arXiv:2301.02197v1 [cond-mat.dis-nn] 5 Jan 2023

用于完整声子预测的虚拟节点图神经网络

Ryotaro Okabe[1,2,†], **Abhijatmedhi Chotrattanapituk**[1,3,†], **Artittaya Boonkird**[1,4], **NinaAndrejevic**[5], **Xiang Fu**[3], **Tommi S. Jaakkola**[3], **QichenSong**[6], **ThanhNguyen**[1,4], **NathanDrucker**[1,7], **Sai Mu**[8], **Bolin Liao**[9], **YongqiangCheng**[10], 和 **Mingda Li**[1,4,*]

[1]麻省理工学院量子测量组，剑桥，马萨诸塞州，美国[2]麻省理工学院化学系，剑桥，马萨诸塞州，美国[3]麻省理工学院电气工程与计算机科学系，剑桥，马萨诸塞州，美国[4]麻省理工学院核科学与工程系，剑桥，马萨诸塞州，美国[5]阿贡国家实验室，莱蒙特，伊利诺伊州，美国[6]哈佛大学化学与化学生物学系，剑桥，马萨诸塞州，美国[7]哈佛大学工程与应用科学学院应用物理，剑桥，马萨诸塞州，美国[8]南卡罗来纳大学物理与天文学系，哥伦比亚，南卡罗来纳州，美国[9]加州大学圣巴巴拉分校材料系，圣巴巴拉，加利福尼亚州，美国[10]橡树岭国家实验室中子散射部光谱科化学光谱组，橡树岭，田纳西州，美国

†这些作者贡献相同*电子邮箱:
mingda@mit.edu

摘要

结构-性质关系在材料科学中发挥着核心作用。理解固态材料中的结构-性质关系，对于通过结构设计优化材料性质至关重要。过去几年，在晶体材料结构与性质的关联研究方面取得了显著进展，例如机器学习方法，尤其是作为晶体结构自然表征方式的图神经网络。然而，仍存在重大挑战，包括如何以复杂晶胞作为输入来预测性质，以及如何处理依赖材料且长度可变的输出。在此，我们提出虚拟节点图神经网络，以应对这些挑战。通过开发三种虚拟节点方法——向量、矩阵和动量依赖矩阵虚拟节点，我们仅以原子坐标为输入，即可直接预测 Γ声子谱和完整色散关系。我们在各种合金系统上验证了声子能带结构，并进一步构建了包含 Materials Project 中超过 146,000 种材料的 Γ声子数据库。我们的工作为复杂材料的声子谱和能带结构提供了快速、高质量的预测途径，并支持为能源应用设计具有优异声子性质的材料。图神经网络的虚拟节点增强方法也为以更高的灵活性设计其他功能性质带来了启示。

引言

结构-性质关系是材料科学中最根本的问题之一[16,21]。结构-性质关系无处不在，深刻影响着材料科学的几乎所有分支，例如结构材料[3], 能量采集与转换以及储能材料[5,17,19], 催化剂[37] 和聚合物[13], 以及量子材料[15]。然而，尽管结构-性质关系对材料设计至关重要，构建具有信息价值的结构-性质关系并非易事。一方面，随着晶胞尺寸增大，稳定结构的数量呈指数增长[22], ，而结构设计工作主要局限于晶胞相对较小的晶体固体。另一方面，由于实验或计算方面的复杂性，某些材料性质难以获取。

近年来，数据驱动和机器学习方法在材料科学中发挥着日益重要的作用，并显著推动了构建结构-性质关系的研究[6,24,38]。多孔材料[1,27], 、纳米合金[10,36], 和晶界[34] 等复杂结构正变得更易于处理，而从机械强度到量子有序等各种性质也能够以更高的置信度进行学习[9,29]。图神经网络（GNN）是一种尤其强大的方法[7]。通过将原子表示为图节点、将原子间键表示为图边，图神经网络

为分子和材料提供了自然的表示方式。对于晶体固体，晶体学对称性可进一步提升图神经网络的性能，目前已有一些对称性增强图神经网络被提出[8,30,35]。不过，仍然存在一些基本挑战。首先，许多材料性质并不能自然地表示为实空间中各个原子的加权聚合，例如倒易空间和能量空间性质。其次，输出性质的长度通常是固定的，例如将热容[4] 表示为单个标量。相比之下，许多材料性质具有各自独特的维度，例如电子能带和声子能带的数量[2], 、具有光学响应的频率范围，以及磁结构的传播矢量等特征。

在这项工作中，我们提出了虚拟节点图神经网络（VGNN），作为一种可普遍适用的图神经网络增强方法。与侧重于减少输入数据量的对称性增强图神经网络不同，VGNN 着重处理维度可变甚至任意的输出性质。我们研究了材料的声子谱和色散关系，因为声子能带难以计算或测量，且计算成本高、实验资源有限。以声子谱为例，我们提出了 VGNN 的三个版本：向量虚拟节点（VVN）、矩阵虚拟节点（MVN）和动量依赖矩阵虚拟节点（k-MVN）。这三种 VGNN 模型都以原子结构为输入，无需预先了解原子间作用力。VVN 是最简单的 VGNN，它以含有m 个原子的晶体结构为输入，并输出 $3m$ 支 Γ 声子能量。MVN 是一种更复杂的 VGNN，对于复杂材料具有更高的准确度，但计算成本略高。最后，k-MVN 是一种能够预测布里渊区内任意k 点处完整声子能带结构的 VGNN。为此，晶体图中包含“虚拟动力学矩阵”，即与声子动力学矩阵相似的矩阵结构[14]。我们不对每种材料直接进行从头算计算，而是利用由其他所有材料组成的训练数据，通过神经网络优化过程学习所有矩阵元素。我们的工作提供了一种高效方法，可直接从复杂材料的原子结构计算区中心声子能量和完整声子能带结构，并能在更大的结构设计空间中优化声子性质。这些预测方法使我们能够获取材料的相关信息作为副产品，例如群速度、热容和态密度。同时，虚拟节点结构也为未来灵活的图神经网络设计提供了启示：可以将中间关键量（例如动力学矩阵）作为关键学习参数，而不必将目标性质（例如声子能带结构）作为输出。

结果

面向图神经网络的虚拟节点增强方法。 图 1 概述了 VGNN 方法，展示了它作为一种通用方法增强图神经网络的方式。对于每个晶胞包含 m 个原子的晶体（图 1a），典型的图神经网络模型会将晶体转换为晶体图，其中每个图节点代表一个原子，每条图边代表原子间键合，如图 1b 所示。每个原子节点对应的节点特征（图 1b 中的灰色数组）会根据邻域节点以及连接节点的边（图 1b 中的灰色箭头）进行更新。经过多层迭代图卷积后，可得到 m 个最终层节点特征，分别表示 m 个原子的特征。通过将最终层节点特征聚合为一个固定大小的输出，即可得到最终的图输出。

图 1c、d 介绍了 VGNN 的总体思路：赋予图神经网络更强的预测灵活性。在传统的真实节点图神经网络基础上，向晶体中加入虚拟原子（图 1c 中的黄色节点），这些原子在对应的图神经网络中成为虚拟节点（图 1d 中的黄色节点）。如图 1d 所示，与真实原子节点之间的双向消息传递（灰色双箭头线）类似，虚拟节点之间的消息传递（黄色双箭头线）也是双向的。另一方面，为保留传统图神经网络的结构，从真实节点到虚拟节点的消息传递（灰黄渐变单箭头线）是单向的。由于虚拟节点的选择具有灵活性，VGNN 能够灵活预测具有任意长度和处于任意空间中的材料相关输出。我们将介绍三种用于声子预测的 VGNN 方法，其预测能力和复杂度逐步提高。

用于 Γ声子预测的向量虚拟节点。 如图 1所示，VGNN 能够根据输入信息灵活调整输出维度。我们首先介绍向量虚拟节点（VVN）方法，这是在输入每个晶胞含有m 个原子的晶体时，获取 $3m$ 条声子支的最简单方法。（详情请参见 方法 部分。）图2 展示了 VVN 方法如何预测 Γ声子谱。由于虚拟节点不会向真实节点传递信息，因此虚拟节点的位置选择更加灵活。在不失一般性的前提下，我们将虚拟节点等间距地排列在晶胞的对角线上。由虚拟节点和真实节点构建晶体图（图 2a）。在每个卷积层更新节点特征后，特征向量经过一个线性层，使虚拟节点特征 V_i,$i \in [1,3m]$ 转换为 $3m$ 个标量，用以表示预测的 Γ声子能量。本文中的图神经网络部分均通过能够识别晶体学对称性的欧几里得神经网络实现。数据准备、神经网络架构和优化方法详见补充信息 I–III。

使用 VVN 进行 Γ声子预测的主要结果如图2b 所示。三行光谱对比图

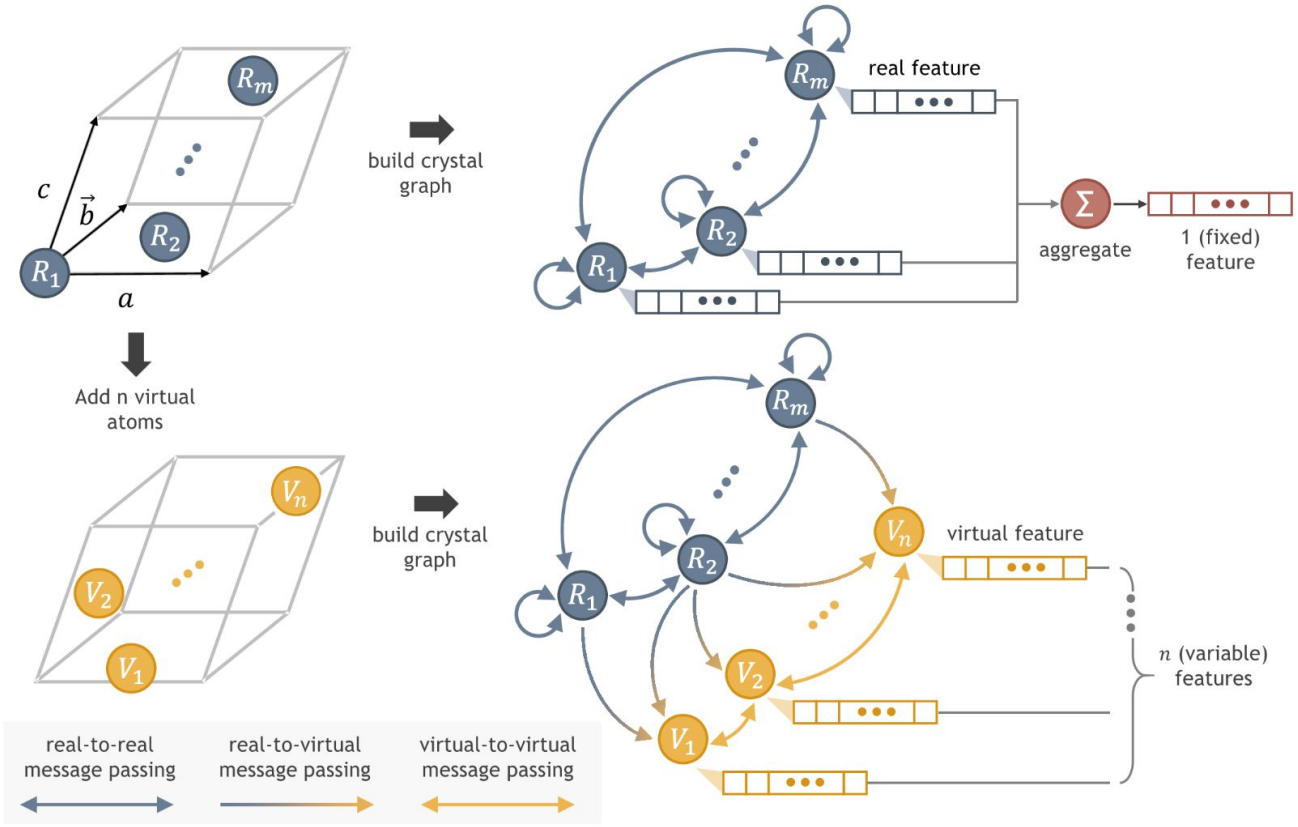

图1. 虚拟节点图神经网络（VGNN）概览。a。每个晶胞含有m个原子的晶体材料的**原子结构**。b。图神经网络将原子结构转换为晶体图。经过多层图卷积（为简洁起见，此处省略）后，最终的节点特征被聚合为单个固定大小的输出特征。**c**。在晶体结构中添加由n 个虚拟原子构成的灵活结构。**d**。将真实节点和虚拟节点共同构成晶体图后，虚拟节点的灵活性使输出选择不必局限于真实节点聚合，还可以具有可变长度并处于不同空间中。

分别是从测试集中每个误差三分位组随机选取的样本（从上到下各行依次对应从高到低的性能三分位组）。前四列取自与训练集相同的数据库，该数据库中的数据来自高质量的密度泛函微扰理论（DFPT）计算[25]，；第五列则包含来自冻结声子数据库的其他测试样例，其晶胞大得多[31]。值得一提的是，尽管对于非常复杂的材料，VNN 预测的声子频率往往偏高（例如第三行第四列的 $Ba_{12}I_{36}Y_4$），但整个布里渊区的声子态密度仍然可以大致相当。然而，随着输入材料变得更加复杂，预测损失会增大，分布也会变得更宽（图2c）。从预测声子频率与真实声子频率的相关性图（图2d）可以看出，大多数数据点都位于对角线上，表明当每个晶胞中的原子数为$m \leq 24$ 时，VNN 预测与 DFPT 计算所得真实值之间具有良好的预测一致性（蓝点）。对于复杂材料，相关性表现可能会下降（橙点）。更多测试结果见补充信息 IV。

用于 Γ 声子预测的矩阵虚拟节点，性能更优。 本节介绍另一种虚拟节点方法——矩阵虚拟节点（MVN）。与 VVN 相比，MVN 方法在 Γ 声子预测方面表现更佳，尤其适用于复杂材料，但计算成本略高。此外，MVN 的结构为下一节将讨论的完整声子能带结构奠定了基础。在 MVN 中，对于每个晶胞含有m 个原子的材料，会生成m 个虚拟晶体副本，每个副本包含m 个虚拟节点，其晶体结构与真实晶体相同（图3a）。由此共得到m^2 个虚拟节点 V_{ij}、i、$j \in [1,m]$，其节点连接关系更为复杂。（更多详情请参见方法 。）

采用这种图构建方案，经过神经网络训练后，虚拟节点V_{ij} 将捕捉R_i与R_j之间连接的本质。因此，在每个卷积层中进行消息传递后，每个虚拟节点的特征

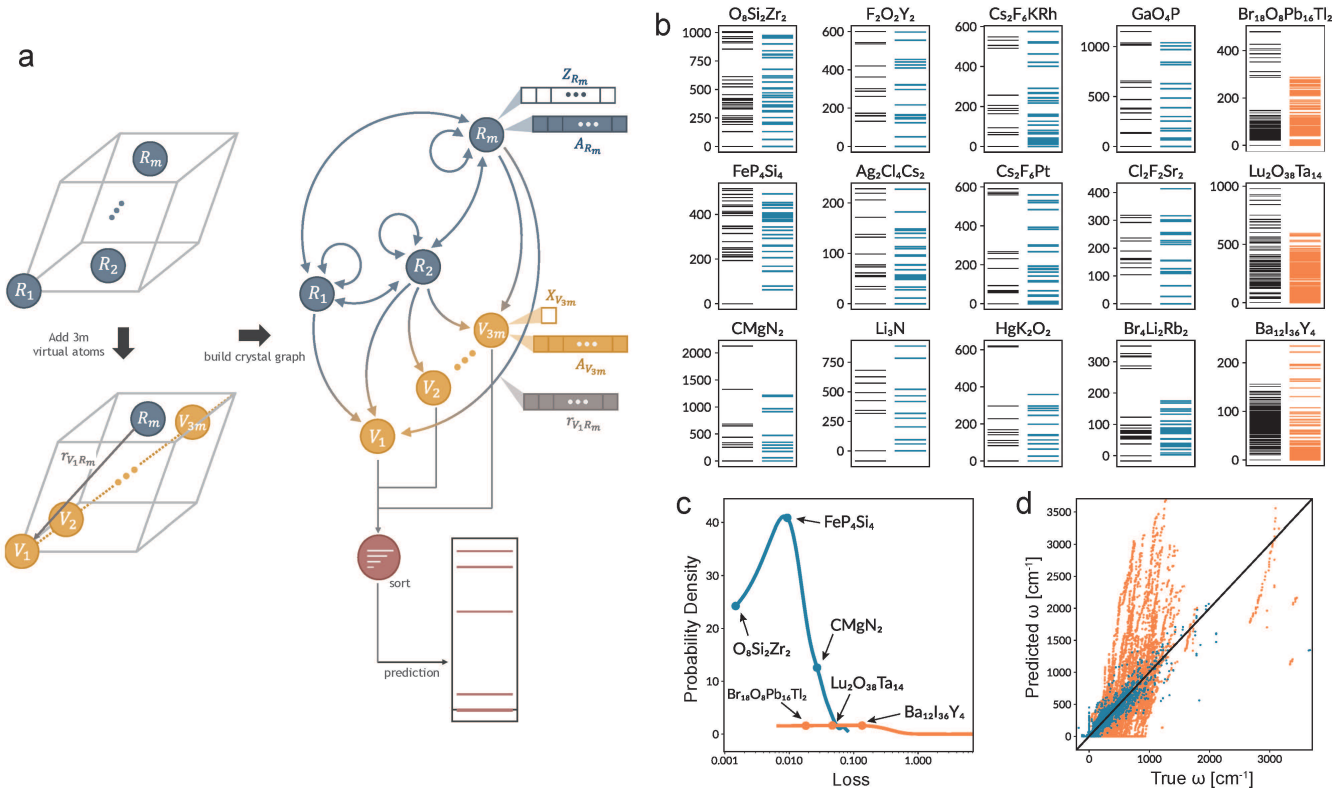

图2. 用向量虚拟节点（VVN）方法预测 Γ-点声子。a. VVN 模型构建与预测示意图。对于每个晶胞含有 m 个原子的材料，沿晶胞的对角向量 $\vec{v} = \vec{a} + b + \vec{c}$ 增设 3m 个虚拟节点。构建图神经网络模型时，我们嵌入了晶体的组成部分。例如，将 m^{th} 个真实原子（A_{R_m}）和 $3m^{th}$ 个虚拟原子（$A_{V_{3m}}$）的原子序数嵌入为各节点的属性。将 m^{th} 个真实原子（Z_{R_m}）的原子质量设为该节点的初始特征。节点 V_1 相对于 R_m 的位置为 $\vec{r}_{V_1 R_m}$，该位置用于嵌入两个节点之间的边属性。模型通过对虚拟节点的标量输出特征排序来预测 Γ-声子谱。b. 测试集中各误差三分位组的光谱预测样本与真实值（黑色）的对比：来自与训练集相同数据库的测试样本（蓝色），以及来自包含复杂材料的不同数据库的测试样本（橙色）。c-d. 分别通过损失函数分布以及真实值与预测平均声子频率之间的相关性图评估测试准确度。测试集（蓝色）的分布图在低损失区间集中度较高，且相关性图中的数据点沿对角线分布，这表明模型至少能够对每晶胞原子数 $m \leq 24$相对较少的简单材料实现高质量的声子预测。对于复杂材料（橙色），损失增大，性能下降。

会进一步转换为一个三乘三矩阵。每个V_{ij} 组合起来，构成超矩阵(i, j) 块 $\tilde{D}$，其形状为$(3m,3m)$。鉴于该矩阵与方程（2）中表示的动力学矩阵具有结构相似性，且 $\vec{k} = 0$，我们通过求解矩阵D的 $3m$ 个特征值，来预测 Γ点声子能量。值得一提的是，尽管该矩阵与动力学矩阵具有相似特征，但其矩阵元素是通过神经网络训练学习得到的，不一定是实际动力学矩阵的矩阵元素。直观地说，图神经网络的边不一定反映真实的化学键，更像是一种原子邻接关系。

使用 MVN 预测的声子汇总于图 3b 中，其结构与图2b 相同，展示了高质量 DFPT 数据库（蓝色）和复杂材料数据库（橙色）的误差三分位组。对于简单材料，MVN 的表现与 VVN 相当（图2c 和图 3c 中的蓝色曲线），但对复杂材料的表现有显著提升。与 VVN 相比，MVN 的预测损失分布更多集中于较低损失区间（图2c 和图 3c 中的橙色曲线），且相关性图中的平均声子频率与真实值更加吻合（图 2d 和图 3d 中的橙色点）。更多结果和相关性图见补充信息 IV。

