GRAPH-TEXT CONTRASTIVE LEARNING OF INORGANIC CRYSTAL STRUCTURE TOWARD A FOUNDATION MODEL OF INORGANIC MATERIALS

A PREPRINT

Keisuke Ozawa, Teppei Suzuki
DENSO IT Laboratory
2-15-1 Shibuya, Shibuya-ku, 150-0002 Tokyo, Japan
E-mail:ozawa.keisuke@core.d-itlab.co.jp

Shunsuke Tonogai, Tomoya Itakura
DENSO CORPORATION
1-1 Showa-cho, Kariya, Aichi 448-8661, Japan

March 28, 2024

ABSTRACT

Developing foundation models for materials science has attracted attention. However, there is a lack of work on inorganic materials due to the difficulty in the comprehensive representation of geometric concepts composing crystals: the local atomic environments, their connections, and the global symmetries. We present a contrastive learning of inorganic crystal structure (CLICS) for embedding the geometric concepts, which contrasts texts representing the contextual patterns of geometries with the crystal graphs. We demonstrate that the geometric concepts are integrally embedded on CLICS feature space, through experiments of concept retrieval from crystal graphs, similar structure search, and few-shot/imbalanced crystal structure classification.

Keywords Materials informatics · Inorganic materials · Crystal structure · Foundation model · Contrastive learning · Language model · Graph neural network

1 Introduction

Crystals have a pivotal role in the functional design of inorganic materials. Data-driven approaches have attracted considerable interest for accelerating materials design and exploration [1, 2]. In particular, high-throughput screening of materials and prediction of their properties are attractive options. Various machine learning techniques have been used to accelerate screening [3–8]. Given the hardest challenge of exploring crystals never seen, generative models have been extensively studied for sampling candidates [9–17]. However, current models frequently generate stuffs that deviate from chemical principles, specifically from the geometric "concepts" such as the symmetry of crystals and local atomic environments. Incorporating some symmetries has been considered [13–16]; however, global symmetries and local atomic environments have not been integrally embedded in a feature space. Such a geometric-concept aware feature space would not only augment existing models but also help scientists' intuition or deduction with unveiling relationships between the concepts, inductively from the ever-increasing data reported in the world.

Foundation models (FMs) have recently emerged on the basis of techniques from natural language processing, computer vision, and any other machine learning communities [18]. They are trained on a large amount of data with multiple modalities of concepts. They have been evolving interactively with large language models (LLMs) [19, 20], which have also brought a significant impact on materials science [21–29]. FMs and LLMs often refer the same as natural language is so common for us, which should also be the case for materials science; however, FMs more specifically connect different modalities and are expected to open up ways of conceptual representation. After the studies of contrastive

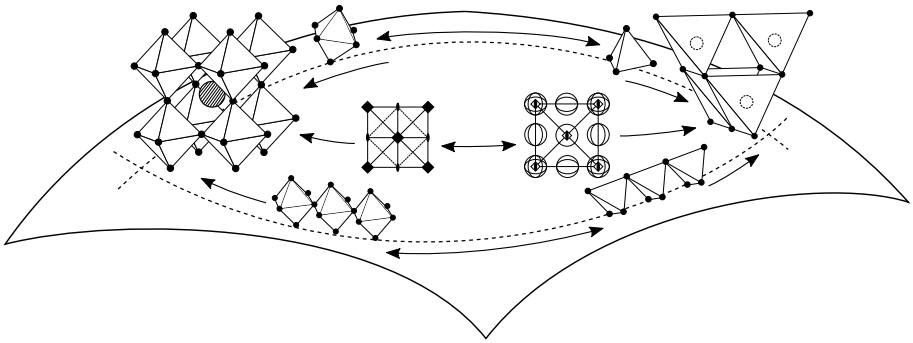

Figure 1: Intuition behind CLICS. Crystals and geometric concepts are contrasted on their shared feature space to learn the underlying concepts of crystal structures. Here, the crystal on the left is positively paired with the octahedron, the connected octahedra, and the tetragonal symmetry. The structure is negative against the tetrahedron, the connected tetrahedra, and the cubic symmetry, which are positively paired with the crystal on the right. The two crystals will be represented in graphs (Section 3). The local geometric environments (e.g., octahedron and tetrahedron on the top), the space groups (e.g., two placed in the middle), and the connected geometries (e.g., corner-sharing octahedra and corner-sharing tetrahedra on the bottom) will be represented in texts (Section 2, 3). Each double-headed arrow represents a negative pair (repulsion).

representation learning [30–32], the seminal work CLIP [33] has widespread multimodal FM and we now find a variety of its folks. It has also inspired materials science community where scientific facts and measurement data are provided along with their specific modalities. Models for applications in organic and drug chemistry have been trained by pairing molecular structures and texts that describe their properties [34–41], which have rich information on molecules. In contrast, relatively few such examples have been reported for inorganic materials, except for studies that used pairs of crystal structures and X-ray diffraction data [42], which aimed to learn the concepts of crystals via the diffracted data, and that used triplet pairs of crystal structures, density of states, and properties [43]. This is considerably because only a few published and curated datasets of inorganic chemistry with multiple modalities are available, especially those written in natural language. More essentially, inorganic materials are so intricate to be characterized only with motifs, which often characterize the functionalities of organic materials.

Although the properties of inorganic materials depend on their compositions and are so intricate to be simply put, they come from the interactions over local atomic coordinations, their connections, and the global crystal structure. We can see typical patterns of local atomic coordinations, where charged ions orderly fill up the space or electron orbits make bonds with certain symmetries in their environments. Crystal structures can take a variety of symmetries and atom coordinates, but the local environments of atoms, their connections, and the global symmetries often arise harmonically. Grasping these geometric concepts is thus fundamental for understanding and designing crystal structures. In machine learning literature, a method was developed for the geometric attribution of local atomic environments [44], and the order parameters were defined [45]. Our study has benefit from these work via Robocrystallographer [46], which was also used in recent work showing that graph neural network (GNN)'s performance could be augmented by using space group information [47], and informing both space groups and local atomic environments to a transformer-based architecture could improve the predictive performance [48]. Another recent study focused on the shapes of motifs in crystals, analyzed the statistics of polyhedra, and developed descriptors for the machine learning of crystals [49]. The concept of motifs was also introduced to improve the prediction performance of a GNN model [50]. A mathematical insight on crystal structures was presented for generating space-filling polyhedra with satisfying some symmetries [51]. We share their motivations but have been independently motivated to develop a way of integrally embedding in our model the foundational concepts of inorganic crystal structure: local atomic environments, their connections, and the global symmetries.

In this study, we present a Contrastive Learning of Inorganic Crystal Structure (CLICS), as a first step demonstration toward a FM of inorganic materials. We consider two different modalities of crystal structures: graph and text. In the framework of contrastive learning [33], we have two models each for graph and text embedding, and train these models simultaneously by contrasting their outputs. A graph defines the arrangement of atoms and totally represents the crystal structure as a chemical entity. On the other hand, a text represents contextual patterns of geometric entities that compose the crystal structure, like the local atomic environments of "a Ti atom is bonded to six O atoms forming a TiO_6 octahedral coordination geometry". The intuition behind CLICS is that the structural part of chemical entity can be decomposed into the geometric concepts, and in turn, these concepts integrally represent the chemical entity. Contrasting these two modalities is expected for the models to learn the underlying concept connecting the chemical and geometric entities (Figure 1). A text itself may be considered as a concept; however, what we call concepts are the geometric entities, their classification, and their relationships including the geometric similarities among them. For

instance, “tetrahedral” and “distorted tetrahedral” are similar to each other, and “four coordinate” and “five coordinate” would have similar meanings. The “tetrahedral” geometry is compatible with “four coordinate”, but has a higher order symmetry. These similarities with respect to the coordination numbers and symmetries do not come only from the texts themselves, but rather from the pairing between graphs and texts. Through the contrastive learning, we expect that different crystal graphs would get closer on the feature space if the corresponding texts are similar, and in turn, different texts, which represent the geometric concepts, will also get closer if the corresponding crystals are structurally similar.

We outline the remainder of this paper. Section 2 describes the dataset, especially focusing on the preparation of texts that describe geometric entities. The criterion in preparing the texts defines with what abstraction level our model learns the concepts. Tips are provided for learning the multiple concepts composing each crystal. Section 3 describes the detailed architectures adopted for our purpose. Section 4 demonstrates CLICS' embedding of geometric concepts, through the experiments of concept retrieval, analysis of the relationships among concepts, similar structure search, and few-shot/imbalanced crystal structure classification tasks.

2 Dataset

We converted CIF files from the Materials Project [52, 53] into crystal graphs. Texts describing the crystal structures were generated using Robocrystallographer [46], which attributes various types of geometries based on modules via Pymatgen [54]: AFLOW for mineral matching [55], Spglib [56] for computing global symmetry with matching space groups, ChemEnv [57] and LocalEnv [44] for attributing local atomic environments (e.g., “octahedral”), and also provides a module for determining connected geometries (e.g., “corner-sharing”).

We generated texts from crystals of 3D, or lower dimensional structured without orientation, which span approximately 90% of the materials on the Materials Project. Robocrystallographer [46] provides texts of its specific sentence structures. For our model to learn the geometric concepts, we introduced the following rules for text preprocessing.

First, named crystal structures (e.g., rock salt), space groups, local atomic environments (e.g., 4-coordinate geometry), and their connections (e.g., corner-sharing octahedra) were separately described (Table S1 shows the count of each concept with an example). We added an unknown class “[UNK]” to the list of space groups as an indefinite label. Crystal systems and points groups were not explicitly considered, but will be partly discussed regarding some hierarchical similarities among space groups. When using the full texts to be contrasted with the graphs, the loss dropped without any performance gain. Then, we inputted randomly selected N sentences from the texts in each epoch. N was set to 2 in the experiments, but using more sentences is possible. Inputting a single sentence is for our model to learn every concept without training only with a part of the concepts that are easy to be paired with the graphs, nor being trained to have one-to-one correspondence between the graphs and texts. Inputting two sentences is for the concepts to be learned in various contexts. Each phrasing was randomized in order to avoid fully memorizing the whole sentence, where we would rather let our model to see, e.g., what kind of element is connected how, in just such a short context though. In the validation phase, we inputted for each crystal another phrase that was not used during the training phase (Table S2).

Second, each chemical composition in each text was replaced with e.g., “This crystal” (or randomly with the phrases in Table S2), because the existence of compositions can make the training obvious: The nodes of a crystal graph have features specific to the atoms, whereas the composition in the corresponding text also has the corresponding word-embedding vectors, where the text encoder we adopted [58] tends to tokenize the compositions into the element symbols. These features can be leaked and linearly converted into one another. In addition, the element symbols were replaced with “[METAL]”, “[NONMETAL]”, or “[METALLOID]”, according to the attribution of each element, and the local coordinations of atoms or clusters were replaced with the word “[POLYHEDRA]”. This denomination may not best represent all types of coordinations such as ligands, but the words replacing chemical entities are arbitrary. These replacements also prevent possible leaks between each graph and text, but the other important aspect is to define the abstraction level of the chemical entities, rather than the geometric ones, which are the focus of this study.

Finally, every integer was replaced with a word (e.g., “4” with “four”), while real values were replaced with “[REAL-VAL]”. The latter replacement was introduced to avoid possible leaks between the values in the CIF files and texts. Sentences that only describe bond lengths and angles were not used.

Some examples generated by Robocrystallographer [46] and the preprocessed texts are in the supplementary information. We randomly split the data into a training set consisting of 126035 and a validation set of 14004.

We did not use the named crystal structures during the training. This is because any crystal structure refers to a specific arrangement of atoms. Thus, local atomic environments, their connections, and the global symmetries are at the bottom of the concepts of crystal structures from a geometric perspective. Instead of learning the named crystal structures, we will demonstrate that CLICS-pretrained model improves the classification accuracy of the named crystal structures even when given a few number of data or imbalanced data, which would be typical situations for materials exploration.

3 Method

We trained our model in the framework of contrastive learning between crystal graphs and texts. The crystal graphs were computed from the CIF files of crystals, and then embedded into vectors by using a GNN: We employed ALIGNN [59], which takes into account the bond angles in addition to the lengths. Aggregating bond angles as well as the lengths is preferable in capturing the geometric concept of crystals. These graphs were contrasted with the texts that were generated by using Robocrystallographer [46] and then preprocessed as described in Section 2. These texts were encoded by using the MatSciBERT [58], which is a fine-tuned model of BERT [60] pretrained on a corpus of materials science literature. We used its original tokenizer, which does not tokenize independent entities like space groups as they are. Adding original tokens can be an option, but we had empirically no distinct performance gain from additional tokens, as well as no remarkable degradation nor improvement using the other language models each pretrained on a less domain specific corpora [60, 61]. Likely CLIP [33], we introduced projection heads after the graph embedding and text embedding to have two copies of a feature space of the same dimension, so that they share the feature space where concepts are learned.

As each crystal is a distinct chemical entity, the positive pair is defined between each crystal and any part of the text describing the crystal. Note that, for each crystal, there are multiple positive pairs from the local atomic environments and their connections. As described in Section 2, randomly sampled sentences were used in each epoch, which allows for our model to learn each concept one by one as well as the contexts by inputting a text with two sentences.

The parameters of ALIGNN [59], MatSciBERT [58], and the projection heads were optimized through a symmetric cross entropy (SCE) loss [33, 62] with the stochastic gradient descent. The loss function for each mini-batch is

$$\mathcal{L}_{\mathrm{SCE}} = -\sum_{i=1}^{B}\left[\log\frac{\exp\left(\langle\mathcal{P}_g(\mathbf{g}_i),\mathcal{P}_t(\mathbf{t}_i)\rangle/\tau\right)}{\sum_j \exp\left(\langle\mathcal{P}_g(\mathbf{g}_i),\mathcal{P}_t(\mathbf{t}_j)\rangle/\tau\right)} + \log\frac{\exp\left(\langle\mathcal{P}_t(\mathbf{t}_i),\mathcal{P}_g(\mathbf{g}_i)\rangle/\tau\right)}{\sum_j \exp\left(\langle\mathcal{P}_t(\mathbf{t}_i),\mathcal{P}_g(\mathbf{g}_j)\rangle/\tau\right)}\right], \tag{1}$$

where $\mathbf{g}_i$ and $\mathbf{t}_i$ denote the feature vectors obtained from ALIGNN [59] and MatSciBERT [58] for an i-th sample in the mini-batch, respectively; $\mathcal{P}_g$ and $\mathcal{P}_t$ denote the projection heads for graphs and texts, which project graph and text embedding vectors into the same dimension (128-dimension in the experiments); τ denotes the trainable temperature parameter [33].

We optimized the parameters for 128 epochs with the AdamW [63] optimizer. Following recent exploration of the parameter tuning recipes [64], we decayed learning rate by a cosine schedule after linearly warming up the learning rates for 5 epochs. We set learning rates for ALIGNN [59], MatSciBERT [58], and the projection heads to 10^{-3}, 10^{-4}, and 10^{-3}, respectively, and the coefficient of weight decay to 10^{-3} in the experiments.

