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

Yuta Suzuki[1,2,†], **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.
[†](Current affiliation: Advanced R&D and Engineering Company, TOYOTA MOTOR CORPORATION, Shizuoka, Japan.)

Appendix A. Additional neighbourhood analysis and comparisons

In this appendix, we provide 1) additional local-neighbour analysis for notable materials. Using the same analysis approach, we also compare our approach with 2) its counterpart baseline methods and 3) existing hand-crafted descriptors[1].

A1. Local neighbourhood analysis for 2D materials and permanent magnets

2D materials are an interesting class of materials whose structural variations can yield diverse functionality. Particularly, 2D ferromagnets are gathering increasing attention from the magnetic materials community[2]. Therefore, we here inspected the neighbourhoods of $Cr_2Ge_2Te_6$, a 2D van der Waals crystalline insulator whose ferromagnetism was discovered in 2017[3]. Interestingly, its immediate neighbour was $CrSiTe_3$ (see the top-50 neighbour list in Table S1). This material was predicted to be a possible 2D compound through data mining and first-principles calculations in 2013[4] and has been studied as a potential 2D ferromagnetic insulator[5]. Similarly, other 2D van der Waals materials that are recently studied for interesting properties were in the neighbourhoods of $Cr_2Ge_2Te_6$. For example, $CrTe_3$ at the 4st neighbour has been studied for antiferromagnetism[6]. Most interestingly, another 2D ferromagnetic material, CrI_3[7], was found at 15th neighbour. Because the existence of 2D ferromagnets had been long questioned, the discoveries of $Cr_2Ge_2Te_6$ and CrI_3 in 2017 have been of immense interest to the magnetic materials community[2]. The structural similarity of two materials is only evident when the structures are visualised with appropriate bonding and polyhedra (see Fig. 7), which often requires a certain level of expertise. Nevertheless, our model places these materials close enough to be classified as 'neighbours', suggesting that the model captures their functionality-level similarity in the structures.

We next examined ferromagnetic materials for permanent magnets. $SmCo_5$ and Sm_2Co_{17} are two major components in the Sm-Co magnets and the structural similarity between them is well known among experts[8]. Because the 2-17 structure of Sm_2Co_{17} is reproduced by simply replacing some Sm ions in the 1-5 structure of $SmCo_5$ with a Co-Co dumbbell, we expect both materials to be closely located in the embedding space. The local-neighbour analysis showed that they were mutually found at about the top 0.5% neighbourhoods of each other. Furthermore, if we look for other 1-5 structures including those sharing same atomic positions but with different chemical compositions, we found $DyGaCo_4$ at the 246th neighbour (top 0.2%) of Sm_2Co_{17} (see Table S2). This result seems satisfactory considering the abundance of binary materials with several atoms in the unit cell located around $SmCo_5$ effectively pushing $SmCo_5$ out of the immediate neighbours of Sm_2Co_{17}.

Also interesting about the neighbours of Sm_2Co_{17} is the presence of other structural families related to permanent magnets. For example, the 2-14-1 family was found at the 60th ($Tb_2Co_{14}B$), 73th ($TbNdCo_{14}B$), 77th ($NdYCo_{14}B$), and 95th ($Nd_2Co_{14}B$). This family is famous for $Nd_2Fe_{14}B$, the main compound of the Nd-Fe-B magnet, which is essential for the modern society. Unlike the aforementioned connection between the 2-17 and 1-5 structures, this family has no such explicit structural connection with Sm_2Co_{17} in a way reasonably understandable to humans. Thus, the embedding seems to capture their functionality-level similarity as important structures of permanent magnets. Furthermore, the 1-12 ($ThMn_{12}$ type) structural families was observed as $DyMn_{12}$ (44th), $GdMn_{12}$ (45th), $SmCo_{12}$ (255th) and many others. The 1-12 families is gaining attention as parent compounds for next-generation permanent magnets[9–11]. The 1-12 and 2-17 structures of $SmCo_{12}$ and Sm_2Co_{17} have the known connection that both can be derived from $SmCo_5$[8,9]. However, without the literature context and proper visualisation, it is difficult for a human analyst to identify such a connection between a hexagonal structure and a tetragonal structure (see Extended Data Fig. 8). This result demonstrates that our model connects materials with similar functionality as neighbours by capturing their structural fingerprints that are obscure for human experts.

Since our embedding was learned solely from crystal structures without any human annotation, it is not constrained by human bias in principle. The additional analyses indeed show the materials relationships that are known in the literature but not evident to non-experts. Our analyses for the socially important, diverse material classes, from superconductors and battery materials to 2D materials and permanent magnets, strengthen the claim that our model recognises various materials concepts from crystal structures.

A2. Comparison with baseline approaches using DNN's latent feature vectors

In the main text, we discussed the two key factors of our approach that supposedly enabled the learning of materials concepts from crystal structures. Specifically, 1) explicit metric optimisation between embeddings via deep metric learning, and 2) cross-modal learning between the two complementary factors (the local structure and periodicity) of crystal structures. To support this hypothesis, we here compare our approach with its counterpart through local neighbourhood analysis. The counterpart methods thus 1) learn embeddings as DNN's latent vectors trained by a surrogate task without explicit metric optimisation, 2) using only a single form of input expression (either crystal structures or XRD patterns). Essentially, the existing material embedding learning methods[12–15] fall into this counterpart methodology with differences in training tasks, input expressions, and encoder architectures.

Among various choices for a surrogate training task, we adopted the prediction of the total total energy by following the existing approach by Xie *et al.*[15]. The total energy is a fundamental physical measure of crystal structures that is closely related to their chemical bonds. Because the chemical bond is a basis for various properties of all materials including inorganic compounds, the approach by Xie *et al.* can be justified based on the idea that if the total energy is predicted accurately from an embedding, it well describes the crystal structure. Similar to Xie *et al.*, our objective in this study is to build a single ML model that can universally recognise various materials concepts. Therefore, we also consider the total energy as an appropriate prediction target that is not directly coupled with specific functionality but is related to diverse characteristics of materials.

For comparison, we prepared two baseline methods by borrowing our two encoders (Fig. 2a). Each baseline adopted either the crystal-structure encoder or the XRD pattern encoder, whose final layer was modified to output a scalar prediction value of the total energy. See also the Appendix D for the detailed network architectures of our encoders. The two models were trained to minimise the mean squared error between the predicted and simulated values of the total total energy. We conducted iterative training for 500 epochs similarly to the procedures given in the *Methods* section. After the training, the latent vectors that are fed to the middle layer (the one before the final layer) were collected as embeddings, which have the same 1024-dimensions as ours.

Tables S3 and S4 list the top-50 neighbours of Hg-1223 and $LiCoO_2$, respectively, comparing our embedding with the two baselines. For these two materials, our embedding successfully captured high-T_c superconductors similar to Hg-1223 and the important three families of lithium-ion battery cathode materials similar to $LiCoO_2$, as discussed in the main text. Since these results were produced by using the inputs and encoder that are essentially the same as those of the crystal-structure-based baseline, this baseline should at least have the potential to produce similar results. Indeed, the list of neighbours of the two baselines suggests conceptual material similarity at the level of roughly comparable to our approach. As Xie *et al.*[15] point out, the total energy is a fundamental materials parameter, so it is not surprising that DNNs could indirectly learn embeddings that capture the concept of materials in a supervised learning framework with labels of total energy. From this result, we confirm that our two encoders were both well enough to learn the material concept. We conclude that the two factors of our approach, namely, explicit metric optimisation on embeddings and cross-modal learning, are comparable to supervised learning with labels by large-scale ab initio calculation.