用于预测完整声子能带结构的动量依赖矩阵虚拟节点。MVN 的结构启发我们

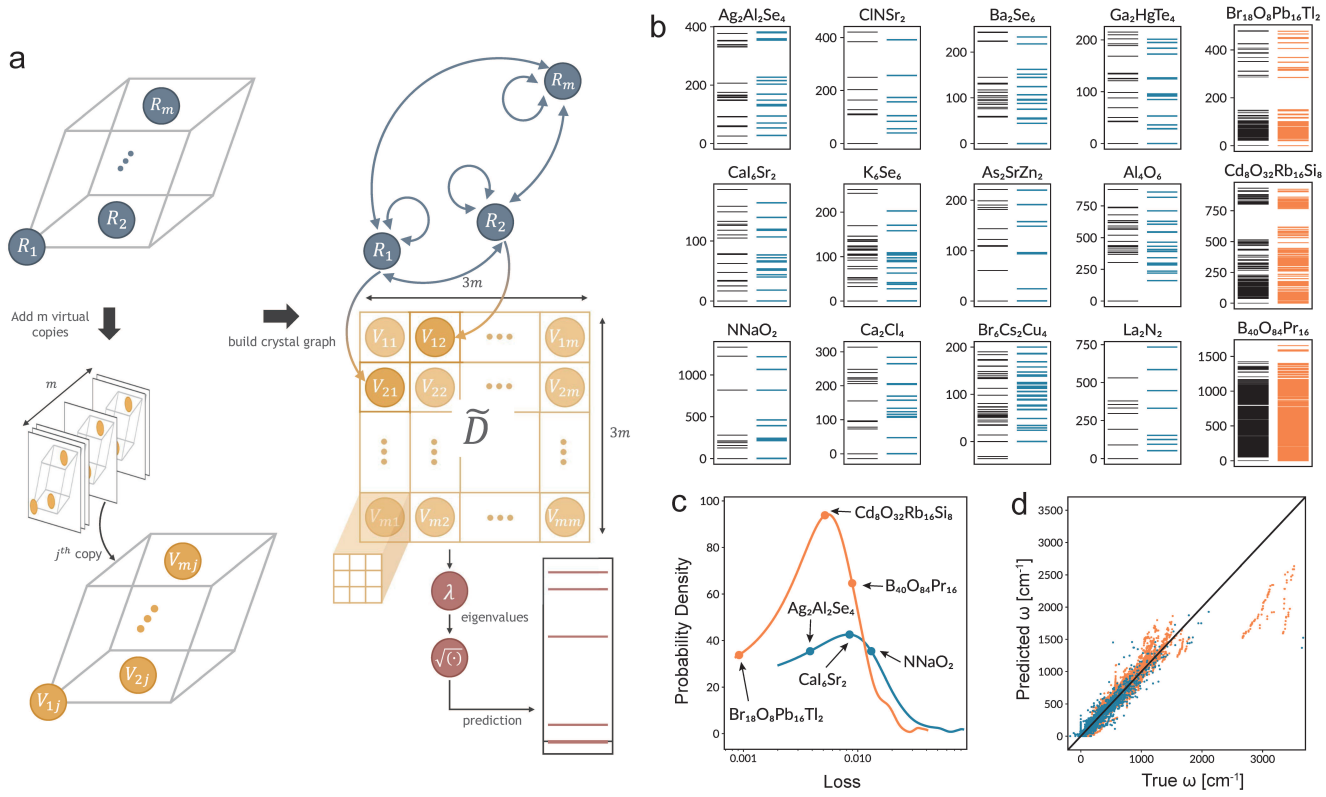

图 3. 用于预测 Γ 点声子的矩阵虚拟节点（MVN）方法。a. 将 m^2 个虚拟节点扩充为m 组虚拟晶体（左），以及消息传递方案和虚拟节点特征的后处理（右）。图例与图 2 相同。与 VVN 不同，后者的每个节点 V_i 都是标量，而这里的每个节点 V_{ij} 都是一个 3×3 矩阵。MVN 中的声子谱通过求解特征值问题获得，而不像 VVN 那样直接输出。**b.** 每个误差三分位组中选取的测试样例。与训练集来自同一数据集的测试样例，以及包含复杂材料的额外测试样例，分别以蓝色和橙色表示。 **c.** 预测损失分布与若干材料样例的比较。 **d.** 以图的 $y=x$ 为参照，绘制平均声子频率的相关性图。对于复杂材料（橙色），MVN 的表现优于 VVN，这一点从损失分布和平均声子频率两方面均可看出。

进一步构建完整的动量依赖虚拟动力学矩阵，并将晶胞平移纳入考量，这种方法称为动量依赖矩阵虚拟节点（k-MVN）。我们按照公式（2）构建虚拟动力学矩阵 $\vec{}$。与 MVN 专注于 Γ 点声子、令 $\mathbf{k}=0$ 不同，在 k-MVN 中，我们定义虚拟动力学矩阵时纳入相位因子 $e^{i\vec{k}\cdot\vec{T}}$，其中 $\vec{T}$ 表示相邻晶胞原点相对于所选参考晶胞 $\vec{T}_0$ 的相对晶胞平移（图 4a）。若纳入 t 个相邻晶胞，每个晶胞的平移为 $\vec{T}_h$，$h\in[0,t-1]$（包括参考晶胞），则会生成 t 组 MVN 型虚拟节点矩阵，共包含 tm^2 个虚拟节点 V hi j·h $\in[0,t-1]$，其中 i、j $\in[1,m]$，用于 k-MVN。为获得声子能带结构，需要将给定 $\vec{T}_h$ 下的每组虚拟节点乘以相位因子 $e^{i\vec{k}\cdot\vec{T}_h}$，并在公式 3 中对每个 $\vec{T}_h$ 下的所有虚拟节点求和。得益于截断半径内的图连通性（见方法），只要能维持晶体图连通性，所需的 t 就很小。实际上，t 取决于材料；对于许多材料，$t=27$（最近邻晶胞）就已足够，而且在所有情况下都不必超过 $t=125$（次近邻晶胞）。直观而言，这种超胞方法类似于采用冻结声子的从头算能带结构计算。为便于训练，训练数据中纳入了选定高对称点处的声子，而不必使用整个布里渊区内的完整声子能量。这样既能显著简化训练过程，又能保持准确性。更多细节请参见方法和补充信息 V。

图4b 展示了声子能带结构的预测结果。这里从同一数据集中选取了 12 种材料用于训练（蓝色），并从用于复杂材料的附加数据集中选取材料（橙色）。尽管一般的声子能带结构十分复杂，*k*-MVN 模型仍能预测声子能带的位置和形状，例如不同光学支之间的带隙。声学声子的色散关系也在 Γ点附近得到了良好生成

左侧三列中，即使我们没有强制要求声学 Γ声子必须无带隙且具有零能量，这一条件称为声学和规则[28]。这或许能为未来预测晶体稳定性提供帮助。尽管高频声子能带的预测性能可能会下降，但大多数预测的声子都与参考结果一致，包括每个晶胞含有超过 40 个原子的复杂材料。

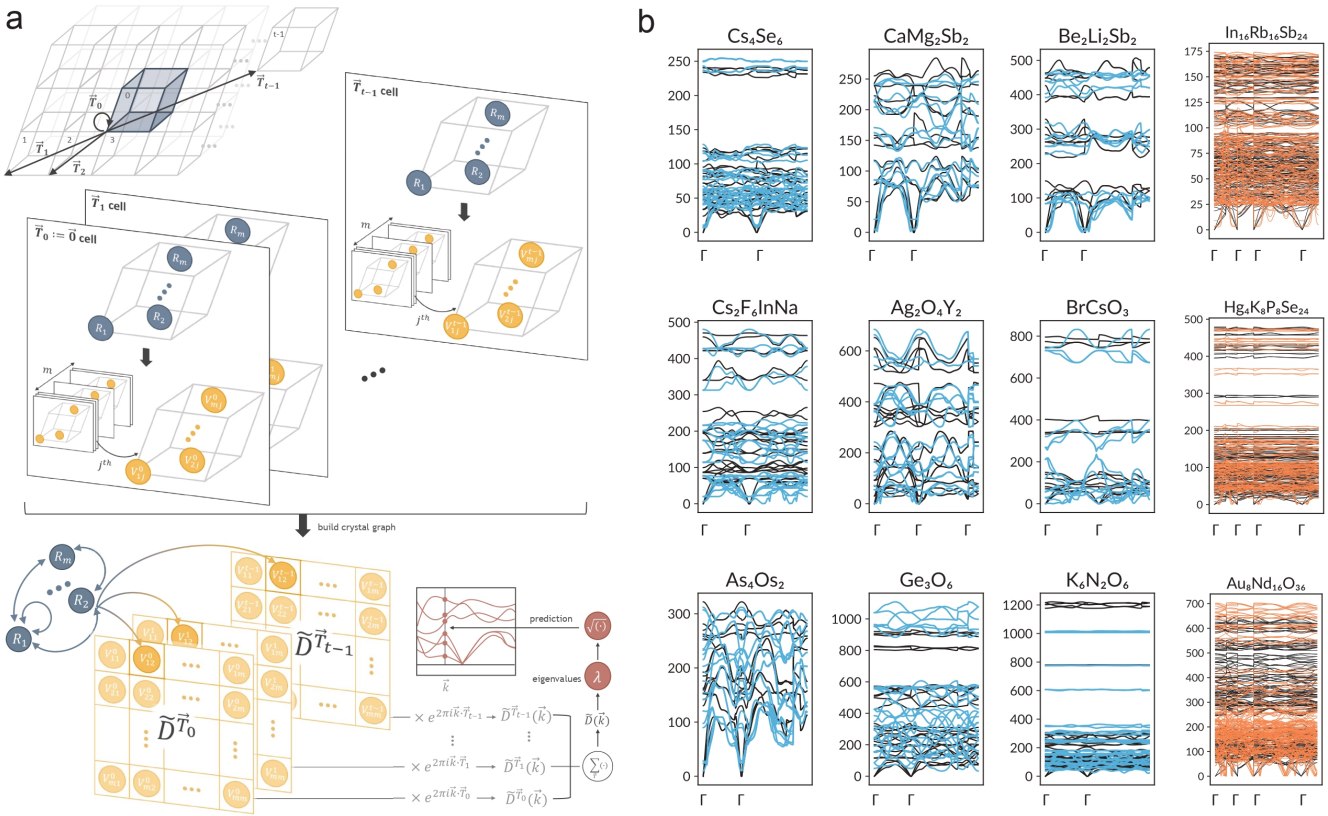

图 4. 用于预测完整声子能带结构的动量依赖矩阵虚拟节点（*k*-MVN）。a.（上）为每个平移矢量 $\vec{T}$ 添加 m^2 个虚拟节点；对于总计 t 个相邻晶胞，共生成 tm^2 个虚拟节点。（下）将相位因子乘以每个平移后的晶胞，即可构建完整的虚拟动力学矩阵。**b.** 高质量 DFPT 数据库（蓝色）和额外复杂材料测试集（橙色）中，测试集各误差三分位组的代表性示例。每个光谱的 Γ-点位置均有标注。

讨论

我们展示了如何利用三种不同类型的虚拟节点——VVN、MVN 和 *k*-MVN——增强对称性感知欧几里得神经网络，从材料的原子坐标直接预测声子。表 1总结了三种虚拟节点方法的比较。VVN 直接从虚拟节点获取声子谱。分配 $3m$ 个虚拟节点，可确保原胞包含m 个原子的晶体，其输出声子能带数始终为 $3m$ 。对于 MVN，首先构建虚拟动力学矩阵（VDM），再将声子能量作为特征值问题求解，而非直接计算声子能量。这一步对于提高复杂材料预测的稳健性至关重要，因为力常数和动力学矩阵等中间量被认为比最终声子能量更“基础”，能够更好地反映原子间相互作用。k-MVN 更进一步，利用晶胞平移生成动量依赖性，从而获得完整的声子能带结构。

如今，从头算 计算方法（如冻结声子法和密度泛函微扰理论）仍是计算声子的最准确方法。尽管如此，由于基于 VGNN 的声子计算无需直接计算每种材料各自的动力学矩阵，因此在保持合理准确度的同时，计算速度显著更快。我们还对硅锗合金、铁钴镍合金及其他高熵合金进行了额外测试，结果与现有文献吻合良好（补充信息 VI）。最后，我们利用 MVN 构建了一个数据库，其中包含 Materials

Project 中列出的超过 140,000 种材料的 $\cdot$-声子谱（补充数据和补充信息 VII）。尽管有些材料的每个晶胞中含有 400 多个原子，使用一套配备八个图形处理器的系统，仍只需不到五小时便可获得所有结果。这种高效率使我们能够在更大的设计空间中进行材料设计、搜索和优化，涵盖合金、界面乃至非晶固体，以获得性能优异、经过设计的声子性质，用于热储存、能量转换与收集以及超导性。与此同时，借助虚拟节点赋予的灵活性，还可以类似地预测传统图神经网络难以预测的其他性质，例如电子能带结构、紧束缚模型以及具有可变能带数量的$k \cdot p$ 有效哈密顿量；还有洛伦兹振子模型中灵活变化的光吸收峰等光学性质，以及传播矢量数量等磁性性质。

表1. 从物理意义和计算成本两方面比较虚拟节点对声子预测的贡献。这里，m、N_{train}、t和n 分别表示每个晶胞中的原子数、训练数据中该数量的平均值、晶胞数量以及一个任意但不大的数。

	VVN	MVN	k-MVN
力常数	-	-	反映在 VDM 中
动力学矩阵	-	VDM	VDM
声子数据	虚拟节点	特征值	特征值
运行时间	$O(m^2)$	$O(m^{2.37})$	$O(t \times m^{2.37})$
存储空间	$O(m)$	$O(m^2)$	$O(t \times m^2)$
向更大系统的推广	否	True	True

参考文献

1. C. Altintas, O. F. Altundal, S. Keskin, and R. Yildirim. 机器学习与金属有机框架相结合，用于气体储存和分离。 *Journal of Chemical InformationandModeling*，61(5):2131–2146，2021年。**2.** S. Baroni, S. Gironcoli, and A. Corso. 基于密度泛函微扰理论的声子及相关晶体性质。现代物理评论，73:515，2001。**3.** Y. Cheng and E. Ma. 金属玻璃的原子级结构与结构—性质关系。 *Progress in Materials Science*，56(4):379–473，2011。**4.** O. Delaire, A. F. May, M. A. McGuire, W. D. Porter, M. S. Lucas, M. B. Stone, D. L. Abernathy, V. Ravi, S. Firdosy, and G. Snyder. la 3- x te 4 的声子态密度和热容。 *Physical Review B*，80(18):184302，2009。**5.** M. S. Dresselhaus, G. Chen, M. Y. Tang, R. Yang, H. Lee, D. Wang, Z. Ren, J.-P. Fleurial, and P. Gogna. 低维热电材料的新方向。 *Advanced Materials*，19(8):1043–1053，2007。**6.** A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain. 材料性质预测方法的基准测试：Matbench测试集和Automatminer基准算法。 *npj Computational Materials*，6(1):1–10，2020。**7.** V. Fung, J. Zhang, E. Juarez, and B. G. Sumpter. 材料化学中图神经网络的基准测试。 *npjComputational Materials*，7(1):1–8，2021。**8.** M. Geiger and T. Smidt. e3nn：欧几里得神经网络。 *arXiv*预印本 *arXiv:2207.09453*，2022年。**9.** K. Guo, Z. Yang, C.-H. Yu, and M. J. Buehler. 人工智能和机器学习在力学材料设计中的应用。 *Materials Horizons*，8(4):1153–1172，2021。**10.** G. L. Hart, T. Mueller, C. Toher, and S. Curtarolo. 合金机器学习。 *NatureReviewsMaterials*，6(8):730–755，2021。

11. Y. Hinuma, G. Pizzi, Y. Kumagai, F. Oba, and I. Tanaka. 基于晶体学的能带结构图路径。计算材料科学，128:140–184，2017年。12. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, 等. 评论：材料项目：加速材料创新的材料基因组方法。《APL Materials》，1(1):011002，2013年。13. J. Jancar, J. Douglas, F. W. Starr, S. Kumar, P. Cassagnau, A. Lesser, S. S. Sternstein, and M. Buehler. 聚合物纳米复合材料结构—性质关系研究中的当前问题。《Polymer》，51(15):3321–3343，2010年。14. L. T. Kong. 通过分子动力学模拟直接测量声子色散。Computer Physics Communica-tions，182(10):2201–2207，2011年。15. N. Kumar, S. N. Guin, K. Manna, C. Shekhar, and C. Felser. 从化学角度看拓扑量子材料。《Chemical Reviews》，121(5):2780–2815，2020年。16. T. Le, V. C. Epa, F. R. Burden, and D. A. Winkler. 多种材料性质的定量构效关系建模。《Chemical Reviews》，112(5):2889–2919，2012年。17. Z. Liu, T. Lu, J. Ye, G. Wang, X. Dong, R. Withers, and Y. Liu. 用于储能应用的反铁电材料：综述。《Advanced Materials Technologies》，3(9):1800111，2018年。18. B. K. Miller, M. Geiger, T. E. Smidt, and F. Noé. 旋转等变卷积在分子性质预测中的相关性。arXiv预印本 arXiv:2008.08461，2020年。19. A. Mishra, M. K. Fischer, and P. Bäuerle. 用于染料敏化太阳能电池的无金属有机染料：从结构—性质关系到设计规则。《Angewandte Chemie International Edition》，48(14):2474–2499，2009年。20. J. Mortensen, J. Blomqvist, I. Castelli, R. Christensen, M. Dułak, J. Friis, M. Groves, B. Hammer, C. Hargus, E. Hermes, 等. 原子模拟环境——用于处理原子的 Python 库。《Journal of Physics. Condensed Matter: an Institute of Physics Journal》，29(27):273002–273002，2017年。21. A. R. Oganov and C. W. Glass. 利用从头算进化技术进行晶体结构预测：原理与应用。《The Journal of Chemical Physics》，124(24):244704，2006年。22. A. R. Oganov, C. J. Pickard, Q. Zhu, and R. J. Needs. 结构预测推动材料发现。Nature Reviews Materials，4(5):331–348，2019年。23. S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder. Python材料基因组学（pymatgen）：用于材料分析的稳健开源Python库。计算材料科学，68:314–319，2013年。24. J. Peng, D. Schwalbe-Koda, K. Akkiraju, T. Xie, L. Giordano, Y. Yu, C. J. Eom, J. R. Lunger, D. J. Zheng, R. R. Rao, 等. 面向可持续性和脱碳的以人为中心和以机器为中心的分子与材料设计。Nature Reviews Materials，页码1–19，2022年。25. G. Petretto, S. Dwaraknath, H. PC Miranda, D. Winston, M. Giantomassi, M. J. Van Setten, X. Gonze, K. A. Persson, G. Hautier, and G.-M. Rignanese. 无机材料的高通量密度泛函微扰理论声子计算。《Scientific Data》，5(1):1–12，2018年。26. A. D. I. PyTorch. PyTorch，2018年。27. D. Schwalbe-Koda, S. Kwon, C. Paris, E. Bello-Jurado, Z. J ensen, E. Olivetti, T. Willhammar, A. Corma, Y. Román-Leshkov, M. Moliner, 等. 通过高通量模拟先验控制沸石相竞争与互生。《Science》，374(6565):308–315，2021年。28. L. Sham. 绝缘晶体晶格动力学中的电子贡献。《Physical Review》，188(3):1431，1969年。29. V. Stanev, K. Choudhary, A. G. Kusne, J. Paglione, and I. Takeuchi. 用于量子材料搜索与发现的人工智能。《Communications Materials》，2(1):1–11，2021年。30. N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley. 张量场网络：面向三维点云的旋转和平移等变神经网络。arXiv电子预印本，页码arXiv:1802.08219，2018年2月。31. A. Togo. 京都大学声子数据库。http://phonondb.mtl.kyoto-u.ac.jp/，2015年。32. A. Togo and I. Tanaka. 材料科学中的第一性原理声子计算。《Scripta Materialia》，108:1–5，2015年。33. A. Togo and I. Tanaka. spglib：用于晶体对称性搜索的软件库。arXiv预印本 arXiv:1808.01590，2018年。

34. M. Wagih、P. M. Larsen 和 C. A. Schuh。学习多晶材料中的晶界偏析能谱。《自然·通讯》，11(1):1–9，2020年。**35.** T. Xie 和 J. C. Grossman。用于准确且可解释地预测材料性质的晶体图卷积神经网络。《物理评论快报》，120(14):145301，2018年。**36.** Y. Yao、Q. Dong、A. Brozena、J. Luo、J. Miao、M. Chi、C. Wang、I. G. Kevrekidis、Z. J. Ren、J. Greeley 等。高熵纳米颗粒：合成—结构—性质关系与数据驱动发现。《*Science*》，376(6589):eabn3103，2022年。**37.** W. Zheng 和 L. Y. S. Lee。用于电催化的金属有机框架：催化剂还是前催化剂？《*ACS*能源快报》，6(8):2838–2843，2021年。**38.** T. Zhu、R. He、S. Gong、T. Xie、P. Gorai、K. Nielsch 和 J. C. Grossman。绘制无机晶体的晶格热导率图谱，并发现用于热电材料的稀土硫族化合物。《能源与环境科学》，14(6):3559–3566，2021年。