4 Results and Discussion

This section is dedicated to see the embeddings by CLICS. In section 4.1, the retrieval accuracies of 228 space groups and 233 local atomic environments (Table S1) are evaluated. Some examples are discussed in detail. The accuracy of retrieving 5166 full patterns of polyhedral connections was not evaluated, because the intra-categories are too fine to be rigorously matched. A detailed analysis for the connected geometries is provided in the supplementary information. Section 4.2 shows the relationships among geometric concepts acquired by CLICS. Section 4.3 and 4.4 demonstrate that CLICS has integrally embedded the geometric concepts, through similar structure search, and by showing a noticeable improvement on few-shot/imbalanced crystal structure classification tasks.

4.1 Retrieval accuracies of space group and local atomic environments

The crystal graphs in the validation set were used to evaluate the top-K ($K = 1, 5$) similarity scores against the embedded vectors of texts that describe space groups and local atomic environments. Note that there are multiple, a non-constant number of local atomic environments each crystal; we thus counted the top-K retrieval as correct if at least one of the local atomic environments is included in the list of the ground truths. In evaluating the local atomic environments, the metric that counts the cases of "at least one of them is included" as correct may be biased; however, this metric seems not to deviate from the qualitative analysis through Table 2–5. Note that, we also observed cases where retrieved concepts do not exactly match the list, but are still considered reasonable in their meanings of geometry.

In Table 1, reduced but comparable retrieval accuracies for space group were achieved compared to the supervised training using ALIGNN [59], the same architecture used for CLICS. We did not perform supervised learning of local atomic environments, which have non-constant numbers of ground truths. For reference, the top-1 and 5 accuracies from CLICS were even higher than those for space group.

Table 1: Classification accuracies of space group and local atomic environments: Top-1 and top-5 accuracies of CLICS (using ALIGNN [59] for the graph encoder) and supervised learning (SL) using ALIGNN [59].

Concept	Top-1 (CLICS)	Top-5 (CLICS)	Top-1 (SL)	Top-5 (SL)
Space group	66.7	84.5	70.1	88.7
Local atomic environments	72.0	95.9	–	–

Table 2–5 show examples of top-5 retrieved space groups and local atomic environments with their similarity scores. Each table caption describes a part of the original text generated by Robocrystallographer [46] for comparison against retrieved concepts. The number after each element symbol indicates an equivalent position of the atom. The pink colored texts highlight correct retrievals. The purple colored texts highlight reasonable retrievals but requiring careful consideration. The green colored boxes highlight the correspondences between words or sentences.

Table 2 is the result for $\mathrm{PrVO_4}$ (mp-19169 on the Materials Project), which has a zircon-type structure and has attracted attention due to its redox activity as photocatalyst [65–67]. Its space group $I4_1/amd$ was retrieved as the second scored one. Both The $I4_1/amd$ and the first scored $I4_122$ have the tetragonal crystal system but with the point groups of D_4 and $D_{4\mathrm{h}}$, respectively. These space groups are thus hierarchically familiar with each other. Note also that the difference between the first and second scores are small compared to those between the second score and the third, forth, and fifth scores. Next, the top-3 scored local atomic environments were retrieved correctly. Pr(1) and V(1) are attributed to [METAL], and O(1) is to [NONMETAL], respectively. Note that CLICS cannot distinguish which retrieval has come from which atom, but only can grasp the concepts within the given abstraction level of elements and predefined geometric attributions [44–46]. The forth and fifth scores are much lower than those of the top-3 correct retrievals, which suggests that the concepts have been embedded distinctively on the feature space.

Table 3 is the result for $\mathrm{Sm_2CoMnO_6}$ (mp-1188690), which has a double perovskite-type structure and its magnetic property was investigated [68–70]. The space group $P2_1/c$ was retrieved first with a distinct score from the others. The top-3 scored local atomic environments were retrieved correctly. This crystal also has corner-sharing octahedra, but they were not found in the top-5 retrievals. This missing depends on what phrases we input to see the embedded geometries. In fact, the corner-sharing octahedral geometries were retrieved via similarities to the full patterns of connected polyhedral geometries, but the accurate number of connections was not retrieved (detailed discussion is provided in the supplementary information). Note that we have two different descriptions for an equivalent site of oxygen atom, O(1). Our dataset was prepared on 2023 August in date via the API of Pymatgen [54], and any update manually confirmed on 2024 March is added in the table caption in the bracket with a dagger. The forth scored local atomic environment (with a dagger) is then correct. The fifth scored one was not found natural.

Table 4 is the result for $\mathrm{FeTa_2O_6}$ (mp-31755), an ordered tapiolite [71], a cation-ordered derivative of the rutile structure [72, 73], of which antiferromagnetism was investigated [72, 74–77]. Its space group $P4_2/mnm$ was found as the first scored one. $P4_2/mnm$ is accompanied with the point group symmetry of $D_{4\mathrm{h}}$, and is hierarchically familiar with the second scored $P4_2nm$ with its point group $C_{4\mathrm{v}}$. Regarding local atomic environments, the top-2 scored geometries were correctly retrieved. The third scored one is incorrect but has a close meaning to the first scored one except for "distorted". The fourth scored one does not appear in the captioned text, but "three coordinate" has similar meaning to "distorted trigonal planar" and "distorted T shaped" (highlighted in green boxes). Finally, although CLICS does not distinguish whether "[METAL]" has come from Fe or Ta, the fifth scored "octahedra" is correct but with a relatively low score. We empirically observed that CLICS may have independently embedded the local atomic environments from the connected geometries. Not so essential, but the embedded concepts are retrieved via the sentences describing the local atomic environments or the full patterns of connected polyhedral geometries. It seems difficult to retrieve up to the detailed number of connections (discussed in the supplementary information).

Table 5 shows the erroneous result for $\mathrm{NaClO_3}$ (mp-630949). CLICS failed to retrieve any of the concepts. Its space group is $P2_1/c$, but was not retrieved as top-5 scored. $P2_1/c$ is monoclinic, while the first and forth scored space groups $P2_12_12$ and $P2_12_12_1$ are both orthorhombic. There is a hierarchical relationship between $P2_1/c$ and $P2_12_12$, both of which are attributed to translationengleiche subgroups of $Pbcm$ [78], but the Laue class of $P2_1/c$ is $2/m$ and that of $P2_12_12$ and $P2_12_12_1$ is mmm. The second scored Cc is certainly monoclinic but with a different symmetry along the principal axis. Also, none of the local atomic environments were retrieved as top-5 scored. Both of trigonal bipyramidal and pentagonal pyramidal geometries have 6-coordinate, but the very "6-coordinate" geometry was not retrieved. From a purely geometric view, the Na(1) atom is in fact bonded in a distorted trigonal planar with three shorter bonds ranging 2.27, 2.47, 2.69 Å, and two vertical bonds, making a distorted trigonal bipyramidal geometry (Figure S1 bottom). However, there is another bond of 2.78 Å closely near to the bond of 2.69 Å, and the geometry deviates from the trigonal bipyramids. From another viewpoint, the Na-centered geometry also seems a distorted pentagonal pyramidal geometry. The fifth scored one seems not even nearly describe any geometry.

Table 2: Top-5 scored concepts for $PrVO_4$ that "crystallizes in the tetragonal $I4_1/amd$ space group. Pr(1) is bonded in a 8-coordinate geometry to eight equivalent O(1) atoms. V(1) is bonded in a tetrahedral geometry to four equivalent O(1) atoms. O(1) is bonded in a 3-coordinate geometry to two equivalent Pr(1) and one V(1) atom."

Space group	Score	Local atomic environment	Score
$I4_122$	1.97	[METAL] is bonded in a tetrahedral geometry.	1.98
$I4_1/amd$	1.96	[METAL] is bonded in a eight coordinate geometry.	1.81
$I\bar{4}m2$	1.85	[NONMETAL] is bonded in a three coordinate geometry.	1.80
$C222$	1.82	[POLYHEDRA] is bonded in a eight coordinate geometry.	1.60
$Pccm$	1.75	[METAL] is bonded in a distorted tetrahedral geometry.	1.58

Table 3: Top-5 scored concepts for Sm_2CoMnO_6 that "crystallizes in the monoclinic $P2_1/c$ space group. Sm(1) is bonded in a 8-coordinate geometry to two equivalent O(1), three equivalent O(2), and three equivalent O(3) atoms. Mn(1) is bonded to two equivalent O(1), two equivalent O(2), and two equivalent O(3) atoms to form MnO_6 octahedra that share corners with six equivalent $Co(1)O_6$ octahedra. Co(1) is bonded to two equivalent O(1), two equivalent O(2), and two equivalent O(3) atoms to form CoO_6 octahedra that share corners with six equivalent $Mn(1)O_6$ octahedra. O(1) is bonded in a 4-coordinate geometry to two equivalent Sm(1), one Mn(1), and one Co(1) atom [O(1) is bonded to two equivalent Sm, one Mn, and one Co atom to form distorted corner-sharing OSm_2MnCo tetrahedra]†. O(2) is bonded in a 5-coordinate geometry to three equivalent Sm(1), one Mn(1), and one Co(1) atom. O(3) is bonded in a 5-coordinate geometry to three equivalent Sm(1), one Mn(1), and one Co(1) atom." The text in [ ]† is added as the difference between our data generated in 2023 August and the description on the Materials Project in 2024 March.

Space group	Score	Local atomic environment	Score
$P2_1/c$	1.84	[NONMETAL] is bonded in a five coordinate geometry.	1.87
$P\bar{1}$	1.58	[METAL] is bonded in a eight coordinate geometry.	1.79
[UNK]	1.35	[NONMETAL] is bonded in a four coordinate geometry.	1.76
$R\bar{3}$	1.33	[NONMETAL] is bonded in a distorted tetrahedral geometry.†	1.67
Pc	1.23	[NONMETAL] is bonded in a distorted trigonal bipyramidal geometry.	1.58

Table 4: Top-5 scored concepts for $FeTa_2O_6$ that "crystallizes in the tetragonal $P4_2/mnm$ space group. Ta(1) is bonded to two equivalent O(1) and four equivalent O(2) atoms to form TaO_6 octahedra that share corners with four equivalent $Ta(1)O_6$ octahedra, corners with four equivalent $Fe(1)O_6$ octahedra, an edgeedge with one $Ta(1)O_6$ octahedra, and an edgeedge with one $Fe(1)O_6$ octahedra. Fe(1) is bonded to two equivalent O(1) and four equivalent O(2) atoms to form FeO_6 octahedra that share corners with eight equivalent $Ta(1)O_6$ octahedra and edges with two equivalent $Ta(1)O_6$ octahedra. O(1) is bonded in a distorted T-shaped geometry to two equivalent Ta(1) and one Fe(1) atom. O(2) is bonded in a distorted trigonal planar geometry to two equivalent Ta(1) and one Fe(1) atom."

Space group	Score	Local atomic environment	Score
$P4_2/mnm$	2.00	[NONMETAL] is bonded in a distorted trigonal planar geometry.	1.85
$P4_2nm$	1.99	[NONMETAL] is bonded in a distorted T shaped geometry.	1.75
$Cmmm$	1.93	[NONMETAL] is bonded in a trigonal planar geometry.	1.73
$I4_1md$	1.85	[NONMETAL] is bonded in a three coordinate geometry.	1.60
$Cmm2$	1.72	[METAL] is bonded in an octahedral geometry.	1.51

Table 5: Top-5 scored concepts for $NaClO_3$ that "crystallizes in the monoclinic $P2_1/c$ space group. Na(1) is bonded in a 6-coordinate geometry to two equivalent O(1), two equivalent O(2), and two equivalent O(3) atoms. O(1) is bonded in a 3-coordinate geometry to two equivalent Na(1) and one O(2) atom. O(2) is bonded in a 3-coordinate geometry to two equivalent Na(1) and one O(1) atom. O(3) is bonded in a distorted trigonal planar geometry to two equivalent Na(1) and one Cl(1) atom. Cl(1) is bonded in a single-bond geometry to one O(3) atom."

Space group	Score	Local atomic environment	Score
$P2_12_12$	1.86	[METAL] is bonded in a trigonal bipyramidal geometry.	1.81
Cc	1.85	[METAL] is bonded in a distorted pentagonal pyramidal geometry.	1.80
$P3_1$	1.73	[METAL] is bonded in a distorted trigonal bipyramidal geometry.	1.80
$P2_12_12_1$	1.70	[METAL] is bonded in a distorted trigonal pyramidal geometry.	1.78
$Fdd2$	1.68	[NONMETAL] is bonded in a distorted water like geometry.	1.78

Table 6: Top-5 similar structures for three example crystals. In each box is shown the composition and the crystal structure: "-*d*" abbreviate "-derived" and "O-" abbreviate "Orthorhombic".

Example	First similar	Second similar	Third similar	fourth similar	Fifth similar
$PrVO_4$	$GdY_3V_4O_{16}$	$DyNbO_4$	EuV_2BiO_8	$CaZrV_2O_8$	$HoPO_4$
Zircon	Zircon-*d*	Zircon	Zircon-*d*	Zircon-*d*	Zircon
Sm_2CoMnO_6	Eu_2RuCoO_6	Nd_2TiCuO_6	La_2CuIrO_6	Sm_2RuCoO_6	$Nd_4Mn_3NiO_{12}$
O-Perovskite-*d*	O-Perovskite-*d*	O-Perovskite-*d*	O-Perovskite-*d*	n/a	O-Perovskite-*d*
$FeTa_2O_6$	$Ta_6Mn_2FeO_{18}$	$AlTiTaO_6$	$LiCu_5F_{12}$	$NbVO_4$	Ti_9SnO_{20}
Rutile-*d*	Hydrophilite-*d*	Rutile-*d*	Hydrophilite-*d*	Rutile-*d*	Hydrophilite-*d*

We have seen examples of a specific phase of each chemical composition. Recent work demonstrated crystal structure classification given only chemical compositions [79, 80]. On the other hand, there exist often different phases for each composition. Although we have avoided the leak of element symbols between crystal graphs and texts, there may be a chance for our model to have connected chemical compositions with crystal structures during the training, and would fail to retrieve the different phase of a crystal of the same composition in the validation set. This was not the case and CLICS retrieved crystal structures given different phases of e.g., $TmCO_4$ and Na_3PS_4, each the same chemical composition, not so surprisingly as we give crystal graphs, not the chemical compositions. The detailed results and discussion are provided in the supplementary information.

4.2 Geometric similarities among local atomic environments on CLICS feature space

CLICS has retrieved the geometric concepts based on the similarity scores on the feature space. It was also found that some concepts often arise incorrectly but with sharing similar geometric attributions. These results suggest that CLICS has acquired with their meanings w.r.t. their geometric realization in the 3-dimensional space. Note that the "similarity" scores between texts would come from two aspects, and we will distinguish relatedness from similarity: Similarity indicates how a concept is shared by two or more crystals, while we use the term relatedness regarding how two concepts arise simultaneously for a crystal.

