优势最为显著。只有在极性非质子–极性非质子混合物类别中，SEGWE 的误差分数略低于 ESE。值得注意的是，对于极性质子–极性非质子和极性质子–极性质子混合物类别，SE 的预测效果优于 SEGWE。

图6展示了四种代表性混合物中，SE、SEGWE 和 ESE 预测的 D_{ij}^{∞} 随温度变化的情况，并同时给出了数据库中相应的实验测试结果。为实现这些温度相关的预测，我们采用了 NIST WebBook[49] 中的相应溶剂黏度关联式（十二烷[50] 和水[51]），或文献中的关联式（十六烷[52] 和乙醇[53]）。

对于甲缩醛 - 十二烷、乙腈 - 乙醇和二氧化碳 - 水这些混合物，ESE 的预测结果与实验测试数据非常吻合，展现出 ESE 对含有未见过的溶质的混合物的 D_{ij}^{∞} 预测具有出色的准确性。相比之下，SEGWE 对这些混合物的无限稀释扩散系数始终预测偏低，而 SE 的低估程度则更为严重。这种系统性偏差表明，SE[5]中观察到的偏低固有偏差，并未被 SEGWE模型所采用的改进完全消除；我们将在补充信息中进一步展示这一点。

对于二氧戊环–十六烷混合物，ESE未能充分预测 D_{ij}^{∞}的温度依赖性，且对其有所低估。然而，在所研究的模型中，ESE仍预测出相对于温度最陡的斜率，从而最接近地反映了实验趋势。对最低温度处实验数据点的高估，可能是因为该温度接近十六烷的熔点，而熔点可能以模型未考虑的方式影响扩散。

