Machine Learning: Science and Technology

PAPER • OPEN ACCESS

Self-supervised learning of materials concepts from crystal structures via deep neural networks

To cite this article: Yuta Suzuki *et al* 2022 *Mach. Learn.: Sci. Technol.* **3** 045034

MACHINE
LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
1 July 2022

REVISED
17 October 2022

ACCEPTED FOR PUBLICATION
11 November 2022

PUBLISHED
29 December 2022

PAPER

Self-supervised learning of materials concepts from crystal structures via deep neural networks

Yuta Suzuki[1,2,6], Tatsunori Taniai[3], Kotaro Saito[2,4], Yoshitaka Ushiku[3] and Kanta Ono[1,2,5,*]

[1] The Graduate University for Advanced Studies (SOKENDAI), Ibaraki, Japan
[2] Institute of Materials Structure Science (IMSS), High Energy Accelerator Research Organization (KEK), Ibaraki, Japan
[3] OMRON SINIC X Corporation, Tokyo, Japan
[4] Randeft, Inc., Tokyo, Japan
[5] Department of Applied Physics, Osaka University, Osaka, Japan
[6] Current affiliation: Advanced R&D and Engineering Company, TOYOTA MOTOR CORPORATION, Shizuoka, Japan.
[*] Author to whom any correspondence should be addressed.

E-mail: ono@ap.eng.osaka-u.ac.jp

Keywords: materials informatics, deep metric learning, crystal structure, self-supervised learning

Supplementary material for this article is available online

Abstract

Material development involves laborious processes to explore the vast materials space. The key to accelerating these processes is understanding the structure-functionality relationships of materials. Machine learning has enabled large-scale analysis of underlying relationships between materials via their vector representations, or embeddings. However, the learning of material embeddings spanning most known inorganic materials has remained largely unexplored due to the expert knowledge and efforts required to annotate large-scale materials data. Here we show that our self-supervised deep learning approach can successfully learn material embeddings from crystal structures of over 120 000 materials, without any annotations, to capture the structure-functionality relationships among materials. These embeddings revealed the profound similarity between materials, or 'materials concepts', such as cuprate superconductors and lithium-ion battery materials from the unannotated structural data. Consequently, our results enable us to both draw a large-scale map of the materials space, capturing various materials concepts, and measure the functionality-aware similarities between materials. Our findings will enable more strategic approaches to material development.

1. Introduction

The diverse properties of the inorganic materials originate from their crystal structures, i.e. the atomic-scale periodic arrangements of elements. How structures determine low-level material properties such as the band gap and formation energy is well studied as the structure-property relationship [1, 2]. On the other hand, the materials science literature often discusses 'superconductors' [3], 'permanent magnets' [4], or 'battery materials' [5], referring to their higher-level properties, or *functionality*. Nevertheless, understanding what structures exhibit such functionality, or understanding the structure-functionality relationship, is a fundamental question in materials science. We call this functionality-level material similarity 'materials concepts'. Traditionally, materials science has sought new materials by experimentally and theoretically understanding specific functionalities of materials in a bottom-up fashion [1–5]. However, this labour-intensive narrowly focused analysis has prevented us from grasping the whole picture of the materials space across various materials concepts. For next-generation material discovery based on the structure-functionality relationship, we argue here the need for a top-down unified view of crystal structures through materials concepts. We pursue this ambition by learning a latent representation space of crystal structures. Thus, this representation space should ideally both (a) recognise materials concepts at scale and

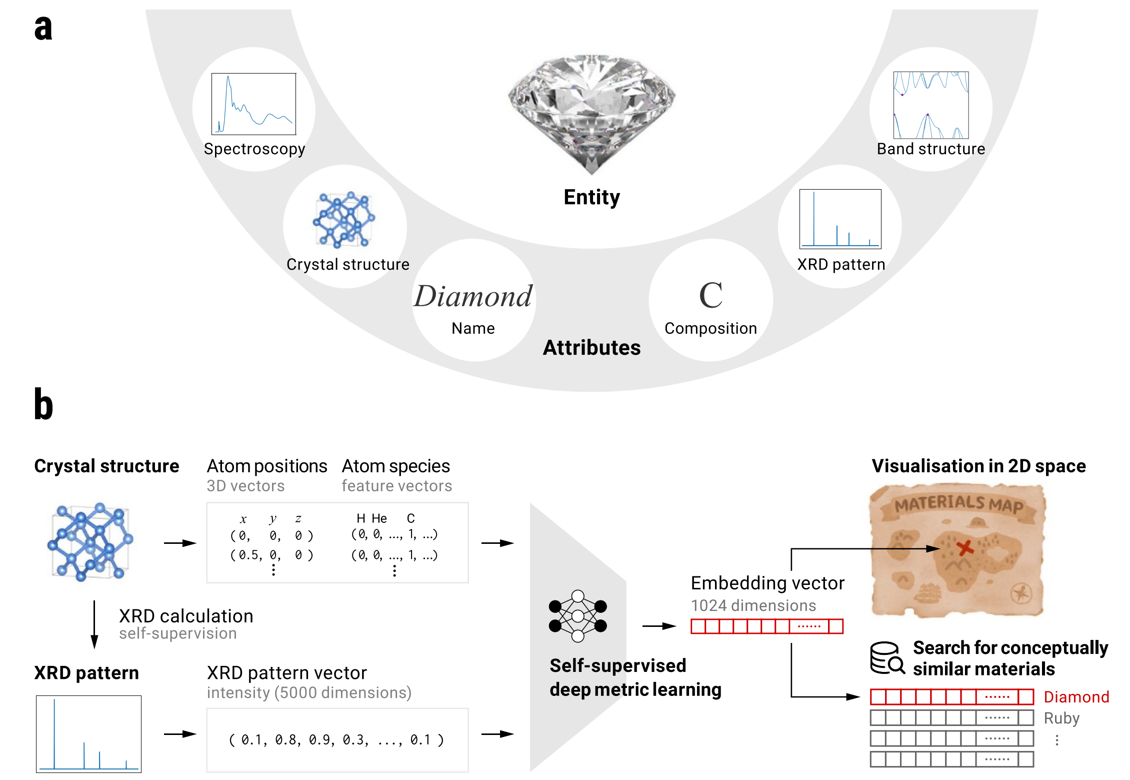

Figure 1. Our strategy for learning materials concepts. (a) Diamond as an example material and its various structural attributes showing different information of diamond in different data forms, or modalities. Since each attribute has its own advantages and disadvantages in expressing a material, using multiple attributes for a material can provide a more comprehensive view of the material. Particularly, the combination of the crystal structure and the x-ray diffraction (XRD) pattern, which we employ in this study, is known to well reflect two complementary structural features of materials, the local structure and the periodicity [1]. (b) Our goal is to represent each material as an abstract constant size vector (embedding) whose distances to other embeddings reflect conceptual (functionality-level) similarities between materials. These embeddings allow us to visualise the materials space intuitively and also to search for conceptually similar materials given a query material. We learn embeddings from pairs of crystal structures and XRD patterns in the framework of deep metric learning. This cross-modal learning approach trains deep neural networks by teaching them that each pair should represent the same material entity. Because the XRD pattern can be theoretically calculated from the crystal structure, this learning can be performed in a self-supervised manner without any explicit human annotations for the materials dataset.

(b) be equipped with a functionality-level similarity metric between materials. We here utilise multi-modal structural attributes of materials to effectively capture structural patterns correlated to material functionality (figure 1). The underlying hypothesis here is that materials concepts are the intrinsic nature of crystal structures, and therefore, deeply analysing the structural similarity between materials will lead to capturing functionality-level similarity.

Figure 3(a) highlights key results by our representation space, which maps the crystal structures of materials to abstract 1024-dimensional vectors. For visualisation, these vectors were reduced to 2D plots in the figure using a dimensionality reduction technique called t-distributed stochastic neighbour embedding (t-SNE) [6]. We target 122 543 inorganic materials registered in the Materials Project (MP) database (amounting to 93%) to capture nearly the entire space of practically known inorganic materials. These crystal structures themselves contain information about their functionalities implicitly. However, they do not explicitly tell us what structural patterns lead to specific material functionalities such as superconductivity due to complicated structure-functionality relationships. Nevertheless, these materials form clusters of various materials concepts in the space (see annotations in figure 3(a)), showing the success of our representation space capturing structural patterns correlated to material functionality.

When analysing diverse relationships entangled with complex features in large-scale data, machine learning (ML) and deep neural networks (DNNs) are key technologies [7, 8]. Indeed, these technologies often surpass human ability in recent materials informatics work. For example, when extracting features from materials data for complex tasks such as physical property prediction, learning-based descriptors [9–18] have been shown to outperform traditional hand-crafted descriptors [19–24].

Likewise, representation learning is gaining attention for understanding human-incomprehensible large-scale materials data [25–29], visualising the materials space [25, 27–30], and generating crystal structures [31–36]. These material representations aim to map the abstract, comprehensive information of each material into a vector called an 'embedding'. Our work has the same purpose as that of embedding learning. However, to date, neither descriptor nor embedding learning explicitly learns the underlying relationships or similarities between materials. Particularly, existing embeddings [25–29] are learned indirectly as latent feature vectors in an internal layer of a DNN by addressing a surrogate training task (e.g. the prediction of physical properties [27, 28], a task of natural language processing (NLP)[26] or its variants [25, 29]). In such an approach, it is unclear from which layer we should obtain the latent vectors or which metric we should use to measure the distance/similarity between them.

Capturing abstract concepts of materials via learning structural similarities between them is analogous to word embedding learning [26] in NLP. Similar to materials concepts, meanings of words in natural languages often reside in complex and abstract notions, which prevent us from acquiring precise definitions for them. Word embeddings then attempt to capture individual word concepts, without being explicitly taught, by absorbing our word notions implicitly conveyed in the contexts provided by a large-scale text corpus. Once optimised, similarities/distances between embeddings express their concepts, e.g. the embedding of 'apple' will be closer to those of other fruits such as 'grape' and 'banana' than 'dog' or 'cat'. Our crystal structure embedding shares a similar spirit with word embedding in that both attempt to capture abstract concepts via learned similarities. More importantly, we exploit a large-scale material database as a *corpus of materials* that implicitly conveys important structural patterns in its *contexts of crystal structures*, as analogously to word embedding. These structure instances of diverse kinds of materials, even without explicit annotations about their properties, should contain tacit but meaningful information about physics and material functionality that can guide the learning of ML models. From an ML perspective, such a learning strategy is called *self-supervised learning* [37], in which the data of interest themselves provide supervision.