When these two factors are combined in our method, they form the training task of cross-modal retrieval, as discussed in the *Methods* section. This task is to ensure that each embedding is uniquely identifiable among others as the nearest neighbour of its paired embedding given as a query. The task of learning uniquely identifiable embeddings can be considered a more direct approach to learning distinctive features and concepts of individual materials, compared to other surrogate tasks used in the existing methods[12–15]. As the analysis in Appendix B reveals that our method successfully carried out the retrieval task, this reasoning from the aspect of the training task could also account for the success of our approach.

A3. Comparison with traditional hand-crafted descriptors

We here provide comparative neighbourhood analysis of traditional hand-crafted descriptors, whose detailed discussions were omitted from the main text. In particular, we examined Ewald Sum Matrix (ESM)[1] and Sine Coulomb Matrix (SCM)[1]. We also investigated other choices such as the Smooth Overlap of Atomic Positions (SOAP)[16, 17] and the Bag of Bonds (BoB)[18]. However, these methods could not scale to our dataset of 122,543 materials, as the dimensions of these descriptors can grow extremely large for a dataset containing a large number of chemical elements. In our preliminary analysis, the descriptors of

SOAP and BoB became 170k and 500k dimensions, respectively, for a random subset containing 5,000 materials (4%) of our dataset. Thus, ESM and SCM were chosen as representative hand-crafted descriptors of crystal structures that were applicable to the dataset scale of interest in this study.

ESM[1] is viewed as an extension of the Coulomb matrix[19] for periodic systems. ESM forms a symmetry matrix whose elements model the electrostatic interaction between atoms, i and j, in the primitive cell of a crystal structure as follows.

$$M_{ij}^{\mathrm{ESM}} = \begin{cases} x_{ij}^{\mathrm{real}} + x_{ij}^{\mathrm{recip}} + x_{ij}^{\mathrm{self}} + x_{ij}^{\mathrm{bg}} & \text{for } i = j \\ 2\left(x_{ij}^{\mathrm{real}} + x_{ij}^{\mathrm{recip}} + x_{ij}^{\mathrm{bg}}\right) & \text{for } i \neq j \end{cases} \tag{S1}$$

Here, x_{ij}^{real} and x_{ij}^{recip} encode the short- and long-range interactions between atoms in the real and reciprocal spaces, respectively, x_{ij}^{self} represents the self-energy correction, and x_{ij}^{bg} is a constant term introducing a uniform background charge to neutralise the system. Note that the formulation in Equation S1 follows the modified ESM definition used in the DScribe library[20], which fixes an issue related to the self-energy and the background-charge correction in the original work[1].

SCM[1] is another variant of the Coulomb matrix for periodic systems. Although ESM computes the correct electrostatic interactions between atoms, this computation can be heavy for large systems. SCM aims to reduce the computational effort by replacing the long-range interaction with a simpler expression[1,20].

For a crystal structure containing N atoms in the primitive cell, these Coulomb matrix variants produce a $N \times N$ matrix whose rows and columns are ordered by the indices of the atoms in the cell. This form is problematic when evaluating the distance between two descriptors, because the descriptor sizes can be inconsistent among materials and the descriptor representations depend on the ordering of atomic indices. To allow the distance evaluation between descriptors, we used the schemes suggested by Himanen *et al.*[20]. Specifically, we computed the eigenvalues of ESM and SCM sorted by their absolute value in descending order, and then applied the zero-padding to the eigenvalue vectors according to their maximum dimension among the dataset. Consequently, ESM and SCM were converted to 444-dimensional vectors, which effectively compress the original matrices that have at most 197k (444^2) dimensions.

Tables 1 and 2 show the top-50 neighbourhoods of Hg-1223 and $\mathrm{LiCoO_2}$, respectively, obtained by ESM and SCM in comparison with our embedding discussed in the main text. Likewise, Tables S1 and S2 show the comparisons for $\mathrm{Cr_2Ge_2Te_6}$ and $\mathrm{Sm_2Co_{17}}$, respectively, discussed in the Appendix A1 above. As shown in these tables, the conceptual similarities of materials captured in our embedding space are not observed in the results of ESM and SCM.

In addition to this superior ability in capturing conceptual material similarity, our method has other advantages in terms of its scalability and representation over existing hand-crafted descriptors. As explained above, existing descriptors such as SOAP[16,17] and BoB[18] tend to suffer from the scalability issues when applied to a large-scale dataset. ESM and SCM could also produce 197k-dimensional descriptors for our dataset if not compressed by eigenvalues. These scalability issues stem from the fact that the dimensions of existing descriptors often vary according to, for example, the number of chemical elements contained in the target dataset as in SOAP and BoB, or the system sizes of individual materials as in ESM and SCM. By contrast, our method can produce embedding vectors of predefined fixed size, regardless of the sizes and scales of input crystal structures and target datasets. This consistent representation is important for ML applications[20]. With the ML-friendly fixed-size (1024-dimensional) vectors, our model was able to uniquely describe more than 10,000 materials, as revealed in Appendix B, while recognising materials concepts in a dataset of over 120,000 materials, as demonstrated in the main text.

Table S1. The top-50 neighbours of $Cr_2Ge_2Te_6$ ($CrGeTe_3$) in comparison with existing descriptors.

