Supporting Information: Digital Features of Chemical Elements Extracted from Local Geometries in Crystal Structures

Andrij Vasylenko[1], Dmytro Antypov[1], Sven Schewe[2], Luke M. Daniels[1], John B. Claridge[1], Matthew S. Dyer[1], Matthew J. Rosseinsky[1,*]

[1]Department of Chemistry, University of Liverpool, Crown Street, L69 7ZD, United Kingdom
[2]Department of Computer Science, University of Liverpool, Ashton Building, L69 3DR, United Kingdom
*rossein@liverpool.ac.uk

Table of contents

Similarity calculations for local structure environments

In calculations of local structure environment of atomic sites we employ local structure order parameters(LoStOPs)[1] to quantify the agreement between a given observed coordination environment and the perfect elementary target environments in terms of angles. The elementary target motifs, such as 'linear', 'water-like', 'tetrahedral', etc. are illustrated in part in Figures 1 and 3 in the main text and in full in the original study[1]. In this approach, atomic site coordination is determined based on the Voronoi tessellation, and rescaling of the solid angle weights (defined by the Voronoi polyhedron) with the site properties, such as electronegativity differences and distance cut-offs. The resemblance between the local coordination environment of a given atomic site with a range of target motifs is calculated as maximum motif resemblance

$$q_a = max([q_{a,j}]) \tag{S.1}$$

with individual motifs resemblance $q_{a,j}$ calculated with one single neighbor *j* as the North pole for resemblance evaluation to motif type *a* around the central site, their values vary smoothly between 0 and 1. For example, for the T-shaped coordination environment, q_T LoStOP is calculated as

$$q_T = \max_{j \in N, k \neq j} \left\{ \sum_{l \neq j}^{N} \exp\left[-\frac{(\theta_{jl} - 90°)^2}{\Delta\theta^2} \right] \cos^2 \varphi_{jkl} \right\}, \tag{S.2}$$

where *N* is the number of nearest neighbours, θ_{jl} is an angle between the North pole neighbour *j*, central atomic site and neighbour *l*, $\Delta\theta$ is a parameter penalising angle difference with $90°$, φ_{jkl} is angle of a prime meridian (Figure S1).

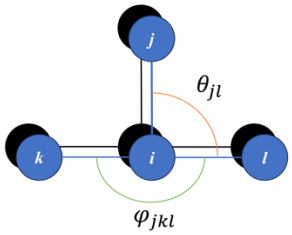

Figure S1. Structural motif similarity on example of the T-shaped motifs (adapted from Ref.[1]). Comparison of the motif for the central atom i (blue circles) with the target T-shaped motif (black circles) in terms of the angles formed by the neighbouring atoms θ_{jl} and φ_{jkl}.

The resemblance values q_a for all atomic sites with 37 target motifs are calculated with the LoStOPs implementations in Matminer[2] for 200809 inorganic crystal structures reported as Crystallographic Information Files (CIFs) in Inorganic Crystal Structure Database (ICSD)[3] (accessed 7.9.2021), which are processed with Pymatgen's CifParser[4]. The calculated values of structural motif similarities for this crystal structural data form the basis for LEAFs and extended matrix of local environments M presented in the main text. For example, for Mg-atom in MgO, the similarity values, s, to 37 target motifs are presented in Figure S2.

Building LEAFs from similarities to local environments

N	Local environment	s	N	Local environment	s	N	Local environment	s
1	Single bond CN1	0.0	14	trigonal pyramidal CN4	0.0	27	q6 CN9	0.0
2	L-shaped CN2	0.0	15	pentagonal planar CN5	0.0	28	q2 CN10	0.0
3	water-like CN2	0.0	16	square pyramidal CN5	0.0	29	q4 CN10	0.0
4	bent 120 degrees CN2	0.0	17	trigonal bipyramidal CN5	0.0	30	q6 CN10	0.0
5	bent 150 degrees CN2	0.0	18	hexagonal planar CN6	0.2	31	q2 CN11	0.0
6	linear CN2	0.0	19	octahedral CN6	1.0	32	q4 CN11	0.0
7	trigonal planar CN3	0.0	20	pentagonal pyramidal CN6	0.5	33	q6 CN11	0.0
8	trigonal non-coplanar CN3	0.0	21	hexagonal pyramidal CN7	0.0	34	Cube octahedral CN12	0.0
9	T-shaped CN3	0.0	22	pentagonal bipyramidal CN7	0.0	35	q2 CN12	0.0
10	square co-planar CN4	0.0	23	body-centered cubic CN8	0.0	36	q4 CN12	0.0
11	tetrahedral CN4	0.0	24	hexagonal bipyramidal CN8	0.0	37	q6 CN12	0.0
12	rectangular see-saw-like CN4	0.0	25	q2 CN9	0.0			
13	see-saw-like CN4	0.0	26	q4 CN9	0.0			