方法

声子数据准备

我们使用从头算密度泛函微扰理论计算数据库，研究谐振模型中的声子色散[25]。该数据集包含材料结构（与从 Materials Project 获取的原胞结构相同[12]）、不可约区内规则点处能量相对于原子扰动的二阶导数，以及 1,521 种无机晶体材料沿高对称路径的声子色散。这些材料的每个晶胞含有 2 至 40 个原子，平均为 7.38 个。在本研究中，我们仅将沿高对称路径的声子色散用作训练数据。色散数据包括分数倒易单位中的波矢 $\vec{k}$ 以及以 cm^{-1}为单位的响应谱。所有模型都将数据随机划分为 90% 的训练集（1,365 种材料）和 10% 的测试集（156 种材料）。此外，我们采用五折交叉验证方案训练模型。

我们还从 Atsushi Togo 的声子数据库[31]中获取了复杂材料（每个晶胞中原子数更多）的声子色散。我们使用 *seekpath*[11,33] 模块获取每种材料的高对称路径。然后，将其与数据库中的*POSCAR*、*FORCE_SET*和 *phonopy*。*config* 文件一同输入 *Phonopy*[32] 的 Python 命令，以计算沿该路径的声子色散。为控制数据质量，我们筛选出最低 Γ 声子带高于 $-0.07\ \mathrm{cm}^{-1}$的材料。我们还对材料进行了筛选，仅保留每个晶胞中原子数超过 40 个的材料。最后，我们从筛选后的 505 种材料中随机选取 156 种（与测试集的数据数量相同，以便于比较），并将其作为复杂材料数据集。

计算环境

我们使用 Python 3.9.13 编写模型，并在配备 CUDA 10.2版本的 GPU集群上进行训练。为便于模型实现和训练，我们使用了一些重要的 Python 模块：*Pymatgen*[23] 和 *ase*[20] 用于处理材料结构文件（。*CIF*），*PyTorch*[26] 用于管理模型训练框架，*e*3*nn*[8] 则用于实现具有欧几里得群等变性的神经网络模型。

虚拟节点图神经网络（VGNN）

我们提出了一种图神经网络（GNN）方案，使其能够根据输入大小具有可变的输出维度。为便于理解，我们将结合声子预测方面的工作来解释该方法。

对于每个晶胞含有m 个原子的材料，我们添加n 个额外的虚拟原子。我们可以根据模型架构调整n 的数量。利用真实原子和虚拟原子，我们将晶体结构转换为周期图，其中实际原子对应m 个真实节点，添加的虚拟原子对应n 个虚拟节点。随后，我们通过边连接节点，以表示消息传递过程。为保留材料的结构信息并限制计算成本，我们采用以下连接规则。首先，如果两个真实节点之间的距离小于指定的截断半径$r_{\max}$，则通过双向边连接这两个真实节点。我们还根据模型描述，在真实节点与虚拟节点之间建立边，但该边的方向由真实节点指向虚拟节点。最后，我们将径向距离向量的信息（例如，从原子b 到原子a的$\vec{r}_{ab}$ ）以径向基函数和球谐函数的形式嵌入相应的边中，作为边属性，分别表示$\vec{r}_{ab}$ 的距离和方向。

由于每个节点代表晶胞中的一个原子，我们将原子序数A 信息作为节点属性嵌入 $\mathcal{A}$，方法是将长度为 118 的 one-hot 表示向量传入嵌入层。对于模型的输入，我们将原子质量Z 信息作为输入节点特征嵌入 $\mathcal{Z}$，方法是将原子质量与原子序数的 one-hot 表示相乘后传入嵌入层。

随后，将构建好的图传入模型的消息传递模块，该模块通过多个卷积和门控激活层[18]对特征进行处理。经过最后一层（仅包含卷积，不含门控激活）后，每个

个n 虚拟节点特征都会被收集起来，并传入后处理模块，该模块输出 $3m$ 条预测的声子支。后处理模块各不相同，后续章节将分别详细介绍每种模型的后处理模块。

模型通过最小化训练数据集中声子与模型预测值之间的均方误差（MSE）损失函数进行优化，计算前先用各材料的最大声子频率对二者进行归一化。完整的网络结构见补充信息。

向量虚拟节点方法（VVN）

VVN 是我们设计的一种 VGNN，用于学习从材料结构预测 Γ-声子谱。对于每个晶胞含有 m 个原子的材料，其声子能带有 $3m$ 条，因此，一种合理的添加虚拟原子的方法是添加 $3m$ 个虚拟原子，每个虚拟原子分别输出一条能带的预测值。因此，当晶体材料的晶胞中有 m 个原子时，我们根据以下方程（$\vec{r}_{V_i}$）指定虚拟节点 $V_i, i \in [1, 3m]$ 的位置。虚拟节点的原子种类可以任意设定；优化后我们采用铁。

$$\vec{r}_{V_i} = \frac{i-1}{3m}(\vec{a}+\vec{b}+\vec{c}). \tag{1}$$

这里 $\vec{a}$、$\vec{b}$、$\vec{c}$ 表示材料的晶胞矢量。换言之，$3m$ 个虚拟原子以相等间距沿从 $(0,\ 0,\ 0)$ 到 $\vec{a}+\vec{b}+\vec{c}$ 的对角线排列。通过保持虚拟节点在实空间中的间距，可以在特征更新过程中引入位置依赖关系。从这个意义上说，方程（1）能够确保虚拟节点彼此分离，并使我们能够将这 $3m$ 个虚拟节点用作网络的输出节点。为了获取整个结构的信息，$3m$ 个虚拟节点分别通过从真实节点指向虚拟节点的有向边与所有真实节点相连。每个卷积层之后，虚拟节点特征会传递到线性层，转换为标量输出，并按大小排序。输出的 $3m$ 个标量表示预测的 Γ-声子。

矩阵虚拟节点方法（MVN）

MVN 是我们设计的一种 VGNN，它受到周期谐振系统动力学矩阵表示的启发，用于学习从材料结构预测 Γ声子谱。给定动量矢量 k，动力学矩阵元素 $\tilde{D}ij(\vec{k})$ 是一个三乘三矩阵，表示原子 R_i 与 R j 之间的三维谐相互作用，可根据公式（2）写成力常数矩阵 $\Phi^{\alpha\beta}$ i j 的傅里叶变换。这里，Z_{R_i} 是原子 R_i 的原子质量，而 $\vec{T}\alpha$ 是 α^{th} 晶胞位置。请注意，对于每个 k 向量，系统有 3m 个自由度和频率，其中 m 是每个晶胞中的原子数。通过求解 $\tilde{D}(\vec{k})$ 的特征值 $\omega^2(\vec{k})$，我们可以得到声子色散关系 $\omega(\vec{k})$；$\tilde{D}(\vec{k})$ 是一个形状为 (3m,3m) 的矩阵，由 m^2 个 $\tilde{D}ij(\vec{k})$ 块组成，其中 i 、 $j \in [1, m]$,

$$\tilde{D}_{ij}(\vec{k}) = \sum_{\alpha,\beta} \frac{\Phi_{ij}^{\alpha\beta}}{\sqrt{Z_{R_i} Z_{R_j}}} e^{i\vec{k}\cdot(\vec{T}_\alpha - \vec{T}_\beta)}. \tag{2}$$

在 MVN 方法中，我们生成一个可用作动力学矩阵的矩阵，如公式（2）所示。在这里，我们专注于 $\vec{k}=0$ 预测 Γ-声子，即 k $=0$。因此，对于每一对相同的原子（例如 R_i 和 R_j），来自各晶胞 $\vec{T}_\alpha - \vec{T}_\beta$ 间隔的贡献会相加，且不包含依赖于 k 的指数相位因子。因此，模型需要预测一个形状为 (3m, 3m) 的矩阵，以表示这种求和。为此，在保留各矩阵元素之间关系的同时，我们生成 m 个虚拟晶体 Cj，$j \in [1, m]$；每个虚拟晶体包含 m 个虚拟节点 Vij, $i \in [1, m]$，它们与真实原子 R_i, $i \in [1, m]$具有相同的原子种类和位置。这里，虚拟节点 Vij 表示从真实节点 R j 到另一个真实节点 R_i 的相互作用项：只要 R j 与 R_i 之间存在边，就从 R j 向 V_{ij} 添加一条有向边。每经过一层卷积层，虚拟节点特征都会传入线性层，并转换为长度为 9 的复值输出向量。对于每个输出特征，我们将其重塑为三乘三矩阵，并按如下方式排列：V_{ij} 的矩阵作为形状为 (3m, 3m) 的 $\tilde{D}$ 超矩阵中的 (i, j) 块。最后，我们求解 $\tilde{D}$ 的 3m 个特征值，将其作为Γ-声子预测。

动量依赖矩阵虚拟节点方法（k-MVN）

k-MVN 是 MVN 模型在 k 非零情况下的推广。与 MVN 情况不同，k-MVN 模型需要预测表示晶胞中原子之间相互作用的矩阵，例如 $\vec{T}\beta$，以及不同晶胞中的原子之间的相互作用矩阵，例如 $\vec{T}\alpha$。由于相位因子仅取决于晶胞位置之差，我们可以将 T 重新定义为这种位置差，并将公式（2）简化为

$$\tilde{D}_{ij}(\vec{k}) = \sum_{\vec{T}} \frac{\Phi_{ij}^{\vec{T}}}{\sqrt{Z_{R_i} Z_{R_j}}} e^{i\vec{k}\cdot\vec{T}} := \sum_{\vec{T}} D_{ij}^{\vec{T}} e^{i\vec{k}\cdot\vec{T}}. \tag{3}$$

经过这一简化后，对于每个 $\vec{T}$，我们按照与 MVN 相同的方式生成 m 个虚拟晶体 $C^{\vec{T}}_{j\in[1,m]}$。不过，在这种情况下，$V_i^{\vec{T}}$j$\vec{}$表示从真实节点 R$_I$ 到另一个真实节点 R$_i$ 的相互作用项，后者位于相对于 R j 的晶胞位置为 T 的晶胞中。换言之，只要存在一条连接 R j 与 R$_i$的边，且该边表示 $\vec{}$r$_i$ − $\vec{}$r j = $\vec{}$r$'_i$ + $\vec{}$T − $\vec{}$r$'_j$，我们就添加一条从 R j 指向 V Tij 的有向边。这里，$\vec{}$r′表示相对于其晶胞的原子位置。由于图神经网络仅考虑原子间距离小于 r$_{max}$的边，因此采用这一方案，模型只能为满足以下条件的有限个 T 生成非零矩阵$\vec{}$

$$\min_{i,j\in[1,m]} |\vec{r}'_i + \vec{T} - \vec{r}'_j| \leq r_{\max}. \tag{4}$$

因此，在 虚拟晶体生成之前，模型还会遍历原子对，以找出所有可行的 $\vec{T}$.

与 MVN 模型类似，我们将虚拟节点特征转换为三乘三矩阵，并将它们合并为一个形状为 $(3m,3m)$ 的矩阵，表示每个 $\vec{}T$ 对应的 $\tilde{}D^{\vec{T}}$。最后，我们对这些矩阵按其相位因子进行加权求和，得到 $\tilde{}D$，并求解$\vec{}$其 $3m$ 个特征值，作为波矢k处的声子谱。

致谢

RO 和 AC 对本研究的贡献相同。RO、AC、AB 和 ML 感谢 M Geiger、S Fang、T Smidt 和 K Persson 提供的宝贵讨论，并感谢美国能源部（DOE）科学办公室（SC）基础能源科学（BES）项目（奖项编号：DE-SC0021940）以及美国国家科学基金会（NSF）“设计材料以革新和工程化我们的未来”（DMREF）计划（奖项编号：DMR-2118448）提供的支持。BL 感谢美国国家科学基金会 DMREF 项目（奖项编号：DMR-2118523）提供的支持。TN、ND 和 ML 获得了美国能源部基础能源科学项目（奖项编号：DE-SC0020148）的部分资助。TN 感谢 Mathworks奖学金和 Sow-Hsin Chen奖学金提供的支持。ML 感谢 1947届职业发展讲席的支持，并感谢 S. Yip 的讨论。

利益冲突

作者声明不存在利益冲突

数据可用性声明

支持本研究结论的数据已在 GitHub 上公开，网址为 https://github.com/RyotaroOKabe/phonon_prediction。采用 MVN 方法生成的 Γ-声子数据库可在 https://osf.io/k5utb/ 获取。

arXiv:2301.02197v1 [cond-mat.dis-nn] 5 Jan 2023

用于完整声子预测的虚拟节点图神经网络：补充信息

Ryotaro Okabe[1,2,†], **Abhijatmedhi Chotrattanapituk**[1,3,†], **Artittaya Boonkird**[1,4], **NinaAndrejevic**[5], **Xiang Fu**[3], **Tommi S. Jaakkola**[3], **QichenSong**[6], **ThanhNguyen**[1,4], **NathanDrucker**[1,7], **Sai Mu**[8], **Bolin Liao**[9], **YongqiangCheng**[10], 和 **Mingda Li**[1,4,*]

[1]量子测量组，麻省理工学院，美国马萨诸塞州剑桥[2]化学系，麻省理工学院，美国马萨诸塞州剑桥[3]电气工程与计算机科学系，麻省理工学院，美国马萨诸塞州剑桥[4]核科学与工程系，麻省理工学院，美国马萨诸塞州剑桥[5]阿贡国家实验室，美国伊利诺伊州莱蒙特[6]化学与化学生物学系，哈佛大学，美国马萨诸塞州剑桥[7]应用物理，工程与应用科学学院，哈佛大学，美国马萨诸塞州剑桥[8]物理与天文学系，南卡罗来纳大学，美国南卡罗来纳州哥伦比亚[9]材料系，加州大学圣巴巴拉分校，美国加利福尼亚州圣巴巴拉[10]化学光谱组，光谱科，中子散射部，橡树岭国家实验室，美国田纳西州橡树岭

[†]这些作者贡献相同[*]电子邮箱：
mingda@mit.edu

目录

I 数据准备 1 **II 虚拟节点图神经网络** 2
II.1 欧几里得神经网络与真实节点图卷积. 2
II.2 虚拟节点增强与卷积. 3
II.3 神经网络架构. 4
III 神经网络训练与优化 6 **IV 性能 Γ-声子预测** 6
IV.1 附加 Γ-声子预测. 6
IV.2 对未见材料预测的普遍有效性. 13
IV.3 平均声子能量的逐元素预测. 15 **V 完整声子能带结构预测** 20 **VI 应用** 24
VI.1 拓扑外尔半金属的声子预测. 24
VI.2 合金系统中的声子预测. 24 **VII Γ-声子数据库** 26

I 数据准备

包含 1521 种半导体无机材料完整声子能带的数据集，是通过密度泛函微扰理论（DFPT）计算得到的[7]，并作为主要训练数据集（简称“主数据库”）。提取了Γ点处的声子能量值（用于向量虚拟节点 VVN 和矩阵虚拟节点 MVN 模型），以及其他高对称点处的声子能量值（用于k-MVN 模型），并随机划分为训练集（90%）和测试集（10%）。高对称点因材料结构不同而异，

并针对原胞在实空间和倒易空间中均采用约化分数坐标。我们对所有数据进行了 consistency check，确保分数坐标中的预期 k 点与布里渊区中期望的k点相匹配。模型测试所用的另一组数据取自京都大学 Dr. Atsuhi Togo 的声子数据库（简称“Togo 数据库”）[?]，其中包含更复杂材料的声子，但同时也包含更多具有虚声子能量的声子。鉴于主数据库的收敛标准更严格，我们使用主数据库进行训练。此外，我们从 Togo 数据库中随机选取了 156 种材料进行测试，这些材料的最低 Γ 声子大于 -0.07 cm^{-1}，且每个晶胞至少含有 40 个原子，用于测试以主数据库训练的模型。Togo 数据库中的 156 种材料与主数据库测试集的数量相同（1521 种材料的 10%）。主数据库测试集和 Togo 数据库的附加测试集中均包含大多数元素。图*S1*展示了主数据库和 Togo 数据库中每个晶胞原子数的分布（正文中的m值）。关于元素的数据集概况见图S2 -S3。

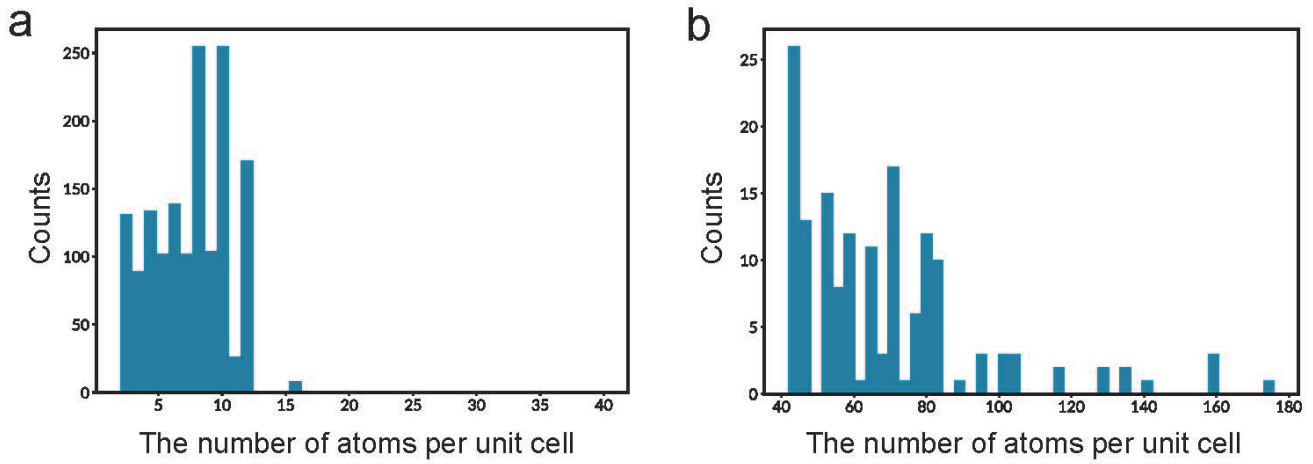

图 S1。每个晶胞中的原子数分布。a 中每个晶胞的原子数。来自主数据库的训练集和测试集 **b**。来自 Togo 数据库的复杂材料附加测试集。**a**。主数据库中的 1521 种材料总计包含每个晶胞 2 到 40 个原子，平均每个晶胞 7.4 个原子。**b**。从 Togo 数据库中随机选取的 156 种真实值质量良好的复杂材料，每个晶胞包含 42 到 174 个原子（平均每个晶胞 69.1 个原子）。

II 虚拟节点图神经网络

II.1 欧几里得神经网络与真实节点图卷积

我们的方法是在具备对称性感知能力的图卷积神经网络框架——欧几里得神经网络（E3NN）内开发的[?]。图 S4 展示了模型的整体架构。与先前的报告相同，该模型将原子质量和晶胞中的原子位置作为输入[?]。这些信息被转换为一个周期图，其中节点表示原子，边控制节点间的消息传递。随后，这些输入经过一系列由非线性层分隔的卷积层，从而赋予模型复杂性。卷积层计算输入特征的张量积，卷积核定义如下：

$$f_i' = \frac{1}{\sqrt{z}} \sum_{j \in \partial(i)} f_j \otimes (h(||r_{ij}||)) Y(r_{ij}/||r_{ij}||) \tag{1}$$

其中，f'_i是原子i的输出特征，而 $\partial(i)$是满足$r_{ij} \leq r_{max}$的邻近原子集合，其中r_{ij}表示从原子 j 指向 i 的相对位置。我们将原子j 的特征与卷积核的张量积求和。该卷积核由学习得到的径向函数（$h(||r_{ij}||)$）和球谐函数$Y(r_{ij}/||r_{ij}||)$组成。归一化因子z为配位数，也就是邻近原子的数量。通过最后一个卷积层后，输出特征会经由筛选虚拟节点和处理层进行处理。计算均方误差（MSE）损失，并将其用于反向传播以优化模型。