Table S14–S19 in the supplementary information show examples of the top-7 similarities among the concepts of local atomic environments. On the CLICS feature space, just one more or less numbered coordinate geometries have similar meanings. Named entities like "distorted square co planar" geometry is similar to "four coordinate" geometry, "distorted hexagonal planar" geometry is similar to "six coordinate" geometry, "distorted body centered cubic" geometry is similar to "eight coordinate" geometry, and so on. CLICS may have grasped the degree of symmetry in addition to coordination number, as "distorted" appeared on top of these concepts. The orders in the top-7 ranked concepts show some relationships among the geometries: e.g., hexagonal planar is similar to pentagonal planar, a hexagonal pyramidal geometry is obtained by adding a vertex perpendicularly to the hexagonal plane, and adding another vertex at the opposite side leads a hexagonal bipyramidal geometry. We also found the relatedness between the [METAL]-centered 8-coordinate geometry and [NONMETAL]-centered 4/5-coordinate geometries. We have faced its origin in e.g. Sm_2CoMnO_6, a family of perovskite structure (Table 3). A more detailed discussion is provided in the supplementary information.

4.3 Similar crystal structure search on CLICS feature space

CLICS has well retrieved space groups, local atomic environments, and roughly the connections of polyhedral geometries. The meanings of geometric concepts have also been acquired. The arrangement of atoms defines crystal structures; in turn, CLICS feature space is expected to have grasped the patterns of the arrangement of atoms, from the named crystal structures to arbitrary structures that are not attributed to any class of structures. To confirm this idea, we demonstrate similar structure search in this section, and also an improved classification performance of named crystal structures by only training the projection heads on top of the CLICS graph encoder in the next section 4.4.

Finding similar crystal structures is an interesting task for possible application to materials design [42, 81–83]. We here demonstrate the similar structure search on the CLICS feature space. Table 6 shows top-5 similar structures in the validation set for $PrVO_4$, Sm_2CoMnO_6, and $FeTa_2O_6$ (displayed in Figure S2). The top-5 similar structures of $PrVO_4$ are all zircon-type or zircon-derived. Those of Sm_2CoMnO_6 are all orthorhombic perovskite-derived. This class of structure is abundant and these similar structures have the same type of compositions. $FeTa_2O_6$ is a tapiolite, a rutile-derived structured. The second and fourth similar crystals have also rutile-derived structures. The first, third, and fifth are hydrophilite-derived structured, which are attributed to a distorted rutile family [71]. In fact, the four corner-connected chains of edge-sharing octahedral geometries were observed for these crystals.

4.4 Few-shot/imbalanced crystal structure classification using the CLICS-pretrained model

We prepared two type of datasets: Few-shot and imbalanced. Each few-shot dataset has the same number of training data. For example, in the 2-shot task, every two crystal structures were randomly selected for training, and the remaining data were used for validation. We have a fixed number of training data for every crystal structure, while the number of validation data varies. On the other hand, each imbalanced dataset has randomly selected training dataset under a fixed percentage against all the data, irrespective of crystal structures. Therefore, in each imbalanced classification task, we have various numbers of data for both training and validation sets.

We compared the classification performance between the CLICS-pretrained model and the supervised model using ALIGNN [59] from scratch. For the CLICS-pretrained model, we only trained the weights of the projection head from scratch and the remained weights were frozen. For the scratch model, all the weights of ALIGNN [59] and the projection head were trained. The output dimensions of the projection heads were both set to the number of crystal structures. For a fair comparison, we varied the batch size and the learning rate and selected the best results for the scratch model, while the batch size and the learning rate were fixed for CLICS-pretrained model predetermined on an imbalanced task with 50 percent training data among all the data.

Figure 2 plots the top-1 and 5 classification accuracies against the number of shot. The CLICS-pretrained model classified the crystal structures accurately with more than twice the chance compared to the scratch model. Figure 3 plots the classification accuracies for the imbalanced tasks, against the ratio of training data in the total number of data. As the same as the few-shot tasks, the CLICS-pretrained model outperformed the scratch model for all the ratios. A detailed analysis of how accurately each crystal structure was classified is discussed in the supplementary information.

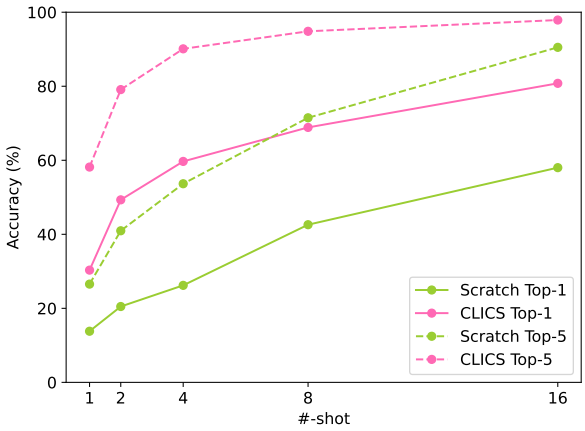

Figure 2: Accuracy plots against the number of training data (1, 2, 4, 8, and 16 shots). The pink and green colored plots compare the CLICS-pretrained model and the scratch model, and the solid and dotted lines show the top-1 and top-5 accuracies.)

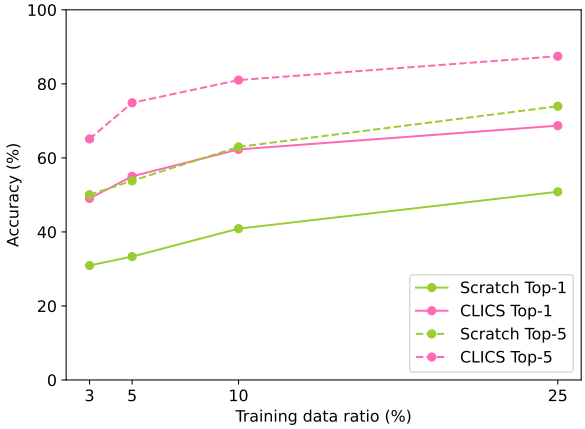

Figure 3: Accuracy plots against the imbalanced training data. The ratios of training data were selected from $3, 5, 10, 25\%$.

5 Conclusions

We have presented a graph-text contrastive learning of inorganic crystal structure (CLICS) for grasping geometric concepts composing crystal structures. CLICS breaks down chemical entities into geometric concepts and integrally embeds them in the feature space. The contextual patterns of the local atomic environments, their connected geometries, and the space groups were represented in texts, and contrasted with the graphs. Throughout experiments, we have demonstrated that CLICS has integrally embedded the geometric concepts and is aware of the meanings of the geometries. Experimental results of similar structure search and classification tasks under challenging situations also suggest that similar crystal structures have got closer based on the geometric concepts. CLICS could be a first step toward developing a geometric foundation model of inorganic materials. Incorporating chemical properties of elements in addition to the geometric concepts is a promising research direction. As possible broader impacts, property prediction would be improved by CLICS-pretraining, and the feature space would also be useful for materials exploration as well as a soft constraint for generative models.

Author contributions

All the authors were motivated for inorganic materials design. K.O. and T.S. launched and conducted this study. T.S. gave the idea of applying contrastive learning to inorganic materials. K.O. conceived to learn the geometric concepts composing crystal structures. T.S. developed the prototypes for contrastive learning and supervised learning of space group. K.O. prepared the dataset including the text preprocessing, adopted the encoders for graphs and texts, and implemented for evaluations. Both K.O. and T.S. contributed the detailed implementations and discussed throughout the experiments. K.O. conducted the experiments, interpreted the results, and wrote the manuscript. T.S. contributed to the Method section and the construction of Introduction section. All the authors discussed throughout the study and reviewed the manuscript.

Competing interests

DENSO CORPORATION owns a patent related to this work, including some possible applications.

References

[1] Krishna Rajan. Materials informatics: The materials "gene" and big data. *Annual Review of Materials Research*, 45:153–169, 2015.

[2] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, 3(1):54, 2017.

[3] Atsuto Seko, Tomoya Maekawa, Koji Tsuda, and Isao Tanaka. Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single-and binary-component solids. *Physical Review B*, 89(5):054303, 2014.

[4] Atsuto Seko, Atsushi Togo, Hiroyuki Hayashi, Koji Tsuda, Laurent Chaput, and Isao Tanaka. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization. *Physical Review Letters*, 115(20):205901, 2015.

[5] Henry C Herbol, Weici Hu, Peter Frazier, Paulette Clancy, and Matthias Poloczek. Efficient search of compositional space for hybrid organic–inorganic perovskites via bayesian optimization. *npj Computational Materials*, 4(1):51, 2018.

[6] Randy Jalem, Kenta Kanamori, Ichiro Takeuchi, Masanobu Nakayama, Hisatsugu Yamasaki, and Toshiya Saito. Bayesian-driven first-principles calculations for accelerating exploration of fast ion conductors for rechargeable battery application. *Scientific Reports*, 8(1):5845, 2018.

[7] Juhwan Noh, Geun Ho Gu, Sungwon Kim, and Yousung Jung. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chemical Science*, 11(19):4871–4881, 2020.

[8] Yunxing Zuo, Mingde Qin, Chi Chen, Weike Ye, Xiangguo Li, Jian Luo, and Shyue Ping Ong. Accelerating materials discovery with bayesian optimization and graph deep learning. *Materials Today*, 51:126–135, 2021.

[9] Asma Nouira, Nataliya Sokolovska, and Jean-Claude Crivello. CrystalGAN: learning to discover crystallographic structures with generative adversarial networks. *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*, 2019.

[10] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *10th International Conference on Learning Representations*, 2022.

[11] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

[12] Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. *12th International Conference on Learning Representations*, 2024.

[13] Yong Zhao, Edirisuriya M Dilanga Siriwardane, Zhenyao Wu, Nihang Fu, Mohammed Al-Fahdi, Ming Hu, and Jianjun Hu. Physics guided deep learning for generative design of crystal materials with symmetry constraints. *npj Computational Materials*, 9(1):38, 2023.

[14] Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. *12th International Conference on Learning Representations, ICLR*, 2024.

[15] Ruiming Zhu, Wei Nong, Shuya Yamazaki, and Kedar Hippalgaonkar. WyCryst: Wyckoff inorganic crystal generator framework. *Available at SSRN 4658842*, 2023.

[16] Youzhi Luo, Chengkai Liu, and Shuiwang Ji. Towards symmetry-aware generation of periodic materials. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Yuan Chiang, Chia-Hong Chou, and Janosh Riebesell. LLaMP: Large language model made powerful for high-fidelity materials knowledge retrieval and distillation. *arXiv preprint arXiv:2401.17244*, 2024.

[18] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan,

Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL `https://crfm.stanford.edu/assets/report.pdf`.

[19] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[20] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[21] Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

[22] Vadim Korolev and Pavel Protsenko. Accurate, interpretable predictions of materials properties within transformer language models. *Patterns*, 4. doi:10.1016/j.patter.2023.100803.

[23] Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023.

[24] Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Is GPT all you need for low-data discovery in chemistry? 2023. doi:10.26434/chemrxiv-2023-fw8n4-v2.

[25] Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using GPT-4. *arXiv preprint arXiv:2311.07361*, 2023.

[26] Santiago Miret and NM Krishnan. Are LLMs ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*, 2024.

[27] Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, et al. ChemLLM: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024.

[28] Xianjun Yang, Stephen D Wilson, and Linda Petzold. Quokka: An open-source large language model chatbot for material science. *arXiv preprint arXiv:2401.01089*, 2024.

[29] Lei Ge, Docherty Ronan, and Samuel J Cooper. Materials science in the era of large language models: a perspective. *arXiv preprint arXiv:2403.06949*, 2024.

[30] Kihyuk Sohn. Improved deep metric learning with multi-class N-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[32] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[34] Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, 2021.

[35] Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature Communications*, 13(1):862, 2022.

[36] Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing activity prediction models in drug discovery with the ability to understand human language. In *International Conference on Machine Learning*, pages 30458–30490. PMLR, 2023.

[37] Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023.

[38] He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*, 2023.

[39] Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. MolCA: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798*, 2023.

[40] Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *ICLR*. OpenReview.net, 2024. URL `https://openreview.net/pdf?id=Tlsdsb6l9n`.

[41] Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. ChatGPT-powered conversational drug editing using retrieval and domain feedback. *arXiv preprint arXiv:2305.18090*, 2023.

[42] Yuta Suzuki, Tatsunori Taniai, Kotaro Saito, Yoshitaka Ushiku, and Kanta Ono. Self-supervised learning of materials concepts from crystal structures via deep neural networks. *Machine Learning: Science and Technology*, 3(4):045034, 2022.

[43] Viggo Moro, Charlotte Loh, Rumen Dangovski, Ali Ghorashi, Andrew Ma, Zhuo Chen, Peter Y Lu, Thomas Christensen, and Marin Soljačić. Multimodal learning for crystalline materials. *arXiv preprint arXiv:2312.00111*, 2023.

[44] Nils ER Zimmermann, Matthew K Horton, Anubhav Jain, and Maciej Haranczyk. Assessing local structure motifs using order parameters for motif recognition, interstitial identification, and diffusion path characterization. *Frontiers in Materials*, 4:34, 2017.

[45] Nils ER Zimmermann and Anubhav Jain. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC advances*, 10(10):6063–6081, 2020.

[46] Alex M Ganose and Anubhav Jain. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications*, 9(3):874–881, 2019.

[47] Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Crysmmnet: multimodal representation for crystal property prediction. In *Uncertainty in Artificial Intelligence*, pages 507–517. PMLR, 2023.

[48] Andre Niyongabo Rubungo, Craig Arnold, Barry P Rand, and Adji Bousso Dieng. LLM-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions. *arXiv preprint arXiv:2310.14029*, 2023.

[49] Yuki Inada, Yukari Katsura, Masaya Kumagai, and Kaoru Kimura. Atomic descriptors generated from coordination polyhedra in crystal structures. *Science and Technology of Advanced Materials: Methods*, 1(1):200–212, 2021.

[50] Huta R Banjade, Sandro Hauri, Shanshan Zhang, Francesco Ricci, Weiyi Gong, Geoffroy Hautier, Slobodan Vucetic, and Qimin Yan. Structure motif–centric learning framework for inorganic crystalline systems. *Science Advances*, 7(17), 2021.

[51] Tomoyasu Yokoyama, Kazuhide Ichikawa, and Hisashi Naito. Crystal structure generation based on polyhedra using dual periodic graphs. *Crystal Growth & Design*, 2023.

[52] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

[53] Shyue Ping Ong, Shreyas Cholia, Anubhav Jain, Miriam Brafman, Dan Gunter, Gerbrand Ceder, and Kristin A Persson. The materials application programming interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles. *Computational Materials Science*, 97:209–215, 2015.

[54] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

[55] Michael J Mehl, David Hicks, Cormac Toher, Ohad Levy, Robert M Hanson, Gus Hart, and Stefano Curtarolo. The AFLOW library of crystallographic prototypes: part 1. *Computational Materials Science*, 136:S1–S828, 2017.

[56] Atsushi Togo and Isao Tanaka. Spglib: a software library for crystal symmetry search. *arXiv preprint arXiv:1808.01590*, 5, 2018.

[57] David Waroquiers, Xavier Gonze, Gian-Marco Rignanese, Cathrin Welker-Nieuwoudt, Frank Rosowski, Michael Gobel, Stephan Schenk, Peter Degelmann, Rute André, Robert Glaum, et al. Statistical analysis of coordination environments in oxides. *Chemistry of Materials*, 29(19):8346–8360, 2017.

[58] Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102, 2022.

[59] Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.

[60] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:4171–4186, 2019.

[61] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, page 3615–3620, 2019.

[62] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330, 2019.