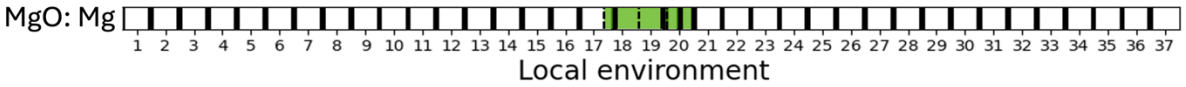

Figure S2. Similarities of Mg local structure environment in MgO to common structural motifs and representation of chemical element (Mg) as a binary 370-long vector. Among 37 considered motifs, 34 motifs are dissimilar to the Mg local environment in MgO (s = 0), and through discretization in 10-bin one-hot encoding, for illustration, can be represented as (1 0 0 0 0 0 0 0 0 0), where only the first bin denoting s =0 contains a value. Six-coordinated motifs (highlighted with green), hexagonal planar (s =0.2), pentagonal pyramidal (s =0.5) and the most similar octahedral (s =1) can be represented as (0 0 1 0 0 0 0 0 0 0), (0 0 0 0 1 0 0 0 0 0), and (0 0 0 0 0 0 0 0 0 1), respectively. Concatenation of these similarities forms a 370-vector for Mg local environment represented as white and black stripes (for 0s and 1s, accordingly). In this vector, the six-coordinated representations are divided with the dashed lines, the 0s are highlighted with green, and the 1s, represented with black, are located in the 2nd ,10th, and the 5th positions, respectively.

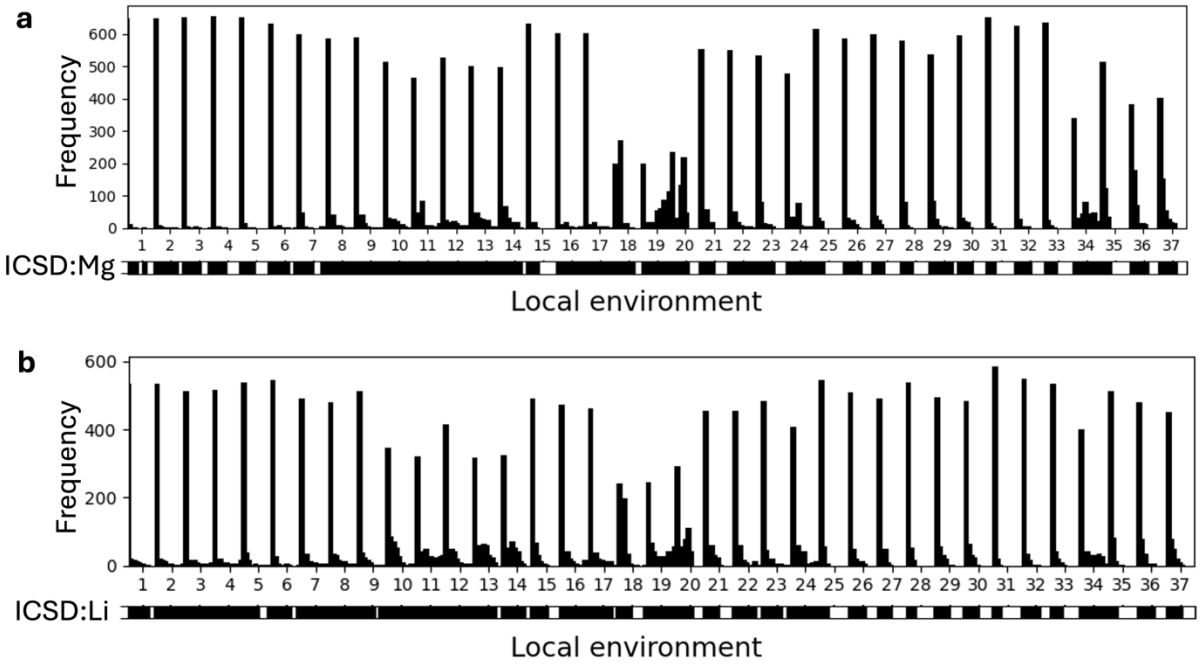

Figure S3. Similarity values for Mg (a) and Li (b) to 37 common local structure motifs collected from all materials reported in ICSD, illustrated as 370-bin histograms and corresponding 370-bit binary strings below. Visual similarity of the Mg and Li binary vectors, presented here with 370 bits for illustration, is further reduced by 1000-bit discretization of structural motifs resulting in 37000-bit elemental vectors, used in this work.