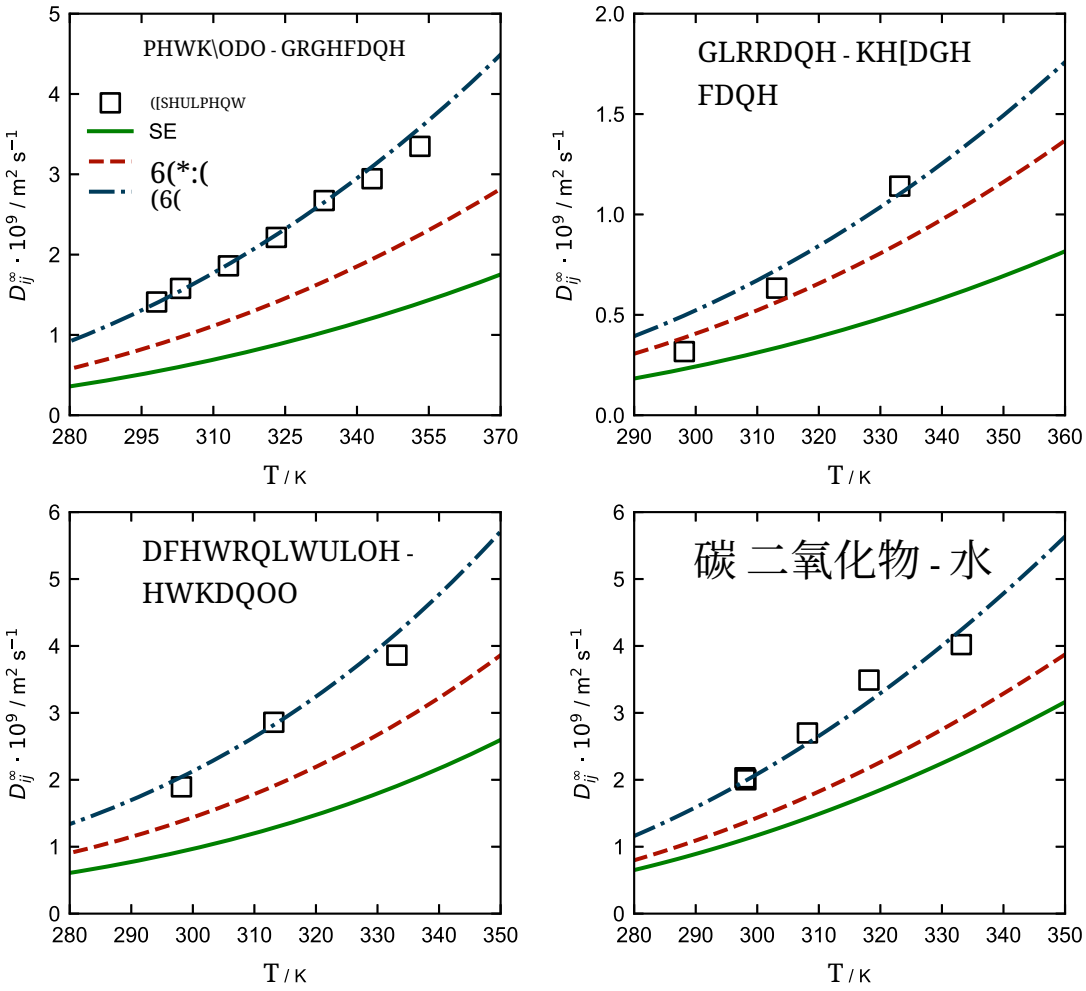


图6：无限稀释扩散系数 D_{ij}^{∞} 随温度 T变化的典型混合物示例。符号表示实验测试数据，曲线表示 SE、SEGWE 和 ESE 的预测结果。

所呈现的结果证明了ESE的预测能力及其在广泛二元混合物中可靠估算无限稀释扩散系数随温度变化的适用性。尽管这些结果是针对含先前未见过的溶质的混合物获得的，但该模型同样非常适合预测未见过的溶剂中以及含未见过的溶质和溶剂的混合物中的D_{ij}^{∞}，正如我们在补充信息中所展示的那样。如图6所示，跨越宽温度范围的预测结果证实了ESE所得结果的物理一致性。

结论

液体混合物中扩散系数的信息对于传质建模和模拟分离过程至关重要。然而，即使是最简单的类型，即单一溶质在纯溶剂中无限稀释时的扩散系数 D_{ij}^{∞}，实验数据也极为匮乏。此外，所有此前用于预测 D_{ij}^{∞} 的模型，包括物理模型、半经验模型和机器学习方法，通常都仅适用于特定类别的溶剂或仅含有实验数据可用组分的混合物，预测准确度较低，或无法给出符合物理规律的预测结果。

在本研究中，我们通过提出增强型斯托克斯–爱因斯坦（ESE）模型来应对这一挑战。这种混合方法将物理斯托克斯–爱因斯坦（SE）模型与一个神经网络相结合，该网络能够学习对SE公式缺陷的系统性修正。因此，ESE能够可靠地预测 D_{ij}^{∞}，同时保留原始SE模型的物理可解释性。除SE模型所需的输入（通常很容易获得）外，ESE仅需容易获取的溶质和溶剂分子结构信息，因此适用范围广泛。我们使用在模型开发和训练期间有意排除的溶质和溶剂的未见测试数据，验证了ESE的预测性能。在所有情况下，ESE在预测 D_{ij}^{∞} 方面均明显优于现有预测方法，包括SEGWE模型。[10]

当前版本的 ESE 使用有机分子和水的数据进行训练，其中组成原子的原子质量不超过氯，摩尔质量低于 1000 g mol^{-1}。虽然可以在此范围之外进行预测，但不建议这样做。不过，这些限制并非模型本身固有的，而是反映了适用实验数据的可获得性。要扩大 ESE 的适用范围，需要使用规模远大于目前可用数据集、更多样且质量更高的数据集。此外，ESE 目前尚未涵盖离子物种的扩散。沿这一方向扩展模型颇具吸引力，但需要重新审视底层物理模型。

除了直接预测 D_{ij}^{∞} 之外，ESE框架的通用性还带来了更多机会。具体而言，它可用于反向预测问题，即根据扩散数据推断未知溶质的性质，例如其摩尔质量 M_i。结合核磁共振（NMR）指纹技术，这种方法对于开发适用于规格不明混合物的合理建模策略尤其具有吸引力。[11–15]

利益冲突

作者声明不存在利益冲突。

数据可用性

最终训练完成的ESE模型可在Zenodo获取，网址为 `https://doi.org/10.5281/zenodo.18787099`。

致谢

我们衷心感谢卡尔蔡司基金会在“过程工程 4.0”和“Halocycles”项目中提供的资金支持，以及德国研究联合会在研究训练组“WERA”（项目编号503479768）、核心设施“LASE-MR”（项目编号537627671）和FJ的埃米·诺特研究组（项目编号528649696）框架下提供的资助。模型训练在RPTU的高性能计算机Elwetrisch上完成，资助编号为RPTU-MLVT。

可获取补充信息

分子描述符的生成；本研究测得的实验数据；超参数研究；与MCM和TCM的比较；补充结果；溶剂划分交叉验证的结果

参考文献

(1) Vignes, A. 二元溶液中的扩散。扩散系数随组成的变化。*Industrial & Engineering Chemistry Fundamentals* **1966**, *5*, 189–199.