[63] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *12th International Conference on Learning Representations*, 2017.

[64] Varun Godbole, George E. Dahl, Justin Gilmer, Christopher J. Shallue, and Zachary Nado. Deep learning tuning playbook, 2023. URL `http://github.com/google-research/tuning_playbook`. Version 1.0.

[65] Enrico Bandiello, Catalin Popescu, Estelina Lora da Silva, Juan Ángel Sans, Daniel Errandonea, and Marco Bettinelli. $PrVO_4$ under high pressure: Effects on structural, optical, and electrical properties. *Inorganic Chemistry*, 59(24):18325–18337, 2020.

[66] Rozita Monsef, Maryam Ghiyasiyan-Arani, Omid Amiri, and Masoud Salavati-Niasari. Sonochemical synthesis, characterization and application of $PrVO_4$ nanostructures as an effective photocatalyst for discoloration of organic dye contaminants in wastewater. *Ultrasonics Sonochemistry*, 61:104822, 2020.

[67] Neetu Yadav, Kovuru Gopalaiah, Jyoti Pandey, and Rajamani Nagarajan. Zircon $PrVO_4$: an efficient heterogeneous catalyst for tandem oxidative synthesis of 2, 3-disubstituted quinoline derivatives. *Dalton Transactions*, 52 (18):5969–5975, 2023.

[68] PR Mandal, RC Sahoo, and TK Nath. A comparative study of structural, magnetic, dielectric behaviors and impedance spectroscopy for bulk and nanometric double perovskite Sm_2CoMnO_6. *Materials Research Express*, 1(4):046108, 2014.

[69] Liaoyu Wang, Weiping Zhou, Dunhui Wang, Qingqi Cao, Qingyu Xu, and Youwei Du. Effect of metamagnetism on multiferroic property in double perovskite Sm_2CoMnO_6. *Journal of Applied Physics*, 117(17):17D914, 2015.

[70] Giuseppe Muscas, K Prabahar, Francesco Congiu, Gopal Datt, and Tapati Sarkar. Nanostructure-driven complex magnetic behavior of Sm_2CoMnO_6 double perovskite. *Journal of Alloys and Compounds*, 906:164385, 2022.

[71] Werner H Baur. The rutile type and its derivatives. *Crystallography Reviews*, 13(1):65–113, 2007.

[72] Toshiaki Osaka and Tadayuki Nakayama. On the structure of tapiolite formed by solid state reactions. *Transactions of the Japan Institute of Metals*, 10(6):437–438, 1969.

[73] TS Ercit. Hidden story of tapiolite. *Mineralogical Magazine*, 74(4):715–730, 2010.

[74] Mikio Takano and Toshio Takada. Magnetic properties of MTa_2O_6 ($M = Fe, Co$ or Ni). *Materials Research Bulletin*, 5(6):449–454, 1970.

[75] SM Eicher, JE Greedan, and KJ Lushington. The magnetic properties of $FeTa_2O_6$. magnetic structure and low-dimensional behavior. *Journal of Solid State Chemistry*, 62(2):220–230, 1986.

[76] EML Chung, MR Lees, GJ McIntyre, C Wilkinson, G Balakrishnan, JP Hague, D Visser, and D McK Paul. Magnetic properties of tapiolite ($FeTa_2O_6$); a quasi two-dimensional (2D) antiferromagnet. *Journal of Physics: Condensed Matter*, 16(43):7837, 2004.

[77] Patrick M Bacirhonde, Nelson Y Dzade, Carmen Chalony, Jeesoo Park, Eun-Suk Jeong, Emmanuel O Afranie, Sunny Lee, Cheol Sang Kim, Do-Hwan Kim, and Chan Hee Park. Reduction of transition-metal columbite-tantalite as a highly efficient electrocatalyst for water splitting. *ACS Applied Materials & Interfaces*, 14(13): 15090–15102, 2022.

[78] Ulrich Müller. *Symmetry relationships between crystal structures: applications of crystallographic group theory in crystal chemistry*, volume 18. OUP Oxford, 2013.

[79] Yong Zhao, Yuxin Cui, Zheng Xiong, Jing Jin, Zhonghao Liu, Rongzhi Dong, and Jianjun Hu. Machine learning-based prediction of crystal systems and space groups from inorganic materials compositions. *ACS omega*, 5(7): 3596–3606, 2020.

[80] Yousef A Alghofaili, Mohammed Alghadeer, Abdulmohsen A Alsaui, Saad M Alqahtani, and Fahhad H Alharbi. Accelerating materials discovery through machine learning: Predicting crystallographic symmetry groups. *The Journal of Physical Chemistry C*, 127(33):16645–16653, 2023.

[81] LM Gelato and E Parthé. Structure tidy–a computer program to standardize crystal structure data. *Journal of Applied Crystallography*, 20(2):139–143, 1987.

[82] AV Dzyabchenko. Method of crystal-structure similarity searching. *Acta Crystallographica Section B: Structural Science*, 50(4):414–425, 1994.

[83] John C Thomas, Anirudh Raju Natarajan, and Anton Van der Ven. Comparing crystal structures with symmetry and geometry. *npj Computational Materials*, 7(1):164, 2021.

[84] Koichi Momma and Fujio Izumi. VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *Journal of Applied Crystallography*, 44(6):1272–1276, 2011.

[85] Takeshi Tahara, Izumi Nakai, Ritsuro Miyawaki, and Satoshi Matsubara. Crystal chemistry of $\mathrm{RE(CO_3)OH}$. *Zeitschrift für Kristallographie*, 222(7):326–334, 2007.

[86] Wei-Guo Yin, Jianjun Liu, Chun-Gang Duan, Wai-Ning Mei, Robert W Smith, and John R Hardy. Superionicity in $\mathrm{Na_3PS_4}$: A molecular dynamics simulation. *Physical Review B*, 70(6):064302, 2004.

[87] Xuyong Feng, Po-Hsiu Chien, Zhuoying Zhu, Iek-Heng Chu, Pengbo Wang, Marcello Immediato-Scuotto, Hesam Arabzadeh, Shyue Ping Ong, and Yan-Yan Hu. Studies of functional defects for fast Na-ion conduction in $\mathrm{Na_{3-y}PS_{4-x}Cl_x}$ with a combined experimental and computational approach. *Advanced Functional Materials*, 29(9):1807951, 2019.

[88] Akitoshi Hayashi, Kousuke Noi, Atsushi Sakuda, and Masahiro Tatsumisago. Superionic glass-ceramic electrolytes for room-temperature rechargeable sodium batteries. *Nature Communications*, 3(1):856, 2012.

[89] Thorben Krauskopf, Sean P Culver, and Wolfgang G Zeier. Local tetragonal structure of the cubic superionic conductor $\mathrm{Na_3PS_4}$. *Inorganic chemistry*, 57(8):4739–4744, 2018.

[90] Katharina Hogrefe, Jana Konigsreiter, Anna Bernroitner, Bernhard Gadermaier, Sharon E Ashbrook, and H Martin R Wilkening. Length-scale-dependent ion dynamics in ca-doped $\mathrm{Na_3PS_4}$. *Chemistry of Materials*, 36(2): 980–993, 2024.

Supplementary information

Example texts generated by Robocrystallographer and the preprocessed texts

We show example texts generated by Robocrystallographer [46] and their preprocessed texts (Section 2). Druing the training phase, e.g., the sentence "This material crystallizes in Pnma space group." was randomly replaced with one of the four templates listed on the top of Table S2, the sentence "Sr(1) is bonded in a 9-coordinate geometry to three equivalent O(1) and six equivalent O(2) atoms." was also randomly replaced with one of the four templates listed in the middle of Table S2, e.g., to "[METAL] has bondings to make a nine coordinate local geometry."

Original text-1:

> "SrOsO3 is Orthorhombic Perovskite structured and crystallizes in the orthorhombic Pnma space group. Sr(1) is bonded in a 9-coordinate geometry to three equivalent O(1) and six equivalent O(2) atoms. There are a spread of Sr(1)-O(1) bond distances ranging from 2.46-2.99 Å. There are a spread of Sr(1)-O(2) bond distances ranging from 2.49-2.80 Å. Os(1) is bonded to two equivalent O(1) and four equivalent O(2) atoms to form corner-sharing OsO6 octahedra. The corner-sharing octahedral tilt angles are 22°. Both Os(1)-O(1) bond lengths are 2.03 Å. All Os(1)-O(2) bond lengths are 2.02 Å. There are two inequivalent O sites. In the first O site, O(1) is bonded in a 4-coordinate geometry to three equivalent Sr(1) and two equivalent Os(1) atoms. In the second O site, O(2) is bonded in a 5-coordinate geometry to three equivalent Sr(1) and two equivalent Os(1) atoms."

Preprocessed text-1:

> "This material crystallizes in Pnma space group. This material has Orthorhombic Perovskite crystal structure. [METAL] is bonded in a nine coordinate geometry to three equivalent [NONMETAL] and six equivalent [NONMETAL] atoms. [METAL] is bonded to two equivalent [NONMETAL] and four equivalent [NONMETAL] atoms to form corner sharing [POLYHEDRA] octahedra. In the first [NONMETAL] site, [NONMETAL] is bonded in a four coordinate geometry to three equivalent [METAL] and two equivalent [METAL] atoms. In the second [NONMETAL] site, [NONMETAL] is bonded in a five coordinate geometry to three equivalent [METAL] and two equivalent [METAL] atoms."

Original text-2:

> "YbSc(BO3)2 is Calcite-derived structured and crystallizes in the trigonal R-3 space group. Yb(1) is bonded to six equivalent O(1) atoms to form YbO6 octahedra that share corners with six equivalent Sc(1)O6 octahedra. The corner-sharing octahedral tilt angles are 58°. All Yb(1)-O(1) bond lengths are 2.36 Å. Sc(1) is bonded to six equivalent O(1) atoms to form ScO6 octahedra that share corners with six equivalent Yb(1)O6 octahedra. The corner-sharing octahedral tilt angles are 58°. All Sc(1)-O(1) bond lengths are 2.13 Å. B(1) is bonded in a trigonal planar geometry to three equivalent O(1) atoms. All B(1)-O(1) bond lengths are 1.38 Å. O(1) is bonded in a distorted trigonal planar geometry to one Yb(1), one Sc(1), and one B(1) atom."

Preprocessed text-2:

> "This material crystallizes in R-3 space group. This material has Calcite-derived crystal structure. [METAL] is bonded to six equivalent [NONMETAL] atoms to form [POLYHEDRA] octahedra that share corners with six equivalent [POLYHEDRA] octahedra. [METAL] is bonded to six equivalent [NONMETAL] atoms to form [POLYHEDRA] octahedra that share corners with six equivalent [POLYHEDRA] octahedra. [METALLOID] is bonded in a trigonal planar geometry to three equivalent [NONMETAL] atoms. [NONMETAL] is bonded in a distorted trigonal planar geometry to one [METAL], one [METAL], and one [METALLOID] atom."

Statistics of text data and templates for text replacement

Table S1 and S2 provide additional details to support Section 2.

Table S1 shows the statistics of our text data. The sentences were categorized into the six types of concepts, and we counted the number of patterns appeared in each category. The 233 local atomic environments were extracted as sentences including the word "geometry". The 20 patterns of connected polyhedra describe the connected geometries without how they are connected to the other polyhedra. The 2780 patterns of polyhedral connections at the second row to last were not used in this study. The 5166 full patterns of polyhedral connections describe how polyhedra are connected each other. We used these full patterns of texts to see whether the connected geometries have been embedded on the CLICS feature space.

Table S2 shows the templates for text replacement. For example, a sentence regarding space group was randomized using the four templates on top of the table, while the fixed template was used to retrieve the concepts. The phrases are not essential in themselves for embedding the geometric concepts, and any contextual embedding would work if they represent the geometric concept like what an element is bonded to, what geometry the bondings form, and what the geometry is connected to.

Table S1: Statistics of text data: count of each concept with an example.

Concept	Count	Example
Named Crystal Structures	382	"This material has Salt crystal structure."
Space groups	228	"This material crystallizes in Pmma space group."
Local atomic environments	233	"[NONMETAL] is bonded in a body centered cubic geometry."
Connected polyhedra	20	"This crystal has bondings to form [POLYHEDRA] tetrahedra."
Polyhedral connections	2780	"This crystal has bondings to form [POLYHEDRA] tetrahedra that share an edge-edge with one [POLYHEDRA] tetrahedra."
Full patterns of — " —	5166	"This crystal has bondings to form [POLYHEDRA] octahedra that share corners with four equivalent [POLYHEDRA] octahedra and corners with two equivalent [POLYHEDRA] pentagonal pyramids."

Table S2: Templates of text replacement for training and fixed inputs used for validation.

Concept	Count	Example
Space group	Training	"This is crystallized in 4mm space group." "This crystal has a space group of 4mm." "It is 4mm space group that this crystal is classified." "This is an inorganic material that has 4mm space group."
	Validation	"This material crystallizes in 4mm space group."
Local atomic environments	Training	"[METAL] has bondings to make a octahedral local geometry." "The bonding of [METAL] makes a local geometry of octahedral." "[METAL] is bonded in a local geometry of octahedral." "The bonding of [METAL] has a octahedral geometry."
	Validation	"[METAL] is bonded in an octahedral geometry."
The other patterns of random replacement	Training	"geometry" → "local geometry" with "is bonded in a" → "has bondings to make a", or "is bonded in a" → "is bonded in a local coordination of"

Visual confirmation of concept retrieval

Figure S1 are the snapshots of the four crystals in Section 4.1, captured from different viewpoints on VESTA [84].

For $PrVO_4$ (Table 2), the geometric concepts of local atomic environments were correctly retrieved.

For Sm_2CoMnO_6 (Table 3), an equivalent site of O atom is bonded in a 5-coordinate geometries. Another equivalent O atom is also certainly bonded in a 4-coordinate geometry: Two Sm atoms are additionally bonded to the O atom at the right bottom, but are not shown the bonds to a Co atom and a Mn atom.

For $FeTa_2O_6$ (Table 4), all the local atomic coordinates are observed. The distorted trigonal planar geometry and the distorted T shaped geometry have similar geometries but CLICS successfully attributes them. Note that the octahedra are connected with sharing their edges, and the edge-sharing octahedral chains are alternately tilted and connected at the vertices of the octahedra. The connections are better visualized along the chains (the right top of Figure S2).

Finally, we have obtained erroneous retrievals for $NaClO_3$ (Table 5). We here visually see why CLICS made the errors. We show Na-centered trigonal bipyramidal geometry that is obtained by deleting the longest Na-O bond. Also, the five O atoms except for the left side nearly make a distorted pentagonal planar geometry. These geometric objects were assumed purely from geometric similarities, but the six Na-O bonds should have been considered. Adding the left side O atom would make a distorted pentagonar pyramidal, as erroneously suggested in Table 5.