	Our embedding		Ewald Sum Matrix		Sine Coulomb Matrix	
No.	Formula	ID	Formula	ID	Formula	ID
Query	CrGeTe3	mp-541449	CrGeTe3	mp-541449	CrGeTe3	mp-541449
1	CrSiTe3	mp-3779	InSiTe3	mp-567931	Fe2Te3	mp-685077
2	Cd2As3Br	mp-28900	CrSiTe3	mp-3779	Ga2Te3	mp-38970
3	Cr4Cu3Te8	mp-675546	Co(PdSe)2	mp-12464	Ni2SbTe2	mp-3250
4	CrTe3	mp-540922	Al2Te3	mp-1228524	Ga2Te3	mp-32580
5	Mg2SiSe4	mp-1192582	Ba4Al	mp-1214528	K3AsI6	mp-1111178
6	In2Ag2GeSe6	mp-505607	K2YCuI6	mp-1112213	K3GaI6	mp-1111270
7	Cr2CuTe4	mp-22625	Rb3ScI6	mp-1114633	Te3As2	mp-484
8	Ba4Cd11Ge12	mp-1214704	Rb2AlInI6	mp-1114521	Sc2Te3	mp-32654
9	CsYZnSe3	mp-574620	K2GaAgI6	mp-1112466	K3YI6	mp-1113611
10	BaCu6Te6S	mp-1228010	K2NaScI6	mp-1111618	K3ScI6	mp-1111693
11	Sc19(RuBr7)4	mp-1219646	Rb2GaAgI6	mp-1113726	CrSiTe3	mp-3779
12	BaCu6Te6Se	mp-1228039	Sc2Te3	mp-32654	RbCrI3	mp-676553
13	Mn2In2Se5	mp-1222074	K3ScI6	mp-1111693	K2RbGaI6	mp-1111285
14	InSe	mp-21405	K2AgMoI6	mp-1112093	Na3YI6	mp-1113485
15	CrI3	mp-1213805	Ti2Te3	mp-1217180	K2YCuI6	mp-1112213
16	In2Si(AgSe3)2	mp-640614	Cs3AlI6	mp-1112654	K2RbAsI6	mp-1111606
17	InAgS2	mp-1097000	Nb(SeI)2	mp-1205627	K2RbAlI6	mp-1111610
18	Rb2Cd3Se4	mp-16818	Yb(Mo3S4)2	mp-2945	K3AlI6	mp-1111183
19	Cd4GeSe6	mp-18163	Ba3LiN	mp-13288	Cs2SnAs2	mp-8934
20	CsYMnSe3	mp-1213646	RbCu2I3	mp-1103650	K2NaScI6	mp-1111618
21	In4Se3N2	mp-1246310	Ba4Pd	mp-1214438	Na3ScI6	mp-1113505
22	RbFe2Te3	mp-15121	Zr10HN8	mp-674456	LiGe3SbTe5	mp-1222357
23	TePdI2	mp-573321	Zr4Mo	mp-1207454	K2NaYI6	mp-1111220
24	Cd2GeAs4	mp-5712	Cs2CoSe2	mp-8770	AlSiTe3	mp-31220
25	CsYCdSe3	mp-11116	Tm(Mo3S4)2	mp-1103493	K3MoI6	mp-1111267
26	HoAgS2	mp-1199297	Dy(Mo3S4)2	mp-1103518	K2LiYI6	mp-1111243
27	TbAgS2	mp-1208370	K2RbGaI6	mp-1111285	Al2Te3	mp-1228524
28	GdAgS2	mp-1200242	Cd2PCl2	mp-31276	K2CuMoI6	mp-1112050
29	Cr2AgTe4	mp-20118	Na6MnTe4	mp-14782	Rb3AsI6	mp-1114618
30	DyAgS2	mp-1200233	K2NaMoI6	mp-1111633	Rb3GaI6	mp-1114499
31	In6S7	mp-555853	Cs2As2Pd	mp-8857	K2RbYI6	mp-1114560
32	Ba(ZnSb)2	mp-14207	KRb2AsI6	mp-1114510	K2NaMoI6	mp-1111633
33	Cd2As3I	mp-27577	Cs2In3	mp-567752	Rb2NaScI6	mp-1114457
34	Mn2ZnTe4	mp-1104014	Zr2Ga3	mp-30686	Rb3YI6	mp-1114639
35	Mg(ScSe2)2	mp-1001019	Rb2LiYI6	mp-1114584	K2LiMoI6	mp-1111254
36	Mg2Al2Se5	mp-29624	SrCaI4	mp-1101345	Rb2YCuI6	mp-1112410
37	AlInSe3	mp-862787	K3MoI6	mp-1111267	Rb3ScI6	mp-1114633
38	K(FeTe)2	mp-1068789	Zr4Zn	mp-1207459	K2RbMoI6	mp-1114406
39	ErAgS2	mp-36029	La4S7	mp-1223154	Rb2CuMoI6	mp-1112459
40	RbIn3S5	mp-542654	K2CuMoI6	mp-1112050	La2Fe2I	mp-30223
41	Rb7(FeTe2)4	mp-1194713	Rb2NaYI6	mp-1114603	Rb3AlI6	mp-1114616
42	Cs5In3As4	mp-582182	K6MnTe4	mp-18246	K2ScAgI6	mp-1112086
43	Cs(SbSe2)2	mp-3312	Rb2YCuI6	mp-1112410	Rb2NaYI6	mp-1114603
44	Mn2SiSe4	mp-17367	SiI3	mp-1078195	K2GaAgI6	mp-1112466
45	Ti5Te8	mp-1208221	Rb3GaI6	mp-1114499	Rb2LiYI6	mp-1114584
46	Ag15P4S16Cl3	mp-560328	Rb2Te	mp-383	Rb2NaMoI6	mp-1114447
47	V3Te4	mp-1028	Sr2CaI6	mp-754710	NbI3O	mp-546285
48	Cr5Te8	mp-1213754	Ti5Sb2Rh	mp-16687	Rb2LiMoI6	mp-1114569
49	YAgS2	mp-1207671	Rb3AsI6	mp-1114618	Ca2InPd2	mp-20792
50	TiCu2Te3	mp-541754	KRb2ScI6	mp-1110633	Na2GaAgI6	mp-1111188

We compare the top-50 neighbours of the 2D ferromagnet $Cr_2Ge_2Te_6$ obtained by using our embedding and existing descriptors[1]. Our embedding well captured 2D materials that are gathering attention as promising new electronic-device materials in the materials science community. As mentioned in the text, $CrSiTe_3$ (No. 1) is a potential 2D ferromagnet similar to $Cr_2Ge_2Te_6$ (query), and $CrTe_3$ (No. 4) are studied for ferroelectricity and antiferromagnetism, respectively. In the lists of ESM and SCM, our first neighbour $CrSiTe_3$ also exists but at lower positions, No. 2 and No. 11, respectively. Note that the other 2D ferromagnet CrI_3 mentioned in the text was in the 15th neighbours by our embedding but was absent in the top-1000 neighbours by ESM and SCM.

Table S2. The top-50 neighbours of Sm_2Co_{17} in comparison with existing descriptors.