In this study, we demonstrate the large-scale self-supervised learning of material embeddings using DNNs. In essence, we follow the principle that the structure determines properties and aim to discover materials concepts purely from crystal structures without explicit human supervision in learning. To this end, we use a collection of crystal structures as the only source of training data and do not provide any annotation regarding specific material properties (e.g. class labels such as 'superconductors' and 'magnets', or property values such as superconducting transition temperature and magnetisation). Furthermore, unlike existing methods for material embedding learning, we explicitly optimise the relationships between embeddings by pioneering the use of *deep metric learning* [38]. Metric learning is an ML framework for learning a measure of similarity between data points. Unlike the common practice of metric learning performed in a supervised fashion using annotated training data [38], we allow our ML model to be learned from the unannotated structural data in a self-supervised fashion.

2. Results

Our key idea for self-supervised learning, illustrated in figure 2(a), is to learn unified embedding representations for paired inputs expressing two complementary structural features characterising materials: the local structure and the periodicity [1]. In our model, the local structure is represented by a graph whose nodes and edges stand for the atoms and their connections. The periodicity is represented by a simulated x-ray diffraction (XRD) pattern, which can be theoretically calculated from the crystal structure using Bragg's law and Fourier transformation [1]. We simultaneously train two DNN encoders by enforcing them to produce consistent embeddings across the two different input forms. This training strategy follows a simple optimisation principle: (a) for a positive pair in which the input crystal structure and XRD pattern come from the same material, the Euclidean distance between two embedding vectors is decreased, and (b) for a negative pair in which these inputs come from different materials, the distance is increased. We implement this principle in the form of a bidirectional triplet loss function, as illustrated in figure 2(b). For the detailed method protocol, see section 5.

By design, we minimise human knowledge of specific materials concepts in both the data source and training process, with the belief that materials concepts are buried in crystal structures. This design principle enhances the significance of the resulting embedding highlighted earlier (figure 3(a)). It captures profound materials relationships through simple data and optimisation operations considering only general and elementary knowledge of materials such as crystallographic data and Bragg's law. The results suggest that materials concepts can be exposed in deeply-transformed abstract expressions unifying the complementary factors, i.e. the local structure and periodicity, of crystal structures.

The following analyses examine the embedding characteristics more carefully to see if the embedding space has the two desired features mentioned above. Specifically, we qualitatively analyse (a) the global embedding distribution using t-SNE visualisation and (b) the local neighbourhoods around some important materials using the learned similarity metric between crystal structures. In the latter, a superconductor (Hg-1223), a lithium-ion battery material ($LiCoO_2$), and some magnetic materials serve as our benchmark

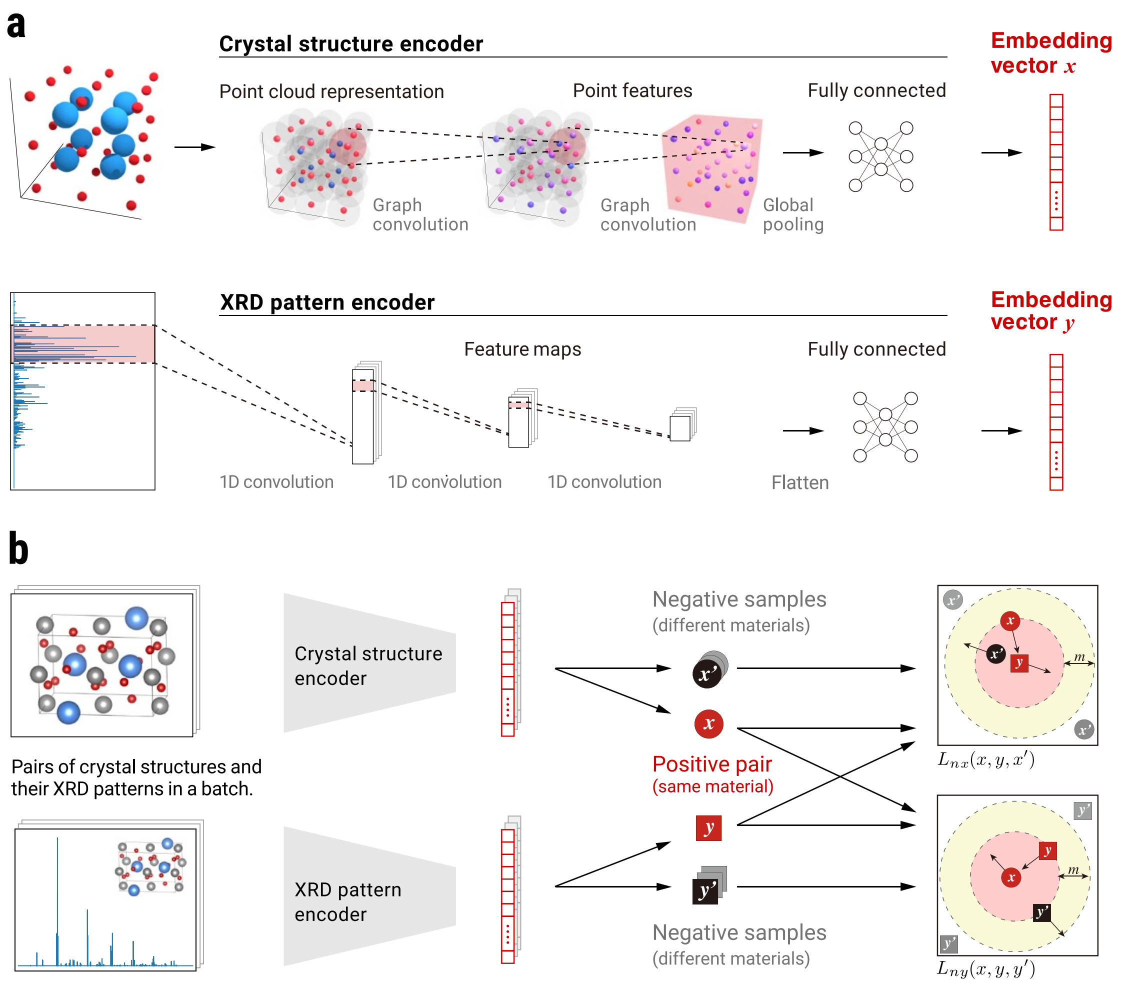

Figure 2. Our self-supervised deep metric learning for materials. (a) The proposed network architecture, including two separate DNNs to encode the crystal structures and XRD patterns into unified embedding vector representations. To account for respective input data forms (figure 1(b)), the crystal-structure encoder employs a DNN for graphs while the XRD pattern encoder employs a 1D convolutional neural network. (b) A schematic view of our bidirectional triplet loss. This triplet loss is used to simultaneously train the two DNN encoders to output embeddings that are close together when the input crystal structures and XRD patterns are paired (red-coloured $\boldsymbol{x}$ and $\boldsymbol{y}$) and far from one another when the inputs are not paired ($\boldsymbol{x}$ and $\boldsymbol{y}$ vs others). More details are given in section 5.

materials because of the high social impacts and the diverse properties yet complex structures of these material classes. These analyses also demonstrate the usefulness of our materials map visualisation and similarity metric for material discovery and development.

2.1. Global distribution analysis

Careful inspection of the embedding space (figure 3(a)) reveals various clusters consistent with our knowledge of materials. Here, we note several interesting examples. A series of clusters corresponding to double perovskites ($A_2BB'X_6$) with different anions, X, exists along the left edge and at the centre of the map, forming a family of materials with the same prototypical crystal structure. This layout suggests that our model captures the structural similarity while properly distinguishing the local atomic environment at each site. At the lower left of the map, well-known 2D materials (transition metal dichalcogenides) form clusters in accordance with their atomic stacking structures [39]. At the top edge lies a cluster of imaginary unstable materials with extremely low-density structures (see also figure 5(a) for more details), representing one of the simplest cases of crystal structures governing physical properties. This cluster of unstable materials is an example showing that our embeddings capture materials characteristics solely from crystal structures without any explicit annotation given for training.

One exciting finding from this map is a cluster of cuprate superconductors at the left edge. This cluster includes the first-discovered copper oxide superconductor, the La–Ba–Cu–O system, and the well-known high-transition-temperature (T_c) superconductors YBCO ($YBa_2Cu_3O_7$ or Y-123), which are located close to La–Ba–Cu–O. These celebrated superconductors share a common structural feature, a CuO_2 plane, that is vital to their superconductivity [3]. The formation of this cluster suggests that our embeddings recognise this