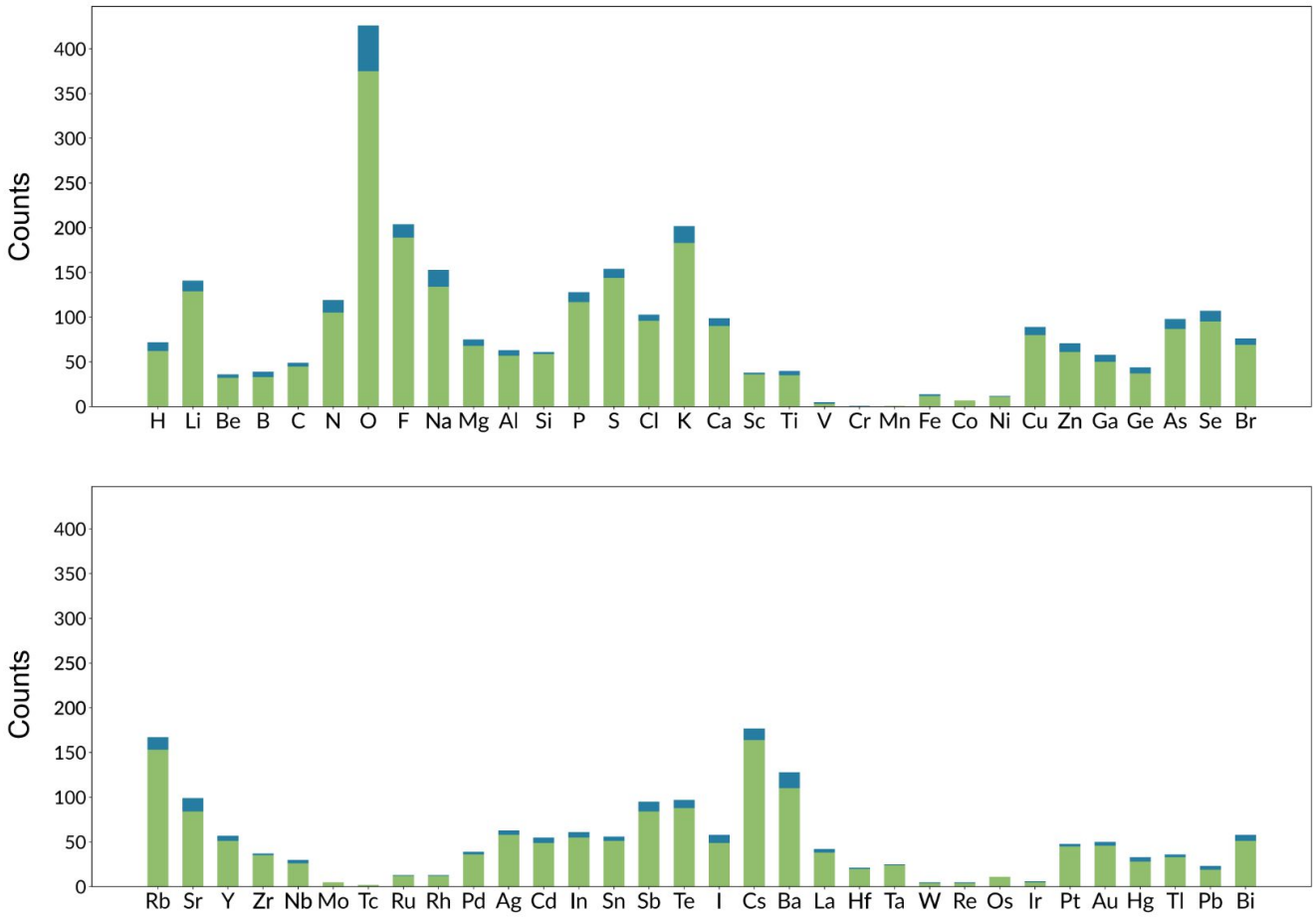

图 S2. 主数据库中各化学元素对应的训练和测试数据。 训练集（绿色）和测试集（蓝色）中各化学元素的出现次数。

II.2 虚拟节点增强与卷积

结合图卷积的参数共享优势和欧几里得神经网络的对称性感知能力，虚拟节点的增广必须满足模型仍能执行正确的、对欧几里得群等变的卷积这一基本要求；虚拟节点必须嵌入与真实节点相同维度的节点属性和输入特征，并且连接虚拟节点与任何其他节点的每条边，都必须嵌入与连接真实节点的边相同维度的边属性。

在我们的声子预测研究中，我们决定通过将表示虚拟节点的虚拟原子添加到晶体结构中来增强虚拟节点，并在构图时采用与不含虚拟原子时相同的方法。由于图构建对虚拟原子和真实原子一视同仁，因此结果必然已满足所有要求。尽管这种增强方法限制了由某些虚拟原子构建虚拟节点的自由度，但我们仍有充足的自由度来设计模型构建，包括所添加虚拟原子的数量、类型和位置。事实上，根据模型的不同，虚拟原子不必是实际的化学元素，例如，虚拟原子的原子序数为 4，但原子质量为 0。

虚拟节点方法允许的另一个自由度是节点连接关系。不过，由于我们希望模型保留真实节点之间的连接结构和消息传递，因此决定将真实节点与虚拟节点之间的边限制为仅从真实节点指向虚拟节点，而真实节点之间的连接则与前文所述相同。因此，在限制真实节点与虚拟节点之间消息传递方向的前提下，我们可以采用其余任意连接组合来设计模型。

总而言之，我们所有的模型（VVN、MVN 和 *k*-MVN）都通过在原始晶胞的特定位置添加一定数量、特定类型的虚拟原子来扩充虚拟节点。随后，它们将虚拟原子视为真实原子来构建图，以保留卷积功能和对称性感知能力。最后，它们纳入连接真实节点的所有边，仅纳入部分从真实节点指向虚拟节点的有向边，并根据模型的设计纳入部分连接虚拟节点的边

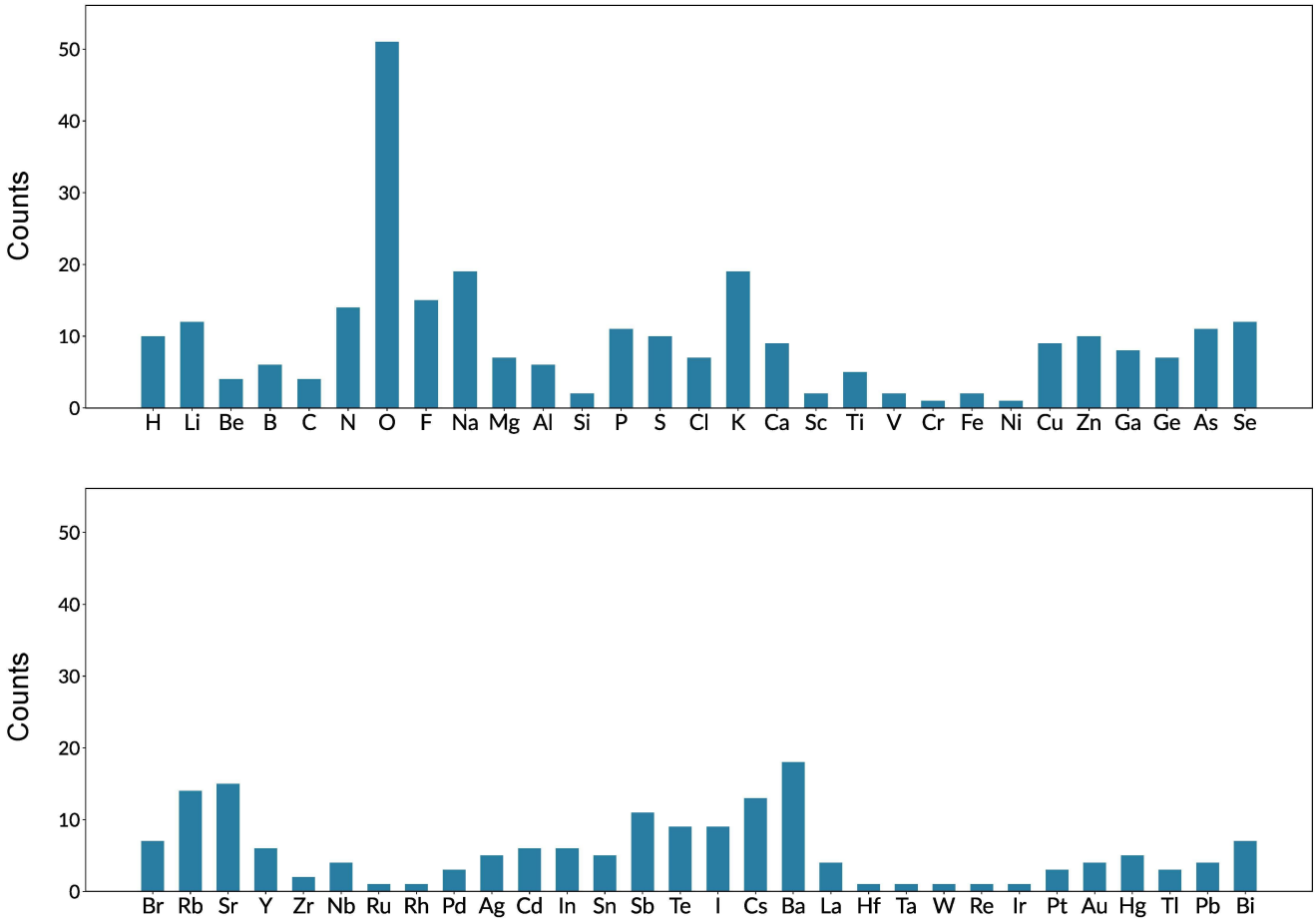

图 S3. Togo 数据库中按元素划分的其他测试数据。 156 种随机选取的材料中，复杂材料测试数据所包含的各化学元素出现次数如图所示。

模型的。

II.3 神经网络架构

通用神经网络架构如图 S4所示。

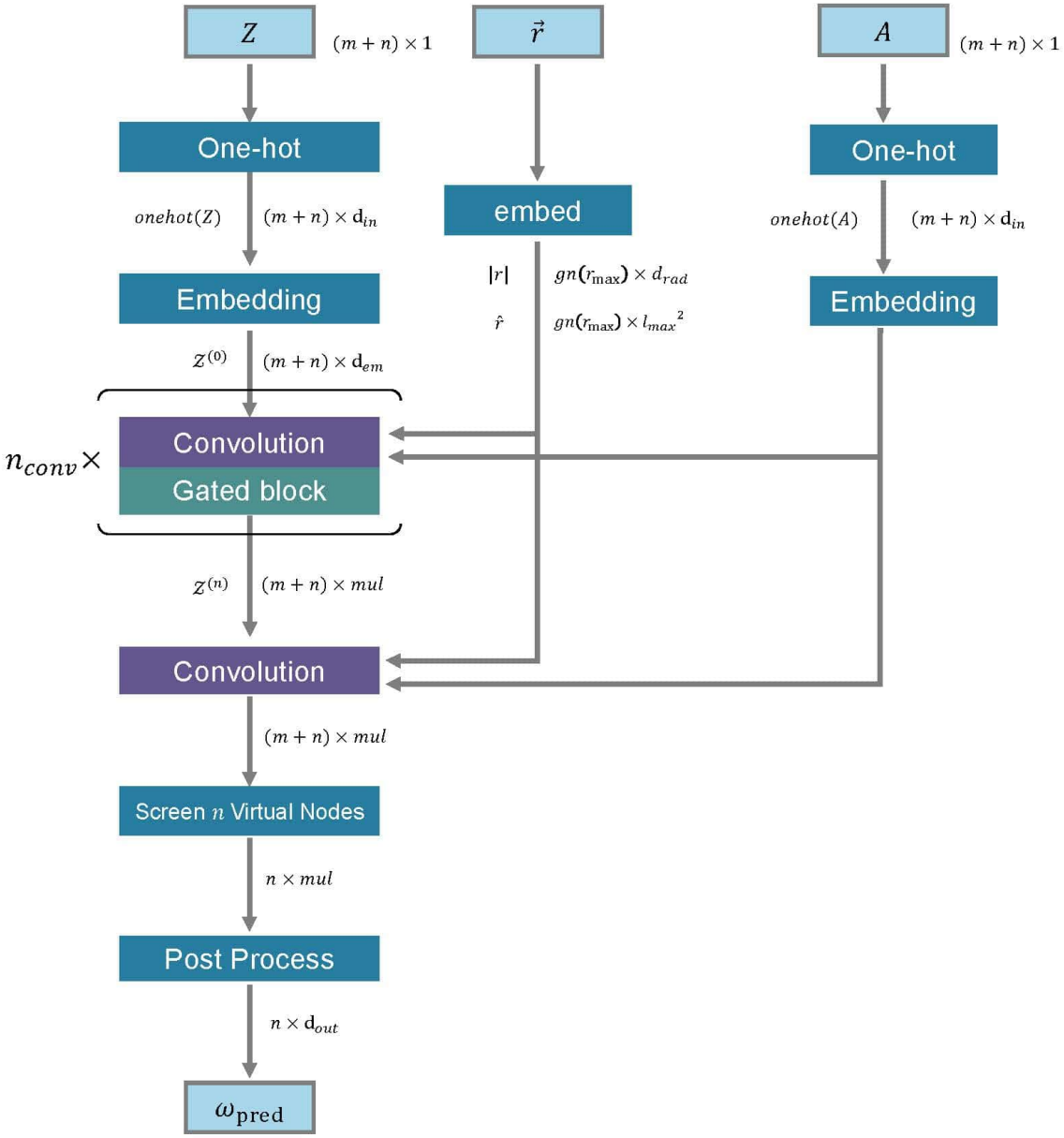

图 S4。 等变神经网络与卷积层的整体架构。 该模型由 n 个卷积层组成，各层之间由非线性层、筛选虚拟节点层和后处理层隔开。

III 神经网络训练与优化

我们对训练参数的取值进行优化。使 VVN、MVN、k-MVN 获得最佳结果的参数集列于表 S1。

表 S1。 VVN 的参数设置

超参数	VVN	MVN	k-MVN
最大截断半径（r_{max}[Å]）	4	4	4
不可约表示的重数 (mul)	16	4	4
逐点卷积层数 (n_{conv})	2	3	2
径向滤波器的基函数数量 (n_{rad})	10	10	10
球谐函数的最大 l 值 (l_{max})	2	2	2
嵌入特征向量长度 (dim_{em})	16	32	32
输出特征向量长度 (dim_{out})	1	18	18
AdamW 优化器学习率	$(5 \cdot 10^{-3}) \times 0.96^k$	$(5 \cdot 10^{-3}) \times 0.96^k$	$(5 \cdot 10^{-3}) \times 0.96^k$
AdamW 优化器权重衰减系数	0.05	0.05	0.05

IV Γ 声子预测性能

IV.1 其他 Γ-声子预测

我们展示了其他 Γ-声子预测。图 S5-S10 展示了使用 VVN 模型、MVN 和 k-MVN 模型，对主数据库中的训练数据和测试数据，以及 Togo 数据库中的测试数据进行直接预测所得的结果及相关性图。

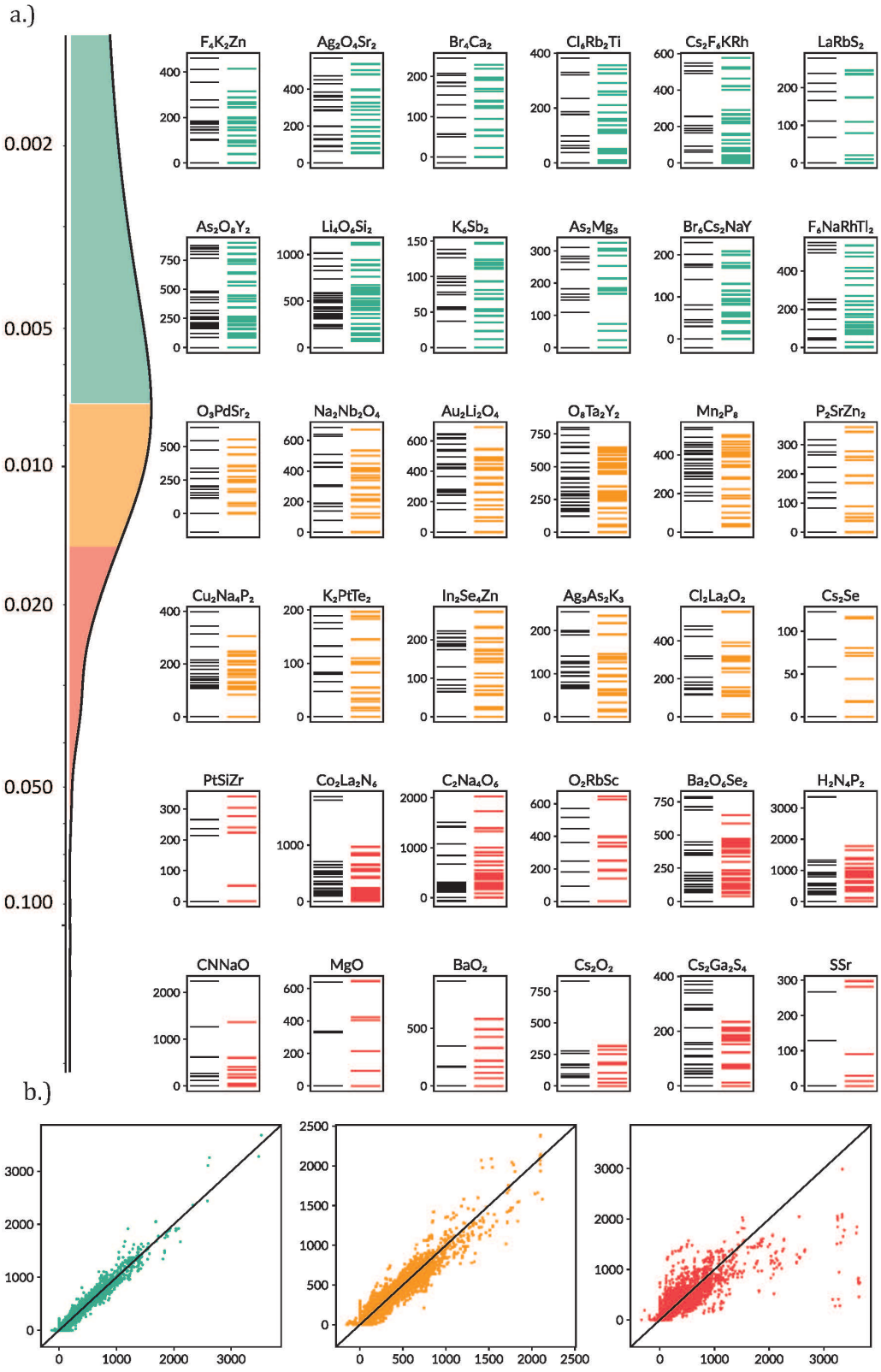

图 S5. VVN 模型：主数据库训练集中的直接预测结果。 **a**。黑线表示 DFPT 计算得到的 Γ-声子（真实值），单位为 $[cm^{-1}]$。彩色（绿、黄、红）线分别表示预测的第一、第二和第三三分位中的Γ-声子。（左）损失分布显示，损失主要集中在误差较低的第一和第二三分位。**b**。各三分位中训练集内所有 Γ-声子的相关性图（预测值与真实值）。

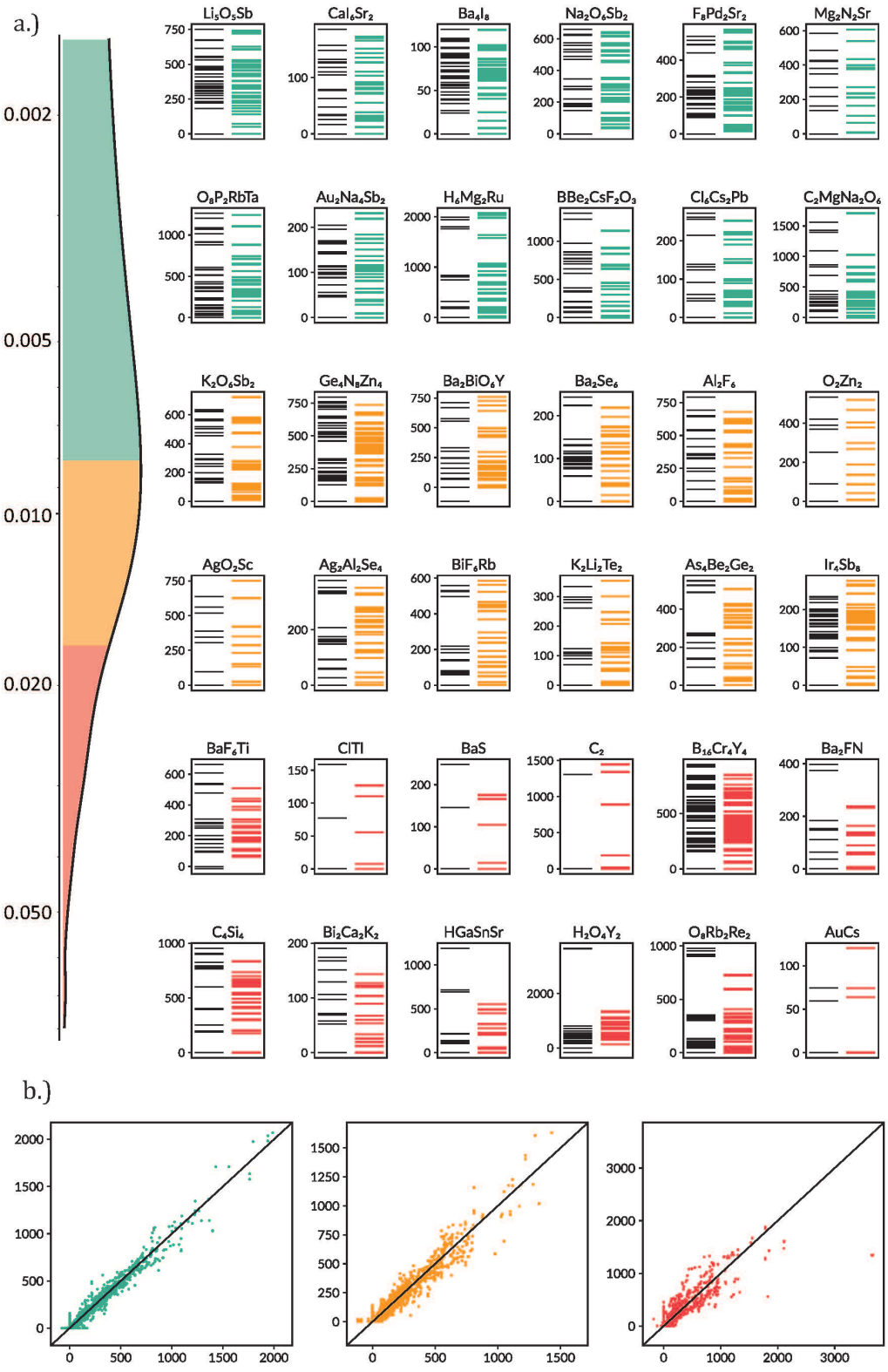

图 S6. VVN 模型：主数据库测试集中的直接预测结果。a. Γ声子的预测结果与第 1、第 2 和第三三分位中的 DFPT 计算结果（分别以绿色、黄色和红色表示）在 $[cm^{-1}]$中进行比较。左侧的损失分布显示，第 1 和第 2 三分位处出现峰值。**b.** 各三分位中测试集内所有Γ声子的相关性图（预测值与真实值）。