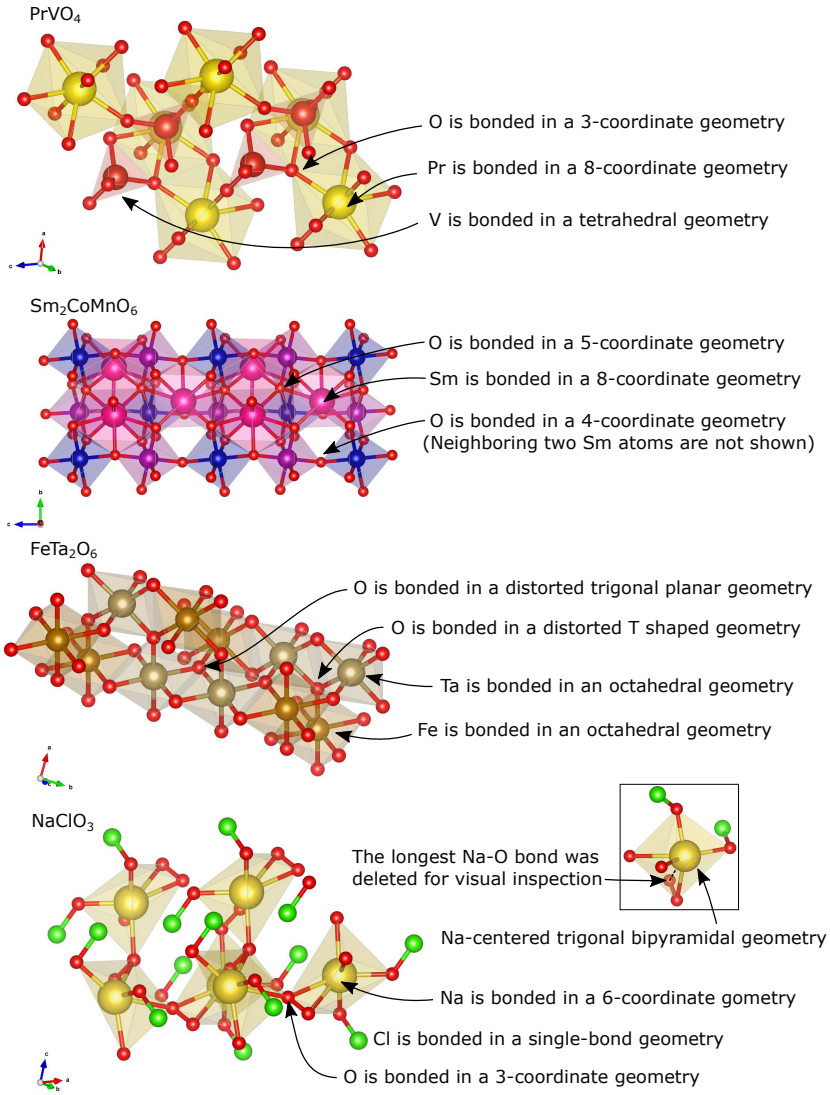

Figure S1: Visual confirmation of concept retrieval. All the structures were visualized using VESTA [84].

Retrieval of connected polyhedral geometries

Table S3: Top-5 scored concepts for SiO_2 that "crystallizes in the hexagonal $P6_3/mmc$ space group. Si(1) is bonded to one O(1), one O(2), one O(3), and one O(4) atom to form corner-sharing SiO_4 tetrahedra. O(6) is bonded in a linear geometry to two equivalent Si(2) atoms. O(4) is bonded in a bent 150 degrees geometry to two equivalent Si(1) atoms. In the sixth O site, O(4) is bonded in a bent 150 degrees geometry to two equivalent Si(1) atoms."

Space group	Score	Local atomic environment	Score
$P6/mcc$	1.89	[METALLOID] is bonded in a hexagonal planar geometry.	2.00
$P6mm$	1.88	[NONMETAL] is bonded in a bent [REALVAL] degrees geometry.	1.86
$P6_3/mmc$	1.78	[NONMETAL] is bonded in a linear geometry.	1.83
$P6_3/mcm$	1.77	[POLYHEDRA] is bonded in a linear geometry.	1.73
$P6/m$	1.76	[METALLOID] is bonded in a distorted pentagonal pyramidal geometry.	1.70

We have briefly discussed missing retrieval of a polyhedron composing connected geometries in Section 4.1 (e.g., Sm_2CoMnO_6 in Table 3). There are cases where a polyhedron in the connection is retrieved simply via the similarity to the local atomic environments (e.g., $FeTa_2O_6$ in Table S12), but for the other cases, intricate texts are required to retrieve the connected geometries. The later cases seem to arise when the structures are specifically characterized by the connected geometries rather than by each single polyhedron. Although it is not essential in itself what phrases would retrieve the concepts, a successful retrieval allows us to see whether the concept is embedded near the target crystal. We here discuss the retrieval of connected geometries in detail.

Table S3 is the erroneous result for SiO_2 (mp-1256614), which is quartz (alpha)-like crystal structured. This crystal is specifically characterized with its rings composed of connected SiO_4 tetrahedra. Its space group $P6_3/mmc$ was retrieved as the third scored one, and both the first scored $P6/mcc$ and $P6_3/mmc$ have the hexagonal crystal system and have the point group of D_{6h}. The second scored space group $P6mm$ has the same crystal system but its point group is C_{6v}, and so the 6-fold rotation axis was suggested by CLICS. The problem is that, although the second and third scored local atomic environments are correct, the first scored one is perfectly wrong and also not similar to any local geometries in the crystal: The only [METALLOID] is Si and is bonded in the common SiO_4 tetrahedra. We found the same errors for e.g. $Ca(SiO_2)_6$, $BaFeSi_4O_{10}$, and $LiZr_2(PO_4)_3$ (Table S9) as their fourth scored local atomic environments, and for other examples with the same type of text including long sentences that describe connected polyhedra. The connected tetrahedra might have been learned as a distinct geometric entity not as a single tetrahedron connected each other.

We then inputted all the full patterns of polyhedral connections (Table S1) and computed the similarities, but none of them was in top-5 of SiO_2.

We then specifically input the phrase "In the [METALLOID] site, [METALLOID] is bonded to one [NONMETAL], one [NONMETAL], one [NONMETAL], and one [NONMETAL] atom to form corner sharing [POLYHEDRA] tetrahedra", and found it as the third scored. Note that, in this analysis, we have also added "trigonal pyramid", "octahedral", and "hexagonal pyramid" that has a [METALLOID] center or a [METAL] center, and changed the details of the phrases according to the number of vertices. Among these 8 connected geometries in addition to the 233 local atomic environments (Table 15), "corner sharing tetrahedra" was retrieved with the third score of 1.86 on top of the others, and "corner sharing trigonal pyramid" with the fifth score of 1.79, which is relatively similar to tetrahedral geometry.

Moreover, we had more reasonable retrievals when inputting texts with the two sentences from the correct space group and the local geometries: "This material crystallizes in $P6_3/mmc$ space group. [NONMETAL] is bonded in a bent [REALVAL] degrees geometry" was the first scored of 2.32, "This material crystallizes in $P6_3/mmc$ space group. [POLYHEDRA] is bonded in a bent [REALVAL] degrees geometry" was the second scored of 2.23, "This material crystallizes in $P6_3/mmc$ space group. In the [METALLOID] site, [METALLOID] is bonded to one [NONMETAL], one [NONMETAL], one [NONMETAL], and one [NONMETAL] atom to form corner sharing [POLYHEDRA] tetrahedra" was the third scored of 2.20, 'This material crystallizes in $P6_3/mmc$ space group. In the [METALLOID] site, [METALLOID] is bonded to one [NONMETAL], one [NONMETAL], one [NONMETAL], and one [NONMETAL] atom to form corner sharing [POLYHEDRA] trigonal pyramids" was the forth scored of 2.16, and "This material crystallizes in $P6_3/mmc$ space group. [POLYHEDRA] is bonded in a linear geometry" was the fifth scored of 2.14, respectively. The erroneous "hexagonal planar" geometry did not appear anymore. Finally, the reason why the hexagonal geometry appeared is considerably because the sentence "In the sixth [NONMETAL] site, [NONMETAL] is bonded in a bent [REALVAL] degrees geometry to two equivalent [METALLOID] atoms" has a relatively close embedding to the sentence "[METALLOID] is bonded in a hexagonal planar geometry", where the word "sixth" may have been placed

near the "hexagonal". Getting more phrase variation on connected geometries would be beneficial, but we have not given experimental verification.

Our analysis suggests that CLICS has learned the concepts of connected polyhderal geometries to some extent but as a distinct concept when the crystal is specifically characterized with the connected polyhedra rather than each single local geometry. This observation also suggests a limitation of CLICS that, although the geometric concepts have been embedded on the feature space and the relationships among the local atomic environments were also acquired, CLICS cannot reach the relationships between each local atomic environment and each connected polyhedral geometry. Additional mechanisms would be required to learn the relationships.

Although a systematic analysis is difficult by using so specific patterns of connected geometries, we additionally show some noticeable results. Unlike the case of SiO_2, we did not input specific phrases directly extracted the corresponding texts, but input the 5166 full patterns of connected polyhedra in Table S1.

For $FeTa_2O_6$ discussed earlier, we have retrieved the octahedral geometry in Table 4, but with a relatively low score. A more important thing is that we cannot see the connected geometries only with this retrieval. When we added the full patterns of connected polyhedral geometries, we retrieved the first scored "This crystal has bondings to form [POLYHEDRA] octahedra that share corners with four equivalent [POLYHEDRA] octahedra and edges with two equivalent [POLYHEDRA] octahedra", which is in fact correct except for the two edge-edge connections are inequivalent and there are the other four connections to octahedra. This lack of details may be due to the abstraction of elements. The second scored "This crystal has bondings to form [POLYHEDRA] octahedra that share corners with eight [POLYHEDRA] octahedra and edges with two [POLYHEDRA] octahedra" is correct even up to the detail. This result may support that CLICS could grasp not only the global symmetry and local atomic environments, but also the connections of polyhedral geometries; however, we should also note that the intricate concepts came to have top scores but instead the other entities like "distorted trigonal planar" geometry was out of top-5 scored ones. This is because so finely detailed sentences are similar to each other with including words that trigger to increase similarity scores.

Sm_2CoMnO_6 was discussed and its local atomic environments were retrieved (Table 3), but the connected geometries were not. When the full patterns of connected polyhedral geometries were added, the five coordinate, four coordinate, and eight coordinate geometries were again retrieved as the first, third, and forth scored ones. In addition, the second scored geometry was "This crystal has bondings to form [POLYHEDRA] octahedra that share corners with two equivalent [POLYHEDRA] octahedra and corners with two equivalent [POLYHEDRA] octahedra". The corner-sharing octahedra were correctly retrieved, but the detailed number of connections was inaccurate.

Additional concept retrieval results for different phases of the same chemical compositions

We compare retrieval results for two different phases of $\mathrm{TmCO_4}$ [85], one of which is orthorhombic (mp-1191329) in the validation set and the other is tetragonal (mp-1200742) in the training set.

Table S4 is the result for the orthorhombic $\mathrm{TmCO_4}$ (mp-1191329). The first scored space group is correct, and the score is distinguishable from the top-2 to 5 scores. Also, the top-4 scored local atomic environments were retrieved correctly.

Table S5 is the result for the tetragonal $\mathrm{TmCO_4}$ (mp-1200742). This is in the training set. Its space group $P4_2/nmc$ was retrieved as the forth scored one. The top-2 to 4 scores are the same up to the two decimal places. The first scored space group $P4_2nm$ has the point group of $C_{4\mathrm{v}}$, and so is considered familiar to the correct one $P4_2/nmc$ with $D_{4\mathrm{h}}$. The top-3 scored local atomic environments are also correct. The $\mathrm{TmO_8}$ hexagonal bipyramids were also retrieved, but we should care that this type of sentences have often caused errorneous results when compared with the 233 of local atomic environments, as discussed using examples of $\mathrm{FeTa_2O_6}$ and $\mathrm{SiO_2}$. The fifth scored one is correct except for "distorted" attribution in addition to the trigonal planar geometry.

Note that $\mathrm{TmCO_4}$ is a carbonate hydroxide with possible phases depending on the ionic radii of the rare earth [85]. CLICS directly encodes crystal graphs without finely attributing elements on the text encoder. Although structural dependency on external conditions should also be challenging, mitigating this technical limitation would be useful for connecting chemical compositions and the crystal structures, a possible extension of CLICS.

These two phases of $\mathrm{TmCO_4}$ are quite different in their crystal structures. We next show the results for different phases of $\mathrm{Na_3PS_4}$ that have similar crystal structures induced by an orientationally disordered transition of phosphate anions [86].

Table S6 is the result for the cubic phase of $\mathrm{Na_3PS_4}$ (mp-985584: Its synthesis was reported in [87, 88]). The space group $I\bar{4}3m$ was found as the second scored. The first scored $I\bar{4}2m$ and the third scored $P4_2mc$ are both tetragonal with their point groups of $C_{4\mathrm{v}}$ and $D_{2\mathrm{d}}$. The space group $P4_2mc$ is considered familiar with $P\bar{4}2_1c$, which is also the space group of the tetragonal phase of $\mathrm{Na_3PS_4}$ (mp-28782). Both space groups are attributed to the same Laue class. They share a certain kind of similar crystal structure [89, 90] despite their different ionic conductivities [89]: Grasping functional difference is beyond the expected ability of CLICS. The first and third scored local atomic environments were retrieved correctly. The second scored one is partly correct since the trigonal pyramids is a kind of polyhedra. The fifth scored one was not found in the structure.

Table S7 is the result for the tetragonal phase of $\mathrm{Na_3PS_4}$ (mp-28782). This is particularly similar to the cubic phase with a structural distortion at the Na sites. Its space group $P\bar{4}2_1c$ was retrieved as the first scored one. The top-2 scored local atomic environments were retrieved correctly, but the following three retrievals were wrong. This error is the same type that is considered to have come from the complex sentence structure and our relatively simple text inputting. Two updated descriptions were found on the Materials Project, which were added in the table caption in [ ]†'s. Regarding the second equivalent Na site, Na(2), there are shorter and longer Na-S bonds. Our dataset was based on the shorter bonds and thus CLICS retrieved the four coordinate geometry (highlighted in the blue colored boxes). Regarding the S(1) atoms with four or [five]† bonds to Na's, the S-Na bond lengths are 2.81, 2.90, 2.96, 2.99, and 3.44 Å. Attributing the four shorter bonds gives our generated dataset.

The connected geometries were missed in the above analysis. We give additional results for $\mathrm{Na_3PS_4}$ with adding all the full patterns of polyhedral connections. For the cubic phase (Table S6), "This crystal has bondings to form distorted [POLYHEDRA] trigonal pyramids that share corners with three equivalent [POLYHEDRA] tetrahedra" was retrieved as the second scored next to the first scored "[NONMETAL] is bonded in a tetrahedral geometry". The other retrievals are just similar to the trigonal pyramids, like tetrahedra, and seem not be improved compared to Table S6. For the tetragonal phase (Table S7), "This crystal has bondings to form distorted [POLYHEDRA] tetrahedra that share corners with four equivalent [POLYHEDRA] pentagonal pyramids and an edge-edge with one [POLYHEDRA] tetrahedra" was retrieved as the first scored, and "This crystal has bondings to form distorted [POLYHEDRA] pentagonal pyramids that share corners with eight equivalent [POLYHEDRA] tetrahedra" as the second scored, both of which grasp the connected geometries to some extent, but not accurately up to the detailed numbers of connections.