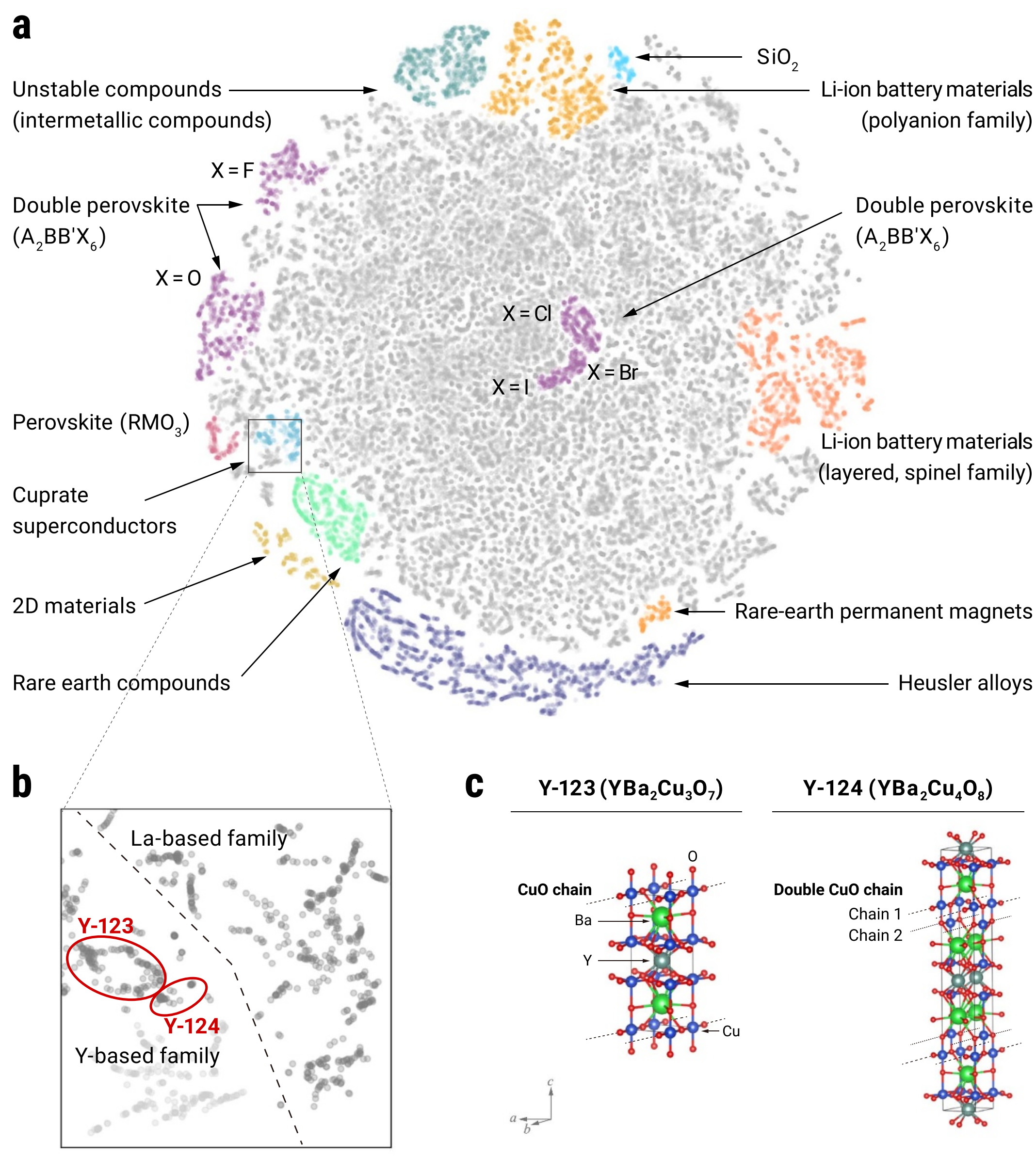

Figure 3. A map of the materials space. (a) A global map of the materials space, plotted via a t-SNE visualisation of the embeddings. Each point corresponds to an individual material that is encoded by the trained crystal-structure encoder. The map was annotated with cluster labels through manual inspection. Our materials map is publicly accessible via an interactive website (https://github.com/quantumbeam/materials-concept-learning). (b) A close-up view of a cluster of cuprate superconductors in the materials space. (c) Crystal structures of Y-123 and Y-124 families, which are closely distributed in (b). The CuO chain in Y-123 and the double CuO chain in Y-124, which is similar to vertically repeating Y-123, are important features of YBCO superconductors.

hallmark structural feature. A closer look at this cluster (figure 3(b)) further reveals the presence of subclusters with structural features linking them. Y-123 and its variant Y-124 have a non-trivial structural similarity related to the CuO chain (see figure 3(c)). In addition, we confirmed other major cuprate superconductors containing Bi, Tl, Pb, or Hg form respective clusters in accordance with their local structures called 'block layers', a key structural concept for understanding the underlying physics of cuprate superconductors [40]. The proximity of these materials on the map further supports the claim that the embeddings capture the structural characteristics and, consequently, the structure-functionality relationships between cuprate superconductors.

These findings naturally lead us to the idea that the map might be able to identify potential superconductors or other beneficial compounds that have not yet been recognised. We leave this idea as an open question and have set up a project website where anyone can dig into the embedding map to search for, or rediscover, potential compounds with preferable functionality.

The t-SNE visualisation also provides a macroscopic perspective on the materials space based on the crystal structure. The simplest indicator of success for this model is the distribution of the elements within

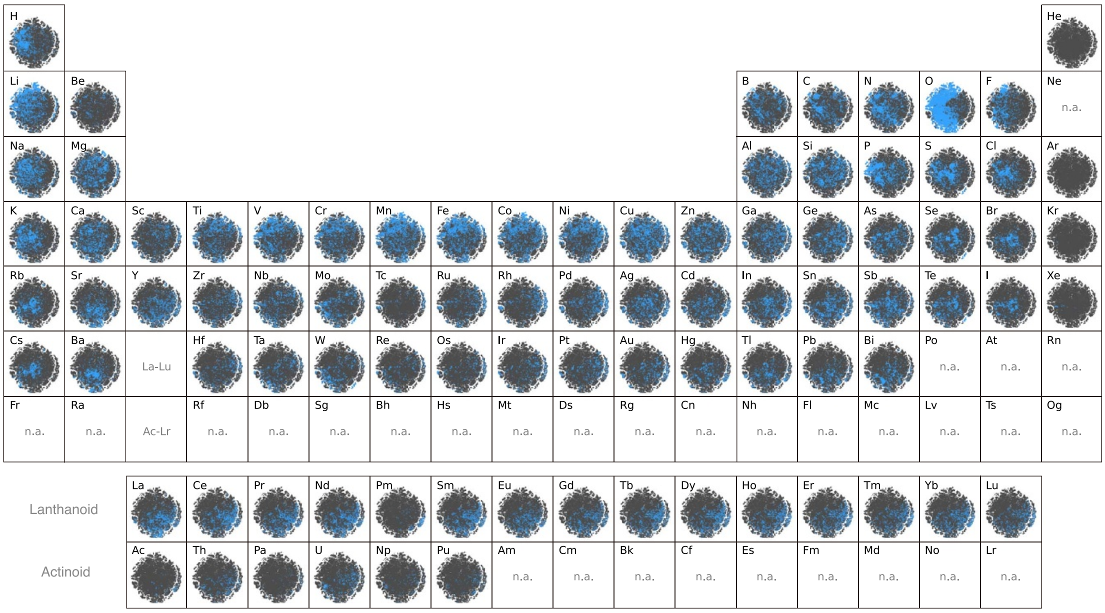

Figure 4. Elemental distributions within the material embedding space. t-SNE plots of the embeddings are laid out on the periodic table, coloured blue or grey according to whether each material contains the corresponding element or not. Similar distributions in the vertical and horizontal directions (groups and periods) of the table indicate that the embeddings successfully capture the similarities of roles between elements in crystal structures. ‘n.a.’ means no material containing the element is found in our dataset.

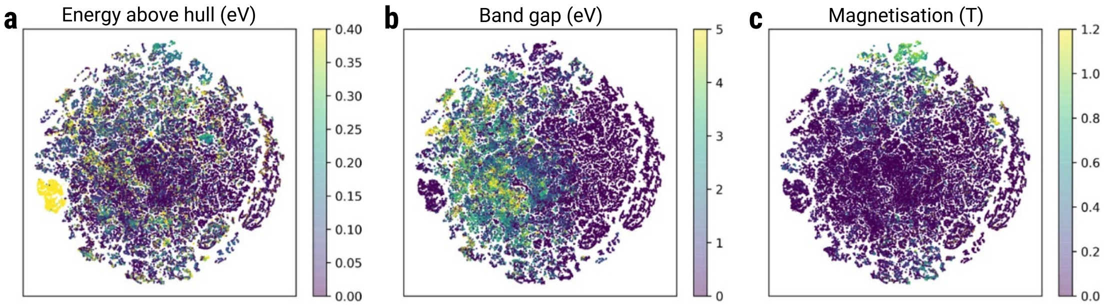

Figure 5. Physical property distributions within the material embedding space. t-SNE plots of the embeddings are coloured according to the physical properties: energy above the hull, band gap, and magnetisation. These plots show clusters of materials with similar physical properties, indicating that the embeddings capture the property similarities between materials. (a) The distribution of energy above the hull (eV). A large value of energy above the hull indicates that a material is unstable. A cluster of unstable compounds containing sparse unsynthesisable crystal structures was identified on the upper left. (b) The distribution of the band gap (eV). The distribution overlap of large-bandgap materials in this figure and oxides in figure 4 demonstrates a well-known connection between the band gap and oxygen. (c) The distribution of magnetisation (T). Materials with large magnetic moments have higher composition ratios of magnetic elements such as Mn, Fe, Co, and Ni and are particularly studied in the rare-earth permanent magnet research. The distinct yellow cluster in the top right of this figure contains intermetallic compounds of the magnetic elements and rare-earth elements (e.g. Ce, Pr, Nd, and Sm), as evident from figure 2 where the distributions of these elements overlap in this area.

the materials map. Because atoms and ions with similar electron configurations compose materials with the same or similar crystal structures, we expect the element distributions to show cluster-like features if our embeddings have been trained successfully. In figure 4, we highlight each element in the map and display all elements in the form of a periodic table. As expected, figure 4 clearly shows similar distributions of blue-coloured clusters in the vertical and horizontal directions. These distributions can be analogously called the ‘alkali metal plateau’, the ‘3d transition metal district’, or ‘rare-earth mountains’ if we follow the map metaphor, indicating that the embeddings succeed in capturing the similarities of roles between elements in crystal structures. Additionally, we noticed that well-known connections between physical properties and elements can also be probed using this plotting technique (see figures 5(b) and (c) for details). Although these visualisations (figures 4 and 5) are intended to confirm expected outcomes rather than showing interesting findings, they demonstrate their potential utility, e.g. for giving researchers new insights or helping them find materials with desired properties.

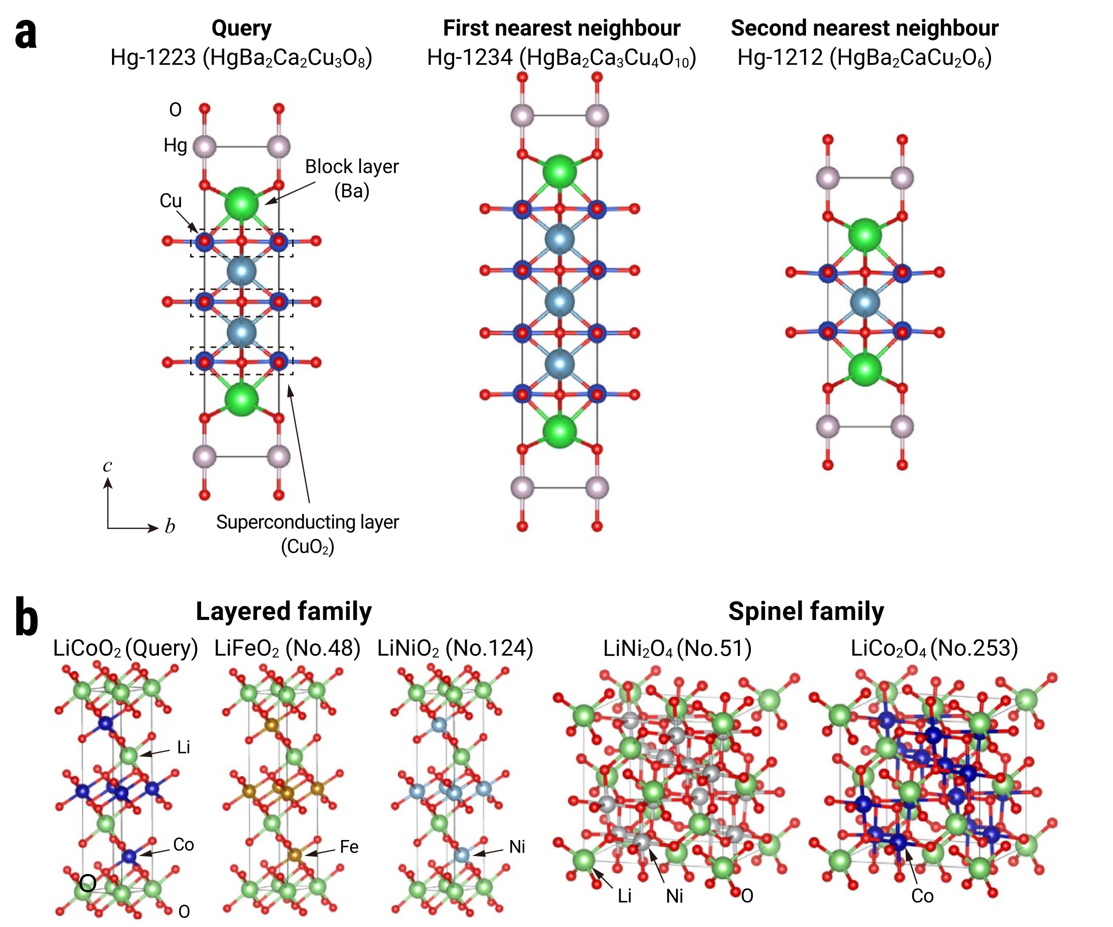

Figure 6. Crystal structures of queries and nearest neighbours. (a) Crystal structures of the Hg-1223 superconductor and its first and second nearest neighbours, Hg-1234 and Hg-1212, in the embedding space. These are all Hg-based copper oxide superconductors with high T_c and are built on two important components, namely, block layers and superconducting layers. Placing these close-kin materials, yet with different numbers of superconducting layers, as close neighbours suggests that the embedding captures the conceptual similarity in their structures. See also table 1 for the top-50 neighbour list of Hg-1223, in which more superconductors with high T_c are found. (b) Crystal structures of the well-known lithium-ion battery cathode material $LiCoO_2$ and its neighbours in the embedding space. The layered and spinel families, which are two major types of the cathode materials, were identified in the neighbourhoods. See also table 2 for the list of the top-50 neighbours, in which more layered-family materials are found.