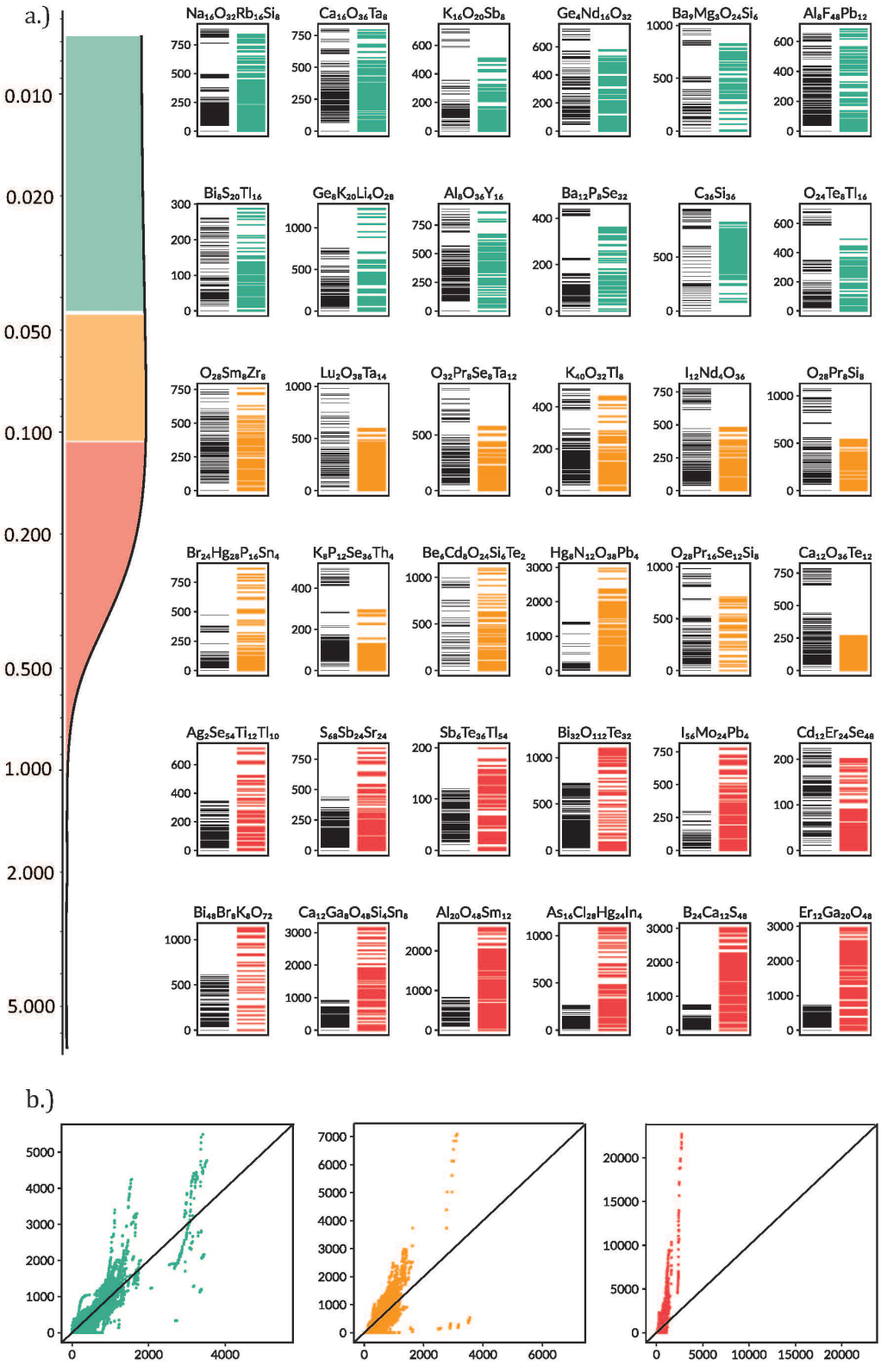

图 S7. VNN 模型：使用 VVN 对 Togo 数据库测试集进行直接预测的结果。a. Γ声子的预测值与 DFPT 计算结果（黑色）对比，分别显示第一、第二和第三三分位（绿色、黄色和红色）中的结果，单位为 [厘米$^{-1}$]。与训练数据（图 S5）和测试数据（图 S6）相比，损失分布在高值范围内更为宽广。 **b.** 测试集中各三分位内所有 Γ声子的相关性图（预测值与真实值）。

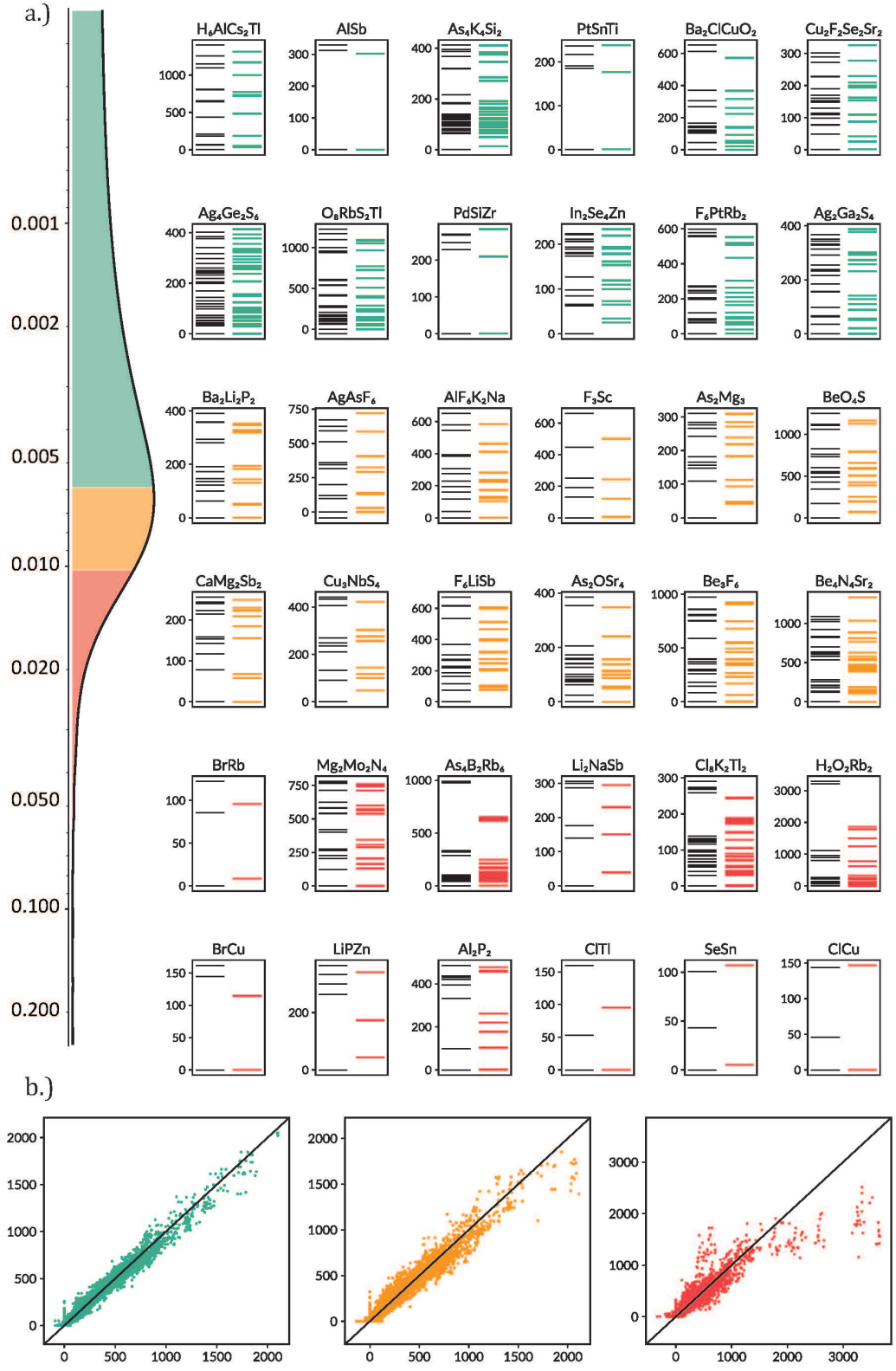

图 S8。 MVN 模型：使用 MVN 对 MainDatabase 训练集中的预测结果进行直接预测。 Γ声子的预测结果与第 1、第 2 和第 3 三分位数中的 DFPT 计算结果（黑色）进行比较（绿色、黄色和红色），见 [cm^{-1}]。左侧的损失分布显示，在第 1 和第 2 三分位数处有一个峰值。**b**。测试集中各三分位数内所有Γ声子的相关性图（预测值与真实值）。

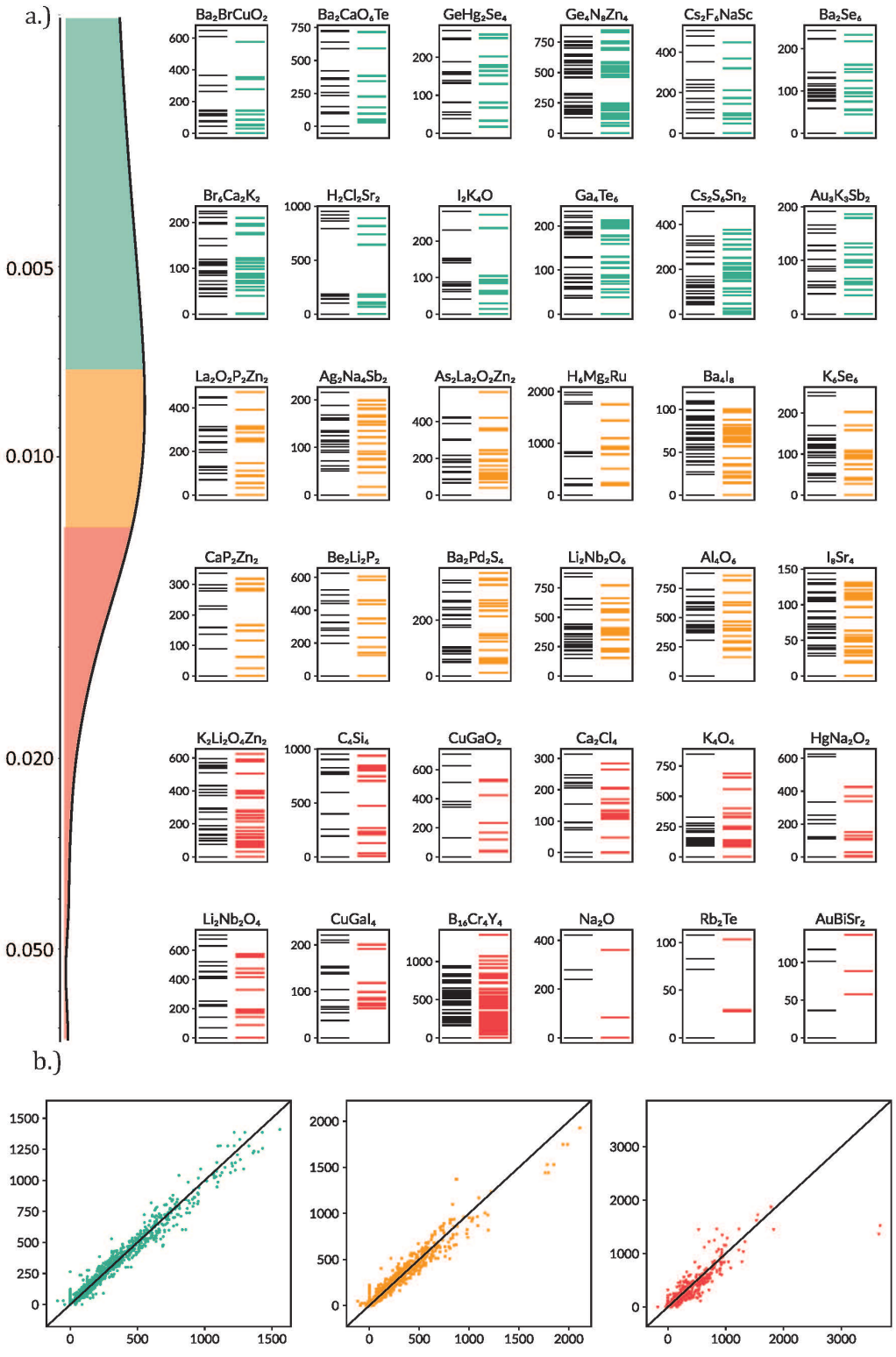

图 S9. MVN 模型：使用 MVN 对主数据库测试集中的数据进行直接预测的结果。a。 Γ-声子的预测结果与 DFPT 计算（黑色）在[cm^{-1}]中的第一、第二和第三三分位（绿色、黄色和红色）进行比较。左侧的损失分布显示，峰值位于第一和第二三分位。**b**。各三分位中测试集内所有 Γ-声子的相关性图（预测值与真实值）。

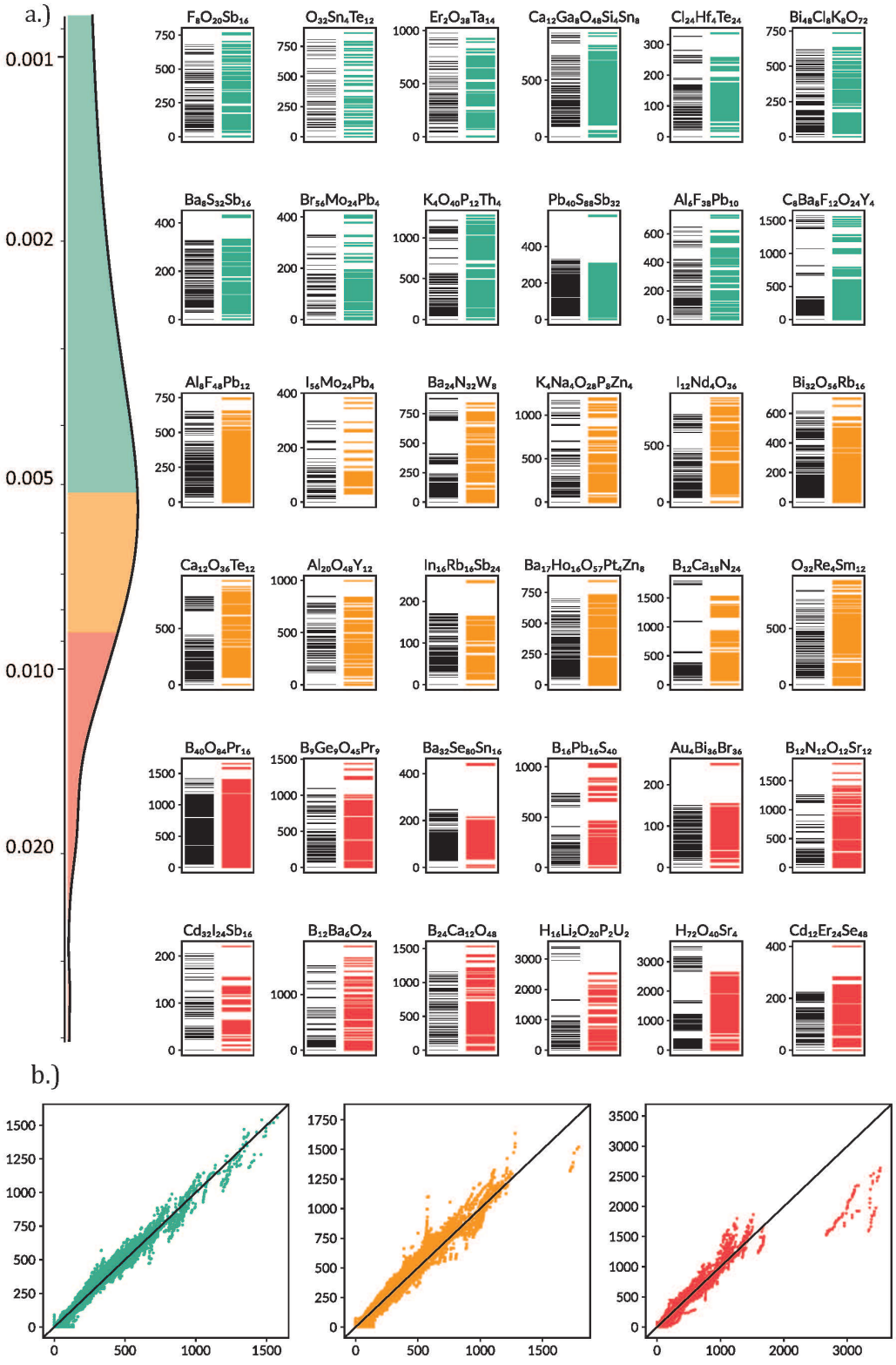

图 S10. MVN 模型：使用 MVN 对 Togo 数据库测试集中的结果进行直接预测。a。 Γ-声子的预测结果与 DFPT 计算结果（黑色）进行比较，分别显示第一、第二和第三三分位（绿色、黄色和红色）中的结果，单位为 $[cm^{-1}]$。左侧的损失分布显示，第一和第二三分位处存在峰值。**b**。测试集中各三分位内所有Γ-声子的相关性图（预测值与真实值）。

IV.2 未见材料预测结果的普遍有效性

在图 S11 中，我们绘制了平均声子频率与晶胞内原子质量均方根的关系。我们将预测模型应用于来自 Materials Project 的 5000 种未见晶体结构，这些结构的每个晶胞中原子位点数 N 分别为 $1 \leq N \leq 20$、(S11c,d) $21 \leq N \leq 40$ 和 (S11e,f) $41 \leq N \leq 80$。对于每种材料，平均声子频率与质量呈双曲关系 $\bar{\omega} = C\ \bar{m}^{-1/2}$，其中常数 C 表示晶体的刚度。刚度的合理分布支持了我们的模型对未知材料的物理有效性。此外，我们通过计算晶体中最小质量 m_{min} 与 $\bar{m}$ 的比值，来表征每种材料中原子质量的非均匀性。高 $m_{min}/\bar{m}$ 的材料同时包含小原子和大原子，往往聚集在较低的 $\bar{\omega}$ 和较高的 $\bar{m}$ 区域。相反，低 $m_{min}/\bar{m}$ 的材料所含原子质量相近，往往聚集在较高的 $\bar{\omega}$ 和较低的 $\bar{m}$ 区域。这些趋势表明，氢和氧等小原子会产生高频声子，这与先前研究的结果一致[?]。值得注意的是，随着 N 增大，VVN 的图中低 $\bar{m}$ 区域的 $\bar{\omega}$ 分布更高且更宽，而 MVN 在同一区域的分布则较窄。这与图 S7 中的结果相符：当输入材料较为复杂时，VVN 预测的 Γ-声子频率会高于真实值。