LEAFs autoencoder training

In the unsupervised setting, LEAFs can be calculated through dimensionality reduction of the matrix of local environments $M = (m_{ij})^{21706 \times 86}$ to $(\tilde{m}_{ij})^{n \times 86}$, where 86 signifies the number of elements, and 21706 is the number of non-zero columns resulting from discretization of the elemental 37 LEAFs values into 1000 bins each as described in the main text. Dimensionality reduction is achieved by training a single latent layer of size *n* shallow autoencoder neural network, while minimising the loss function – the reconstruction error, which, in this context, is the Euclidean distance between the decoded output vectors and the original input vectors, which constitute M matrix. The best training results for *n* = 59 are achieved with the loss function calculated as the mean squared error as presented in Figure S4 and fixed learning rate 1e-5, demonstrating convergence of the loss on the held-out test data after 500 epochs.

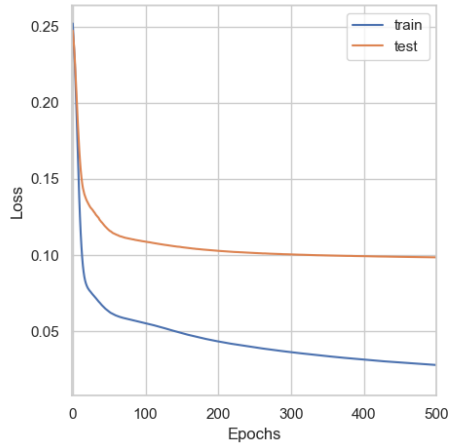

Figure S4. Training of the LEAF shallow autoencoder with mean squared error loss.

Elemental similarity

Chemical elements represented as vectors with LEAFs can be compared with cosine similarity. When arranged by increasing atomic number, the two-dimensional cosine map unveils similarity trends between the elements. Such maps can also qualitatively highlight differences in various elemental descriptors as proposed in Ref.[5] (Figure S5). High cosine similarity between chemical elements represented with LEAFs can serve a reliable basis for element substitution as defined in Eq. 1 in the main text, while retaining the structure of the host material (Table 1 in the main text and Figure S6). The partition function, Z, in Eq. 1, is calculated as a sum over all cosine similarity values in

$$Z = \sum_{i,j} e^{cos(a_i, a_j)}$$

Figure S5: , where indices i,j label different elements in a pair, and summation is performed over all elemental pairs.

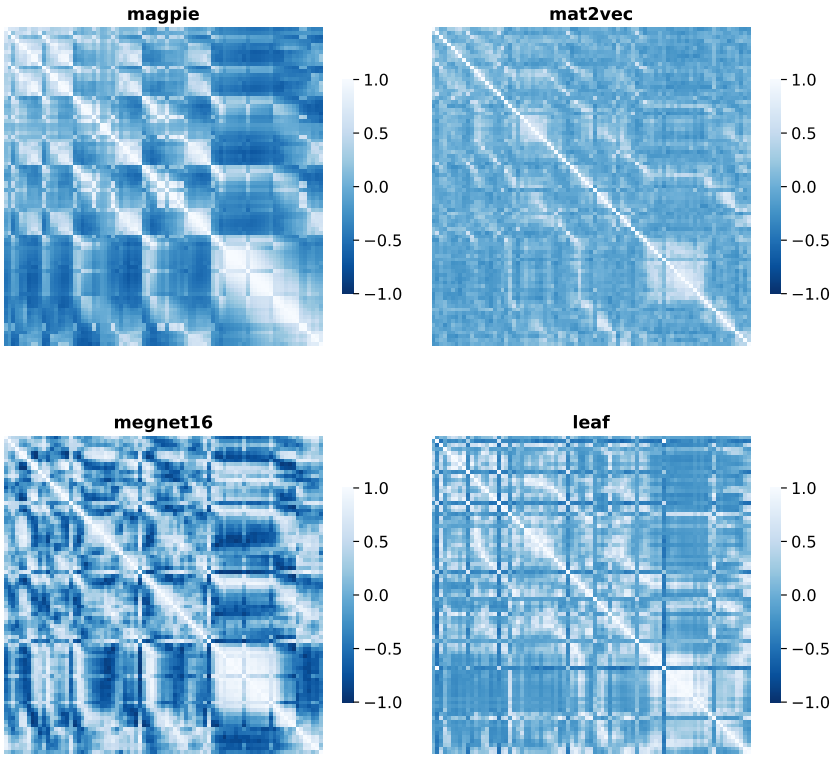

Figure S5. Cosine similarity between chemical elements described with various elemental features (adapted from Ref [6]) highlights different grouping trends arising, depending on the features used.