2.2. Local neighbourhood analysis

We next examine the local neighbourhoods of several benchmark areas to verify whether the learned metric recognises functionality-level material similarity. Since the embeddings were optimised with the Euclidean distance, we also used this metric to determine the neighbourhoods.

As the first example, we analysed the neighbourhoods of Hg-1223, a superconductor with the highest known T_c (134 K) at ambient pressure [42]. To our surprise, the first and second nearest neighbourhoods correspond to its close kin Hg-1234 and Hg-1212, which also have high T_c values (125 K and 90 K) but different block layers [40] from those of Hg-1223 (see figure 6(a)). Further investigation identified major Tl-based high-T_c superconductors, such as Tl-2234 ($T_c = 112$ K), Tl-2212 ($T_c = 108$ K), and Tl-1234 ($T_c = 123$ K)[43], and many other superconductors occupying the top-50 neighbourhoods (see table 1). The connection between the crystal structures and T_c values involves non-trivial mechanisms that are not immediately evident from the crystal structures [3, 40]. The results suggest that our model effectively bridges this gap with the help of learned structure-functionality relationships that are deeply buried in the 1024-dimensional space.

Next, we examined lithium-ion battery materials, which substantially support our lives of today. This technology has been developed through the discovery of new materials and the understanding of their structure-composition-property-performance relationships and is now bottlenecked by the cathodes (positive electrodes) in terms of the energy density and production cost [5]. We therefore studied the neighbourhoods of $LiCoO_2$, the first yet most dominant cathode material [5]. Impressively, two of the three leading cathode material groups, namely, the layered, spinel families [5] (see figure 6(b) for visualisations), were identified in the neighbourhoods. Specifically, similar to $LiCoO_2$, a family of layered $LiMO_2$, with M

Table 1. The top-50 neighbours of Hg-1223 in comparison with hand-crafted descriptors.

	Our embedding		Ewald sum matrix		Sine Coulomb matrix	
No.	Formula	ID	Formula	ID	Formula	ID
Query	$Ba_2Ca_2Cu_3HgO_8$	mp-22601	$Ba_2Ca_2Cu_3HgO_8$	mp-22601	$Ba_2Ca_2Cu_3HgO_8$	mp-22601
1	$Ba_2Ca_3Cu_4HgO_{10}$	mp-1228579	$Sr_4TlFe_2O_9$	mp-1218464	$Tl(CuTe)_2$	mp-569204
2	$Ba_2CaCu_2HgO_6$	mp-6879	$Ba_2La_2Ti_2Cu_2O_{11}$	mp-1214655	$CaLa_2BiO_6$	mvc-15176
3	$Ba_6Ca_6Cu_9Hg_3O_{25}$	mp-1228760	CeY_4Mg_5	mp-1226574	$PtC_4S_2(IO)_2$	mp-1102535
4	$Sr_2CaCu_2(BiO_4)_2$	mp-1218930	$Ba_6Nb_2Ir(ClO_6)_2$	mp-558113	Ba_2FeReO_6	mp-31756
5	$Ba_{10}Ca_5Cu_{10}Hg_5O_{31}$	mp-1229139	$Ba_6Ru_2Pt(ClO_6)_2$	mp-554949	$Hg(SbO_3)_2$	mp-754065
6	$SrCa_2Cu_2(BiO_4)_2$	mp-1208800	$Ba_2Nd_2Ti_2Cu_2O_{11}$	mp-557043	Ba_2CuWO_6	mp-505618
7	$Ba_8Ca_4Cu_8Hg_4O_{25}$	mp-1228371	$Ba_4ScTi_4BiO_{15}$	mp-1228157	$CaLa_2WO_6$	mvc-15479
8	$Ba_2Ca_3Tl_2(CuO_3)_4$	mp-556574	La_3ZnNi_3	mp-18573	Ba_2YTaO_6	mp-12385
9	$Ba_2Mg_3Tl_2(WO_3)_4$	mvc-129	Zr_4WC_5	mp-1215364	$TlCdTe_2$	mp-998919
10	$Ba_2TlV_2O_7$	mvc-2978	Nd_3GaCo_3	mp-1103877	$TlCuPd_2$	mp-1096374
11	$Sr_2YCu_2(BiO_4)_2$	mp-1208863	$Y_4Ti_6Bi_2O_{21}$	mp-1216208	$LaTlAg_2$	mp-867817
12	$Sr_2LaCu_2HgO_6$	mp-1208803	Sm_3HoS_4	mp-1219190	In_3Au	mp-973498
13	$Ba_2CaTl_2(CuO_4)_2$	mp-573069	$Ba_3Bi(BO_2)_9$	mp-1200141	$CeTlAg_2$	mp-867298
14	$Ba_4CaCu_6(HgO_8)_2$	mvc-15237	$AgRhO_2$	mp-997106	Cs_2WBr_6	mp-541753
15	$Ba_4Ca_4Cu_6Hg_2O_{17}$	mp-1228265	$YbSm_3S_4$	mp-1215523	$TlIn_3$	mp-1187742
16	$Ba_2AlTlCo_2O_7$	mvc-2977	Ca_4Cd_3Au	mp-1227562	In_3Pt	mp-1184857
17	$Sr_8Pr_4Cu_9(HgO_8)_3$	mp-1218674	$InAg_4$	mp-1223819	Ca_4Cd_3Au	mp-1227562
18	$Ba_6Ca_3Cu_6Hg_3O_{19}$	mp-1228161	$Sr_4ZrTi_3O_{12}$	mp-1218457	Cd_3Pt	mp-1183641
19	$Ba_8Ca_8Tl_7(Cu_4O_{13})_3$	mp-1204270	$Ce_3Ni_2Ge_7$	mp-1213875	Ag_3Au	mp-1183214
20	$Ba_4Ca_4Tl_3Cu_6O_{19}$	mp-542197	$Ba_2YTlV_2O_7$	mvc-2994	Mn_4BiSb_3	mp-1221739
21	$Ba_6Ca_6Tl_5Cu_9O_{29}$	mp-680433	Te_3Au	mp-1217358	$NdTlAg_2$	mp-974782
22	$Ba_2AlTlCo_2O_7$	mp-1266279	$Nd_3Cu_4(P_2O)_2$	mp-1209832	HgI_3	mp-973601
23	$Ba_2Ca_2Tl_2Ni_3O_{10}$	mvc-3067	$Ba_4Zn_4B_{14}Pb_2O_{31}$	mp-1194514	$TlCdIn_2$	mp-1093975
24	$Ba_2Ca_2Tl_2Cu_3O_{10}$	mp-653154	$Ba_6Na_2Nb_2P_2O_{17}$	mp-556637	$CePd_2Pt$	mp-1226474
25	$Ba_2Ca_2Tl_2Co_3O_{10}$	mvc-3021	$Ba_2Tb_2Ti_2Cu_2O_{11}$	mp-505223	$PmHgRh_2$	mp-862913
26	$Sr_2CaCu_2(BiO_4)_2$	mp-555855	$Sc_2TlCu_3S_5$	mp-1209018	$Sr_2LaCu_2HgO_6$	mp-1208803
27	$Ba_4Tl_2Cu_2HgO_{10}$	mp-561182	$Eu(GaGe_2)_2$	mp-1225812	$NdPd_2Pb$	mp-1186317
28	$Ba_6Ca_{12}Cu_{15}Hg_3O_{37}$	mp-1229082	$AgTe_3$	mp-1229041	$PmTlRh_2$	mp-862967
29	$BaCuReO_5$	mvc-7248	Sm_3GaCo_3	mp-1105102	Cd_3Ir	mp-1183645
30	$Ba_2Ca_3Tl_2(FeO_3)_4$	mvc-145	Nb_4Rh	mp-1220441	$HgPd_3$	mp-1184658
31	$Sr_{10}Cu_5Bi_{10}O_{29}$	mp-667638	$La_3(Al_2Si_3)_2$	mp-1211155	$SnPd_2Au$	mp-1095757
32	$Ba_2Ca_3TlCu_4O_{11}$	mp-1228589	Ce_2In_8Pt	mp-1103614	$PmTlAg_2$	mp-862966
33	$Ba_2Ca_3Tl_2(CuO_3)_4$	mp-556733	$CaNb_2Bi_2O_9$	mp-555616	$Rb_2LaAuCl_6$	mp-1113498
34	La_2B_3Br	mp-568985	Ce_2In_8Ir	mp-1207157	VAg_3HgO_4	mp-1216423
35	$BaTl(SbO_3)_2$	mvc-10727	Tc_6BiO_{18}	mp-1101632	In_2SnHg	mp-1097125
36	$Sr_{10}Cu_5Bi_{10}O_{29}$	mp-652781	Sb_3Au	mp-1219474	$PrBiPd_2$	mp-976884
37	$Ba_2Tl_2Zn_2Cr_3O_{10}$	mvc-3164	Sr_4LaCl_{11}	mp-1218463	$TlIn_3$	mp-1216611
38	$Ba_2Ca_2Tl_2Fe_3O_{10}$	mvc-3027	$LaBiS_2O$	mp-1078328	Cd_2AgPt	mp-1096169
39	$Ba_2Ti_3Tl_2O_{10}$	mvc-2939	$HfNb_4CN_4$	mp-1224363	$Rb_2CeAuCl_6$	mp-1113397
40	$Sr_2TaAlCu_2O_7$	mp-1251503	MoN	mp-1078389	In_2SnPb	mp-1223808
41	$Ba_2Mg_3Tl_2(SnO_3)_4$	mvc-10576	$YZnGe$	mp-13160	Cd_2AgPt	mp-1183537
42	$Sr_2AlTlCo_2O_7$	mp-1252241	$Pr_3(Al_2Si_3)_2$	mp-571302	Ag_2PdAu	mp-1096329
43	$Ba_2AlTlV_2O_7$	mp-1265780	$Sr_2(BiPd)_3$	mp-1207133	Ag_3AuS_2	mp-34982
44	$Ba_2CaTl_2(CuO_4)_2$	mp-6885	$Na_3HoTi_2Nb_2O_{12}$	mp-676988	$PmRh_2Pb$	mp-862958
45	$Sr_2LaCu_2(BiO_4)_2$	mp-1209034	$Sr_2YCu_2BiO_7$	mvc-280	$PmPd_2Pb$	mp-862950
46	$Ba_2AlTlV_2O_7$	mvc-3002	$Na_3DyTi_2Nb_2O_{12}$	mp-689927	$Sr_2PrTlCu_2O_7$	mp-1208792
47	$Ba_2Mg_3Tl_2(FeO_3)_4$	mvc-28	$Rb_3NaRe_2O_9$	mp-1209462	$InAg_2Au$	mp-1093943
48	$Sr_2DyCu_2(BiO_4)_2$	mp-1209149	$Sr_3Fe_2Ag_2S_2O_5$	mp-1208725	Ag_2SnBiS_4	mp-1229127
49	Ba_2CuHgO_4	mp-6562	$Ba_2Pr(CuO_2)_3$	mp-1228546	Sb_3Au	mp-29665
50	$Ba_2Tl_2W_3O_{10}$	mvc-3144	$Ce_3(Al_2Si_3)_2$	mp-29113	$PmAg_2Pb$	mp-862876