	Our embedding		Ewald Sum Matrix		Sine Coulomb Matrix	
No.	Formula	ID	Formula	ID	Formula	ID
Query	Sm2Co17	mp-1200096	Sm2Co17	mp-1200096	Sm2Co17	mp-1200096
1	Gd2Co17	mp-1201816	Ho2Fe5Co12	mp-1197249	Sm2Ni17	mp-1203310
2	PrErCo17	mp-1220026	Tb2Co17	mp-1199370	Sm4Ga3Fe31	mp-1219432
3	Ce2Co17	mp-2216	Ho2Co12Ni5	mp-1204922	Sm4Fe31Co3	mp-1219400
4	Tb2Co17	mp-1199370	NdEr3Fe34	mp-1220311	Sm4CrFe33	mp-1219321
5	SmGdCo17	mp-1219295	Dy2Ni17	mp-1197654	Eu2Ni17	mp-1201182
6	Dy2Co17	mp-569638	Yb2Fe17	mp-1195706	Sm4ZrFe33	mp-1219455
7	Eu2Ni17	mp-1201182	Ho2Ni17	mp-1202187	Sm4TiFe33	mp-1219364
8	Ce2VCo16	mp-1227655	Tm2Ga2Fe15	mp-1203778	Sm4Cr3Fe31	mp-1219348
9	YbPrCo17	mp-1215870	Lu2Fe17	mp-1195842	TbNd3Fe34	mp-1217543
10	Ce2Co16Cu	mp-1227675	Dy2Mn12Ga5	mp-1237201	Gd2Co17	mp-1201816
11	Nd2Ni17	mp-570596	Lu2Ni17	mp-1202260	Nd2Ni17	mp-570596
12	PrSmCo17	mp-1219785	Lu2Co17	mp-1204082	Gd2Ni17	mp-580102
13	YbPr3Co34	mp-1215883	Tm2Co17	mp-1196360	Gd2Fe17	mp-1196805
14	SmYCo17	mp-1219047	Tm2Ga3Fe14	mp-1197720	Pr3DyFe34	mp-1219904
15	Sm2Ni17	mp-1203310	Tm2Fe15Si2	mp-1200417	Nd3ErFe34	mp-1220953
16	CeYCo17	mp-1226612	Er2Fe17	mp-1724	Sm4V20(CuO4)15	mp-1219719
17	Nd2Co17	mp-356	Yb2Ni17	mp-1199108	Tb3SmFe34	mp-1217679
18	Gd2Ni17	mp-580102	Tm2Ni17	mp-11527	Tb3NdFe34	mp-1217666
19	Ho2Co17	mp-1023	Ho2Fe17N3	mp-1212403	Sm2ZrCo16	mp-1219324
20	Ho2Co12Ni5	mp-1204922	Lu2Mn17C3	mp-1211163	Pr3ErMn6(FeCo13)2	mp-1220144
21	Er2Co12Ni5	mp-1203663	Tb2Fe17H3	mp-1208578	Tb2Co17	mp-1199370
22	Er2Co17	mp-2531	Er2Mn17C3	mp-1213058	Tb2Fe17	mp-1194635
23	Tb2Ni17	mp-569945	Ho2Mn17C3	mp-1212558	Sm4Fe31Si3	mp-1219345
24	Tb2Ga3Co14	mp-1217733	Dy2Al2Fe15	mp-1196052	Tb2Ni17	mp-569945
25	Sm2ZrCo16	mp-1219324	Ce2Co17H3	mp-1213920	Pr2Mn12Co5	mp-1232416
26	Ho2Fe5Co12	mp-1197249	Er2Fe17H3	mp-1213007	CePr3Fe34	mp-1227066
27	TbCo9Si2	mp-1191366	Tm2Fe17H3	mp-1208090	Pr4AlFe33	mp-1219956
28	Pr2Co16Cu	mp-1219957	Tm2Fe17C3	mp-1208084	YbPr3Co34	mp-1215883
29	Pr2Cr2Co15	mp-1219992	Dy2Fe17H3	mp-1213248	Ce2Co16Cu	mp-1227675
30	Sm2Fe4Co13	mp-1219231	Tm2Al2Fe15	mp-1198100	Sm4Fe27Co7C2	mp-1219288
31	La4TaCo33	mp-1224958	Ho2Fe17	mp-1196975	Ce2Co17	mp-2216
32	Dy2Ni17	mp-1197654	Tb2Fe17	mp-1194635	Ce2Fe17	mp-1195962
33	SmMn5Co7	mp-1219042	Dy2Fe17	mp-1196404	Pr2Zn17	mp-976812
34	Tm2Co17	mp-1196360	Ho2Fe17C	mp-1224658	NdErFe17	mp-1220296
35	Yb4ZrCo33	mp-1216133	Yb2Co17	mp-1199900	Dy2Fe17	mp-1196404
36	Yb2Co17	mp-1199900	SmEr3Fe34	mp-1219139	Dy2Co17	mp-569638
37	La2VCo16	mp-1223090	Gd2Fe17	mp-1196805	Ce2VCo16	mp-1227655
38	Sm2Ga2Co15	mp-1188906	Tm2Fe17	mp-30640	Dy2Ni17	mp-1197654
39	Er2Ni17	mp-30608	Ho2Co17	mp-1023	Tb2Zn17	mp-30880
40	ErCo9Si2	mp-1191958	Er2Co17	mp-2531	Dy2Ga3Fe14	mp-1203342
41	Ce2Co17H3	mp-1213920	Er2Co12Ni5	mp-1203663	Sm4Fe34C3	mp-1219344
42	Ce4AlCo25	mp-1227640	Dy2Co17	mp-569638	Nd4Fe29Si5	mp-1220603
43	Sm2Co16Ag	mp-1219201	Er2Al3Fe14	mp-1199551	Eu2Ni12P5	mp-1213550
44	DyMn12	mp-20656	Er2Ni17	mp-30608	Ce2ZrCo16	mp-1227870
45	GdMn12	mp-639892	Ce2Co17	mp-2216	Ce2Zn17	mp-978252
46	SmCo9Si2	mp-17623	AuSCl7	mp-556587	Dy2Mn12Ga5	mp-1237201
47	Y2Co14Cu3	mp-1199930	NaHo(PO3)4	mp-1195468	Gd4Fe34C3	mp-1225869
48	NdCo9Si2	mp-1191853	Gd2Ni17	mp-580102	TbMn5Ge3	mp-623463
49	Ce2Si2Ni15	mp-1202894	Dy2Zn17	mp-570071	Nd2Ni12P5	mp-1210070
50	Y2Co17	mp-570718	Tb2Ni17	mp-569945	Pr4Fe29Si5	mp-1220120

We compare the top-50 neighbours of the Sm_2Co_{17} permanent magnet obtained by using our embedding and existing descriptors[1]. In the above lists, the similarity of the R_2M_{17} family, with different rare-earth metals R and transition metals M, was captured by all of the three methods. Our embedding further captured another major permanent magnet family RM_{12}, the candidate for parent compounds for next-generation permanent as the neighbours at No. 44–45.

Table S3. The top-50 neighbours of Hg-1223 in comparison with latent vectors obtained via a surrogate task (total energy prediction).