LEAFs' performance in crystal structure type multi-class classification

Accuracy and MCC performance metrics for this test presented in Table 1 in the main text can be derived from the detailed confusion matrices (Figure S6), in which on-diagonal elements depict the numbers of correctly classified structure types and off-diagonal numbers depict classification errors. LEAFs demonstrate the highest values for on-diagonal numbers and the smallest values for off-diagonal number, illustrating the best performance in classification.

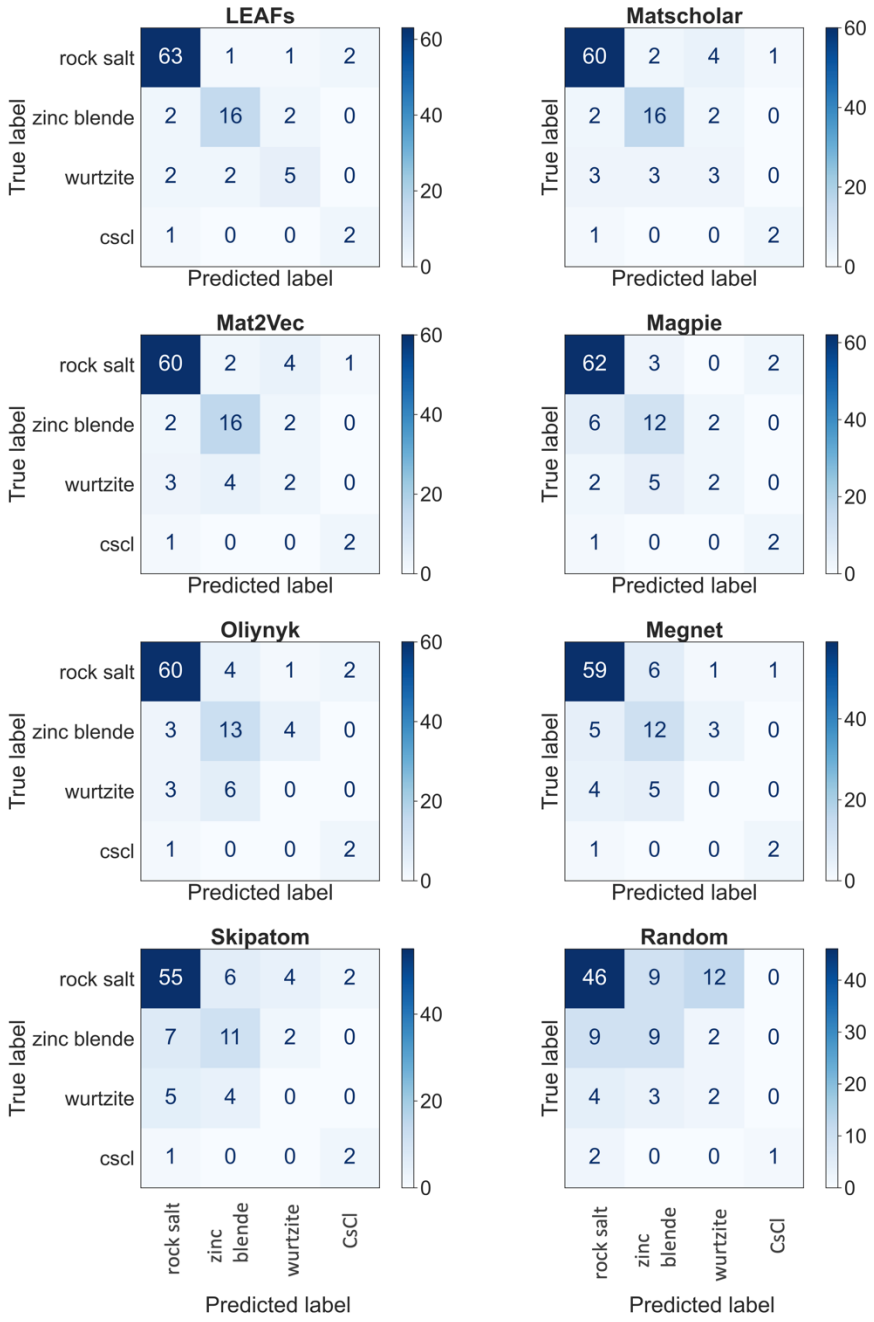

Figure S6. Confusion matrices obtained in classifying crystal structure types for compositions described with different elemental feature sets (adapted from Ref. [6]).