We compare the top-50 neighbours of the Hg-1223 superconductor obtained by using our embedding and two hand-crafted descriptors (Ewald sum matrix and sine Coulomb matrix) [22]. The query material, Hg-1223 ($HgBa_2Ca_2Cu_3O_8$), has the highest known T_c (134 K) at ambient pressure. Quite impressively, the neighbour list obtained by our embedding seems to be completely filled with superconductors, including the well-known Hg-1224 (No. 1) and Hg-1212 (No. 2) as well as Tl-based high-T_c superconductors such as Tl-2234 (No. 8), Tl-1234 (No. 32), and Tl-2212 (No. 44). By contrast, the lists obtained by the two existing descriptors contain irrelevant materials rather than superconductors. These results clearly show that our approach captures the conceptual similarity between superconductors, which is undetectable by the existing descriptors. See also the SI (appendix A3) for the detailed procedures of the descriptor computations and more discussions.

being transition metals, were found within the top-10 neighbourhoods of $LiCoO_2$ (see table 2), including important battery materials $LiNiO_2$ families. Spinels as another important family were found as $LiNi_2O_4$ at the 51th neighbour and $LiCo_2O_4$ in the 200s neighbours. The polyanion family, the remaining one of three major cathode families, were not placed in the vicinity of $LiCoO_2$ but formed a distinctive cluster at the top edge in figure 3(a). Interestingly, all of these materials were developed by the group of Nobel laureate John Goodenough [5]. This fact suggests that the embeddings capture conceptual similarity among the battery materials that previously required one of the brightest minds of the time to be discovered.

Note that our method properly links substituted materials and the original material without being confused by ad hoc supercell expression (e.g. $Li_4Co_3NiO_8 = LiCo_{0.75}Ni_{0.25}O_2$). This advantage is particularly noticeable in comparison with embeddings constructed using conventional features (table 2). This result indicates that our approach can recognise the essential structural features without being affected by superficial differences (i.e. the number of atoms or the size of the unit cell).

Additionally, we analysed the vicinities of magnetic materials, including 2D ferromagnets, which are attracting much attention for their interesting properties [41], and commercial samarium–cobalt (Sm–Co) permanent magnets. Again, the embeddings capture meaningful similarity in these material classes, as shown in figures 7 and 8, which is often not evident to non-specialists (see appendix A1 in the supplementary information (SI) for more discussions and detailed results).

These in-depth analyses across diverse materials consistently support the conclusion that our ML model recognises similar functionalities of materials behind different structures without being explicitly taught to do so. We anticipate that when a material with beneficial properties is found, we may be able to screen for new promising candidates based on the conceptual similarities captured in this embedding space.

2.3. Performance validation as a materials descriptor

Here we provide quantitative insight into characteristics of embeddings. Particularly, we analyse the performance of predicting material properties using trained embeddings as input. As we are more interested in predicting functional material properties, we conducted a binary classification task of materials concepts, in which an ML model predicts whether a material belongs to a particular material class or not.

We expect that our embeddings contain the information of materials concepts. If so, we can rapidly screen materials with a desired concept from a material database by combining the embeddings with an ML model. However, properly labelling materials with their concepts requires experiments or consideration by experts, and thus the number of available labelled data for a given concept is likely to be limited. Therefore, as a benchmark and a use case for our embeddings, we evaluated the materials concept classification in the settings of few training data.

As benchmark materials, we used superconductors and thermoelectric materials for their complex and interesting properties. We used the Crystallography Open Database (COD) as the data source. The number of positive data used for training was 469 for superconductors and 286 for thermoelectric materials. Embeddings of these materials were obtained by the crystal structure encoder trained on the MP dataset via deep metric learning, and were used as input to a random forest classifier. As a baseline for comparison, we used latent feature vectors of crystal graph convolutional neural network (CGCNN) trained for total energy prediction, as in appendix C. We evaluated the prediction performance with leave-p-groups-out cross-validation while varying the training data size. Here, both the training and testing splits were made to contain balanced positive and negative samples.

As shown in figure 9, the classifier using our embeddings obtained good classification performance for both superconductors and thermoelectric materials. In particular, when the number of training data is very small (around 10), our method shows significantly better performance than the baseline. We will more discuss these results in the next section.

3. Discussions

As assumed, materials concepts were exposed spontaneously in an abstract space. As we confirmed in the numerical evaluations of the training task of metric learning (appendix B in SI), this space was shown to successfully unify the two complementary factors of crystal structures. We hypothesise that these remarkable properties of our embeddings were made possible by the following two key features of our method that are distinctive from the existing material embedding methods [25–29]. First, we used deep metric learning, which directly optimises the spatial arrangements of the embedding vectors via a loss in terms of the Euclidean distances between them. This procedure is critically different from the existing methods [25–29],

Table 2. The top-50 neighbours of $\mathrm{LiCoO_2}$ in comparison with hand-crafted descriptors.