	Our embedding		Crystal structure encoder		XRD pattern encoder	
No.	Formula	ID	Formula	ID	Formula	ID
Query	Ba2Ca2Cu3HgO8	mp-22601	Ba2Ca2Cu3HgO8	mp-22601	Ba2Ca2Cu3HgO8	mp-22601
1	Ba2Ca3Cu4HgO10	mp-1228579	Ba6Ca6Cu9Hg3O25	mp-1228760	Ba6Ca6Cu9Hg3O25	mp-1228760
2	Ba2CaCu2HgO6	mp-6879	Ba2Ca3Cu4HgO10	mp-1228579	Ba4Ca4Cu6Hg2O17	mp-1228265
3	Ba6Ca6Cu9Hg3O25	mp-1228760	Ba2CaCu2HgO6	mp-6879	La21Fe8Sb7C12	mp-582023
4	Sr2CaCu2(BiO4)2	mp-1218930	Ba8Ca4Cu8Hg4O25	mp-1228371	La2BiN	mp-1078349
5	Ba10Ca5Cu10Hg5O31	mp-1229139	Ba4Ca4Cu6Hg2O17	mp-1228265	La11(MnC6)3	mp-1195612
6	SrCa2Cu2(BiO4)2	mp-1208800	Ba10Ca5Cu10Hg5O31	mp-1229139	La21Mn8Sn7C12	mp-1201735
7	Ba8Ca4Cu8Hg4O25	mp-1228371	Ba6Ca3Cu6Hg3O19	mp-1228161	Ba6Ca6Tl5Cu9O29	mp-680433
8	Ba2Ca3Tl2(CuO3)4	mp-556574	Ba6Ca12Cu15Hg3O37	mp-1229082	CeS	mp-20560
9	Ba2Mg3Tl2(WO3)4	mvc-129	Ba6Ca15Cu18Hg3O43	mp-1229281	Ba8Ca8Tl7(Cu4O13)3	mp-1204270
10	Ba2TlV2O7	mvc-2978	Ba4Ca8Cu11CO20	mp-1228570	Rb2P	mp-1101799
11	Sr2YCu2(BiO4)2	mp-1208863	BaCa2Cu3O5	mp-1214453	Sr2CaCu2(BiO5)2	mp-1218932
12	Sr2LaCu2HgO6	mp-1208803	Ba2Ca3Tl2(CuO3)4	mp-556733	ZnAgF3	mp-998537
13	Ba2CaTl2(CuO4)2	mp-573069	Sr3La(CuO2)4	mp-1218623	Cs2MnCl4	mp-1025252
14	Ba4CaCu6(HgO8)2	mvc-15237	CaCuO2	mp-554775	Sr11(SiN5)2	mp-1246141
15	Ba4Ca4Cu6Hg2O17	mp-1228265	SrCa3(CuO3)2	mp-1218400	LiNdTiO4	mp-10520
16	Ba2AlTlCo2O7	mvc-2977	Ba3CaLa2Cu6O13	mp-1228590	Ba2CaCu2HgO6	mp-6879
17	Sr8Pr4Cu9(HgO8)3	mp-1218674	SrCaCuO3	mp-1218361	CeAl3Pt	mp-1226648
18	Ba6Ca3Cu6Hg3O19	mp-1228161	Ba6Ca6Tl5Cu9O29	mp-680433	TlBSe3	mp-29959
19	Ba8Ca8Tl7(Cu4O13)3	mp-1204270	Ba8Ca8Tl7(Cu4O13)3	mp-1204270	RbLa2Ti2NbO10	mp-1219633
20	Ba4Ca4Tl3Cu6O19	mp-542197	Ba2Ca3TlCu4O11	mp-1228589	Ba4Ca4Tl3Cu6O19	mp-542197
21	Ba6Ca6Tl5Cu9O29	mp-680433	Ba2Pr(CuO2)3	mp-1214585	Gd3MnAlS7	mp-1191013
22	Ba2AlTlCo2O7	mp-1266279	Ba2Nd(CuO2)3	mp-614981	Nd3GaCoS7	mp-1192335
23	Ba2Ca2Tl2Ni3O10	mvc-3067	SrCa(CuO2)2	mp-1218417	Sr11(GeN5)2	mp-1245458
24	Ba2Ca2Tl2Cu3O10	mp-653154	Ba2Ca3Tl2(CuO3)4	mp-556574	Sr3RuN3	mp-1029750
25	Ba2Ca2Tl2Co3O10	mvc-3021	Sr2Ca2Ga(CuO3)3	mp-1209020	La21Mn8Sb7C12	mp-1203312
26	Sr2CaCu2(BiO4)2	mp-555855	Ba4Ca4Tl3Cu6O19	mp-542197	Rb2SbBr6	mp-568477
27	Ba4Tl2Cu2HgO10	mp-561182	Ba2Y(CuO2)3	mp-1021507	MnTlCuSe2	mp-1221565
28	Ba6Ca12Cu15Hg3O37	mp-1229082	Ba4Pr2Cu6O13	mp-1228176	BaTb2O4	mp-18258
29	BaCuReO5	mvc-7248	Sr8Pr4Cu9(HgO8)3	mp-1218674	Tl5NO5	mp-1101007
30	Ba2Ca3Tl2(FeO3)4	mvc-145	Ba8CaY3(CuO2)12	mp-1228323	LaNb2O7	mp-1079978
31	Sr10Cu5Bi10O29	mp-667638	Sr3Ca(CuO3)2	mp-1218473	TbOF	mp-14093
32	Ba2Ca3TlCu4O11	mp-1228589	Sr16Cu8O23	mp-759634	SrNdMnO4	mp-1217982
33	Ba2Ca3Tl2(CuO3)4	mp-556733	Ca3Cu2(ClO2)2	mp-23095	LiEu4C3(IN2)3	mp-638276
34	La2B3Br	mp-568985	Ba2Sm(CuO2)3	mp-622576	Yb2Be2GeO7	mp-1207637
35	BaTl(SbO3)2	mvc-10727	Ba2NdCu2HgO7	mp-1214587	NaNdTiO4	mp-20980
36	Sr10Cu5Bi10O29	mp-652781	Sr2CaCu2(BiO4)2	mp-1218930	Sr2EuCu2(BiO4)2	mp-1208972
37	Ba2Tl2Zn2Cr3O10	mvc-3164	Ba10Sm5(Cu5O11)3	mp-1229115	SrAgTeF	mp-1080438
38	Ba2Ca2Tl2Fe3O10	mvc-3027	Ba2CuO3	mp-8790	Ba10BrN5Cl4	mp-1228725
39	Ba2Ti3Tl2O10	mvc-2939	Ba4Nd2Cu6O13	mp-1228184	Ba2Ca3Tl2(CuO3)4	mp-556574
40	Sr2TaAlCu2O7	mp-1251503	Ba3SrSm2(CuO2)6	mp-1228212	Ce4Cu3(SO)4	mp-1226848
41	Ba2Mg3Tl2(SnO3)4	mvc-10576	Sr9Nd3Cu12(PbO4)8	mp-1218827	AgCNO	mp-561891
42	Sr2AlTlCo2O7	mp-1252241	Ba2CuC(NO)2	mp-1021669	Pr2BC	mp-1078268
43	Ba2AlTlV2O7	mp-1265780	Sr6Pr3Cu6O17	mp-1218599	Sr3TiN3	mp-1245686
44	Ba2CaTl2(CuO4)2	mp-6885	Ca2CuO3	mp-5869	SrLaMnO4	mp-1218183
45	Sr2LaCu2(BiO4)2	mp-1209034	Ba2Ca2Tl2Cu3O10	mp-653154	La21Fe8Sn7C12	mp-607917
46	Ba2AlTlV2O7	mvc-3002	Sr4Cu2O7	mp-766217	La16Ni8O33	mp-867595
47	Ba2Mg3Tl2(FeO3)4	mvc-28	Ba6Sm3Al(Cu2O5)4	mp-1228395	La20Mn8Te7C12	mp-1223565
48	Sr2DyCu2(BiO4)2	mp-1209149	SrCuO2	mp-5787	Bi3PbWClO8	mp-1227592
49	Ba2CuHgO4	mp-6562	Ba4La2Cu6O13	mp-1228239	Sm10As8Au3O10	mp-1194552
50	Ba2Tl2W3O10	mvc-3144	BaSrSm(CuO2)3	mp-1227431	Rb2LaNb2ClO7	mp-1209483

Our embedding is compared with its two counterpart baseline methods through the neighbourhoods of the Hg-1223 superconductor. These baselines used either our crystal-structure encoder or XRD pattern encoder to learn embeddings as latent vectors in the DNNs, which were trained to predict the total energy. As also discussed in Tables 1, our embedding successfully captured high-T_c superconductors similar to Hg-1223. The majority of the two baseline neighbourhood lists are also occupied by high-T_c superconductors. It is worth noting that our approach, using a self-supervised learning framework, achieves embeddings comparable to those obtained using total energy labels, even though no annotations by ab initio calculations or experts. See also Table S4 for another comparison.

Table S4. The top-50 neighbours of $LiCoO_2$ in comparison with latent vectors obtained via a surrogate task (total energy prediction).