(2) Kooijman, H. A.; Taylor, R. 多组分液体系统中扩散系数的估算。《工业与工程化学研究》**1991**,*30*, 1217–1222. (3) Großmann, O.; Bellaire, D.; Hayer, N.; Jirasek, F.; Hasse, H. 298 K 下液相无限稀释扩散系数数据库及用于预测这些系数的矩阵补全方法。*Digital Discovery* **2022**, *1*, 886–897.

¨(4) 爱因斯坦，A. 关于分子运动论所要求的悬浮于静止液体中的粒子热运动。*Annalen der Physik***1905**，*322*，549–560。

(5) Evans, R.; Deng, Z.; Rogerson, A. K.; McLachlan, A. S.; Richards, J. J.; Nilsson, M.; Morris, G. A. 扩散序核磁共振谱的定量解读：我们能否合理解释小分子扩散系数? *Angewandte Chemie* **2013**，*125*，3281–3284。

(6) Taylor, R.; Krishna, R. 多组分传质；John Wiley & Sons，1993。

(7) Wilke, C. R.; Chang, P. 稀溶液中扩散系数的关联式。《*AIChE*期刊》**1955**，*1*，264–270。

(8) Reddy, K.; Doraiswamy, L. 估算液体扩散率。*Industrial & EngineeringChemistry Fundamentals* **1967**, *6*, 77–79.

(9) Tyn, M. T.; Calus, W. F. 稀二元液体混合物中的扩散系数。*Journalof Chemical and Engineering Data* **1975**，*20*，106–109。(10) Evans, R.; Dal Poggetto, G.; Nilsson, M.; Morris, G. A. 改进小分子扩散系数的解释。*Analytical Chemistry* **2018**，*90*，3987–3994。(11) Specht, T.; M¨unnemann, K.; Hasse, H.; Jirasek, F. 从核磁共振谱中识别与定量结构基团的自动化方法

使用支持向量分类。化学信息与建模**2021**，*61*，143–155。

(12) Specht, T.; Arweiler, J.; St¨uber, J.; M¨unnemann, K.; Hasse, H.; Jirasek, F. 混合物的自动化核磁共振指纹识别。化学中的磁共振**2023**，*62*，286–297。

(13) Specht, T.; M¨unnemann, K.; Hasse, H.; Jirasek, F. 基于核磁共振波谱定义和量化伪组分的合理方法。《物理化学化学物理》**2023**，*25*，10288–10300。

(14) Specht, T.; Hasse, H.; Jirasek, F. 规格不明混合物的预测性热力学建模及其在概念性流体分离过程设计中的应用。工业与工程化学研究**2023**，*62*，10657–10667。

(15) Wagner, J.; Romero, Z.; M¨unnemann, K.; Specht, T.; Jirasek, F.; Hasse, H. 使用核磁共振指纹识别和基团贡献状态方程对规格不明混合物进行热力学建模。流体相平衡**2025**，*596*，114446。

(16) Wagner, J.; M¨unnemann, K.; Specht, T.; Hasse, H.; Jirasek, F. 用于未知混合物自动核磁共振指纹识别的深度集合模型。*Digital Discovery* **2026**，预先发表的文章。

(17) Khajeh, A.; Rasaei, M. R. 使用定量构效关系方法预测无限稀释条件下酸在水中的扩散系数。结构化学 **2011**，*23*，399–406。

(18) Abbasi, A.; Eslamloueyan, R. 使用基于定量构效关系方法的多层感知机和自适应神经模糊推理系统网络测定烃类混合物的二元扩散系数。化学计量学与智能实验室系统**2014**，*132*，39–51。

(19) Mariani, V.; Pulga, L.; Bianchi, G. M.; Cazzoli, G. 贝叶斯神经网络方法——

用于预测液相扩散系数。《国际传热与传质杂志》**2020**,*161*, 120309.

(20) Aniceto, J. P. S.; Zˆezere, B.; Silva, C. M. 极性和非极性流体中无限稀释条件下二元扩散系数的预测模型。《*Materials*》**2021**, *14*, 542.

(21) Aniceto, J. P.; Zˆezere, B.; Silva, C. M. 利用机器学习模型预测水相体系中的扩散系数。《分子液体杂志》**2024**, *405*, 125009.

(22) Beigzadeh, R.; Rahimi, M.; Shabanian, S. R. 开发用于预测液体中二元扩散系数的前馈神经网络多层模型。流体相平衡**2012**, *331*, 48–57.