	Our embedding		Ewald sum matrix		Sine Coulomb matrix	
No.	Formula	ID	Formula	ID	Formula	ID
Query	$\mathrm{LiCoO_2}$	mp-22526	$\mathrm{LiCoO_2}$	mp-22526	$\mathrm{LiCoO_2}$	mp-22526
1	$\mathrm{Li_{14}MgCo_{13}O_{28}}$	mp-769537	$\mathrm{LiNiO_2}$	mp-25587	$\mathrm{LiCoO_2}$	mp-1222334
2	$\mathrm{Li_4Co_3NiO_8}$	mp-867537	$\mathrm{Co(HO)_2}$	mp-24105	$\mathrm{CoHO_2}$	mp-27913
3	$\mathrm{Li_3Fe(CoO_3)_2}$	mp-761602	$\mathrm{LiFeO_2}$	mp-1222302	$\mathrm{LiCoF_2}$	mp-1097040
4	$\mathrm{Li_3(CoO_2)_4}$	mp-850808	$\mathrm{LiNiO_2}$	mp-25316	LiCoN	mp-1246462
5	$\mathrm{Li_3MnCo_3O_8}$	mp-774219	$\mathrm{Li_2NiO_2}$	mp-19183	$\mathrm{Li_2CoN_2}$	mp-1247124
6	$\mathrm{Li_{20}(CoO_2)_{21}}$	mp-532301	$\mathrm{MgMnN_2}$	mp-1247154	$\mathrm{Be_5Co}$	mp-1071690
7	$\mathrm{Li_3CrCo_3O_8}$	mp-849768	$\mathrm{Li_2CaCd}$	mp-1096283	$\mathrm{Be_3Co}$	mp-1183423
8	$\mathrm{Li_3MnCo_3O_8}$	mp-758163	$\mathrm{NiO_2}$	mp-25210	$\mathrm{Be_2Co}$	mp-1227342
9	$\mathrm{Li_8FeCo_9O_{20}}$	mp-764865	LiFeOF	mp-775022	CoCN	mp-1245659
10	$\mathrm{Li_3Co_2NiO_6}$	mp-765538	$\mathrm{MnO_2}$	mp-1221542	$\mathrm{Li_3Co}$	mp-976017
11	$\mathrm{Li_3CrCo_3O_8}$	mp-759149	$\mathrm{Co(HO)_2}$	mp-625939	$\mathrm{Li_2CoO_2}$	mp-755133
12	$\mathrm{Li_3TiCo_3O_8}$	mp-757214	$\mathrm{Co(HO)_2}$	mp-625943	$\mathrm{Li_2CoO_2}$	mp-755297
13	$\mathrm{Li_4MgCo_3O_8}$	mp-754576	$\mathrm{Li_2CuO_2}$	mp-1239022	$\mathrm{Be_{12}Co}$	mp-1104193
14	$\mathrm{Li_5Co_2Ni_3O_{10}}$	mp-769553	$\mathrm{CoO_2}$	mp-1062939	$\mathrm{CoO_2}$	mp-1181781
15	$\mathrm{Li(CoO_2)_2}$	mp-552024	$\mathrm{NaCoO_2}$	mp-1221066	$\mathrm{CoO_2}$	mvc-13108
16	$\mathrm{Li_{14}Co_{13}O_{28}}$	mp-777836	$\mathrm{NiO_2}$	mp-634706	$\mathrm{Co(HO)_2}$	mp-626708
17	$\mathrm{Li_3(NiO_2)_5}$	mp-762165	$\mathrm{MgMnO_2}$	mp-1080243	$\mathrm{Co(HO)_2}$	mp-625939
18	$\mathrm{Li_2CoO_2F}$	mp-764063	$\mathrm{LiCuF_2}$	mp-753098	$\mathrm{Co(HO)_2}$	mp-625943
19	$\mathrm{Li_2(CoO_2)_3}$	mp-758539	$\mathrm{Ni(HO)_2}$	mp-625074	$\mathrm{Co(HO)_2}$	mp-24105
20	$\mathrm{Li_5Fe_2Co_3O_{10}}$	mp-769566	$\mathrm{CrO_2}$	mp-1009555	$\mathrm{CoO_2}$	mp-1062939
21	$\mathrm{Li_2CoNi_3O_8}$	mp-752703	$\mathrm{CoHO_2}$	mp-27913	$\mathrm{CoO_2}$	mp-1062643
22	$\mathrm{Li_{10}Fe_3Co_7O_{20}}$	mp-760848	$\mathrm{NaLi_2As}$	mp-1014873	$\mathrm{CoO_2}$	mp-556750
23	$\mathrm{Li_7Co_5O_{12}}$	mp-771155	$\mathrm{LiNiO_2}$	mp-25411	$\mathrm{CoH_3}$	mp-1183678
24	$\mathrm{Li_3(NiO_2)_4}$	mp-755972	$\mathrm{Li_2CoO_2}$	mp-755133	CoH	mp-1206874
25	$\mathrm{Li_9Ni_{15}O_{28}}$	mp-759153	$\mathrm{LiCuO_2}$	mp-754912	$\mathrm{CoO_2}$	mp-1063268
26	$\mathrm{Li_{20}Co_{21}O_{40}}$	mp-685270	$\mathrm{CrN_2}$	mp-1014264	CoN	mp-1008985
27	$\mathrm{Li_7(NiO_2)_{11}}$	mp-768079	MgCr	mp-973060	CoN	mp-1009078
28	$\mathrm{Li_2(NiO_2)_3}$	mp-762391	$\mathrm{Ni(HO)_2}$	mp-1180084	$\mathrm{FeHO_2}$	mp-755285
29	$\mathrm{Li_4Co_2Ni_3O_{10}}$	mp-778996	$\mathrm{Co(HO)_2}$	mp-626708	$\mathrm{LiFeO_2}$	mp-1222302
30	$\mathrm{Li_2Co_3NiO_8}$	mp-757851	$\mathrm{Ni(HO)_2}$	mp-27912	$\mathrm{LiFeO_2}$	mp-19419
31	$\mathrm{LiCoNiO_4}$	mp-754509	VO	mp-19184	$\mathrm{CoBO_3}$	mp-1183397
32	$\mathrm{Li_4(NiO_2)_7}$	mp-774600	$\mathrm{Be_4AlFe}$	mp-1227272	LiFeOF	mp-775022
33	$\mathrm{Li(CoO_2)_2}$	mp-774082	$\mathrm{FeO_2}$	mp-1062652	$\mathrm{LiNiO_2}$	mp-25411
34	$\mathrm{Li(CoO_2)_2}$	mp-752807	$\mathrm{LiCoF_2}$	mp-1097040	$\mathrm{NiHO_2}$	mp-1067482
35	$\mathrm{Li_8(NiO_2)_{11}}$	mp-758772	$\mathrm{LiFeO_2}$	mp-19419	$\mathrm{Li_4Co(OF)_2}$	mp-850355
36	$\mathrm{Li_3CoNi_3O_8}$	mp-774300	$\mathrm{Na_2NiO_2}$	mp-752558	$\mathrm{NiHO_2}$	mp-999337
37	$\mathrm{Li_2CoNi_3O_8}$	mp-1178042	$\mathrm{Li_2CuO_2}$	mp-4711	$\mathrm{LiNiO_2}$	mp-25587
38	$\mathrm{Li_7(NiO_2)_8}$	mp-690528	$\mathrm{Li_2CoO_2}$	mp-755297	$\mathrm{LiNiO_2}$	mp-25316
39	$\mathrm{Li_{10}Co_3Ni_7O_{20}}$	mp-769555	MgCr	mp-1185858	$\mathrm{LiFeO_3}$	mp-1185320
40	$\mathrm{Li_7Ni_{13}O_{24}}$	mp-758593	$\mathrm{Sc_2CO}$	mp-1219429	LiFeN	mp-1245817
41	$\mathrm{Li_9Co_7O_{16}}$	mp-1175506	$\mathrm{MnBO_3}$	mp-1185996	$\mathrm{CoNF_3}$	mp-1213745
42	$\mathrm{Li_3Cr(CoO_3)_2}$	mp-761831	VN	mp-1001826	$\mathrm{LiNiO_3}$	mp-1185261
43	$\mathrm{Li_2Co_3NiO_8}$	mp-778768	$\mathrm{NiHO_2}$	mp-999337	$\mathrm{Li_4FeN_2}$	mp-28637
44	$\mathrm{Li_2FeCo_3O_8}$	mp-1177976	CrO	mp-19091	LiNiN	mp-29719
45	$\mathrm{Li_4Co_3(NiO_4)_3}$	mp-777850	$\mathrm{Ni(HO)_2}$	mp-625072	$\mathrm{Be_3Fe}$	mp-983590
46	$\mathrm{Li_3Al_2CoO_6}$	mp-1222591	VN	mp-925	$\mathrm{NiO_3}$	mp-1209929
47	$\mathrm{Li(NiO_2)_2}$	mp-752531	$\mathrm{GaH_6N_2F_3}$	mp-1224894	$\mathrm{Be_5Fe}$	mp-1025010
48	$\mathrm{LiFeO_2}$	mp-19419	$\mathrm{Fe(HO)_2}$	mp-626680	$\mathrm{Li_2FeO_2}$	mp-755094
49	$\mathrm{Li_4AlNi_3O_8}$	mp-1222534	CrN	mp-1018157	$\mathrm{Be_{12}Fe}$	mp-1104104
50	$\mathrm{Li_3CoNi_3O_8}$	mp-757871	VN	mp-1018027	$\mathrm{FeB_2}$	mp-569376

We compare the top-50 neighbours of $\mathrm{LiCoO_2}$ obtained by using our embedding and two hand-crafted descriptors (Ewald sum matrix and sine Coulomb matrix) [22]. The query material, $\mathrm{LiCoO_2}$, is one of the most crucial lithium-ion battery cathodes. In the list of our embedding, the many neighbours of $\mathrm{LiCoO_2}$ are occupied by $\mathrm{LiCo_{1-x}M_xO_2}$ families with the same layered structure as $\mathrm{LiCoO_2}$ but partly substituted with different transition metals M. Since these partial substitutions are represented as supercells, the system's apparent size is larger than the original unit cells. Our approach is unaffected by these apparent differences and can recognise the essential similarities. While most of our list is filled with lithium oxides, the other two lists obtained by the existing descriptors do not suggest this consistent trend. These results suggest that our model recognises the concept of lithium-ion battery cathodes, which is not captured by the existing descriptors. See also the SI (appendix A3) for the detailed procedures of the descriptor computations and more discussions.

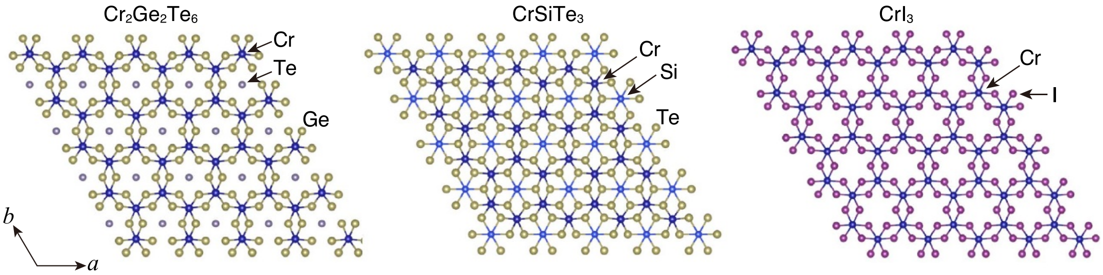

Figure 7. Crystal structures of the 2D ferromagnet $Cr_2Ge_2Te_6$ and its neighbours in the embedding space. The double discoveries of 2D ferromagnets in 2017, after long questioning their existence, are gathering great interest from the magnetic materials community [41]. When we analysed the neighbourhoods of one of these 2D ferromagnets, $Cr_2Ge_2Te_6$ (mp-541449), our embedding space successfully captured $CrSiTe_3$ (mp-3779), a compound known as a potentially 2D-ferromagnetic insulator, as the first neighbour and even the other 2D ferromagnet CrI_3 (mp-1213805) as the 15th neighbours among 122 543 materials. More detailed results and discussions are given in the SI (appendix A1).

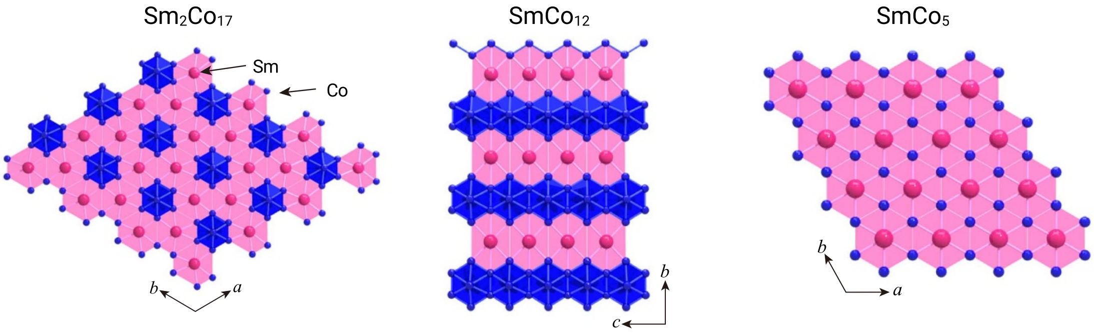

Figure 8. Crystal structures of the Sm_2Co_{17} permanent magnet and its neighbours in the embedding space. Here we highlight two compounds in the neighbourhood list of Sm_2Co_{17}: $SmCo_5$ and $SmCo_{12}$. Sm_2Co_{17} and $SmCo_5$ are the two major components in Sm–Co magnets often used in high-temperature environment, whereas $SmCo_{12}$ is one of the compounds with the so-called 1–12 structure that has been drawing attention for its potential for permanent magnets. In the neighbourhoods of Sm_2Co_{17} (mp-1200096) in our embedding space, we found $SmCo_{12}$ (mp-1094061) as the 255th neighbour and $SmCo_5$ (mp-1429) around the top 0.5% neighbourhoods. It is well known in the community that the crystal structures of Sm_2Co_{17}, $SmCo_5$, and $SmCo_{12}$ have close connections with each other [4]. However, without the literature context and proper visualisation, it is difficult for a human analyst to recognise these connections. More detailed results and discussions are given in the SI (appendix A1).

which learn embeddings indirectly as DNN's latent vectors. Although these latent vectors should encode essential information about materials, the explicit metric optimisation of embeddings is equally important for map creation and similarity learning. Second, our self-supervised learning is enabled by exploiting two forms of inputs expressing complementary structural characteristics: a set of atoms in the unit cell with their connections as the local characteristics and the XRD pattern, which is essentially a Fourier transformed crystal structure [1], as the periodic characteristics. Representation learning is known to be generally more well-informed when diverse multi-modal data are used for training [44]. In contrast to approaches that rely on single forms of materials data expression [25–29], our model benefits from learning across two forms of expression, or cross-modal learning.