	Our embedding		Crystal structure encoder		XRD pattern encoder	
No.	Formula	ID	Formula	ID	Formula	ID
Query	LiCoO2	mp-22526	LiCoO2	mp-22526	LiCoO2	mp-22526
1	Li14MgCo13O28	mp-769537	Li14MgCo13O28	mp-769537	Li14MgCo13O28	mp-769537
2	Li4Co3NiO8	mp-867537	Li4Co3NiO8	mp-867537	Li9Co7O16	mp-1175469
3	Li3Fe(CoO3)2	mp-761602	Li2CoO2F	mp-764063	Li9Co7O16	mp-1175381
4	Li3(CoO2)4	mp-850808	Li3MnCo3O8	mp-758163	Li20(CoO2)21	mp-532301
5	Li3MnCo3O8	mp-774219	Li3(CoO2)4	mp-850808	Li3(CoO2)4	mp-850808
6	Li20(CoO2)21	mp-532301	Li10Fe3Co7O20	mp-760848	CrCo3O8	mp-754623
7	Li3CrCo3O8	mp-849768	Li3MnCo3O8	mp-774219	Li2CoNi3O8	mp-752703
8	Li3MnCo3O8	mp-758163	Li20(CoO2)21	mp-532301	Li14Co13O28	mp-777836
9	Li8FeCo9O20	mp-764865	Li3Fe(CoO3)2	mp-761602	Li(NiO2)2	mp-774941
10	Li3Co2NiO6	mp-765538	Li4MgCo3O8	mp-754576	MnCo3O8	mp-773602
11	Li3CrCo3O8	mp-759149	Li8FeCo9O20	mp-764865	Li4Co3NiO8	mp-867537
12	Li3TiCo3O8	mp-757214	Li4FeCo3O8	mp-765603	Li7Co5O12	mp-1174196
13	Li4MgCo3O8	mp-754576	Li10FeCo9O20	mp-764262	Li4MgCo3O8	mp-754576
14	Li5Co2Ni3O10	mp-769553	Li20Co21O40	mp-685270	Li2FeCo3O8	mp-867710
15	Li(CoO2)2	mp-552024	Li14Co13O28	mp-777836	Li2(CoO2)3	mp-758539
16	Li14Co13O28	mp-777836	Li3Mn(CoO3)2	mp-761633	Li3MnCo3O8	mp-774219
17	Li3(NiO2)5	mp-762165	Li2(CoO2)3	mp-758539	Li4Co2Ni3O10	mp-778996
18	Li2CoO2F	mp-764063	Li5Fe2Co3O10	mp-769566	Li8FeCo9O20	mp-764865
19	Li2(CoO2)3	mp-758539	Li7Co5O12	mp-771155	Li7Co5O12	mp-771155
20	Li5Fe2Co3O10	mp-769566	Li7Co5O12	mp-1174196	Co3NiO8	mp-752738
21	Li2CoNi3O8	mp-752703	Li4Mn3(CoO4)3	mp-755918	Li7Si2(NiO4)3	mp-756986
22	Li10Fe3Co7O20	mp-760848	Li3CrCo3O8	mp-849768	Li9Co7O16	mp-1175409
23	Li7Co5O12	mp-771155	Li2NbCo3O8	mp-757558	Li2VCo3O8	mp-754294
24	Li3(NiO2)4	mp-755972	Li4Mn3Co5O16	mp-754275	Li2CrCo3O8	mp-761748
25	Li9Ni15O28	mp-759153	Li2MnCo3O8	mp-761940	Li2FeCoO4	mp-1222775
26	Li20Co21O40	mp-685270	Li3TiCo3O8	mp-757214	LiAlO2	mp-8001
27	Li7(NiO2)11	mp-768079	Li2MnCo3O8	mp-757572	Li9Co7O16	mp-1175506
28	Li2(NiO2)3	mp-762391	Li9Co7O16	mp-1175469	Li7Ni5O12	mp-755638
29	Li4Co2Ni3O10	mp-778996	Ca(CoO2)2	mp-17544	MnCoO4	mp-752945
30	Li2Co3NiO8	mp-757851	Li3Co2NiO6	mp-765538	Mn3NiO8	mp-775810
31	LiCoNiO4	mp-754509	Li20Co21O40	mp-705640	Li9Ni15O28	mp-759153
32	Li4(NiO2)7	mp-774600	Li9Co7O16	mp-1175409	Li2VSi3O8	mp-766402
33	Li(CoO2)2	mp-774082	Li2VCo3O8	mp-757835	Li7Ni13O24	mp-758593
34	Li(CoO2)2	mp-752807	Li2(CoO2)3	mp-758725	Li(NiO2)2	mp-752531
35	Li8(NiO2)11	mp-758772	Li5Co3(NiO5)2	mp-755076	Li5Co2Ni3O10	mp-769553
36	Li3CoNi3O8	mp-774300	Li8Fe3Co7O20	mp-764985	Li3CrCo3O8	mp-849768
37	Li2CoNi3O8	mp-1178042	Li2VCo3O8	mp-754294	Li9Co7O16	mp-1175418
38	Li7(NiO2)8	mp-690528	Li2(CoO2)3	mp-705847	Mn3NiO8	mp-757044
39	Li10Co3Ni7O20	mp-769555	YHfRh2	mp-1097261	Li11Ni13O24	mp-758517
40	Li7Ni13O24	mp-758593	Li9Co7O16	mp-1175381	Li(NiO2)2	mp-25388
41	Li9Co7O16	mp-1175506	Li3V2(O2F)2	mp-764429	Li9Si2Ni5O16	mp-867679
42	Li3Cr(CoO3)2	mp-761831	YZrTc2	mp-1096721	Li5Fe2Co3O10	mp-769566
43	Li2Co3NiO8	mp-778768	Li3V2(O2F)2	mp-760200	Li10Fe3Co7O20	mp-760848
44	Li2FeCo3O8	mp-1177976	Li9Co7O16	mp-1175506	Li3(CoO2)4	mp-759191
45	Li4Co3(NiO4)3	mp-777850	Li3(CoO2)5	mp-774507	Li2Co3NiO8	mp-755696
46	Li3Al2CoO6	mp-1222591	Be(CoO2)2	mp-757006	Li7Ni5O12	mp-756913
47	Li(NiO2)2	mp-752531	VMoN3	mp-1246912	NiO2	mp-25210
48	LiFeO2	mp-19419	Li3CrCo3O8	mp-759149	Li10CoNi9O20	mp-759912
49	Li4AlNi3O8	mp-1222534	Li3Cr(CoO3)2	mp-761831	NiGe3O8	mp-543103
50	Li3CoNi3O8	mp-757871	Mg(CoO2)2	mp-756442	Li3Fe(CoO3)2	mp-761602

Our embedding is compared with its two counterpart baseline methods through the neighbourhoods of the $LiCoO_2$ lithium-ion battery cathode. See Table S3 for baseline procedures. All three embeddings successfully captured materials of the layered family and lithium oxides similar to $LiCoO_2$ (more discussed in Table 2). Note that our embedding achieved comparable performance to baselines, although ours exploits only crystal structure information and does not require manual annotations.

Appendix B. Performance validation as metric learning

As mentioned in the *Methods* section, our training task for metric learning is essentially the retrieval across two data expressions. That is, when a query embedding from one expression is given, we expect that its paired embedding is uniquely identified among the database of embeddings from the other expression via the nearest-neighbour search. Therefore, when validating our trained ML models, we evaluated the performance in terms of top-*k* retrieval accuracy, *i.e.*, the probability of including the requested embedding in the top-*k* nearest neighbours. We used the top-*k* accuracy with XRD pattern queries as the primary evaluation, because the retrieval in this direction should be more difficult than in the other due to the information loss in converting crystal structures to XRD patterns.