(23) Jirasek, F.; Alves, R. A. S.; Damay, J.; Vandermeulen, R. A.; Bamler, R.; Bortz, M.; Mandt, S.; Kloft, M.; Hasse, H. 热力学中的机器学习：通过矩阵补全预测活度系数。《物理化学快报》**2020**,*11*, 981–985.

(24) Jirasek, F.; Bamler, R.; Mandt, S. 融合物理方法与数据驱动预测方法预测物理化学性质。《化学通讯》 **2020**, *56*, 12407–12410。

(25) Damay, J.; Jirasek, F.; Kloft, M.; Bortz, M.; Hasse, H. 通过矩阵补全预测变化温度下无限稀释活度系数。《工业与工程化学研究》 **2021**, *60*, 14564–14578。

(26) Jirasek, F.; Bamler, R.; Fellenz, S.; Bortz, M.; Kloft, M.; Mandt, S.; Hasse, H. 通过机器学习使混合物热力学模型具备预测能力：对相互作用的矩阵补全。《化学科学》 **2022**, *13*, 4854–4862。

(27) Hayer, N.; Jirasek, F.; Hasse, H. 通过矩阵补全预测亨利定律常数。《*AIChE*期刊》 **2022**, *68*, e17753.

(28) Jirasek, F.; Hasse, H. 将机器学习与物理知识相结合，用于流体混合物的热力学建模。《化学与生物分子工程年度评论》**2023**, *14*,31–51.

(29) Jirasek, F.; Hayer, N.; Abbas, R.; Schmid, B.; Hasse, H. 通过矩阵补全预测混合物基团贡献模型的参数。《物理化学化学物理》**2023**，*25*，1054–1062。

(30) Hayer, N.; Hasse, H.; Jirasek, F. 通过矩阵补全预测温度依赖的亨利定律常数。《物理化学杂志*B*》**2024**，*129*，409–416。

(31) Hoffmann, M.; Hayer, N.; Kohns, M.; Jirasek, F.; Hasse, H. 通过矩阵补全预测混合物中的成对相互作用。《物理化学化学物理》**2024**，*26*，19390–19397。

(32) Gond, D.; Sohns, J.-T.; Leitte, H.; Hasse, H.; Jirasek, F. 用于预测二元混合物性质的分层矩阵补全。《计算机与化学工程》**2025**，*199*，109122。

(33) Hayer, N.; Specht, T.; Arweiler, J.; Gond, D.; Hasse, H.; Jirasek, F. 使用量子化学描述符进行基于相似性的插补，以预测活度系数。《物理化学化学物理》**2025**, *27*, 4307–4315.

(34) Zenn, J.; Gond, D.; Jirasek, F.; Bamler, R. 在预测物理化学性质时平衡分子信息与经验数据。*Digital Discovery* **2025**,*4*, 683–693.

(35) Hayer, N.; Wendel, T.; Mandt, S.; Hasse, H.; Jirasek, F. 通过机器学习推进热力学基团贡献方法：UNIFAC 2.0。《化学工程杂志》**2025**, *504*, 158667.

(36) Romero, Z.; M¨unnemann, K.; Hasse, H.; Jirasek, F. 通过主动学习改进扩散系数预测。《物理化学杂志*B*》**2025**, *129*, 9219–9228.

(37) Romero, Z.; M¨unnemann, K.; Hasse, H.; Jirasek, F. 混合物的扩散系数张量补全预测。2026; https://arxiv.org/abs/2602.23142.

(38) Weininger, D. SMILES：一种化学语言和信息系统。1. 方法学与编码规则简介。《化学信息与计算机科学杂志》**1988**,*28*, 31–36.

(39) RDKit：开源化学信息学。 https://www.rdkit.org, 最后访问日期：2024年12月13日。

(40) Bellaire, D.; Großmann, O.; M¨unnemann, K.; Hasse, H. 二氧化碳和甲烷在水、乙醇、环己烷、甲苯、甲醇和丙酮中的无限稀释扩散系数：脉冲场梯度核磁共振与分子动力学模拟研究。《化学热力学杂志》**2022**,*166*, 106691.

(41) Mross, S.; Schmitt, S.; Stephan, S.; M¨unnemann, K.; Hasse, H. 烷烃与聚甲醛二甲醚混合物中的扩散系数。《工业与工程化学研究》**2024**, *63*, 1662–1669.

(42) 多特蒙德数据库。 https://www.ddbst.org, 2024.

(43) Bellaire, D.; Kiepfer, H.; M¨unnemann, K.; Hasse, H. 含环己烷、乙醇、丙酮和甲苯的二元及三元混合物自扩散系数的脉冲场梯度核磁共振与分子动力学模拟研究。《化学与工程数据杂志》**2020**,*65*, 793–803.