These examples suggest that CLICS has learned crystal structures without directly connecting them to the chemical compositions, even for similarly structured phases.

Table S4: Top-5 scored concepts for $TmCO_4$ that "crystallizes in orthorhombic $P2_12_12_1$ space group. Tm(1) is bonded in a 9-coordinate geometry to two equivalent O(1), two equivalent O(2), two equivalent O(4), and three equivalent O(3) atoms. C(1) is bonded in a trigonal planar geometry to one O(2), one O(3), and one O(4) atom. O(1) is bonded in a bent 120 degrees geometry to two equivalent Tm(1) atoms. O(2) is bonded in a distorted single-bond geometry to two equivalent Tm(1) and one C(1) atom. O(3) is bonded in a distorted single-bond geometry to three equivalent Tm(1) and one C(1) atom. O(4) is bonded in a distorted single-bond geometry to two equivalent Tm(1) and one C(1) atom.

Space group	Score	Local atomic environment	Score
$P2_12_12_1$	1.92	[NONMETAL] is bonded in a distorted single bond geometry.	1.83
$P6_1$	1.69	[NONMETAL] is bonded in a bent [REALVAL] degrees geometry.	1.77
$P6_5$	1.66	[NONMETAL] is bonded in a trigonal planar geometry.	1.74
$P2_1$	1.66	[METAL] is bonded in a nine coordinate geometry.	1.74
$Pna2_1$	1.62	[NONMETAL] is bonded in a trigonal non coplanar geometry.	1.66

Table S5: Top-5 scored concepts for $TmCO_4$ that "crystallizes in the tetragonal $P4_2/nmc$ space group. Tm(1) is bonded to two equivalent O(1), two equivalent O(2), and four equivalent O(3) atoms to form a mixture of distorted corner and edge-sharing TmO_8 hexagonal bipyramids. C(1) is bonded in a trigonal planar geometry to one O(2) and two equivalent O(3) atoms. O(1) is bonded in a 3-coordinate geometry to two equivalent Tm(1) and one C(1) atom. O(2) is bonded in a bent 150 degrees geometry to two equivalent Tm(1) atoms. O(3) is bonded in a distorted single-bond geometry to two equivalent Tm(1) and one C(1) atom."

Space group	Score	Local atomic environment	Score
$P4_2nm$	1.92	[NONMETAL] is bonded in a distorted single bond geometry.	1.80
$P4_222$	1.90	[METAL] is bonded in a distorted hexagonal bipyramidal geometry.	1.77
$P4_2/nnm$	1.90	[NONMETAL] is bonded in a bent [REALVAL] degrees geometry.	1.76
$P4_2/nmc$	1.90	[METAL] is bonded in a hexagonal bipyramidal geometry.	1.76
$Fddd$	1.86	[NONMETAL] is bonded in a distorted trigonal planar geometry.	1.74

Table S6: Top-5 scored concepts for Na_3PS_4 that "crystallizes in the cubic $I\bar{4}3m$ space group. Na(1) is bonded in a 4-coordinate geometry to four equivalent S(1) atoms. P(1) is bonded in a tetrahedral geometry to four equivalent S(1) atoms. S(1) is bonded to three equivalent Na(1) and one P(1) atom to form a mixture of corner and edge-sharing SNa_3P trigonal pyramids."

Space group	Score	Local atomic environment	Score
$I\bar{4}2m$	1.97	[NONMETAL] is bonded in a tetrahedral geometry.	2.09
$I\bar{4}3m$	1.87	[NONMETAL] is bonded in a [POLYHEDRA] geometry.	1.92
$P4_2mc$	1.84	[METAL] is bonded in a four coordinate geometry.	1.92
$P\bar{4}3n$	1.81	[NONMETAL] is bonded in a distorted tetrahedral geometry.	1.89
$P\bar{4}3m$	1.79	[POLYHEDRA] is bonded in a tetrahedral geometry.	1.85

Table S7: Top-5 scored concepts for Na_3PS_4 that "crystallizes in the tetragonal $P\bar{4}2_1c$ space group. Na(1) is bonded to six equivalent S(1) atoms to form distorted NaS_6 pentagonal pyramids that share corners with eight equivalent $Na(1)S_6$ pentagonal pyramids, corners with two equivalent $P(1)S_4$ tetrahedra, edges with two equivalent $Na(1)S_4$ pentagonal pyramids, and edges with two equivalent $P(1)S_4$ tetrahedra. Na(2) is bonded in a 4-coordinate geometry to four equivalent S(1) atoms [Na(2) is bonded in a 8-coordinate geometry to eight equivalent S atoms. There are four shorter (2.90 Å) and four longer (3.44 Å) Na-S bond lengths]†. P(1) is bonded to four equivalent S(1) atoms to form PS_4 tetrahedra that share corners with four equivalent $Na(1)S_6$ pentagonal pyramids and edges with four equivalent $Na(1)S_6$ pentagonal pyramids. S(1) is bonded in a 5-coordinate geometry to one Na(2), three equivalent Na(1), and one P(1) atom [S(1) is bonded in a 6-coordinate geometry to five Na and one P atom]†." The text in [ ]† is added as the difference between our data generated in 2023 August and the description on the Materials Project in 2024 March.

Space group	Score	Local atomic environment	Score
$P\bar{4}2_1c$	2.02	[METAL] is bonded in a four coordinate geometry.	1.86
$P\bar{4}2c$	1.95	[NONMETAL] is bonded in a five coordinate geometry.	1.85
$P4/ncc$	1.88	[NONMETAL] is bonded in a distorted trigonal bipyramidal geometry.	1.77
$P4cc$	1.83	[METAL] is bonded in a distorted rectangular see saw like geometry.	1.76
$I222$	1.83	[NONMETAL] is bonded in a four coordinate geometry.	1.74

Some additional results for the retrieval of space group and local atomic environments

Table S8–S13 provide additional results that support discussion in Section 4.1. We select examples various symmetries of crystals including successful retrievals and failure.

In Table S8, the space group and local atomic environments were correctly retrieved. The fifth scored distorted body centered cubic geometry did not appear in the text, but the Ca-centered 8-coordinate geometry is certainly similar to a body-centered cubic geometry.

In Table S9, the geometric concepts were correctly retrieved except for the octahedra and tetrahedra, both of which are in the connected geometries. The "distorted hexagonal planar" geometry seems the same type of error for SiO_2. When we evaluated the similarities additionally to the full patterns of connected geometries, "This crystal has bondings to form distorted [POLYHEDRA] octahedra that share corners with four equivalent [POLYHEDRA] tetrahedra and faces with two equivalent [POLYHEDRA] octahedra" came to the first, and the next was "This crystal has bondings to form [POLYHEDRA] octahedra that share corners with six equivalent [POLYHEDRA] tetrahedra and faces with two equivalent [POLYHEDRA] pentagonal pyramids". The latter certainly retrieves the Zr-centered octahedra connected at their corners with six P-centered tetrahedra. The face-sharing connection seems not correct.

In Talbe S10, the space group was retrieved as third scored, and the top-2 local atomic environments are also correct. The third and forth scored "pentagonal planar" geometry are wrong, but seems to have confusing bondings.

Table S11 and S12 also retrieved their space groups and local atomic environments except for the connected octahedra in $\mathrm{CaCuO_2}$. The space group of $\mathrm{Be_4TeO_7}$ was not retrieved in Table S13, but the second and forth scored space groups are familiar to the correct one. In additional test, "This crystal has bondings to form [POLYHEDRA] octahedra that share corners with three equivalent [POLYHEDRA] tetrahedra and corners with six equivalent [POLYHEDRA] tetrahedra" was retrieved as the first scored one, and "This crystal has bondings to form [POLYHEDRA] tetrahedra that share corners with three equivalent [POLYHEDRA] octahedra and corners with three equivalent [POLYHEDRA] square pyramids" was retrieved as the forth scored one. They retrieved the octahedra connected at the corners and the tetrahedra, but the accurate number of connections were not retrieved. Note that "tetrahdera that share corners with three equivalent octahdera" is correct but not perfectly.

Table S8: Top-5 scored concepts for $\mathrm{CaCuSi_4O_{10}}$ that "crystallizes in the tetragonal $P4/ncc$ space group. Ca(1) is bonded in a 8-coordinate geometry to four equivalent O(1) and four equivalent O(2) atoms. Cu(1) is bonded in a rectangular see-saw-like geometry to four equivalent O(1) atoms. Si(1) is bonded to one O(1), one O(3), and two equivalent O(2) atoms to form corner-sharing $\mathrm{SiO_4}$ tetrahedra. O(2) is bonded in a distorted bent 150 degrees geometry to one Ca(1) and two equivalent Si(1) atoms. O(3) is bonded in a linear geometry to two equivalent Si(1) atoms. O(1) is bonded in a distorted trigonal planar geometry to one Ca(1), one Cu(1), and one Si(1) atom."

Space group	Score	Local atomic environment	Score
$P4/ncc$	2.22	[METAL] is bonded in a rectangular see saw like geometry.	1.84
$P4cc$	2.03	[NONMETAL] is bonded in a distorted trigonal planar geometry.	1.79
$I422$	2.01	[METAL] is bonded in a eight coordinate geometry.	1.78
$P422$	1.99	[NONMETAL] is bonded in a distorted bent [REALVAL] degrees geometry.	1.77
$P42_12$	1.99	[METAL] is bonded in a distorted body centered cubic geometry.	1.75

Table S9: Top-5 scored concepts for $\mathrm{LiZr_2(PO_4)_3}$ that "crystallizes in the trigonal $R\bar{3}c$ space group. Li(1) is bonded in a 6-coordinate geometry to six equivalent O(2) atoms. Zr(1) is bonded to three equivalent O(1) and three equivalent O(2) atoms to form $\mathrm{ZrO_6}$ octahedra that share corners with six equivalent $\mathrm{P(1)O_4}$ tetrahedra. P(1) is bonded to two equivalent O(1) and two equivalent O(2) atoms to form $\mathrm{PO_4}$ tetrahedra that share corners with four equivalent $\mathrm{Zr(1)O_6}$ octahedra. O(1) is bonded in a bent 150 degrees geometry to one Zr(1) and one P(1) atom. O(2) is bonded in a 3-coordinate geometry to one Li(1), one Zr(1), and one P(1) atom."

Space group	Score	Local atomic environment	Score
$R\bar{3}c$	2.05	[NONMETAL] is bonded in a three coordinate geometry.	1.81
$P6_122$	1.86	[METAL] is bonded in a six coordinate geometry.	1.80
$P\bar{6}c2$	1.82	[NONMETAL] is bonded in a bent [REALVAL] degrees geometry.	1.80
$P6_522$	1.79	[METAL] is bonded in a distorted hexagonal planar geometry.	1.60
$P6/mcc$	1.75	[NONMETAL] is bonded in a distorted T shaped geometry.	1.58

Table S10: Top-5 scored concepts for $MgPd_2$ that "crystallizes in the orthorhombic $Pnma$ space group. Mg(1) is bonded in a 10-coordinate geometry to five equivalent Pd(1) and five equivalent Pd(2) atoms. There are two inequivalent Pd sites. In the first Pd site, Pd(2) is bonded in a 5-coordinate geometry to five equivalent Mg(1) atoms. In the second Pd site, Pd(1) is bonded in a 5-coordinate geometry to five equivalent Mg(1) atoms."

Space group	Score	Local atomic environment	Score
$Cmcm$	1.80	[METAL] is bonded in a five coordinate geometry.	1.85
$P2_1/m$	1.77	[METAL] is bonded in a ten coordinate geometry.	1.82
$Pnma$	1.75	[METAL] is bonded in a distorted pentagonal planar geometry.	1.79
$Pbam$	1.64	[METAL] is bonded in a pentagonal planar geometry.	1.78
$C2/m$	1.62	[METAL] is bonded in a four coordinate geometry.	1.74

Table S11: Top-5 scored concepts for $CePd_2Al_3$ that "crystallizes in the hexagonal $P6/mmm$ space group. Ce(1) is bonded in a distorted hexagonal planar geometry to six equivalent Pd(1) atoms. Pd(1) is bonded in a 9-coordinate geometry to three equivalent Ce(1) and six equivalent Al(1) atoms. Al(1) is bonded in a 4-coordinate geometry to four equivalent Pd(1) atoms."

Space group	Score	Local atomic environment	Score
$P6/mmm$	2.00	[METAL] is bonded in a distorted hexagonal planar geometry.	1.92
$P6mm$	1.96	[METAL] is bonded in a hexagonal planar geometry.	1.90
$Cmmm$	1.81	[METAL] is bonded in a nine coordinate geometry.	1.85
$I4_1/amd$	1.60	[METAL] is bonded in a four coordinate geometry.	1.83
[UNK]	1.54	[METAL] is bonded in a eighteen coordinate geometry.	1.77

Table S12: Top-5 scored concepts for $CaCuO_2$ that "crystallizes in the tetragonal $P4/mmm$ space group. Ca(1) is bonded in a body-centered cubic geometry to eight equivalent O(1) atoms. Cu(1) is bonded in a square co-planar geometry to four equivalent O(1) atoms. O(1) is bonded to four equivalent Ca(1) and two equivalent Cu(1) atoms to form a mixture of face, corner, and edge-sharing OCa_4Cu_2 octahedra."

Space group	Score	Local atomic environment	Score
$P4/mmm$	1.86	[METAL] is bonded in a square co planar geometry.	1.93
$I4/mmm$	1.63	[METAL] is bonded in a body centered cubic geometry.	1.84
[UNK]	1.51	[METAL] is bonded in a rectangular see saw like geometry.	1.67
$P4_2/mmc$	1.50	[METAL] is bonded in a linear geometry.	1.56
$P4mm$	1.48	[NONMETAL] is bonded in a linear geometry.	1.48

Table S13: Top-5 scored concepts for Be_4TeO_7 that "crystallizes in the cubic $F\bar{4}3m$ space group. Be(1) is bonded to one O(1) and three equivalent O(2) atoms to form BeO_4 tetrahedra that share corners with three equivalent $Te(1)O_6$ octahedra and corners with six equivalent $Be(1)O_4$ tetrahedra. Te(1) is bonded to six equivalent O(2) atoms to form TeO_6 octahedra that share corners with twelve equivalent $Be(1)O_4$ tetrahedra. O(1) is bonded in a tetrahedral geometry to four equivalent Be(1) atoms. O(2) is bonded in a trigonal planar geometry to two equivalent Be(1) and one Te(1) atom."

Space group	Score	Local atomic environment	Score
$P6_3mc$	1.92	[NONMETAL] is bonded in a [POLYHEDRA] geometry.	1.89
$I\bar{4}3m$	1.90	[NONMETAL] is bonded in a tetrahedral geometry.	1.85
$R3m$	1.87	[NONMETAL] is bonded in a trigonal planar geometry.	1.85
$P\bar{4}3m$	1.84	[NONMETAL] is bonded in a trigonal non coplanar geometry.	1.68
$P321$	1.77	[NONMETAL] is bonded in a distorted tetrahedral geometry.	1.62

Detailed results of text similarity analysis among local atomic environments

The following results support the discussion in Section 4.2.