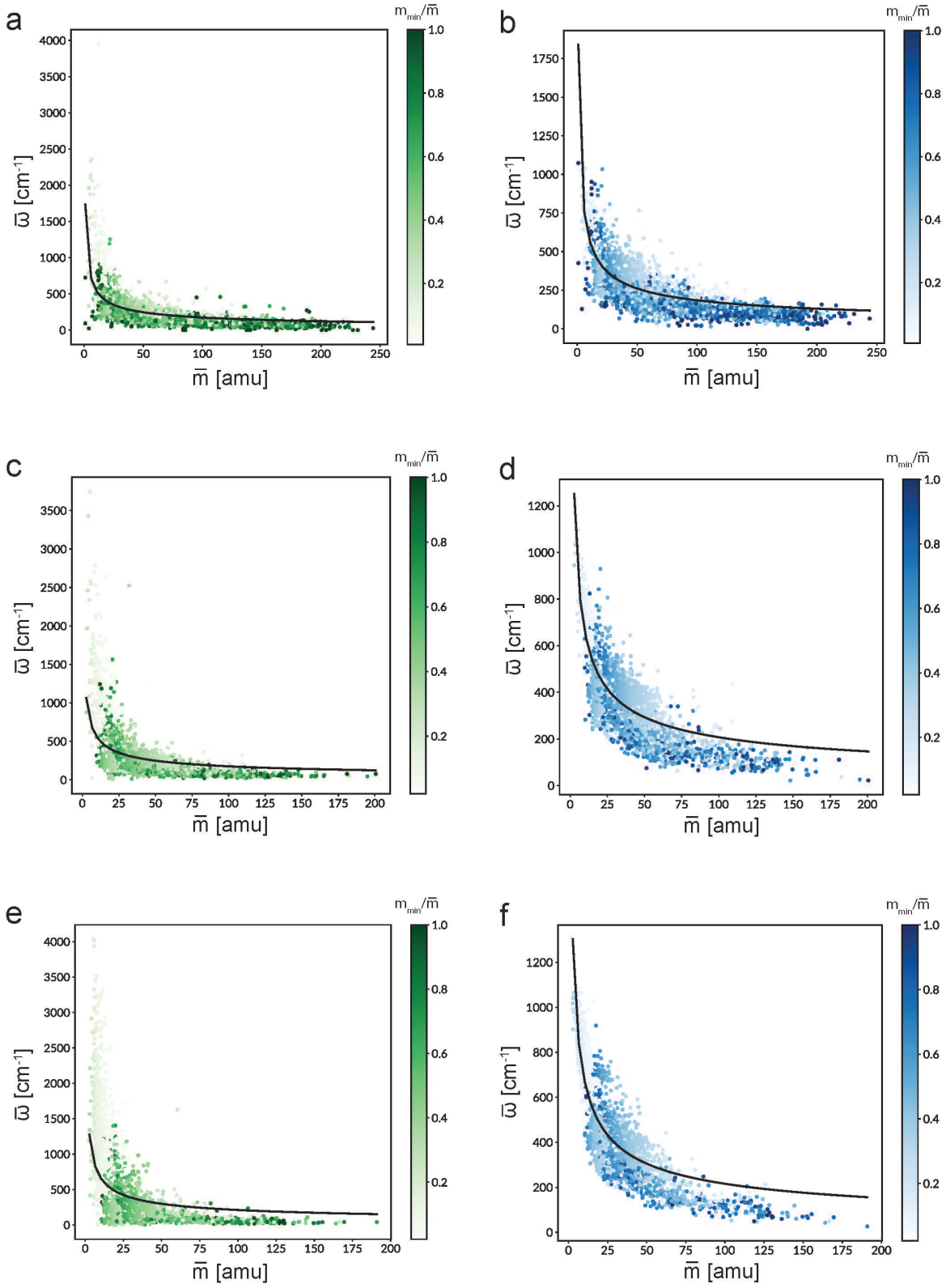

图 S11. 对未见晶体结构的模型预测进行评估。使用 (a.,c.,e.) VVN 和 (b.,d.,f.) MVN，绘制预测的 Γ声子$\bar{\omega}$ 平均频率与平均原子质量 $\bar{m}$ 的关系。随机抽取 5000 个每个晶胞含 N 个原子的结构；(a.,b.) $1 \leq N \leq 20$ (c.,d.) $21 \leq N \leq 40$ (e.,f.) $41 \leq N \leq 80$。黑色实线表示拟合双曲关系 $\bar{\omega} = C\ \bar{m}^{-1/2}$ 的最小二乘拟合线。常数 C 根据拟合估算为：a. C=1743 b. C=1847 c. C=1760 d. C=2073 e. C=2113 f. C=2157。点的颜色表示 $m_{min}/\bar{m}$ 的大小。

IV.3 平均声子能量的逐元素预测

在图 S12-S17中，我们展示了使用 VVN 和 MVN 对训练和测试数据集中出现的所有元素的 γ-声子所绘制的相关性图。背景颜色表示预测损失值（由低到高依次为黄色至蓝色）。我们可以看到，虽然这种关系并不严格，但对于部分元素，预测准确率与周期性存在一定关联（例如从 ^{5}B到 ^{9}F）。此外，与图 S2-S3相比，通常可以发现，图 S12-S17中预测损失较高的元素，其出现频率往往低于误差较小的元素。