(44) Stone, M. 交叉验证选择与统计预测评估。《英国皇家统计学会杂志*B*辑：统计方法学》**1974**, *36*, 111–133。

(45) Loshchilov, I.; Hutter, F. 解耦权重衰减正则化。2017； https://arxiv.org/abs/1711.05101。

(46) Dietterich, T. G. 多分类器系统；施普林格柏林海德堡，2000；第1–15页。

(47) Hoffmann, M.; Specht, T.; Hayer, N.; Hasse, H.; Jirasek, F. MLPROP——一个用于机器学习热物性预测的交互式网页界面。化学工程技术 **2025**,

(48) Paszke, A. 等。PyTorch：一种命令式风格、高性能的深度学习库。2019；https://arxiv.org/abs/1912.01703。

(49) Linstrom, P. NIST化学网络手册，NIST标准参考数据库69。1997；http://webbook.nist.gov/chemistry/。

(50) Lemmon, E. W.; Huber, M. L. 正十二烷的热力学性质。能源与燃料 **2004**，*18*，960–967。

(51) Huber, M. L.; Perkins, R. A.; Laesecke, A.; Friend, D. G.; Sengers, J. V.; Assael, M. J.; Metaxa, I. N.; Vogel, E.; Mareˇs, R.; Miyagawa, K. 水的黏度新国际公式。物理与化学参考数据杂志 **2009**，*38*，101–125。

(52) Klein, T.; Yan, S.; Cui, J.; Magee, J. W.; Kroenlein, K.; Rausch, M. H.; Koller, T. M.; Fr¨oba, A. P. 通过表面光散射测定正己烷、正辛烷、正癸烷和正十六烷在最高573 K下的液体黏度和表面张力。《化学与工程数据杂志》 **2019**，*64*，4116–4131。

(53) Gon˛calves, F.; Trindade, A.; Costa, C.; Bernardo, J.; Johnson, I.; Fonseca, I.; Fer- reira, A. 乙醇的压力-体积-温度、黏度和表面张力：新测量数据及文献数据评估。《化学热力学杂志》**2010**，*42*，1039–1049。

补充信息：用于提升液体扩散系数预测的混合机器学习

Jens Wagner、Zeno Romero、Kerstin M¨unnemann、Sebastian Schmitt、Thomas Specht、Hans Hasse 和 Fabian Jirasek*

德国*RPTU*凯泽斯劳滕工程热力学实验室（*LTD*）

电子邮箱：fabian.jirasek@rptu.de

电话： +49(0)631 - 205 4685

分子描述符的生成

使用 RDKit，根据溶质和溶剂的 SMILES 字符串自动生成分子描述符。[1] 表 S.1 列出了用于获取分子信息以生成分子描述符的 RDKit 函数。非氢原子和卤素原子的数量是根据使用 `rdkit.Chem.rdchem.Atom.GetAtoms` 获取的分子原子列表确定的。

表 S.1：用于获取分子信息以生成分子描述符的 RDKit[1] 函数。

函数	分子信息
`rdkit.Chem.Descriptors.ExactMolWt`	摩尔质量
`rdkit.Chem.Lipinski.RingCount`	分子中的环数
`rdkit.Chem.Lipinski.NumHeteroatoms`	分子中的杂原子数
`rdkit.Chem.Lipinski.NumHAcceptors`	分子中的氢键受体数量
`rdkit.Chem.Lipinski.NumHDonors`	分子中的氢键供体数量

对于水，表 S.1 中的函数会错误地将氢键受体和供体的值均计算为零。为纠正这一问题，我们手动为水设置了 $r_{\mathrm{Acc}} = 0.500$ 和 $r_{\mathrm{Don}} = 0.500$。

本研究测得的实验数据

通过 ^{1}H PFG NMR 实验测量了无限稀释扩散系数 D_{ij}^{∞}，实验在$T = 298.15$ K 下进行，采用低溶质浓度的混合物，并将测量结果外推至溶质的无限稀释状态。实验流程与我们先前的工作中详细描述的完全相同。[2] 这些实验所用化学品的详情见表 S.2。使用超纯水系统（Omnia系列，stakpure）制备去离子纯化水。新实验的数值数据 D_{ij}^{∞} 见表 S.3。

表 S.2：S　所用化学品的供应商及纯度（由供应商注明）　.