For model validation and hyperparameter tuning, we randomly split the dataset into training (64%), validation (16%), and test subsets (20%). We tuned hyperparameters such as the DNN architectures, learning rate, batch size based on the retrieval accuracy on the validation subset. Once appropriate hyperparameters were chosen, we trained our model on the whole dataset and obtained the results reported in the main text.

Table S5 reports retrieval accuracies evaluated on the test set of 24,508 materials in terms of the top-1, top-5, and top-10 metrics, comparing our final settings (bold type) with other hyperparameter settings. Notably, our model achieved the remarkably high top-1 accuracy of 65.969%, considering its chance rate of 0.0041% (the probability by the random selection among 24,508 materials). From this result we can conclude that our model successfully composed unified expressions of the two complementary factors (the local structure and periodicity) of crystal structures.

Table S5. Retrieval accuracy evaluations on the test set (XRD pattern queries).

		Retrieval accracy on test-set (%)		
Settings		Top-1	Top-5	Top-10
Proposed		65.969	97.977	99.768
Embedding dimension	**1024**	**65.969**	**97.977**	**99.768**
	512	60.491	96.260	99.274
	256	62.437	96.965	99.404
	128	62.090	97.724	99.710
Batch size	1024	**66.467**	97.810	99.706
	512	65.969	**97.977**	**99.768**
	256	63.616	97.504	99.653
	128	54.091	91.756	97.341

We evaluated the retrieval accuracy on the test set of 24,508 materials as an indicator for the success of training. In the top row, we show the top-1, top-5, and top-10 retrieval accuracies by the proposed settings. Our top-1 score is remarkably high, given the chance rate of 0.0041% (the probability by the random selection). From the second row, we compare results of hyperparameter search, in which the proposed settings and best scores are highlighted.

Appendix C. Performance validation as a materials descriptor

This appendix aims to provide more insight into characteristics of embeddings for interested readers. Particularly, we analyse the performance of prediction of material properties using trained embeddings as input.

We trained DNNs with the proposed deep metric learning approach to export embeddings from materials data in Materials Project. For the prediction tasks, we used 80% of the embeddings for training and the remaining 20% for testing. Random forest[21] was used as the machine learning model for the prediction, and four regression (density, total energy, bandgap, magnetization) and one classification (space group) tasks were performed. As a baseline for comparison, we used the output of a middle layer of CGCNN[22] trained to predict total energy.

Although the proposed embedding was not designed to predict material properties, it competed with the baseline on the prediction tasks of total energy and magnetisation, and excelled on the density and space group predictions (Fig. S1). The density and space group predictions cannot be solved without information of crystal structures such as the unit cell size and periodicity. These results indicate that multi-modal learning successfully led to embeddings that reflect both local structures and periodicity of crystal structures.

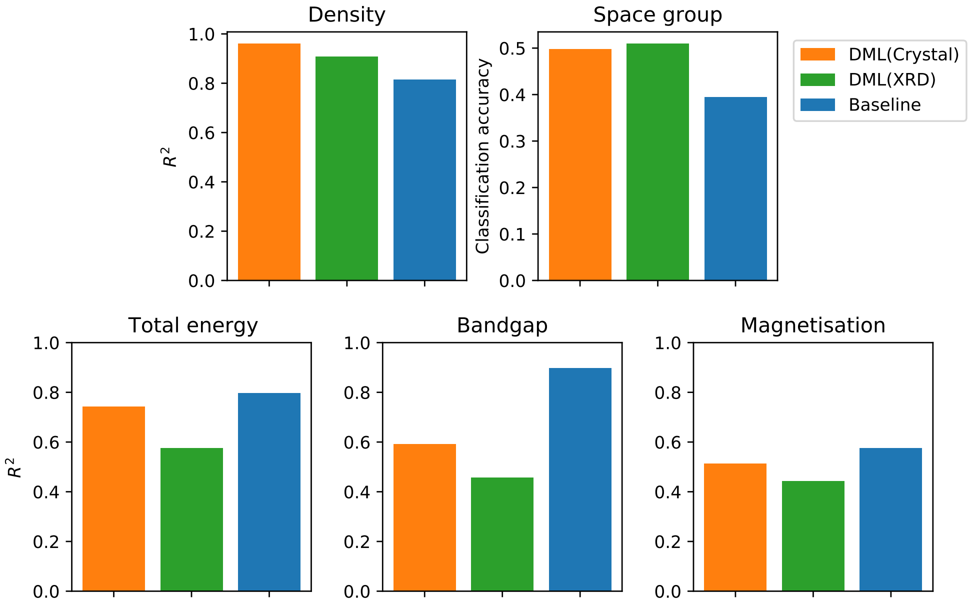

Figure S1. The prediction performance comparison of materials properties using various embeddings. We evaluated prediction performance for materials properties using our embeddings and baseline (middle layer output of a DNN trained to predict total energy) as the materials features.

Appendix D. Detailed network architectures

We summarise the network architectures for the crystal-structure encoder and the XRD pattern encoder in Table S6 and Table S7, respectively. Our ML codes are also available at https://github.com/quantumbeam/materials-concept-learning. The GraphConv operations in Table S6 are defined as

$$\mathbf{x}_i' = \mathrm{GraphConv}_i(\mathbf{x}, \mathbf{e}) = \sum_{j \in \mathcal{N}(i)} \sigma(\mathbf{z}_{ij}\mathbf{W}_f) \odot g(\mathbf{z}_{ij}\mathbf{W}_s) \tag{S2}$$

where $z_{ij} = [x_i, x_j, e_{ij}]$ denotes the concatenation of central node features, neighboring node features, and edge features. σ and g denote the sigmoid and softplus, in which Batch Normalization is inserted before the activation functions.

Table S6. The network architecture of the crystal-structure encoder (CGCNN).

Layers			Output shape
1	Input	Atom features	(64, *N*)
2		Edge features	(41, *E*)
3	Initial transform	Linear ([1])	(64, *N*)
4	Graph convolution	GraphConv ([3], [2])	(64, *N*)
5		BatchNorm	(64, *N*)
6		Add ([3])	(64, *N*)
7		Softplus	(64, *N*)
8	Graph convolution	GraphConv ([7], [2])	(64, *N*)
9		BatchNorm	(64, *N*)
10		Add ([7])	(64, *N*)
11		Softplus	(64, *N*)
12	Graph convolution	GraphConv ([11], [2])	(64, *N*)
13		BatchNorm	(64, *N*)
14		Add ([11])	(64, *N*)
15		Softplus	(64, *N*)
16	Global pooling	Mean ([15])	(64, 1)
17	Fully-connected layers	Linear*	1024
18		Linear*	1024
19		Linear*	1024
20		Linear	1024

Our crystal-structure encoder borrows the network architecture from Crystal Graph Convolution Neural Network (CGCNN)[22] (the top part of Fig. 2 (a)), a deep neural network for the property prediction from crystal structures. Each crystal structure is represented as a set of atoms in the unit cell and their connections, *i.e.*, a graph of atoms. Each atom is represented as a 64-dimensional vector encoding its elemental properties such as the group and period numbers of the atom. When multiple species occupy one atomic site (i.e., when structures have site mixing), a mixture of multiple atomic feature vectors is assigned. Edge features are defined between atoms within a radius of 8 Å, and each is represented as a 41-dimensional vector encoding the distance between two atoms. These inputs are encoded through three GraphConv layers. This architecture can encode a set of arbitrary number of unordered atoms into a fixed-size feature vector in a fashion invariant to permutations of atoms and translations and rotations of the Cartesian coordinate system. This invariance is essential for our crystal-structure inputs. The Linear* layers are followed by the batch normalisation[23] and ReLU activation layers.