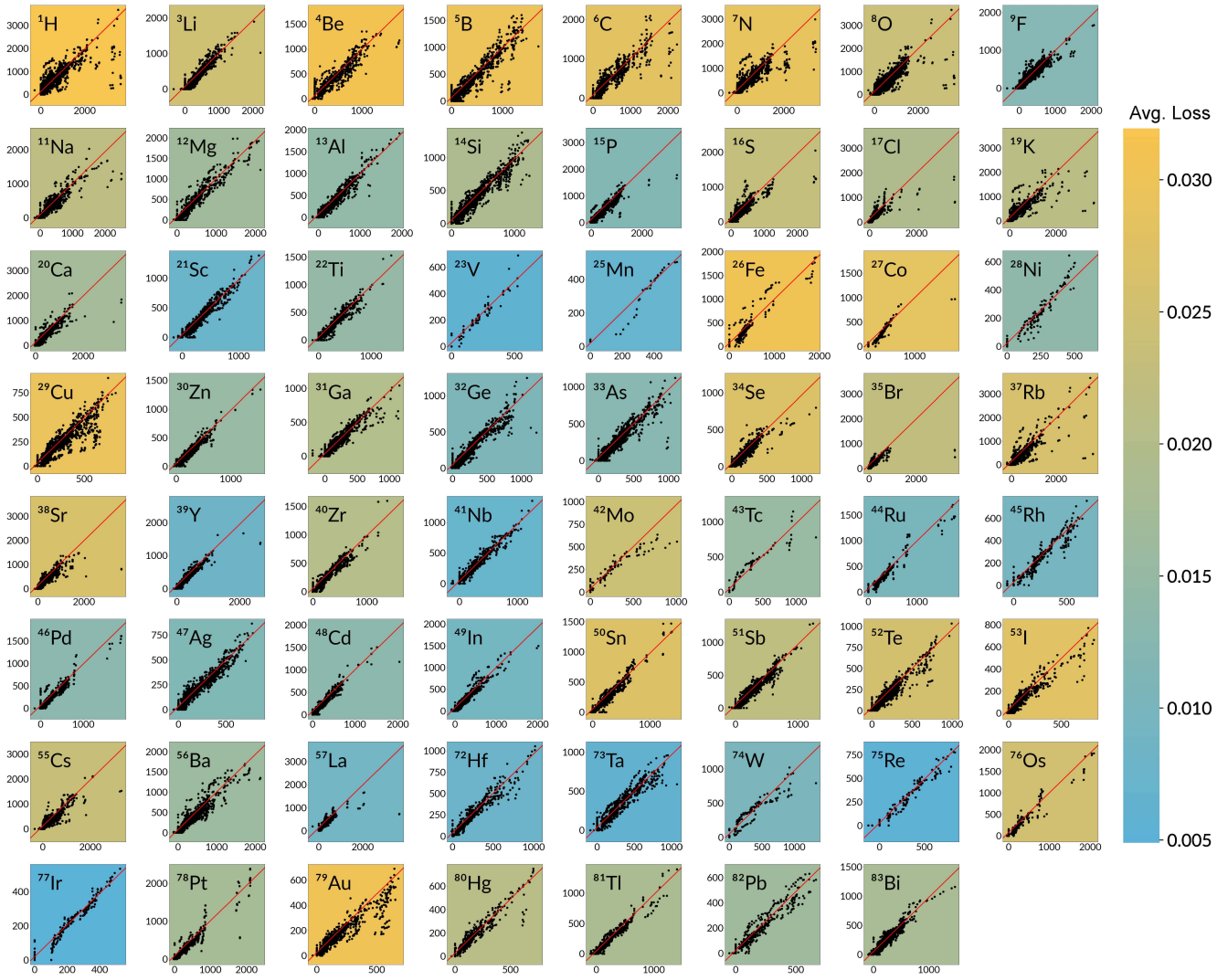

图 S12. 使用 VVN 得到的训练集内预测结果的逐元素相关性图训练集中包含 63 种元素的相关性图。

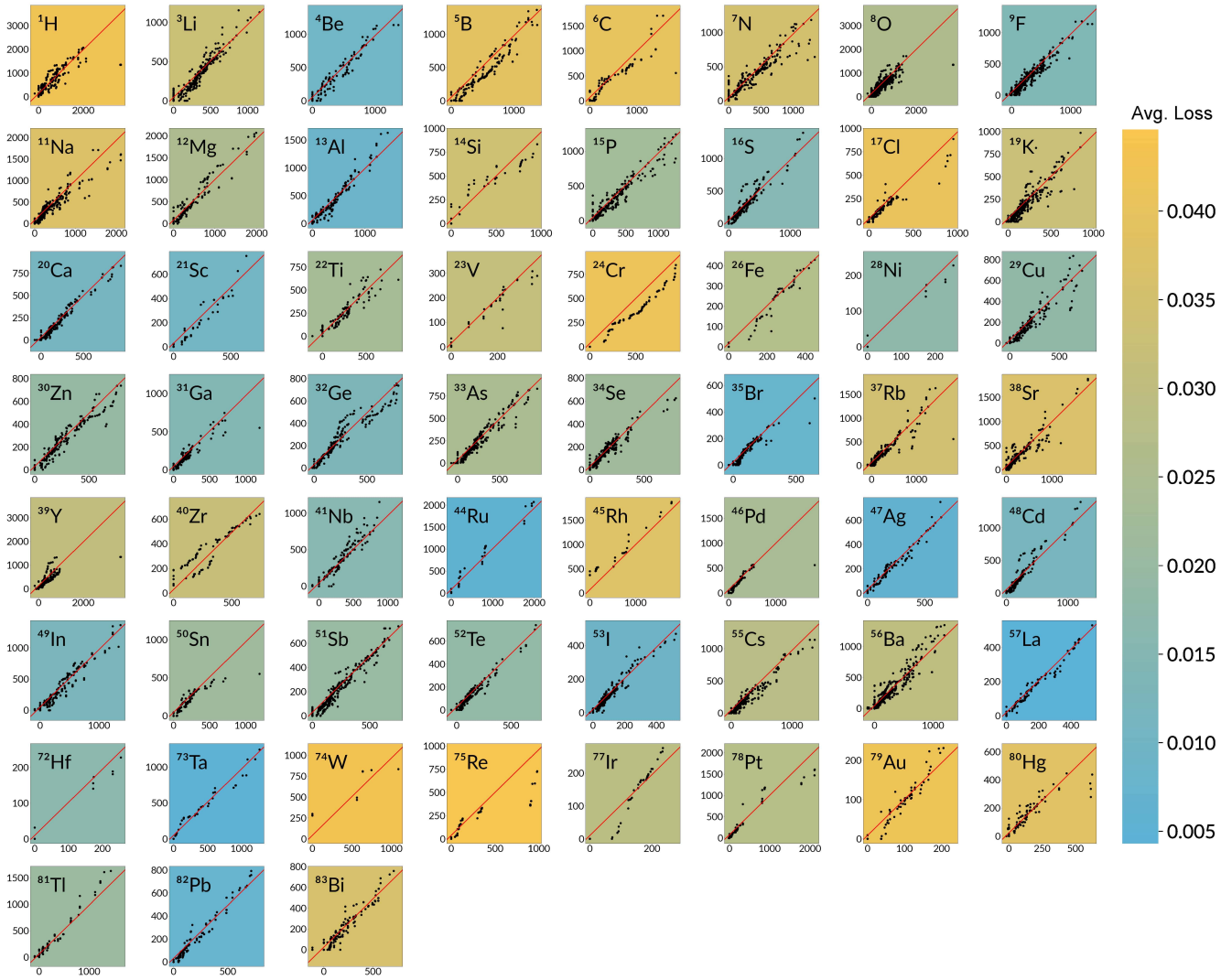

图 S13. 使用 VVN 对测试集内预测结果绘制的逐元素相关性图。测试集中包含 59 个元素的相关性图。

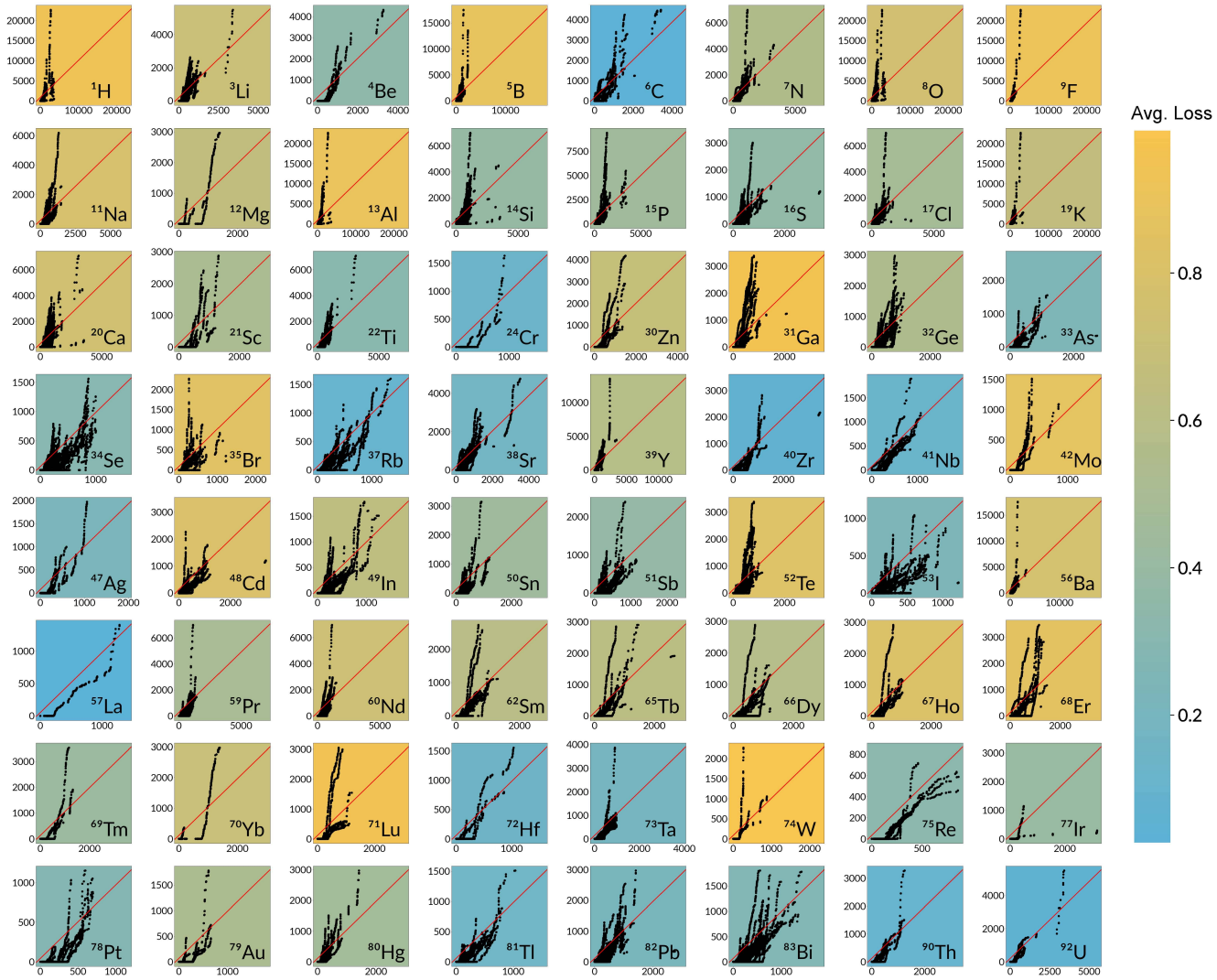

图 S14. 使用 VVN 对 Togo 数据库训练集中的预测结果绘制的逐元素相关性图。复杂材料测试集中含有 64 种元素的相关性图。

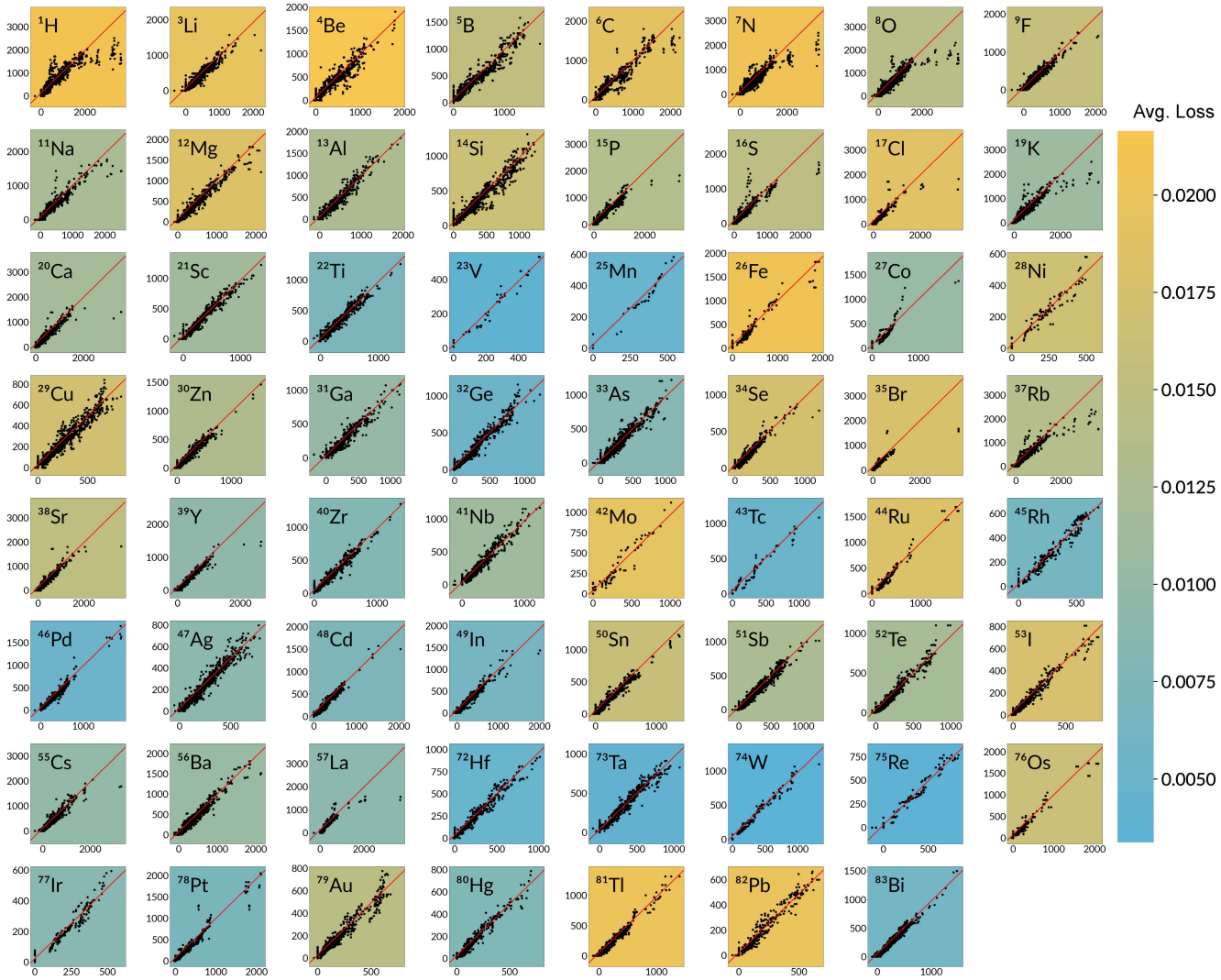

图 S15. 使用 MVN 得到的训练集内预测结果的逐元素相关性图。训练集中约有 63 种元素的相关性图。所绘制的声子频率单位为 $[cm^{-1}]$。

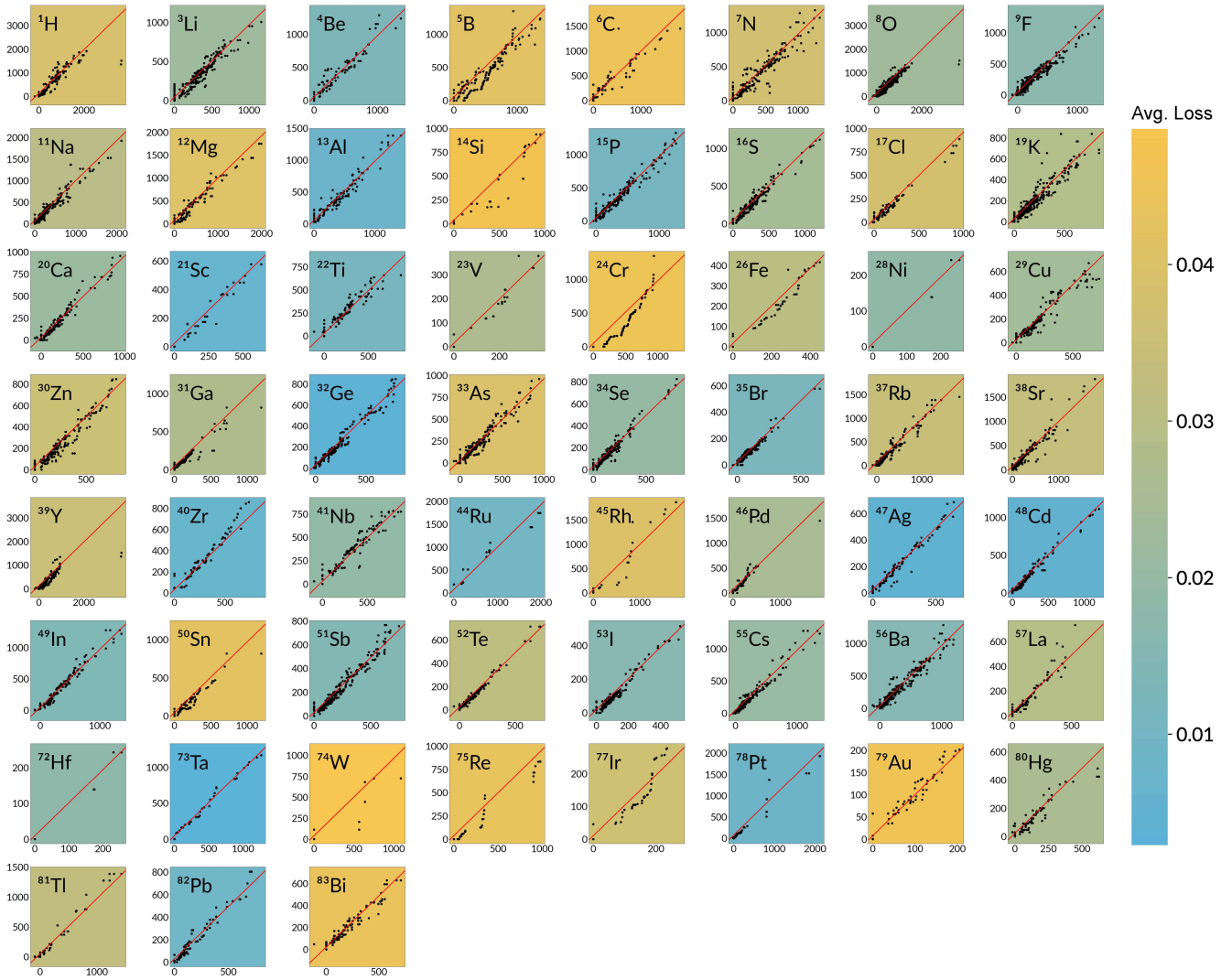

图 S16. 使用 MVN 对测试集中的预测结果绘制的逐元素相关性图。测试集中存在约 59 个元素的相关性图。所绘制的声子频率单位为 $[cm^{-1}]$。

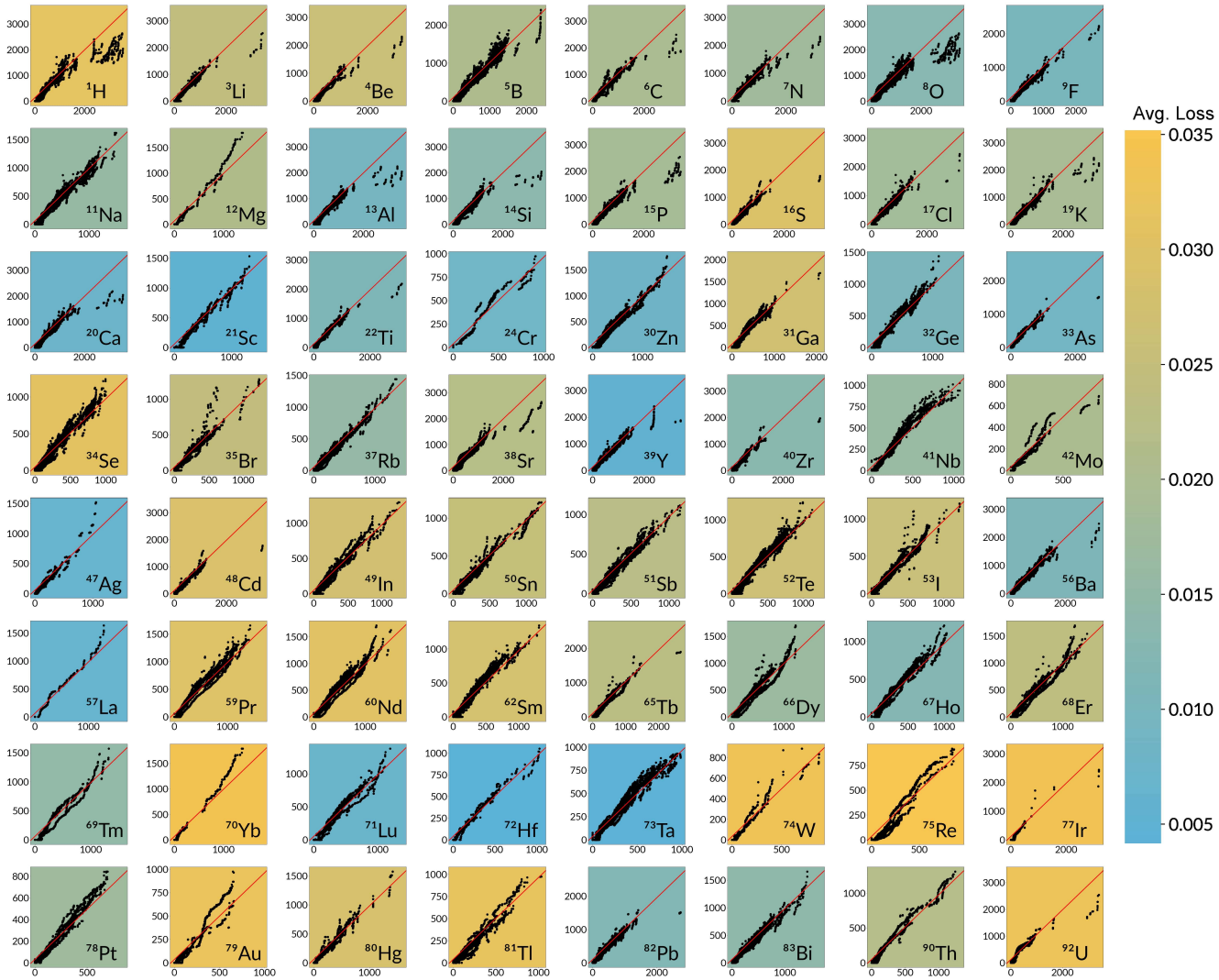

图 S17. 使用 MVN 对 Togo 数据库测试集中的预测结果进行逐元素相关性图分析。复杂材料测试集中约有 64 种元素的相关性图。绘制的声子频率单位为 [厘米$^{-1}$]。

V 完整声子能带结构预测

我们使用*k*-MVN 模型，对主数据库训练和测试数据中的完整声子色散能带，以及 Togo 数据库中的复杂材料进行预测，并展示更多结果。请注意，DFPT 计算中的部分声子值为负，而模型经过设置，只会给出正的声子频率。尽管对某些材料的声子能带预测与 DFPT 计算并不完全一致，但模型很好地捕捉到了能带的某些特征，例如平均频率和带隙。

图S18-S20 提供　　更多使用依赖*k*的矩阵虚拟节点模型预测完整声子能带的结果。

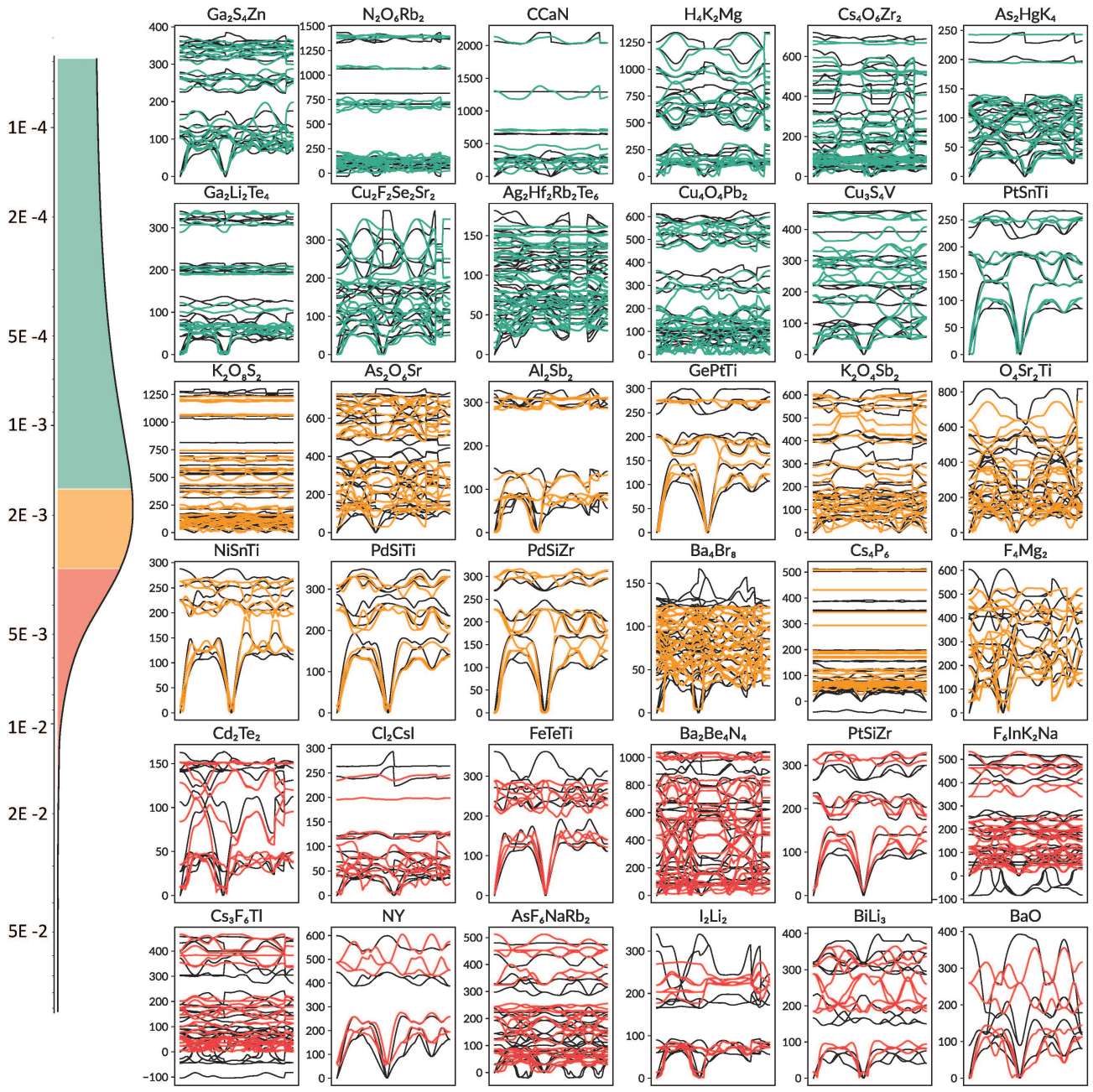

图 S18。 使用*k*-MVN 对主数据库训练集中的声子能带结构进行直接预测的结果。每条能带在布里渊区内包含 100 个 k 点，具体范围因材料的结构而异。通过密度泛函微扰理论计算得到的能带以黑色标示，绿色、黄色和红色线条分别表示第 1 至第 3 四分位区间内预测的声子能带。

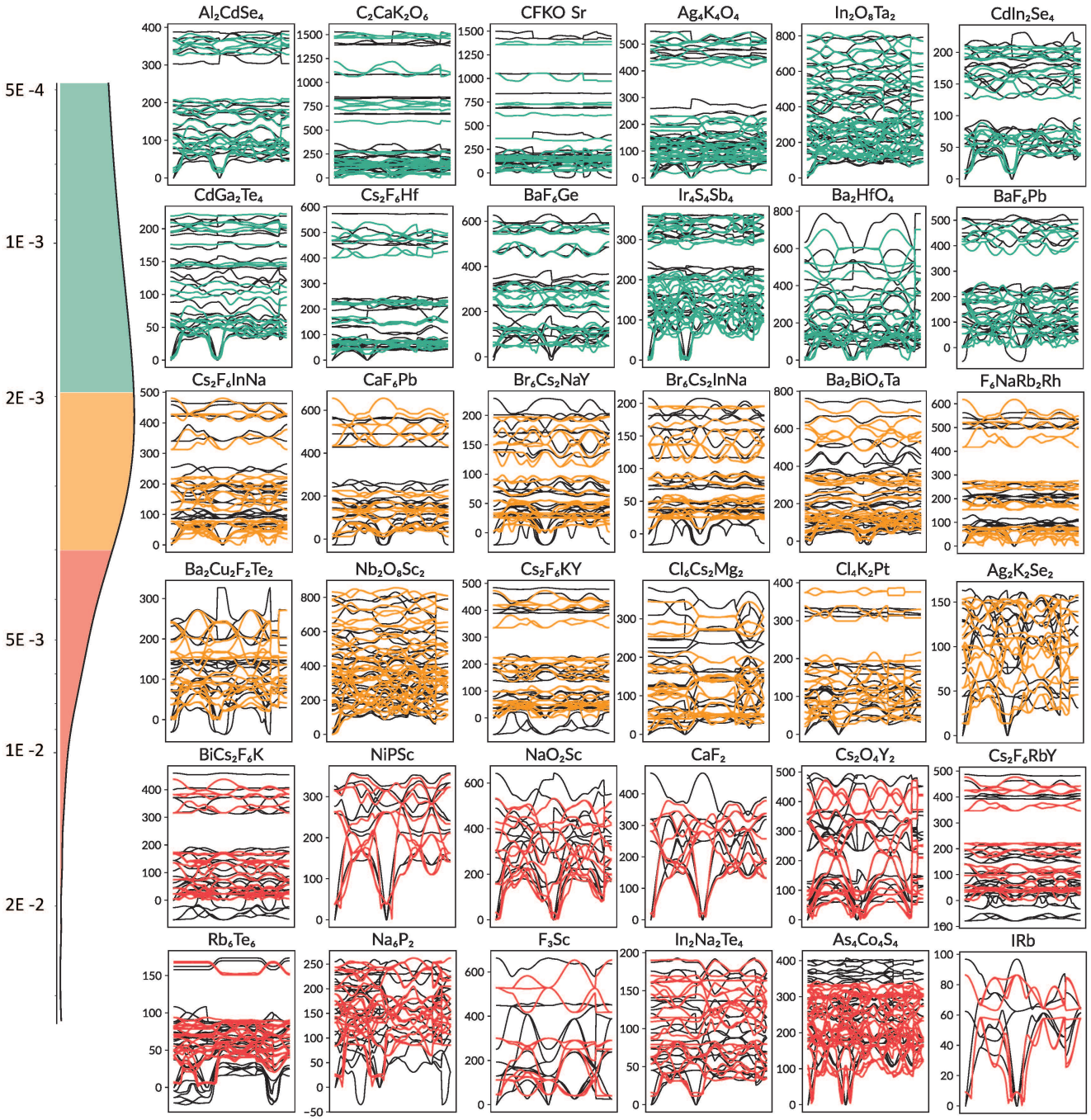

图 S19. 使用*k*-MVN 对主数据库测试集中的声子能带结构进行直接预测的结果。每条能带包含布里渊区中的 100 个 k 点，具体范围因材料的结构而异。通过密度泛函微扰理论计算得到的能带以黑色标示，绿色、黄色和红色线条分别表示第 1 至第 3 四分位区间内预测的声子能带。

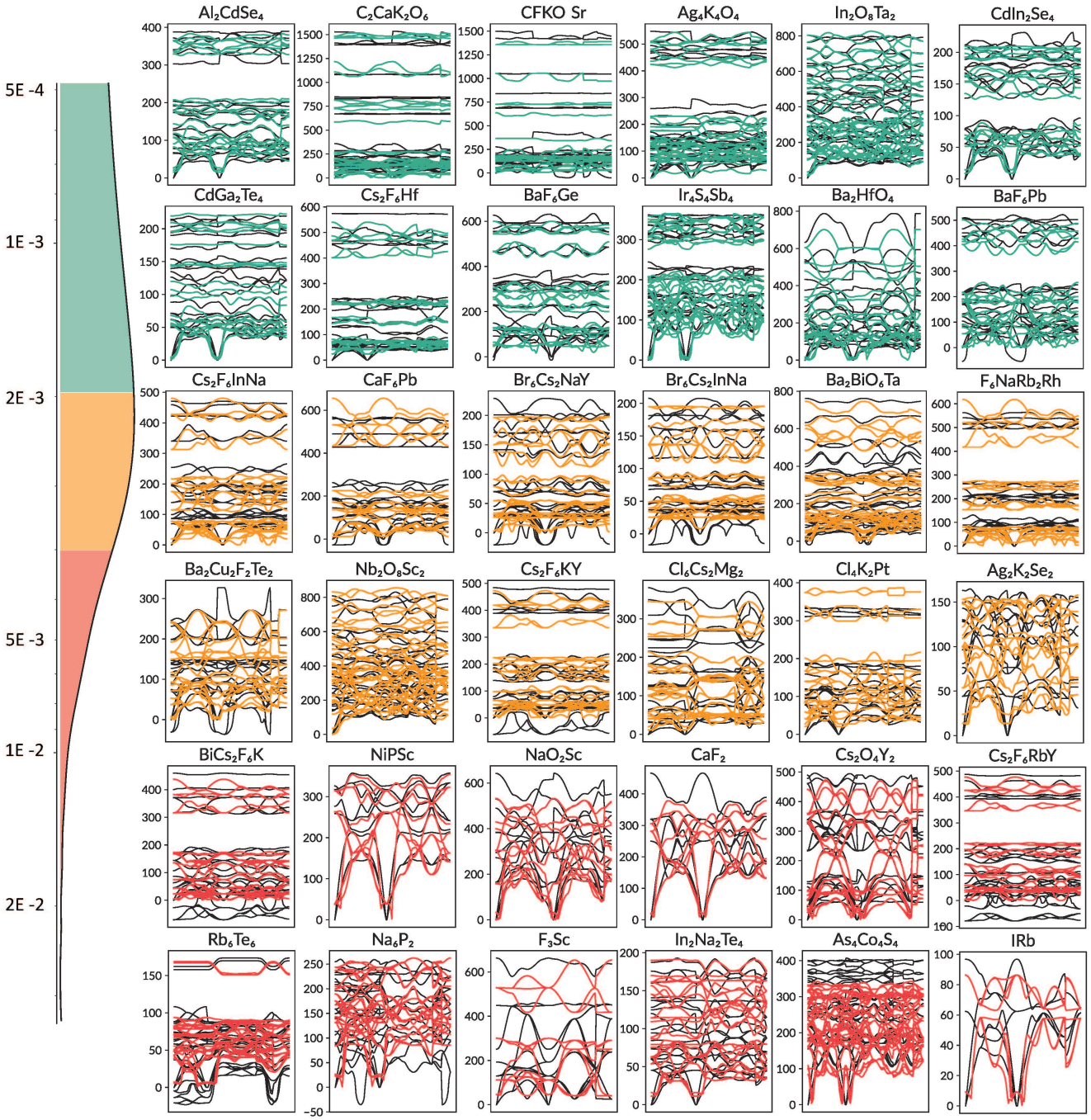

图 S20. 使用*k*-MVN 对 Togo 数据库中的声子能带结构进行直接预测的结果。 每条能带在布里渊区内包含 100 个 k 点；不同材料的布里渊区因其结构而异。通过密度泛函微扰理论计算的能带以黑色标示，绿色、黄色和红色线条分别表示第 1 至第 3 四分位区间内预测的声子能带。

VI 应用

VI.1 拓扑外尔半金属的声子预测

将 k-MVN 模型得到的 LaAlGe Γ-声子预测结果与实验结果进行比较[?]。LaAlGe（一种 II 型外尔半金属）的声子色散通过非弹性中子和 X 射线技术沿高对称点和外尔点测量。