化学品	化学式	供应商	纯度 %
乙腈	C_2H_3N	Sigma-Aldrich	$\geq$ 99.90
癸烷	$C_{10}H_{22}$	TCI	$\geq$ 99.00
二甘醇二甲醚	$C_6H_{14}O_3$	TCI	$\geq$ 99.00
二甲基亚砜	C_2H_6OS	默克	$\geq$ 99.90
2,6-二甲基吡啶	C_7H_9N	Sigma-Aldrich	$\geq$ 99.90
L(+)-抗坏血酸	$C_6H_8O_6$	Roth	$\geq$ 99.00
扁桃酸	$C_8H_8O_3$	Sigma-Aldrich	$\geq$ 99.90

表 S.3：使用脉冲场梯度核磁共振波谱测得的 D_{ij}^{∞} 在 $T = 298.15$ K 下的无限稀释扩散系数。[a]

溶质 i	溶剂 j	D_{ij}^{∞} 10^9 m^{2}s^{-1}
2,6-二甲基吡啶	水	0.71
L(+)-抗坏血酸	水	1.13
扁桃酸	水	0.76
乙腈	癸烷	3.93
乙腈	二甘醇二甲醚	1.87
二甲基亚砜	癸烷	1.92
二甲基亚砜	二甘醇二甲醚	1.45

[a]扩散系数的相对不确定度$\Delta D_{ij}^{\infty} = 0.02 \cdot 10^9$ m^2 s^{-1}。温度的不确定度 $\Delta T = 0.1$ K。

超参数研究

表 S.4 概述了本研究中调整的 ESE 超参数，即 AdamW 优化器的权重衰减 λ、初始学习率、批量大小，以及神经网络的层数和节点数；表中还列出了各变体在所有折上的平均验证损失（MSRE）表现。验证损失最低的模型 1 被确定为最佳超参数配置，并在整篇稿件中使用。不过，表 S.4 的结果表明，ESE 的性能对于超参数的变化具有稳健性。

表 S.4：采用不同超参数测试的 ESE 变体及相应的平均验证损失（MSRE）。 λ 是 AdamW 优化器的权重衰减。

模型编号	λ	初始学习率	批量大小	层	节点	平均验证损失
1	$1 \cdot 10^{-3}$	0.001	4	2	32 和 16	0.029
2	$1 \cdot 10^{-4}$	0.001	4	2	32 和 16	0.030
3	$1 \cdot 10^{-2}$	0.001	4	2	32 和 16	0.030
4	$1 \cdot 10^{-3}$	0.0001	4	2	32 和 16	0.029
5	$1 \cdot 10^{-3}$	0.01	4	2	32 和 16	0.029
6	$1 \cdot 10^{-3}$	0.001	2	2	32 和 16	0.030
7	$1 \cdot 10^{-3}$	0.001	8	2	32 和 16	0.031
8	$1 \cdot 10^{-3}$	0.001	4	1	32 和 16	0.032
9	$1 \cdot 10^{-3}$	0.001	4	3	32 和 16	0.031
10	$1 \cdot 10^{-3}$	0.001	4	2	24 和 12	0.032
11	$1 \cdot 10^{-3}$	0.001	4	2	40 和 20	0.031

ESE 与 MCM 和 TCM 的比较

为确保 ESE 与 MCM[3] 和 TCM[4] 方法的公平比较，我们使用数据集中在 298.15 K、313.15 K 和 333.15 K 下落入其适用域的 $N = 391$ 个数据点子集，重新训练并重新评估了 MCM 和 TCM 模型，并采用原始文献中所述的相同超参数设置和留一法测试程序。对于 MCM，在每个温度下分别训练独立模型，并将其结果合并以评估该方法。对于 ESE，我们使用了稿件中所述的相同超参数设置和程序。然而，我们通过留出与单个体系相关的全部数据（而非与单个溶质相关的全部数据）来评估模型，以匹配 MCM 和 TCM 所使用的评估协议。

图 S.1 展示了 MCM、TCM 和 ESE 预测结果的 ARE 和平方相对误差（SRE），表明 ESE 的预测性能优于其他模型。请注意，此比较中纳入的 $N = 391$ 数据点仅约占 ESE 更广适用域内可用 D_{ij}^{∞} 数据的 39%。