Table S7. The network architecture of the XRD pattern encoder (1D CNN).

Layers		Output shape
Input	X-ray diffraction patterns (2theta 10°-110°, 0.02° step)	(1, 5000)
1D convolution	kernel size 50, stride 5, padding 10	(80, 995)
1D convolution	kernel size 50, stride 5, padding 5	(80, 200)
Average pooling	kernel size 3, stride 2	(80, 99)
1D convolution	kernel size 3, stride 3, padding 0	(80, 33)
Average pooling	kernel size 3, stride 3, padding 0	(80, 11)
Flatten		880
Fully-connected layers	Linear	1024
	Linear	1024
	Linear without batch normalization and activation	1024

Our XRD pattern encoder uses a standard feed-forward 1D convolutional neural network architecture (the bottom part of Fig. 2 (a)) designed following existing studies on XRD pattern encoding[24]. Similar to the crystal-structure encoder, each convolution/linear layer except for the final layer is followed by the batch normalisation and ReLU activation layers. Although the previous work did not use the batch normalisation, it was essential to stabilise the training of our model, as discussed in the *Methods* section.

Appendix E. Re-discovery of superconductors in COD

We conducted an additional analysis on our embedding to show whether our method can re-discover superconductors that are known by the literature but not included in the training data. This analysis simulates the screening of new material candidates by an ML model built on a database of known materials.

We borrowed test materials from the *Performance validation as a materials descriptor* in the main text, which provides the crystal structures of 469 superconductors collected from Crystallography Open Database (COD) (see also *Data acquisition for the concept classification tasks* in the *Methods* section). These materials were further filtered to ensure there was no overlap with the Materials Project (MP) dataset used to train our model, resulting in 357 superconductors. We then obtained the embeddings of their crystal structures and mapped them in the t-SNE visualization in Fig. 3 (a), to see if they correlate with the cluster of superconductors from MP.

Fig. S2 below compares the distributions of the embeddings from MP and COD. Despite the fact that the model does not know these COD's superconductors, they are most intensively concentrated around the superconductor cluster in the MP's training materials, suggesting that the model successfully re-discovered superconductors in COD.

Note here that all of the 357 superconductor materials from COD are structures having site mixing. On the other hand, our model was trained on the MP dataset consisting of only structures without site mixing. Despite this difference between the training (MP) and testing (COD) datasets, our model outputs reasonable embeddings for COD's superconductor structures. This is practically important because structures obtained through experiments often have site mixing.

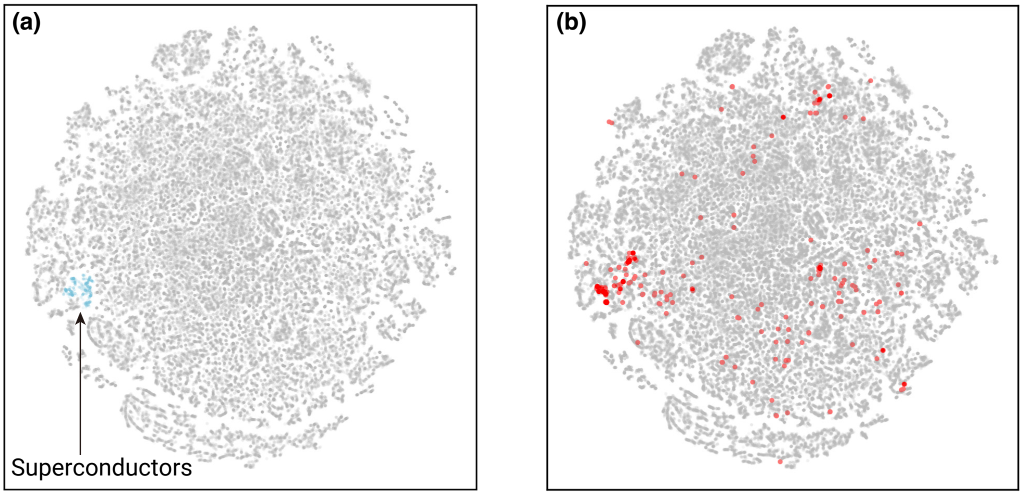

Figure S2. A comparison of superconductor clusters in MP and COD. a, A t-SNE visualisation of the embeddings of the MP dataset (the same as Fig. 3 (a) in the main text) in which a superconductor cluster is identified via manual inspection. **b**, The distribution of superconductor materials registered in COD (red points) overlaid on the materials of MP (gray points). Superconductors in COD are most intensively concentrated around the manually identified superconductor cluster in MP (**a**).

References

1. Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **115**, 1094–1101 (2015).
2. Burch, K. S., Mandrus, D. & Park, J.-G. Magnetism in two-dimensional van der waals materials. *Nature* **563**, 47–52 (2018).
3. Gong, C. *et al.* Discovery of intrinsic ferromagnetism in two-dimensional van der waals crystals. *Nature* **546**, 265–269 (2017).
4. Lebègue, S., Björkman, T., Klintenberg, M., Nieminen, R. M. & Eriksson, O. Two-dimensional materials from data filtering and ab initio calculations. *Phys. Rev. X* **3**, 031002 (2013).
5. Ito, N. *et al.* Spin seebeck effect in the layered ferromagnetic insulators crsite 3 and crgete 3. *Phys. Rev. B* **100**, 060402 (2019).
6. McGuire, M. A. *et al.* Antiferromagnetism in the van der waals layered spin-lozenge semiconductor crte3. *Phys. Rev. B* **95**, 144421 (2017).
7. Huang, B. *et al.* Layer-dependent ferromagnetism in a van der waals crystal down to the monolayer limit. *Nature* **546**, 270–273 (2017).
8. Coey, J. M. *Magnetism and Magnetic Materials* (Cambridge university press, 2010).
9. Coey, J. M. Perspective and prospects for rare earth permanent magnets. *Engineering* **6**, 119–131 (2020).
10. Körner, W., Krugel, G. & Elsässer, C. Theoretical screening of intermetallic thmn 12-type phases for new hard-magnetic compounds with low rare earth content. *Sci. reports* **6**, 1–9 (2016).
11. Krugel, G., Körner, W., Urban, D. F., Gutfleisch, O. & Elsässer, C. High-throughput screening of rare-earth-lean intermetallic 1-13-x compounds for good hard-magnetic properties. *Metals* **9**, 1096 (2019).
12. Zhou, Q. *et al.* Learning atoms for materials discovery. *Proc Natl Acad Sci USA* **115**, E6411 (2018).
13. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
14. Ryan, K., Lengyel, J. & Shatruk, M. Crystal structure prediction via deep learning. *J. Am. Chem. Soc.* **140**, 10158–10168 (2018).
15. Xie, T. & Grossman, J. C. Hierarchical visualization of materials space with graph convolutional neural networks. *The J. Chem. Phys.* **149**, 174111 (2018).
16. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
17. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).
18. Hansen, K. *et al.* Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).
19. Rupp, M., Tkatchenko, A., Müller, K.-R. & Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. review letters* **108**, 058301 (2012).
20. Himanen, L. *et al.* DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
21. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
22. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
23. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 448–456 (2015).
24. Park, W. B. *et al.* Classification of crystal structure using a convolutional neural network. *IUCrJ* **4**, 486–494 (2017).