The results of the materials concept classification (figure 9) clearly support these hypotheses. Remind that the baseline method (CGCNN [12]) learns embeddings as latent vectors in a DNN with only crystal structures as input, whereas our method uses the same DNN but trains it along with another DNN for XRD patterns in cross-modal deep metric learning. Thus, the performance advantage of our method directly indicates the benefit of the proposed cross-modal deep metric learning approach. We believe that our method using both crystal structures and XRD patterns helped the ML model to capture local motifs and lattice more effectively, which contributed to better learning of structural patterns correlated to material functionality and thus better recognition of materials concepts. We expect that incorporating more diverse structure representations of materials such as electronic structure into our multi-modal learning framework will further benefit the representation learning of materials. We leave such extensions as future work.

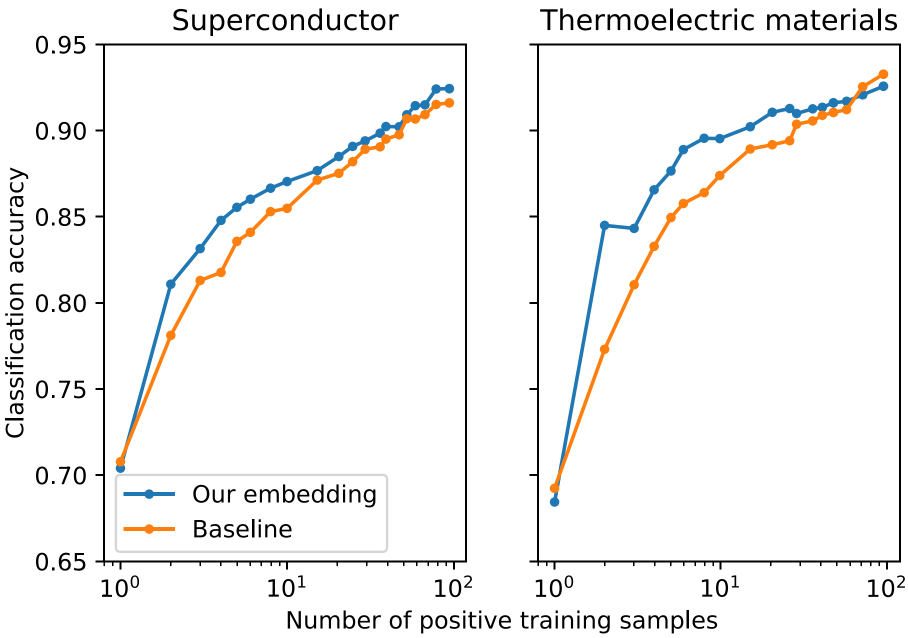

Figure 9. The prediction performance of materials concepts. For superconductors and thermoelectric materials, embeddings obtained by our DML approach show higher performance especially when the size of the training dataset is very small. The embeddings of the baseline were latent vectors of CGCNN trained to predict total energy from crystal structures, as done by Xie *et al* [12].

To provide more insight into the difference between our method and the baseline (CGCNN), we further analysed the performance of these methods for physical property prediction (see appendix C in SI for details). Similarly to the materials concept classification, we trained random forest models to predict materials properties, such as total energy, space group, and density, from learned embeddings. Our embedding outperformed the baseline in predicting density and space group and performed comparably in total energy and magnetisation (figure S1 in SI). This result confirms that our embeddings indeed capture lattice information in crystal structures more effectively than the single-modal baseline using only crystal structures. Performing comparably in total energy prediction is also notable, because the embeddings of the baseline are trained to specifically predict total energy itself using rich supervision from density functional theory calculations while our embeddings are not.

A major interest in the proposed method given its good predictive power is whether it has potential utility for new material discovery. To investigate such possibilities, we conducted a simple test to see if our model can re-discover superconductors known in the literature but not included in the training dataset. To this end, we borrowed the COD's superconductors from the concept classification (figure 9) and, after removing overlaps with the MP's training dataset, we mapped their embeddings in the MP's embedding distribution presented in figure 3. As shown in appendix E, these COD's superconductors are most intensively concentrated around the superconductor cluster in the MP's training materials, despite the fact that these COD's materials are novel to the model. This result suggests a screen method of new candidate materials by using our model trained on a database of known materials.

Another notable strength of our method over existing material embedding methods is that it does not require costly annotations and can be trained using only primitive structural information (i.e. crystal structures and their XRD patterns). This makes our method applicable to a wide range of datasets. Even when annotations are available, our self-supervised approach will benefit many users as a means of pre-training. Pre-training is a general ML technique performed on a large-scale dataset to help an ML model for other tasks where annotated training data are limited [45]. Our self-supervised learning is suitable for this purpose, because it can be performed given only crystal structure data and can thus utilise various material databases at scale.

When compared to classic material descriptors such as the Coulomb matrix variants [22], our method has advantages in terms of its scalability and ability to capture high-level material properties. See tables 1, 2 and the SI (appendix A3) for analysis results and more discussions.

Since the focus of our study was on learning material similarity from unannotated structural data, the resulting map requires manually interpreting clusters on the basis of our knowledge of materials concepts.

Interestingly, a word2vec model [26] has been applied to text symbols appearing in the materials science literature, thus learning relationships such as the connections between 'Fe' and 'metal' and between 'Sm–Co' and 'magnet'. Use of this technique may further automate the interpretation of our results with literal knowledge.

4. Conclusions and broader impacts

In summary, we have demonstrated the self-supervised learning of material embeddings solely from crystal structures using DNNs. Careful inspection of the embedding space, in terms of both the global distribution and local neighbourhoods, has confirmed that the space recognises functionality-level material similarity or materials concepts. Our techniques for the materials space visualisation and the similarity evaluation between crystal structures will be useful for discovering new underlying relationships among materials and screening for new promising material candidates. Since these techniques are not strongly affected by human bias, they could give rise to a new view of materials that can stimulate efforts to break through our knowledge barriers.

Our result is also applicable to material retrieval systems that can search for conceptually similar materials in a database given a query material. This approach will enable us to rediscover materials that have never been recognised to have preferable properties.

Furthermore, constructing a functionality-aware representation space of crystal structures is a first step towards the *inverse design of materials* [8, 46], a grand challenge of materials informatics. This workflow would allow us to design materials in the functionality space and inversely map the functionality attributes to synthesisable crystal structures with the desired properties. We hope that this study will pave the way for breakthroughs in the ML-assisted discovery and design of materials.

5. Methods

5.1. Data acquisition and pre-processing

We used the Materials Project as the data source for this study. We collected data for up to quintet systems, excluding monatomic crystals, on 8 July 2020, using the Material Project APIs, which resulted in a total of 122 543 materials (93% of the source collection) as our targets. We additionally queried thermodynamic stability material attributes on 14 October 2020. We used VESTA [47] for crystal structure visualisation. We calculated the XRD patterns using pymatgen [48]. The x-ray wavelength was set to 1.54 184 Å (Cu $K_{\alpha 1}$), and the 2θ angle ranged from 10° to 110° with a step size of 0.02°; thus, 5000-dimensional vectors of 1D-structured XRD patterns were produced. To ease the learning process, the intensity scale of each XRD pattern was normalised by setting the maximum intensity to 1.

5.2. Neural network architecture

As illustrated in figure 2(a), we used two types of DNNs as embedding encoders. For the crystal-structure encoding, we need to convert a set of arbitrary number of atoms (i.e. the atoms in the unit cell) into a fixed-size embedding vector in a fashion invariant to the permutation of atom indices. For this purpose, we used CGCNNs [12]. As input to CGCNN, the 3D point cloud of the atoms in the unit cell is transformed into a graph of atoms whose edge connections are defined by their neighbours within a radius of 8 Å. The atoms in the graph are represented as atom feature vectors and are transformed into a single fixed-size feature vector via three graph convolution layers and a global pooling layer. For the XRD patterns, we used a standard feed-forward 1D convolutional neural network designed following existing studies on XRD pattern encoding [49]. At the end of each network, we used three fully connected layers to output 1024-dimensional embedding vectors. Since one of these encoders is supervised by the output of the other in our self-supervised learning approach, training them simultaneously tends to be unstable compared to standard supervised learning. To stabilise the training process, we found that batch normalisation [50] is essential after every convolutional/linear layer in both networks except for the final linear output layers. We discuss this further in the SI (appendix B). Further details of our network architecture are provided in the SI (tables S6 and S7 in appendix D) and our ML model codes.

5.3. Training procedures

In each training iteration, we processed a batch of N input material samples. Let $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ be a pair of embedding vectors produced for the ith crystal structure in a batch and its XRD pattern, respectively. For each positive pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, we randomly drew two kinds of negative samples $\boldsymbol{x}_i'$ and $\boldsymbol{y}_i'$, representing a crystal structure and an XRD pattern, respectively, from the batch to form two triplet losses:

$$L_{nx}^{(i)}(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{x}_i') = \max(0, \|\boldsymbol{x}_i - \boldsymbol{y}_i\| - \|\boldsymbol{x}_i' - \boldsymbol{y}_i\| + m), \tag{1}$$

$$L_{ny}^{(i)}(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{y}_i') = \max(0, \|\boldsymbol{x}_i - \boldsymbol{y}_i\| - \|\boldsymbol{x}_i - \boldsymbol{y}_i'\| + m), \tag{2}$$

where the negative sample $\boldsymbol{x}_i'$ was chosen from $\{\boldsymbol{x}_k\}_{k\neq i}$ to produce a positive-valued loss, $L_{nx}^{(i)} > 0$, and $\boldsymbol{y}_i'$ was chosen similarly from $\{\boldsymbol{y}_k\}_{k\neq i}$ (see also figure 2(b) for illustrations). Here, $m > 0$ is a hyperparameter called the margin. Equation (1) essentially requires that for each embedding $\boldsymbol{y}_i$, its negative samples $\boldsymbol{x}_i'$ are cleared out of the area surrounding $\boldsymbol{y}_i$ having the radius of the positive-pair distance $\|\boldsymbol{x}_i - \boldsymbol{y}_i\|$ (red circle in the top-right part of figure 2(b) plus the margin m (yellow area in the figure). Equation (2) is defined similarly. These losses are thus to ensure, given an embedding as a query, that its paired embedding is retrievable as the query's nearest neighbour. Note that the choice of the margin m is quite flexible because its value is relevant only to the scales of the embeddings, which are unnormalised and arbitrarily learnable. Here, $m = 1$. Our bidirectional triplet loss was then computed as the average of the losses for all samples in the batch as follows:

$$L = \frac{1}{2N}\sum_{i=1}^{N}(L_{nx}^{(i)} + L_{ny}^{(i)}). \tag{3}$$

This expression is similar to but simpler in form than a loss expression previously used in cross-modal retrieval [51].