Table S14 is the top-7 similar concepts against the [METAL]-centered four coordinate geometry. CLICS encodes "three" and "five" coordinates similarly to "four" coordinate, and "distorted rectangular see saw like", "distorted see saw like", and "distorted square co planar" geometries are also similar to "four coordinate" geometry. In fact, these geometries have four coordinates. Note that just the "tetragonal" and "see saw like", and "square co planar" geometries are not in the top-7. They have four coordinate, but also have higher symmetries, which are not endowed for the geometric concepts appeared in Table S14. The concepts that CLICS has learned depends on the dataset and the rules that generated the dataset. The symmetries of local atomic environments come from the order parameter [45, 46]. Compared to the results from CLICS, the result from MatSciBERT [58], which is also used for initialization of CLICS, ranked almost randomly numbered coordinate geometries with with high, almost the same scores.

In Table S15, CLICS encodes "six" coordinate geometry similarly to "five" and "seven" coordinate geometries, and next "distorted hexagonal planar" geometry, which has coordinations with six atoms. There does not appear the octahedral geometry, as such a geometric concept with a high symmetry has its own identity apart from the coordination number. Note that MatSciBERT [58] scored "four coordinate" geometry the first, with a slightly higher score than the second scored ones and others. "five" and "seven" coordinate is considered more similar to "six coordinate", compared to "four coordinate", and thus thus the similarity obtained by CLICS is considered to have grasped the geometric concepts rather than the language-only pretrained text encoder.

In Table S16, CLICS encodes the "eight coordinate" geometry most similarly to "seven" and "nine" coordinate geometries. This is also interpretable even when compared to MatSciBERT [58]'s result that "six" and "four" coordinate geometries come next to "seven coordinate" geometry. The third scored "distorted body centered cubic" geometry has coordination with eight atoms. In addition, there are two remarkable concepts "[NONMETAL] is bonded in a four/five coordinate geometry", despite that similarity is measured against the [METAL]-centered bondings. As we have seen, the double perovskite $\mathrm{Sm_2CoMnO_6}$ (Table 3) has in fact these three geometries. Perovskite-type structures often have [METAL]-centered eight bondings and four and/or five nonmetal-centered bondings. Therefore, we consider that CLICS has learned a relatedness between the concepts that [NONMETAL]-centered four or five coordination geometries and [METAL]-centered eight coordination geometry often arise together.

In Table S17, the [NONMETAL]-centered tetrahedral geometry is most similar to the nonmetal centered distorted tetrahedral geometry. Succeeding it follows the other objects centered tetrahedral geometries. The forth scored "distorted trigonal pyramidal geometry" is in fact similar to tetrahedron but a vertex is reduced. A bit interesting thing is that the [NONMETAL]-centered tetrahedral geometry is more similar to the [METALLOID]-centered one rather than the [METAL]-centered one. This result may reflect the similarity of among the attributions of elements. The sixth scored "trigonal non coplanar" is obtained by reducing another vertex from the forth scored "distorted trigonal pyramidal geometry" except for the distortion. The seventh scored "distorted trigonal bipyramidal geometry" is obtained conversely by adding a vertex to the "distorted trigonal pyramidal geometry". In contrast, although MatSciBERT's similarities include some meaningful concepts, they seem ranked almost randomly based purely on the similarities as texts. The following two examples suggest likewise.

The top-7 ranked concepts were also similar in their geometric meanings to the [METAL]-centered octahedral geometries. The third scored [METAL]-centered square pyramidal geometry is obtained by reducing a vertex from an octahedron, and reducing another vertex at the opposite side gives the forth scored square co planar geometry. The fifth scored "linear geometry" seems not so natural, but an atom-centered octahedron has three mutually perpendicular linear bonds. The sixth scored trigonal bipyramidal geometry is obtained by reducing a vertex from the octahedron. The seventh scored cuboctahedral seems not natural.

Finally, the [METALLOID]-centered hexagonal planar geometry is similar to the distorted one. The pentagonal planar geometry is obtained by reducing a vertex, and its distorted one is ranked next, the hexagonal pyramidal geometry by adding a vertex to the planar geometry comes to the next with a distortion, the bipyramidal geometry with another vertex comes to the next. These geometric concepts are certainly similar.

Table S14: CLICS' top-7 similarity/relatedness to "[METAL] is bonded in a **four** coordinate geometry".

CLICS' top-7 retrieved concepts	Score
[METAL] is bonded in a three coordinate geometry.	0.918
[METAL] is bonded in a distorted rectangular see saw like geometry.	0.907
[METAL] is bonded in a five coordinate geometry.	0.900
[METAL] is bonded in a two coordinate geometry.	0.899
[METAL] is bonded in a distorted see saw like geometry.	0.849
[METAL] is bonded in a distorted square co planar geometry.	0.844
[METAL] is bonded in a distorted trigonal non coplanar geometry.	0.824
MatSciBERT's top-7 retrieved concepts	**Score**
[METAL] is bonded in a six coordinate geometry.	0.9998
[METAL] is bonded in a three coordinate geometry.	0.9998
[METAL] is bonded in a five coordinate geometry.	0.9996
[METAL] is bonded in a seven coordinate geometry.	0.9993
[METAL] is bonded in a two coordinate geometry.	0.9990
[METAL] is bonded in a nine coordinate geometry.	0.9986
[METAL] is bonded in a eight coordinate geometry.	0.9986

Table S15: CLICS' top-7 similarity/relatedness to "[METAL] is bonded in a **six** coordinate geometry".

CLICS' top-7 retrieved concepts	Score
[METAL] is bonded in a five coordinate geometry.	0.915
[METAL] is bonded in a seven coordinate geometry.	0.885
[METAL] is bonded in a distorted hexagonal planar geometry.	0.873
[METAL] is bonded in a distorted pentagonal planar geometry.	0.866
[METAL] is bonded in a pentagonal planar geometry.	0.848
[METAL] is bonded in a eight coordinate geometry.	0.830
[NONMETAL] is bonded in a four coordinate geometry.	0.830
MatSciBERT's top-7 retrieved concepts	**Score**
[METAL] is bonded in a four coordinate geometry.	0.9998
[METAL] is bonded in a five coordinate geometry.	0.9996
[METAL] is bonded in a three coordinate geometry.	0.9995
[METAL] is bonded in a seven coordinate geometry.	0.9995
[METAL] is bonded in a nine coordinate geometry.	0.9990
[METAL] is bonded in a eight coordinate geometry.	0.9990
[METAL] is bonded in a two coordinate geometry.	0.9987

Table S16: CLICS' top-7 similarity/relatedness to "[METAL] is bonded in a **eight** coordinate geometry".

CLICS' top-7 retrieved concepts	Score
[METAL] is bonded in a seven coordinate geometry.	0.884
[METAL] is bonded in a nine coordinate geometry.	0.882
[METAL] is bonded in a distorted body centered cubic geometry.	0.873
[METAL] is bonded in a ten coordinate geometry.	0.861
[NONMETAL] is bonded in a five coordinate geometry.	0.846
[NONMETAL] is bonded in a four coordinate geometry.	0.838
[METAL] is bonded in a twelve coordinate geometry.	0.837
MatSciBERT's top-7 retrieved concepts	**Score**
[METAL] is bonded in a seven coordinate geometry.	0.9990
[METAL] is bonded in a six coordinate geometry.	0.9989
[METAL] is bonded in a four coordinate geometry.	0.9986
[METAL] is bonded in a nine coordinate geometry.	0.9985
[METAL] is bonded in a three coordinate geometry.	0.9982
[METAL] is bonded in a five coordinate geometry.	0.9981
[METAL] is bonded in a nineteen coordinate geometry.	0.9977

Table S17: CLICS' top-7 similarity/relatedness to "[NONMETAL] is bonded in an **tetrahedral** geometry".

CLICS' top-7 retrieved concepts	Score
[NONMETAL] is bonded in a distorted tetrahedral geometry.	0.919
[POLYHEDRA] is bonded in a tetrahedral geometry.	0.860
[METALLOID] is bonded in a tetrahedral geometry.	0.837
[NONMETAL] is bonded in a distorted trigonal pyramidal geometry.	0.832
[METAL] is bonded in a tetrahedral geometry.	0.826
[NONMETAL] is bonded in a trigonal non coplanar geometry.	0.817
[NONMETAL] is bonded in a distorted trigonal bipyramidal geometry.	0.811
MatSciBERT's top-7 retrieved concepts	**Score**
[NONMETAL] is bonded in an octahedral geometry.	0.9987
[METAL] is bonded in a tetrahedral geometry.	0.9977
[NONMETAL] is bonded in a single bond geometry.	0.9977
[NONMETAL] is bonded in a square pyramidal geometry.	0.9974
[NONMETAL] is bonded in a distorted hexagonal planar geometry.	0.9973
[NONMETAL] is bonded in a distorted linear geometry.	0.9973
[NONMETAL] is bonded in a distorted T shaped geometry.	0.9973

Table S18: CLICS' top-7 similarity/relatedness to "[METAL] is bonded in an **octahedral** geometry".

CLICS' top-7 retrieved concepts	Score
[METAL] is bonded in a distorted octahedral geometry.	0.917
[METALLOID] is bonded in an octahedral geometry.	0.856
[METAL] is bonded in a square pyramidal geometry.	0.831
[METAL] is bonded in a square co planar geometry.	0.816
[METAL] is bonded in a linear geometry.	0.806
[METAL] is bonded in a trigonal bipyramidal geometry.	0.801
[METAL] is bonded in a distorted cuboctahedral geometry.	0.799
MatSciBERT's top-7 retrieved concepts	**Score**
[METAL] is bonded in a tetrahedral geometry.	0.9990
[METAL] is bonded in a square pyramidal geometry.	0.9985
[NONMETAL] is bonded in an octahedral geometry.	0.9981
[METAL] is bonded in a distorted octahedral geometry.	0.9977
[METAL] is bonded in a distorted tetrahedral geometry.	0.9974
[METAL] is bonded in a distorted hexagonal planar geometry.	0.9971
[METAL] is bonded in a distorted linear geometry.	0.9971

Table S19: CLICS' top-7 similarity/relatedness to "[METALLOID] is bonded in a **hexagonal planar** geometry".

CLICS' top-7 retrieved concepts	Score
[METALLOID] is bonded in a distorted hexagonal planar geometry.	0.934
[METALLOID] is bonded in a pentagonal planar geometry.	0.837
[METALLOID] is bonded in a distorted pentagonal planar geometry.	0.834
[METALLOID] is bonded in a distorted hexagonal pyramidal geometry.	0.818
[METALLOID] is bonded in a distorted hexagonal bipyramidal geometry.	0.811
[METALLOID] is bonded in a distorted cuboctahedral geometry.	0.811
[METALLOID] is bonded in a distorted pentagonal pyramidal geometry.	0.808
MatSciBERT's top-7 retrieved concepts	**Score**
[METALLOID] is bonded in a trigonal planar geometry.	0.9982
[METALLOID] is bonded in a square pyramidal geometry.	0.9981
[METALLOID] is bonded in a pentagonal planar geometry.	0.9980
[METALLOID] is bonded in a square co planar geometry.	0.9980
[METALLOID] is bonded in a single bond geometry.	0.9977
[METALLOID] is bonded in a linear geometry.	0.9973
[METAL] is bonded in a trigonal planar geometry.	0.9972

Visual demonstration of similar structure search on CLICS' feature space

Figure S2 shows the top-5 similar structures of $PrVO_4$, Sm_2CoMnO_6, and $FeTa_2O_6$ discussed in Section 4.3. The snapshots of these structures were captured on VESTA [84]. All the structures in each column are certainly similar to each structure on the top. Note that these crystals were searched only from the validation set, but still such a variety of similar structures were found.

For $PrVO_4$, all the structures listed in this column are structured as zircon-type or zircon derivatives. have the zigzag chains of 8-coordinate geometries shared on their edges. The zigzag chains are also bonded to the tetrahedral geometries. $DyNbO_4$ (mp-768303) and $HoPO_4$ (mp-4104) have the same structure, but $HoPO_4$ appear as less similar to the others. Instead, $GdY_3V_4O_{16}$ (mp-1224509) is most similar on CLICS. $GdY_3V_4O_{16}$, EuV_2BiO_8 (mp-1225151), and $CaZrV_2O_8$ (mp-1226968) have similar structures to $PrVO_4$ but with more kind of elements, and thus with different crystal systems or space groups. We consider that these similarities have come from the abstraction level of elements, with which we have attributed the metal and nonmetal elements. P atoms in $HoPO_4$ is nonmetal, and thus different in chemical concept from $PrVO_4$ and $DyNbO_4$. On the other hand, $GdY_3V_4O_{16}$, EuV_2BiO_8, and $CaZrV_2O_8$ have metal atoms (Bi may be attributed to a metalloid), and thus these structures are similar under this abstraction level. This result suggests that CLICS could be controlled via how chemical entities are attributed.

For Sm_2CoMnO_6, all the structures have similar compositions and structures. All of them are structured in orthorhombic perovskite derivatives except for Sm_2RuCoO_6, which also seems to have quite a similar structure.

Finally, the rutile-derived structure of $FeTa_2O_6$ is characterized with the edge-sharing octahedral geometries [71], the chains of which are also connected at the corners of the octahedra alternately in a tilted angle. All the structures found on the CLICS' feature space are certainly characterized with such chains of octahedra.

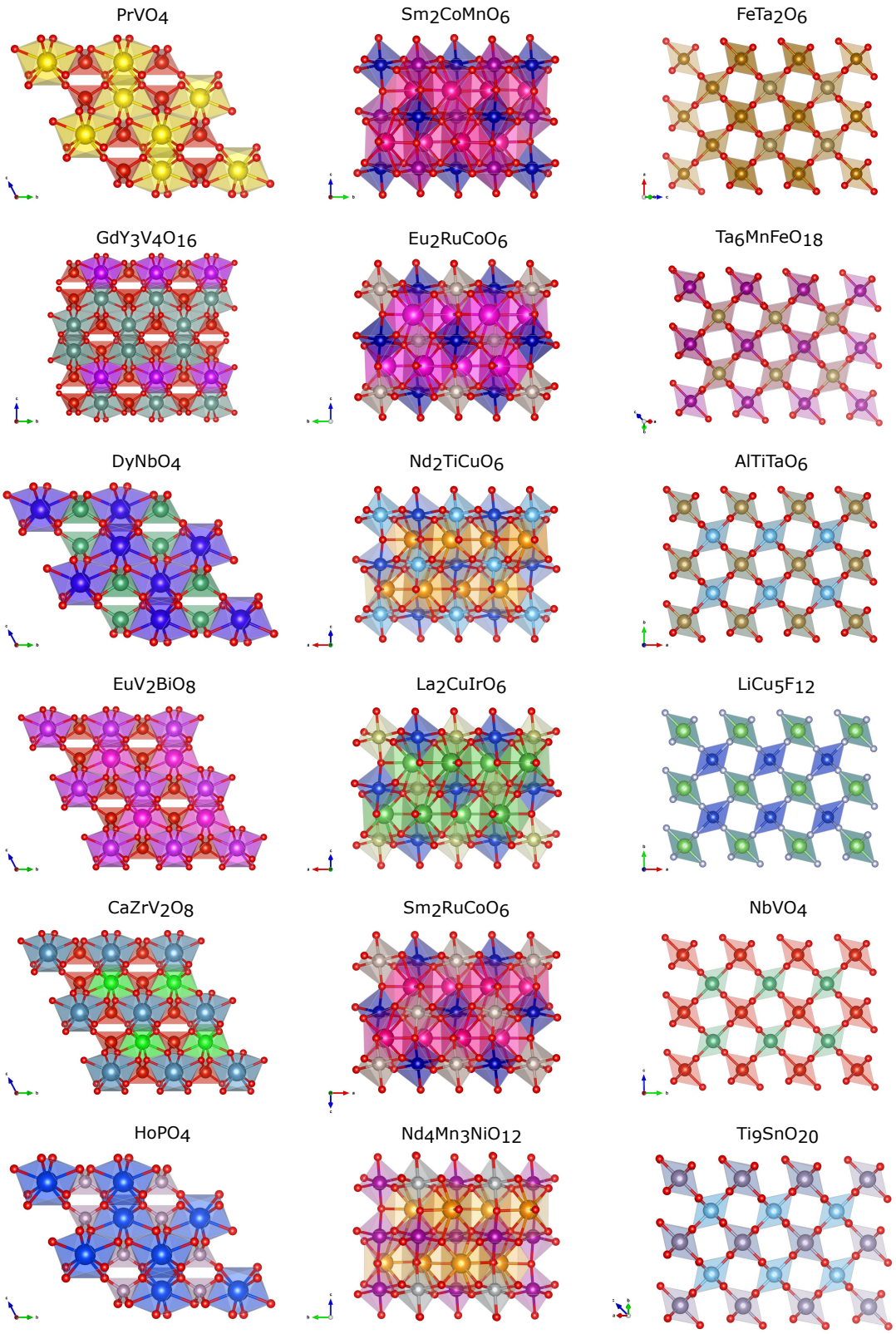

Figure S2: Top-5 similar crystal structures searched on the CLICS feature space. Each column shows the top-5 similar structures to each structure placed on top of the column. All the structures were visualized using VESTA [84] and each aligned on its representative direction.