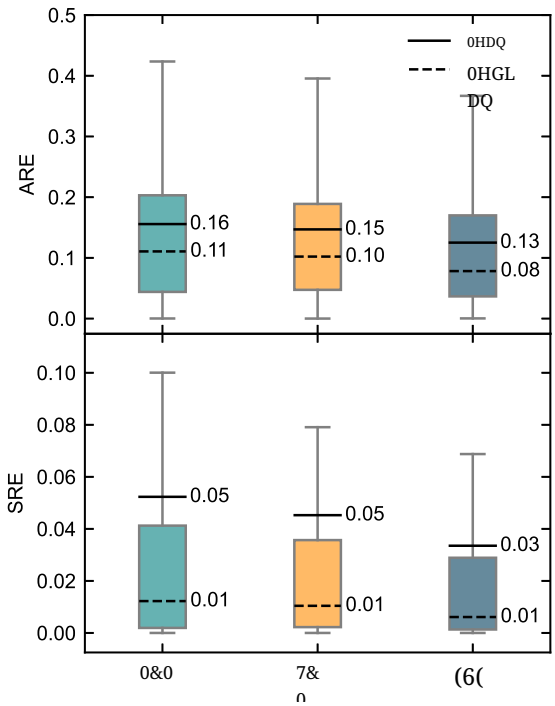


图 S.1：MCM、[3] TCM、[4] 和 ESE 在 MCM 和 TCM 适用范围内的数据子集上预测的无限稀释扩散系数的 ARE 和平方相对误差箱线图。箱体宽度表示四分位距，须线长度为四分位距的 1.5 倍。为便于观察，未显示离群值。

补充结果

图 S.2 展示了 SE、SEGWE 和 ESE 的测试数据所预测 D_{ij}^{∞} 与相应实验值的 parity 图。对于 SE，预测结果 $D^{\infty,\ \mathrm{SE}}_{ij}$ 系统性地低估了实验值，这与 Evans 等人[5] 的发现一致。对于 SEGWE，这种低估预测的程度有所减轻，但仍存在轻微偏差，尤其是在实验值较高时 $D^{\infty,\ \exp}_{ij}$。相比之下，ESE 的预测结果没有表现出相对于实验值的系统性偏差，且预测误差通常较小。