We optimised the loss function using stochastic gradient descent with a batch size N equal to 512. Using the Adam optimiser [52] with a constant learning rate of 10^{-3}, we conducted iterative training for a total of 1000 epochs for all target materials in the dataset. The training took approximately one day using a single NVIDIA V100 GPU. For details regarding our strategies for validating the trained models and tuning the hyperparameters (e.g. choices of the embedding dimensionality and training batch-size), see appendix B and table S5 in the SI.

5.4. Data acquisition for the concept classification tasks

For the materials concept classification, we collected the crystal structure data of superconductors and thermoelectric materials from COD. To collect positive samples for each category, we retrieved material entries containing certain keywords in their paper titles as positive samples. Specifically, the entries including 'superconductor' or 'superconductivity' in their titles were regarded as superconductors, and the entries including 'thermoelectric' or 'thermoelectricity' were regarded as thermoelectric materials. The same number of material entries without these keywords were randomly collected and used as negative samples.

Data availability statement

The materials data retrieved from the Materials Project, the trained embeddings of these materials, and the trained ML model weights are available at the figshare repository [53]. The list of the target materials used in this study, the lists of the neighbourhood search results, and interactive web pages for exploring the materials map visualisation and analysing local neighbourhoods are available in the GitHub repository (https://github.com/quantumbeam/materials-concept-learning).

Acknowledgment

This work is partly supported by JST-Mirai Program, Grant Nos. JPMJMI21G2 and JPMJMI19G1. Y S is supported by JST ACT-I Grant Number JPMJPR18UE.

Author contributions

Y S, Y U, and K O conceived the idea for the present work. Y S and T T carried out the numerical experiments. All authors discussed the results and wrote the manuscript together.

Competing interests

The authors declare no conflicts of interest associated with this manuscri.

ORCID iDs

Yuta Suzuki https://orcid.org/0000-0002-0019-4832
Tatsunori Taniai https://orcid.org/0000-0003-3361-4861
Kotaro Saito https://orcid.org/0000-0003-1281-8099
Yoshitaka Ushiku https://orcid.org/0000-0002-9014-1389
Kanta Ono https://orcid.org/0000-0002-3285-9093

References

[1] De Graef M and McHenry M E 2012 *Structure of Materials: An Introduction to Crystallography, Diffraction and Symmetry* (Cambridge: Cambridge University Press)
[2] Callister W D 2007 *Materials Science and Engineering: An Introduction* (Wiley)
[3] Anderson P W 1997 *The Theory of Superconductivity in the High-Tc Cuprate Superconductors* (Princeton, NJ: Princeton University Press)
[4] Coey J M 2010 *Magnetism and Magnetic Materials* (Cambridge: Cambridge University Press)
[5] Manthiram A 2020 A reflection on lithium-ion battery cathode chemistry *Nat. Commun.* **11** 1–9
[6] van der Maaten L and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 2579–605
[7] Lookman T, Alexander F J and Rajan K 2015 *Information Science for Materials Discovery and Design* (Berlin: Springer)
[8] Butler K T, Davies D W, Cartwright H, Isayev O and Walsh A 2018 Machine learning for molecular and materials science *Nature* **559** 547–55
[9] Kajita S, Ohba N, Jinnouchi R and Asahi R 2017 A universal 3D voxel descriptor for solid-state material informatics with deep convolutional neural networks *Sci. Rep.* **7** 16991
[10] Schütt K T, Sauceda H E, Kindermans P J, Tkatchenko A and Müller K R 2018 SchNet—a deep learning architecture for molecules and materials *J. Chem. Phys.* **148** 241722
[11] Ziletti A, Kumar D, Scheffler M and Ghiringhelli L M 2018 Insightful classification of crystal structures using deep learning *Nat. Commun.* **9** 2775
[12] Xie T and Grossman J C 2018 Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties *Phys. Rev. Lett.* **120** 145301
[13] Chen C, Ye W, Zuo Y, Zheng C and Ong S P 2019 Graph networks as a universal machine learning framework for molecules and crystals *Chem. Mater.* **31** 3564–72
[14] DeFever R S, Targonski C, Hall S W, Smith M C and Sarupria S 2019 A generalized deep learning approach for local structure identification in molecular simulations *Chem. Sci.* **10** 7503–15
[15] Gasteiger J, Giri S, Margraf J T and Günnemann S 2020 Fast and uncertainty-aware directional message passing for non-equilibrium molecules *Machine Learning for Molecules Workshop (NeurIPS)*
[16] Choudhary K and DeCost B 2021 Atomistic Line Graph Neural Network for improved materials property predictions *npj Comput. Mater.* **7** 1–8
[17] Chen C and Ong S P 2022 A universal graph deep learning interatomic potential for the periodic table *Nat. Comput. Sci.* **2** 718–28
[18] Omee S S, Louis S-Y, Fu N, Wei L, Dey S, Dong R, Li Q and Hu J 2022 Scalable deeper graph neural networks for high-performance materials property prediction *Patter* **3** 100491
[19] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 Fast and accurate modeling of molecular atomization energies with machine learning *Phys. Rev. Lett.* **108** 058301
[20] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev.* B **87** 184115
[21] Hansen K, Biegler F, Ramakrishnan R, Pronobis W, von Lilienfeld O A, Müller K-R and Tkatchenko A 2015 Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space *J. Phys. Chem. Lett.* **6** 2326–31
[22] Faber F, Lindmaa A, von Lilienfeld O A and Armiento R 2015 Crystal structure representations for machine learning models of formation energies *Int. J. Quantum Chem.* **115** 1094–101
[23] Isayev O, Fourches D, Muratov E N, Oses C, Rasch K, Tropsha A and Curtarolo S 2015 Materials cartography: representing and mining materials space using structural and electronic fingerprints *Chem. Mater.* **27** 735–43
[24] Isayev O, Oses C, Toher C, Gossett E, Curtarolo S and Tropsha A 2017 Universal fragment descriptors for predicting properties of inorganic crystals *Nat. Commun.* **8** 15679
[25] Zhou Q, Tang P, Liu S, Pan J, Yan Q and Zhang S-C 2018 Learning atoms for materials discovery *Proc. Natl Acad. Sci. USA* **115** E6411
[26] Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson K A, Ceder G and Jain A 2019 Unsupervised word embeddings capture latent knowledge from materials science literature *Nature* **571** 95–98
[27] Ryan K, Lengyel J and Shatruk M 2018 Crystal structure prediction via deep learning *J. Am. Chem. Soc.* **140** 158–10
[28] Xie T and Grossman J C 2018 Hierarchical visualization of materials space with graph convolutional neural networks *J. Chem. Phys.* **149** 174111
[29] Schwaller P, Probst D, Vaucher A C, Nair V H, Kreutter D, Laino T and Reymond J-L 2021 Mapping the space of chemical reactions using attention-based neural networks *Nat. Mach. Intell.* **3** 144–52
[30] Choubisa H, Askerka M, Ryczko K, Voznyy O, Mills K, Tamblyn I and Sargent E H 2020 Crystal site feature embedding enables exploration of large chemical spaces *Matter* **3** 433–48
[31] Hoffmann J, Maestrati L, Sawada Y, Tang J, Sellier J M, and Bengio Y 2019 Data-driven approach to encoding and decoding 3-D crystal structures (arXiv:1909.00949 [cond-mat, physics: physics,stat])
[32] Noh J, Kim J, Stein H S, Sanchez-Lengeling B, Gregoire J M, Aspuru-Guzik A and Jung Y 2019 Inverse design of solid-state materials via a continuous representation *Matter* **1** 1370–84
[33] Noh J, Gu G H, Kim S and Jung Y 2020 Machine-enabled inverse design of inorganic solid materials: promises and challenges *Chem. Sci.* **11** 4871–81
[34] Court C J, Yildirim B, Jain A and Cole J M 2020 3-D inorganic crystal structure generation and property prediction via representation learning *J. Chem. Inf. Model.* **60** 4518–35

[35] Long T, Fortunato N M, Opahle I, Zhang Y, Samathrakis I, Shen C, Gutfleisch O and Zhang H 2021 Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures *npj Comput. Mater.* **7** 1–7

[36] Ren Z *et al* 2022 An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties *Matter* **5** 314–35

[37] Doersch C and Zisserman A 2017 Multi-task self-supervised visual learning *Proc. of the IEEE Int. Conf. on Computer Vision* pp 2051–60

[38] Kaya M and Bilge H S 2019 Deep metric learning: a survey *Symmetry* **11** 1066

[39] Manzeli S, Ovchinnikov D, Pasquier D, Yazyev O V and Kis A 2017 2D transition metal dichalcogenides *Nat. Rev. Mater.* **2** 17033

[40] Tokura Y and Arima T 1990 New classification method for layered copper oxide compounds and its application to design of new high T_c superconductors *Jpn. J. Appl. Phys.* **29** 2388–402

[41] Burch K S, Mandrus D and Park J-G 2018 Magnetism in two-dimensional van der Waals materials *Nature* **563** 47–52

[42] Schilling A, Cantoni M, Guo J D and Ott H R 1993 Superconductivity above 130 K in the Hg–Ba–Ca–Cu–O system *Nature* **363** 56–58

[43] Ihara H, Sugise R, Hirabayashi M, Terada N, Jo M, Hayashi K, Negishi A, Tokumoto M, Kimura Y and Shimomura T 1988 A new high-T_c $TlBa_2Ca_3Cu_4O_{11}$ superconductor with T_c >120K *Nature* **334** 510–1

[44] Ngiam J, Khosla A, Kim M, Nam J, Lee H and Ng A Y 2011 Multimodal deep learning *Proc. Int. Conf. Mach. Learn. (ICML)* pp 689–96

[45] Devlin J, Chang M, Lee K and Toutanova K 2018 BERT: pre-training of deep bidirectional transformers for language understanding *CoRR* (arXiv:1810.04805)

[46] Zunger A 2018 Inverse design in search of materials with target functionalities *Nat. Rev. Chem.* **2** 1–16

[47] Momma K and Izumi F 2011 VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data *J. Appl. Crystallogr.* **44** 1272–76

[48] Ong S P *et al* 2013 Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis *Comput. Mater. Sci.* **68** 314–9

[49] Park W B *et al* 2017 Classification of crystal structure using a convolutional neural network *IUCrJ* **4** 486–94

[50] Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift *Proc. Int. Conf. Mach. Learn. (ICML)* pp 448–56

[51] Wu Y, Wang S and Huang Q 2017 Online asymmetric similarity learning for cross-modal retrieval *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)* pp 3984–93

[52] Kingma D P and Ba J 2015 Adam: a method for stochastic optimization *Int. Conf. Learn. Repres. (ICLR)*

[53] Suzuki Y, Taniai T, Saito K, Ushiku Y and Ono K 2022 Self-supervised learning of materials concepts from crystal structures via deep neural networks *figshare* Dataset (https://doi.org/10.6084/m9.figshare.21717824)