Detailed results of few-shot and imbalanced classification tasks

We provide detailed results to support the discussion in Section 4.4.

Figure S3 shows the top-1 classification accuracies trained only with two data each crystal structure. We displayed the bars for structures with 11 or more validation data for visibility. The CLICS-pretrained model outperformed the scratch model for most crystal structures, but the scratch model was better partly for caswellsilverite-derived, molybdenum carbide max phase-derived, and magnesium tetraboride-like structures. Some structures from indium-derived, indium-like, beta Cu_3Ti-like, hausmannite-derived, and enargite-like were classified inaccurately by using either model. The accuracies for ilmenite-derived and fluorite structures were also unsatisfactory, but only the CLICS-pretrained model successfully classified some of them. The CLICS-pretrained model classified crystals from the fluorite-derived more accurately. The structural difference between fluorite and its derivatives on their definitions may affect the results.

Figure S4 shows a more detailed result that also displays the count of each crystal structure. On top right is shown for some fewer counts. The CLICS-pretrained classified better regardless the abundance of each structure.

Figure S5 shows the detailed counts for the imbalanced task using 5% data for training. The CLICS-pretrained model was better again compared to the scratch model. There is a difference from the few-shot case: The difference in each classification accuracy was not so significant for abundant structures (from heusler to spinel-derived), while the difference became more significant for less abundant structures. This is probably because the scratch model was trained under the bias of imbalanced data.

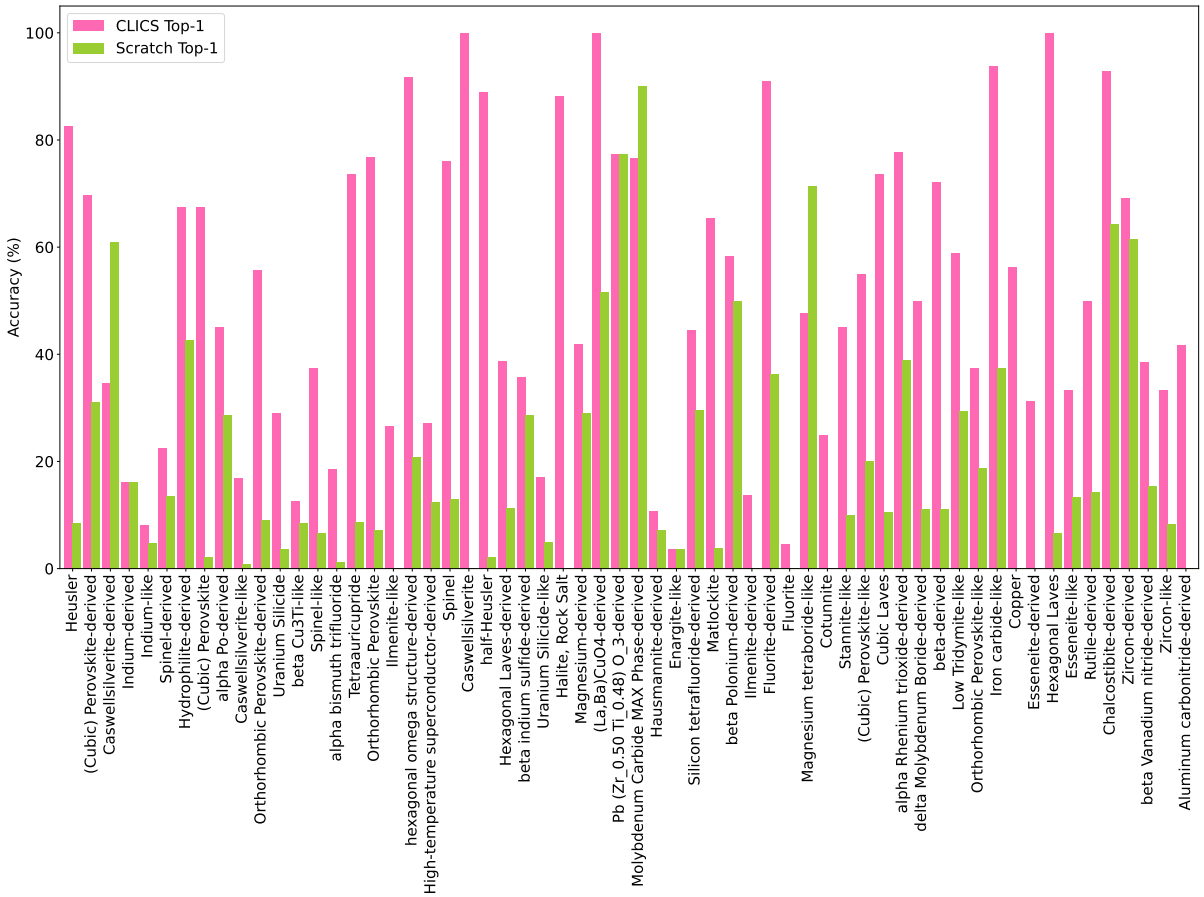

Figure S3: 2-shot classification accuracy (top-1) for each crystal structures with 11 or more validation data. The pink and green colored bars show accuracies of the CLICS-pretrained model and the scratch model.

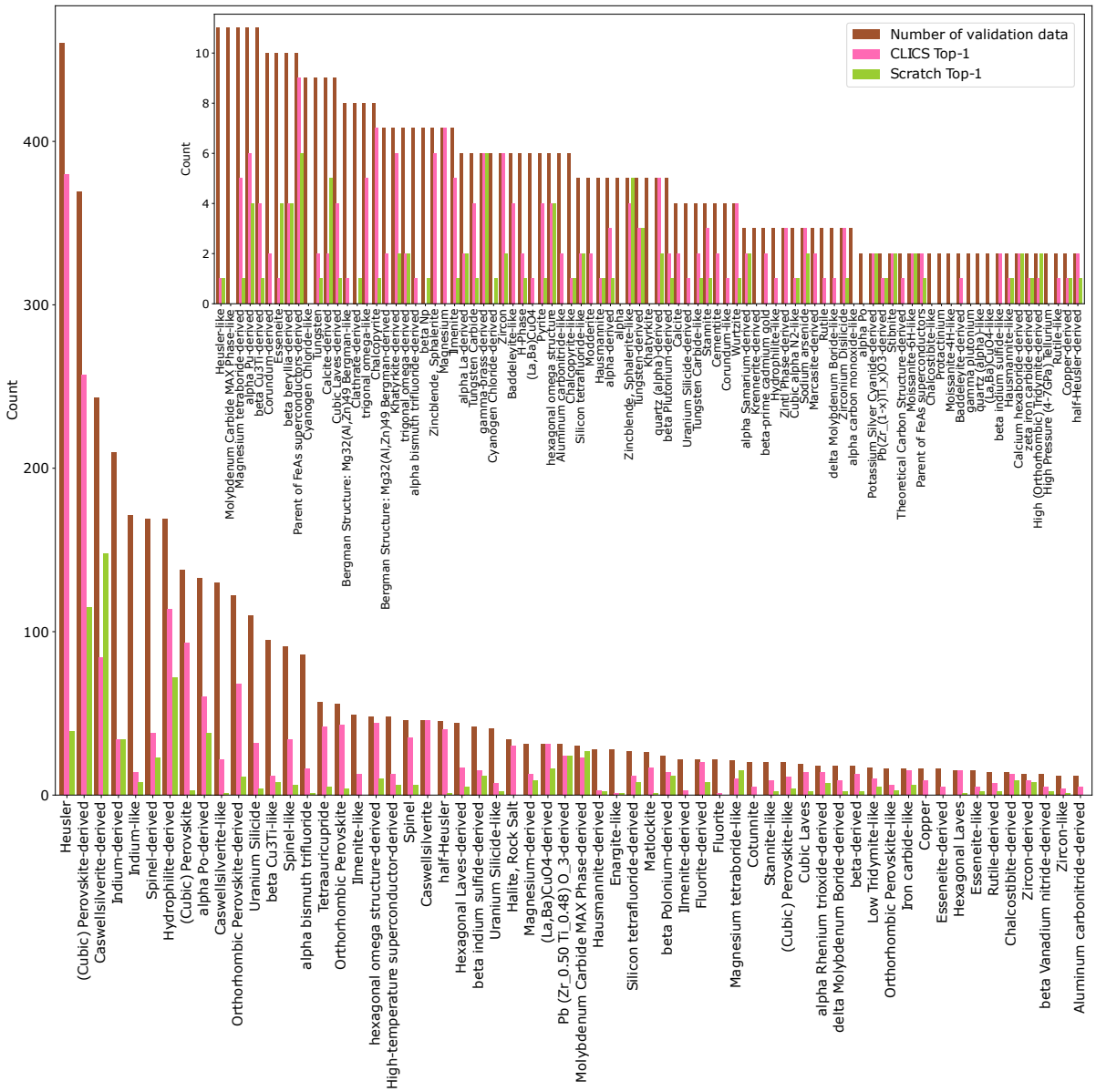

Figure S4: Top-1 counts for all the crystal structures in the 2-shot classification task. The pink and green colored bars show counts that the CLICS-pretrained model and the scratch model correctly classified as top-1, and the brown colored bars show the number of each crystal structure appeared in the validation set. Some fewer crystal structures are displayed on top right for visibility.)

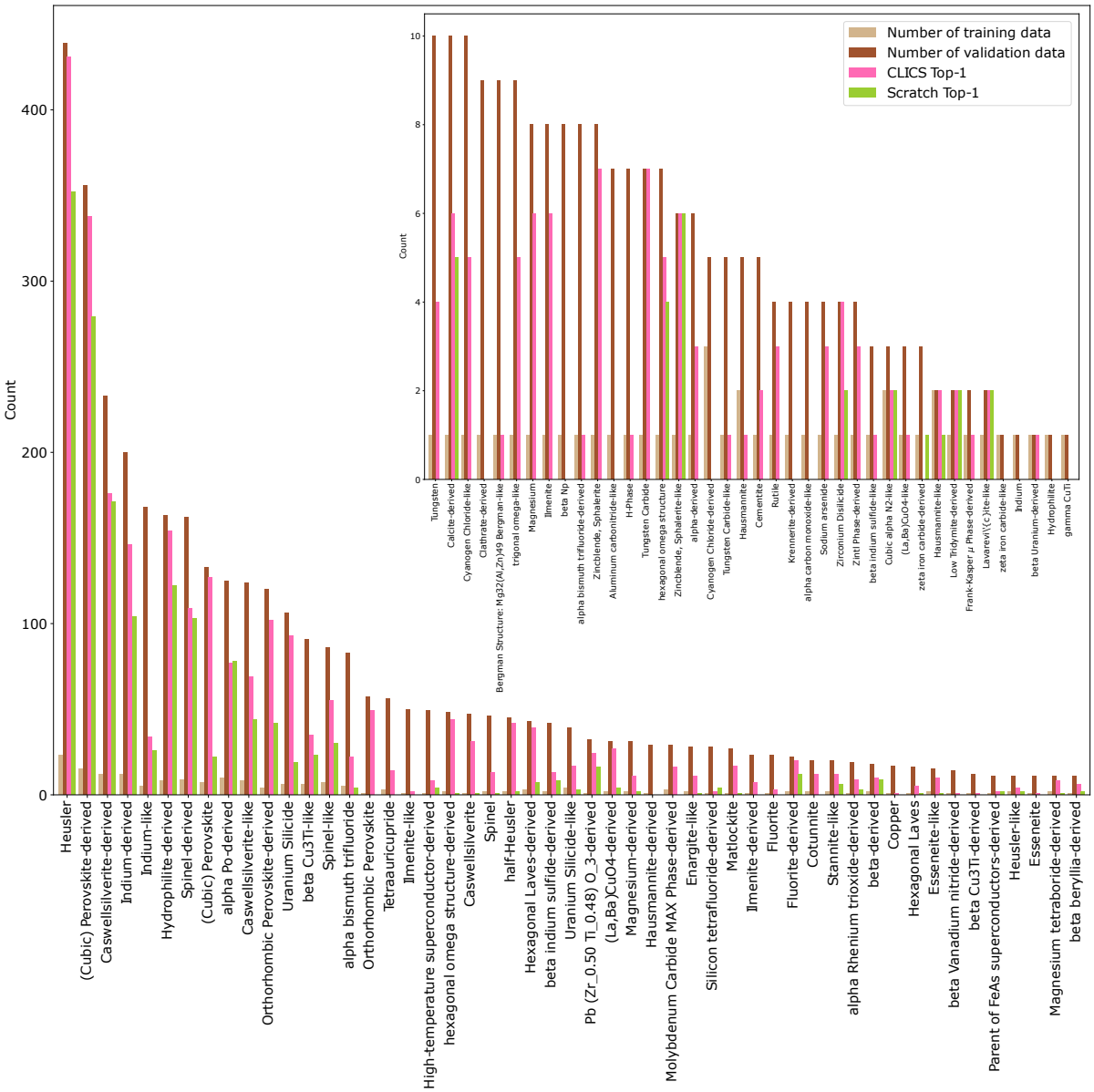

Figure S5: Top-1 counts for all the crystal structures in an imbalanced classification task (5% data for training). The pink and green colored bars show counts that the CLICS-pretrained model and the scratch model correctly classified as top-1; the brown bars show the numbers of each crystal structure appeared in the validation set; the pale brown bars for those in the training set. Some fewer crystal structures are displayed on top right for visibility.)