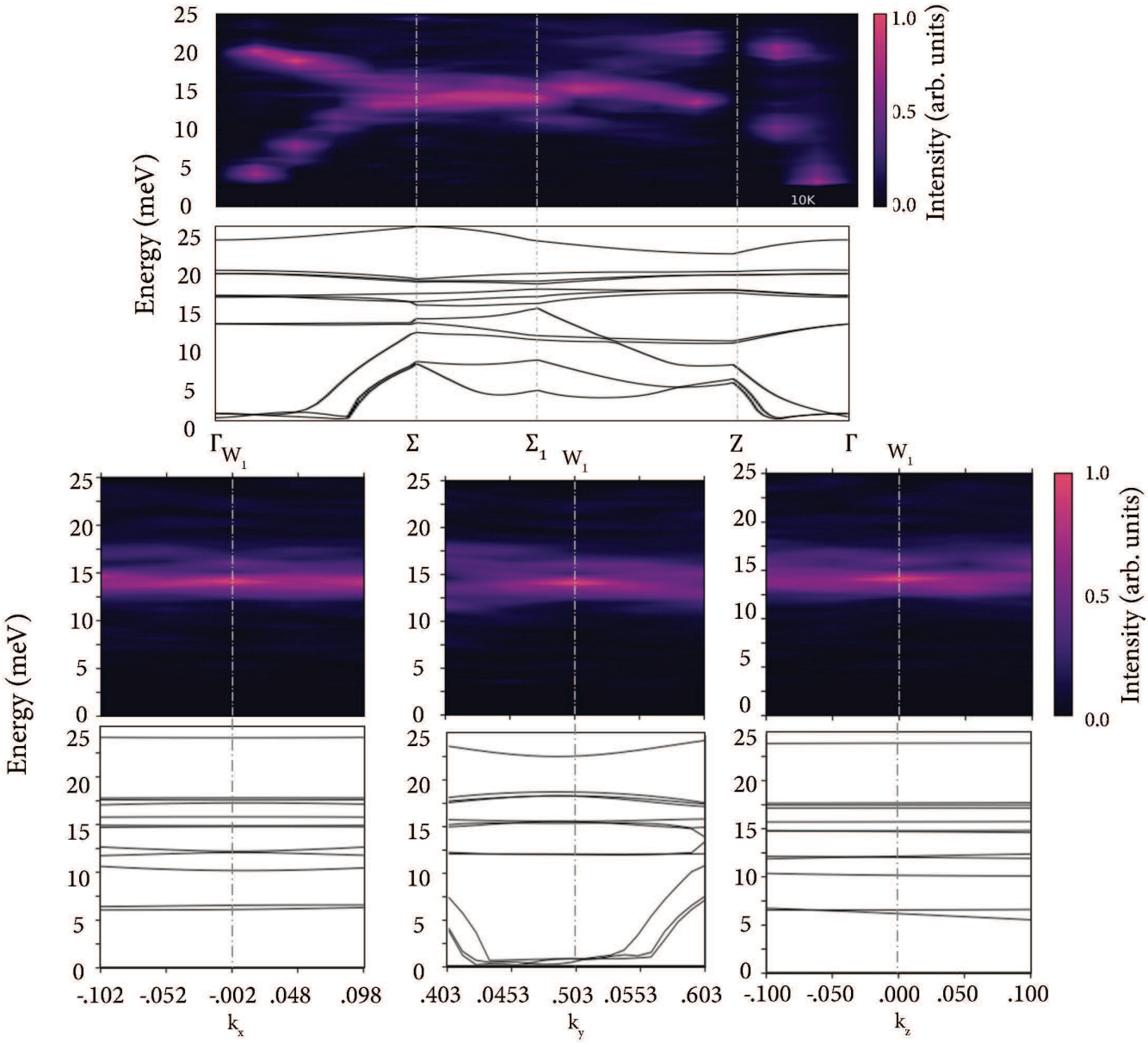

图 S21. 实验与机器学习获取的 LaAlGe 声子谱对比（上）LaAlGe 沿高对称点的声子能带。（下）布里渊区内靠近一个外尔点 W1 处沿 x、y、z 方向的声子能带结构。绘制的声子频率单位为 [毫电子伏]。

VI.2 合金系统中的声子预测

我们的预测模型最重要的应用之一是预测合金系统的声子能带。晶态合金存在一种无序，其原子位置可以呈周期性。然而，由于同一位置上的原子元素分布具有概率性，因此该系统并非周期性系统。因此，在嵌入各位置的原子质量时，我们需要考虑原子组成，这正是虚晶近似（VCA）[?] 所依据的。例如，给定一种组成为$V_pW_{1-p}(0 \leq p \leq 1)$的二元合金，

输入合金编码向量 $\mathscr{X}$可以采用以下形式

$$\mathscr{X} = [0, ..., pm_V, ..., (1-p)m_W, ..., 0] \tag{2}$$

其中，双热编码 pm_V和 $(1-p)m_W$分别位于与 V和W的原子序数对应的向量索引处，并按组成比例加权。根据方程（2）的定义，只需设置 $p = 1$ （或 $p = 0$），即可将嵌入特征直接化简为纯相 V 的独热编码（或 W 的独热编码），并可直接推广到更复杂的合金。

我们通过 SiGe 二元合金展示了这种方法的优势。我们始终假设 SiGe 合金的立方晶胞包含八个原子，边长为 5.466 $^{\circ}$A；它是晶体硅与锗之间的中间组成。作为 kMVN 模型的输入，我们将晶胞转换为包含两个原子的原胞。作为真实标签，我们使用 VCA近似计算得到的声子能带。

图S22 展示了 SiGe 合金的声子能带预测结果。真实值（图S22a）和预测结果（图S22b）沿高对称点呈现出相似的能带结构。黄色和蓝色分别表示硅和锗的组成比例。较轻的原子通常对应较高的声子频率。我们的结果符合这一规律：随着硅的组成比例增大，合金的声子频率也随之升高。

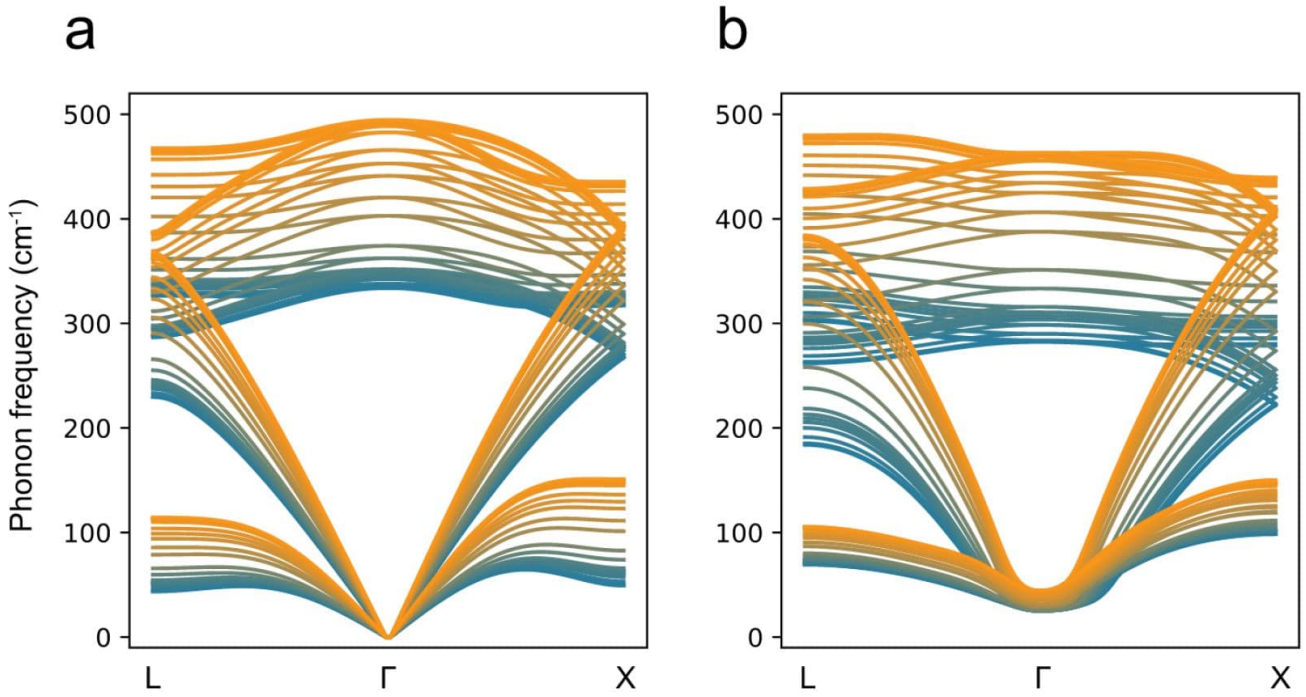

图 S22。 不同组成的 SiGe 合金的声子谱。 黄色和蓝色分别表示纯 Si 和 Ge，中间颜色表示组成比例不同的合金。我们展示了 $p = 0.00$、0.01、0.05、0.10、0.13、0.15、0.17、0.20、0.30、0.40、0.60、0.70、0.80、0.85、0.90、0.96、0.98、0.99、0.995 和 1.00（从蓝色到橙色）时 Si_pGe_{1-p} 的声子预测。我们展示了沿 L、 Γ 和 X点的声子。**a**。真实值。**b**。预测。

接下来，我们展示二元合金 NiCo、NiFe 以及三元合金 NiFeCo 的声子预测结果。对于每种材料，我们随机生成成分并重复 200 次，观察声子预测的表现。作为真实标签（图 S23 中的黑线），我们采用 DFPT近似计算得到的声子能带。作为 DFPT 计算的输入，NiCo 和 NiFe 分别采用含有 64 个原子的超胞，NiFeCo 则采用含有 108 个原子的超胞。图 S23展示了这三种合金的声子预测结果。红、蓝、黄三种颜色分别表示 Ni、Co 和 Fe 的成分。光谱按预测损失从低到高从左至右排列：(a) NiCo、(b) NiFeCo 和 (c) NiFe。这一趋势可由合金组分的原子质量解释。Ni 和 Co 的原子质量几乎相同（分别为 58.69u 和 58.93u），而 Fe 的原子质量较低（55.85 u）。与 NiFe 相比，NiCo 和 NiFeCo 中各元素的原子质量差异较小，因此原子质量编码更加稳定。

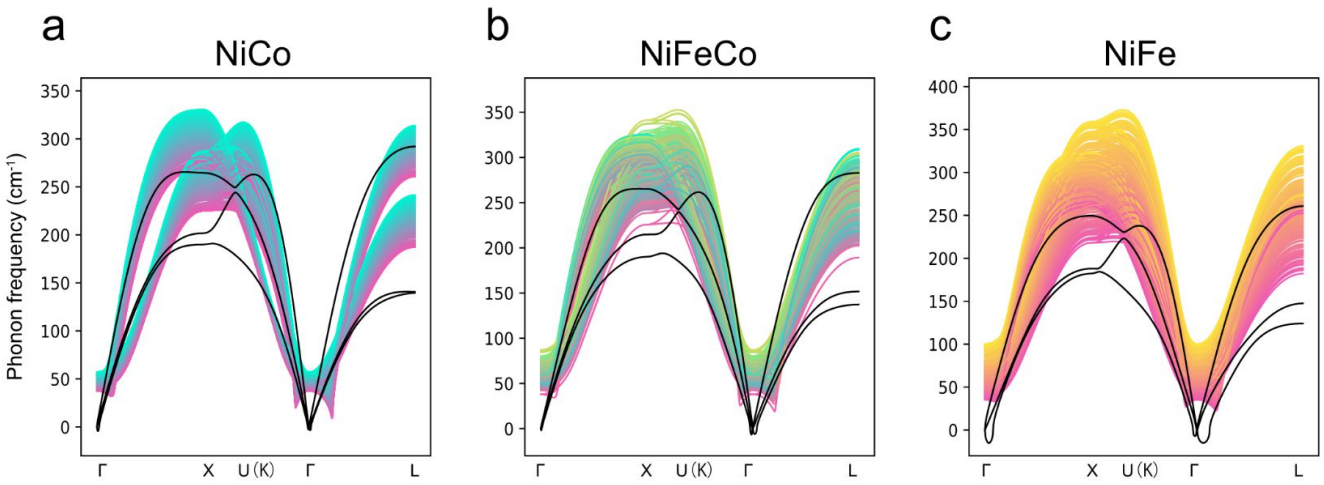

图 S23. 不同组成的高熵合金的声子谱。 红色、蓝色和黄色分别表示纯镍、钴和铁，中间色则表示组成比例不同的合金。每幅图中的黑线是通过密度泛函微扰理论计算得到的真实值。**a**。NiCo。**b**。NiFeCo。**c**。NiFe。

我们已展示，我们的 *k*-MVN 模型能够处理二元和三元合金的声子预测。这里我们进一步展示其对组分更多的合金的处理能力。我们从二元合金开始，接着是三元合金、四元合金，最后是五元合金。图.S24 展示了 MoTaNbWV 和 VWNbTaTi 及其低组分合金的预测光谱。

在高熵合金系统中，力常数和质量都会发生波动。因此，我们需要对这些属性取平均，以用原胞或超胞表示无序态。在图。S24a-d 中，MoTaNbWV 的组元数量从 2 增加到 5。图。S24e-h 展示了 VWNbTaTi 的类似属性。我们的预测遵循公式（2）中的原子质量嵌入，该嵌入根据组态比例对各组元的质量进行加权。即使不纳入力常数平均，我们的模型也能够生成与采用特殊准随机结构（SQS）得到的模拟结果相似的声子谱[?]。图。S24e-h 展示了 VWNbTaTi 的类似属性。

VII Γ-声子数据库

在先前的工作中，展示了 1521 种半导体无机晶体的完整声子能带结构及其衍生量[?]。鉴于机器学习方法能够实现高质量的声子预测，尤其是 MVN 方法适用于复杂材料，我们在此介绍一个新的声子数据库，其中包含截至 2022年 Materials Project（材料项目，MP）中全部 146,323 种材料的声子谱，这些数据由 MVN 方法计算得出。考虑到区中心 Γ声子的重要性，以及它们可通过拉曼散射等更易获取的设备进行测量，我们在此将数据库限定为 Γ声子，并将完整的声子数据库留待未来工作。这个新的 Γ声子数据库为分析许多化合物的晶格动力学提供了可能。它可作为一种有用的工具，与晶体材料的拉曼散射结果进行比较。采用 MVN 方法生成的Γ声子数据库可通过以下网址获取：https://osf.io/k5utb/。

整个数据库以字典形式存储，其键为 MP ID 编号。每种材料对应的字典值也为字典，其中包含每个晶胞中的原子数、化学式等基本信息。表S2列出了各字典中的值。我们通过质点弹簧模型对结果的有效性进行了检验，检验结果汇总于图S25。由于声子频率是整组原子间力常数与构成材料的不同元素质量相互作用的结果，数据围绕双曲线拟合呈现一定的离散分布。可以观察到，组分的质量存在一些相关趋势。质量非均匀的体系（以较小的比值$m_{min}/\bar{m}$表示）往往落在与质量分布较均匀的体系不同的双曲线上。

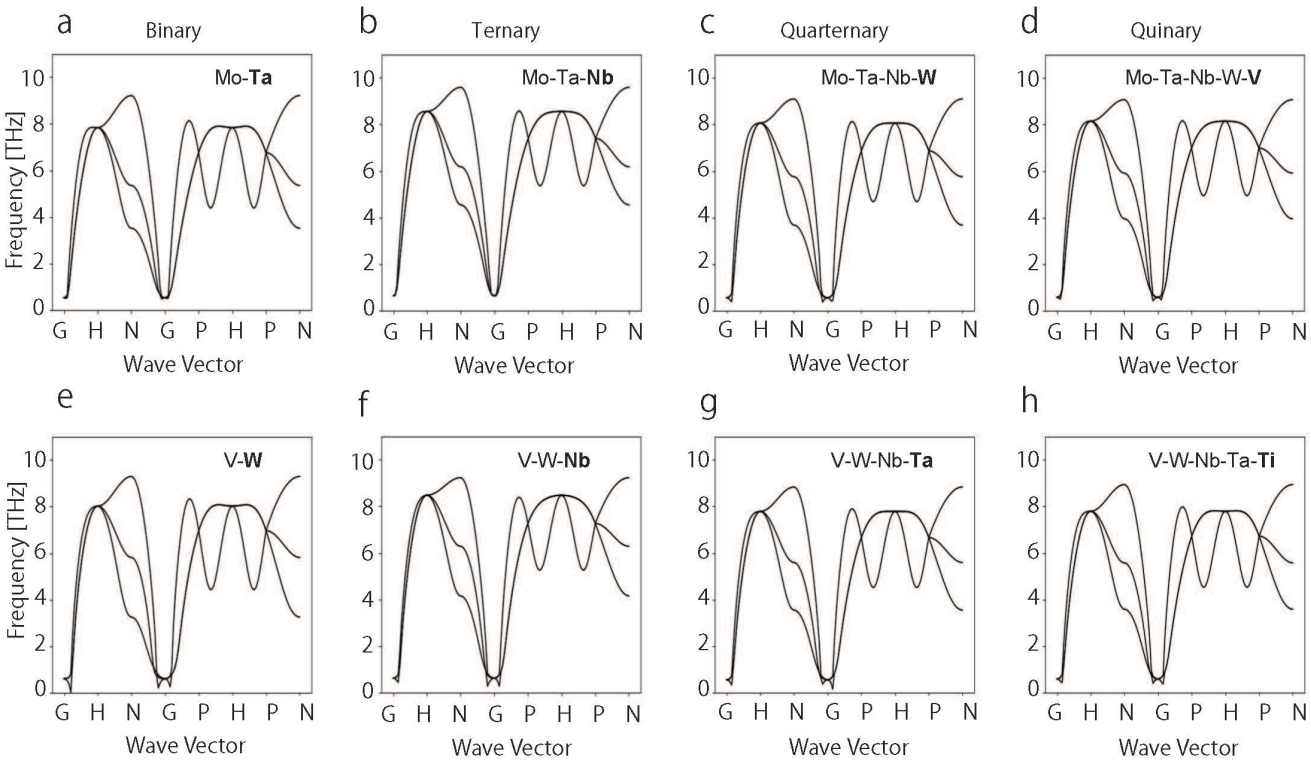

图 S24. 从二元合金到五元高熵合金的声子谱随着组成元素增加而预测的声子谱：从二元合金到五元高熵合金。每种合金从左到右添加的元素以粗体显示。绘制的声子频率单位为 [太赫兹]。**a**。MoTa。**b**。MoTaNb。**c**。MoTaNbW。**d**。MoTaNbWV。**e**。VW。**f**。VWNb。**g**。VWNbTa。**h**。VWNbTaTi。

表 S2. 数据库中每种材料的数据组成

Key	数据类型	描述
材料 ID	字符串	MP ID 编号
晶胞位点数	整数	每个晶胞中的原子数
化学式	字符串	化学式
空间群	整数	空间群编号
结构	字符串	CIF 格式的材料结构
光谱	list	Γ 声子频率的值 [厘米$^{-1}$]

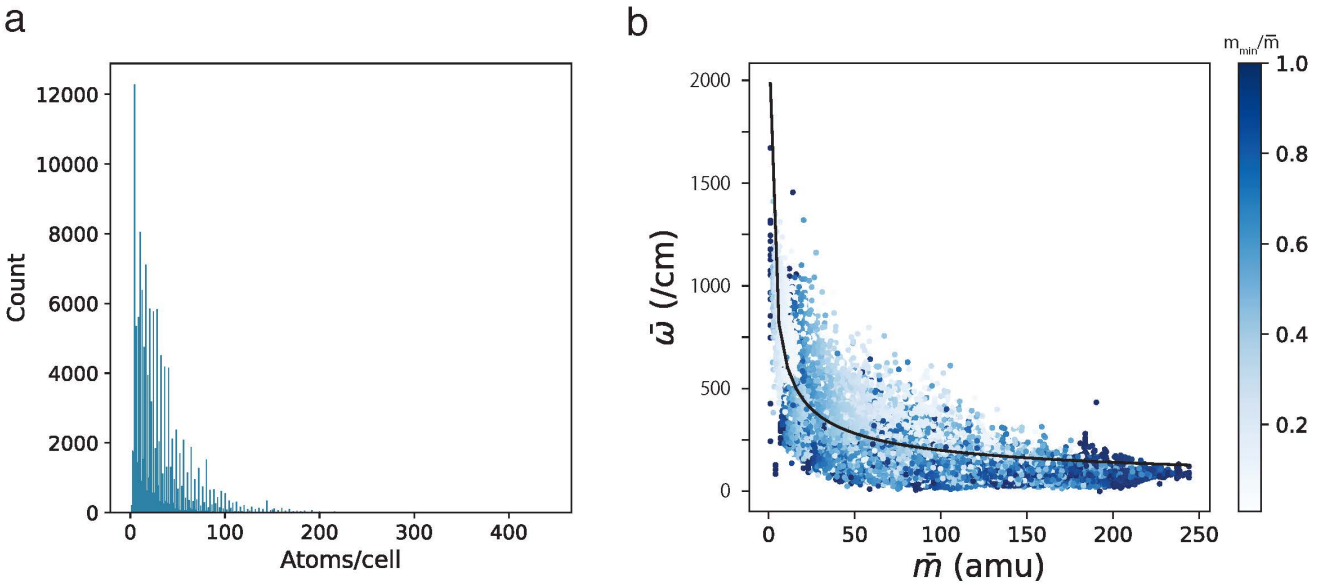

图 S25. 我们的 Γ-声子数据库的概况。a. 数据集中每个晶胞的原子数。该数据集包含 146323 个原子，每个晶胞的原子数范围为 1 至 444。**b.** 预测的 Γ-声子$\bar{\omega}$平均频率与平均原子质量 $\bar{m}$ 的关系。点的颜色表示 $m_{min}/\bar{m}$ 的大小。黑色实线表示对双曲关系 $\bar{\omega} = C\ \bar{m}^{-1/2}$ 进行最小二乘拟合的结果。拟合估算出的常数 C 为 1993。点的颜色表示比值 $m_{min}/\bar{m}$。