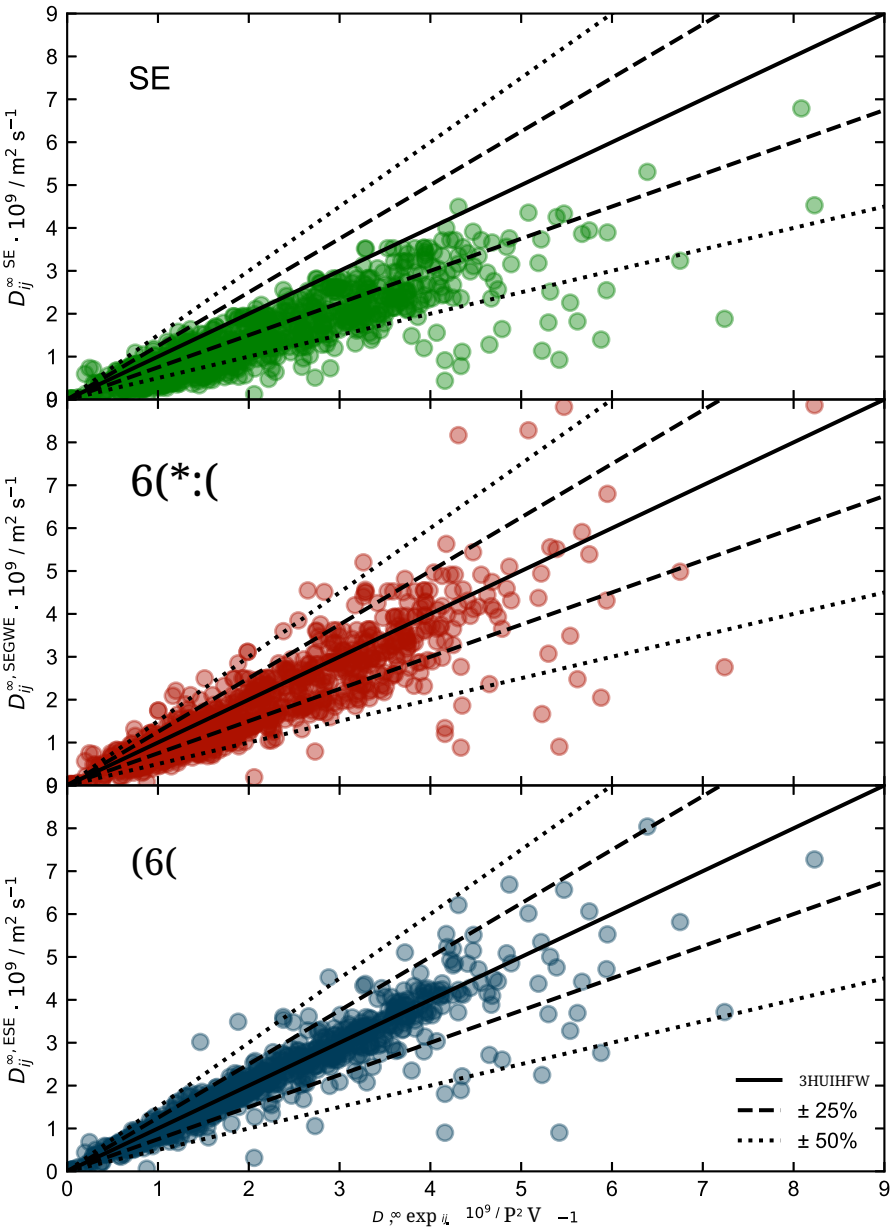


图 S.2：将采用 SE、SEGWE 和 ESE 得到的无限稀释扩散系数预测值（pred） D_{ij}^{∞}与相应的实验测试数据（exp）进行比较。

溶剂划分交叉验证的结果

图S.3将ESE在训练和评估中采用溶剂划分交叉验证（CV）所得结果的ARE和平方相对误差，与稿件中采用溶质划分交叉验证的结果进行了比较。当在含未见过的溶剂而非未见过的溶质的混合物上测试时，ESE的预测性能下降。结果表明，针对含未知溶剂的混合物预测 D_{ij}^{∞} 是一项更具挑战性的任务，这可以解释为我们的数据集中独特溶剂的数量少于溶质（42种而非209种），从而使模型更难学习不同溶剂对 D_{ij}^{∞} 的影响。然而，即便在这种情况下，ESE仍明显优于SEGWE模型，表明其适用于预测含未见组分混合物中的 D_{ij}^{∞}。

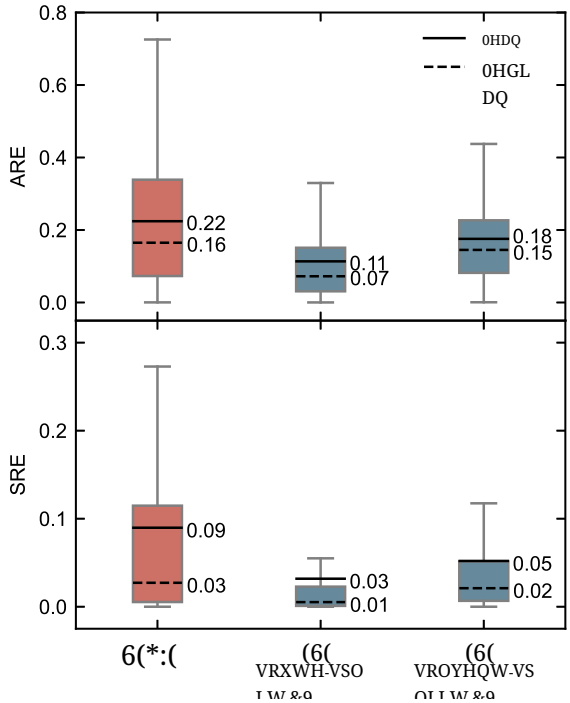


图 S.3：SEGWE 和 ESE 预测的无限稀释扩散系数的 ARE 和平方相对误差（SRE）箱线图。对于 ESE，将在未见过的溶质上获得的结果（溶质划分交叉验证，参见稿件）与在未见过的溶剂上获得的结果（溶剂划分交叉验证）进行比较。箱体宽度表示四分位距，须线长度为四分位距的 1.5 倍。为使图示清晰，未显示离群值。

参考文献

(1) RDKit：开源化学信息学。 https://www.rdkit.org，最后访问：2024年12月13日。 (2) Wagner, J.; Romero, Z.; M¨unnemann, K.; Specht, T.; Jirasek, F.; Hasse, H. 利用核磁共振指纹识别和基团贡献状态方程对规格不明混合物进行热力学建模。流体相平衡 **2025**, *596*, 114446. (3) Romero, Z.; M¨unnemann, K.; Hasse, H.; Jirasek, F. 通过主动学习改进扩散系数预测。《物理化学杂志*B*》 **2025**, *129*, 9219–9228. (4) Romero, Z.; M¨unnemann, K.; Hasse, H.; Jirasek, F. 利用张量补全预测混合物中的扩散系数。 2026; https://arxiv.org/abs/2602.23142. (5) Evans, R.; Deng, Z.; Rogerson, A. K.; McLachlan, A. S.; Richards, J. J.; Nilsson, M.; Morris, G. A. 扩散序核磁共振谱的定量解释：我们能否合理阐释小分子扩散系数？ *Angewandte Chemie* **2013**, *125*, 3281–3284.