LEAFs distance for composition comparison

Chemical elements represented with LEAFs have been demonstrated to cluster according to chemical trends in Fig. 2a in the main text. In contrast, the analogous t-SNE map for chemical elements represented with random value vectors of the same size as LEAFs does not show any meaningful grouping in Fig. S7.

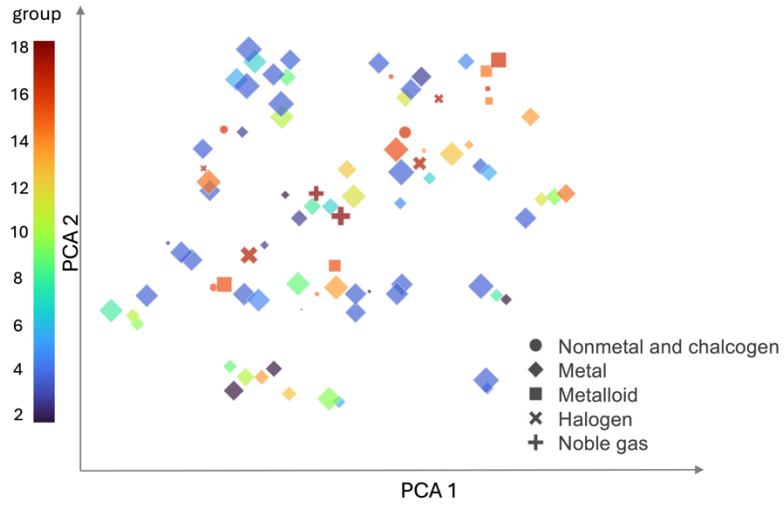

Figure S7. t-distributed Stochastic Neighbour Embedding (t-SNE) map of chemical elements represented with 37-bit vectors of random numbers contrasts with the emerging chemical trends in the analogous t-SNE map produced for elements represented with 37-bit LEAFs (Figure 3a in the main text).

Moreover, clustering of materials regarding their structure type observed in Fig. 2b in the main text and in Figure S8 indicates that distances in the LEAFs-represented chemical space capture structural relationships and can be used as a metric for mapping. For example, the metric for measuring similarity between the compositions can be expressed as cosine similarity between the corresponding compositional representations with LEAFs (Eq. 2 in the main text):

$$S(a_{Li_3PO_4}, a_{Li_7La_3Zr_2O_{12}}) = e^{cos(a_{Li_3PO_4}, a_{Li_7La_3Zr_2O_{12}})}, \quad \text{(S.3)}$$

with the exponent magnifying the differences in the multi-dimensional space.

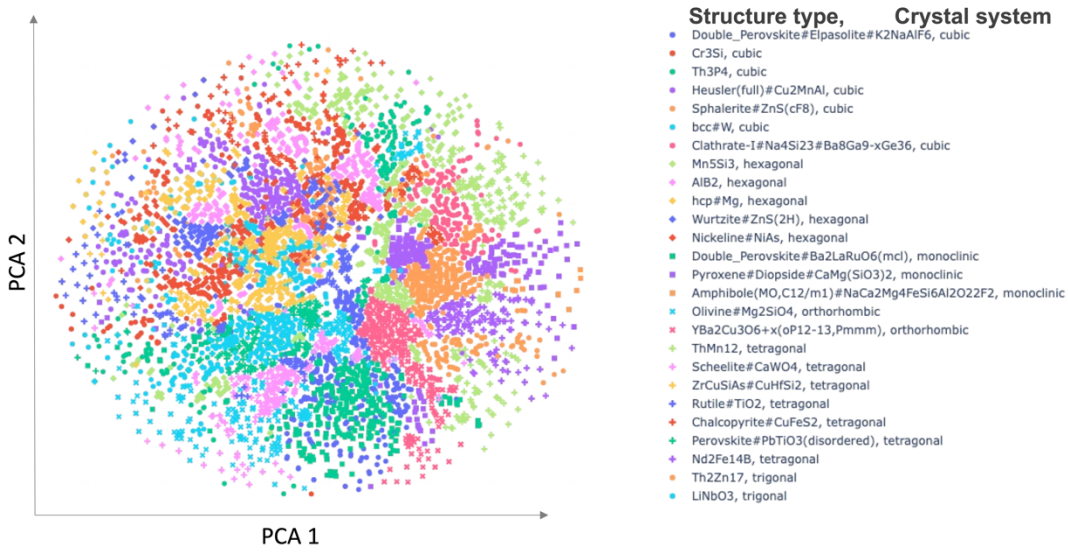

Figure S8. t-SNE map of the LEAFs-represented compositions forming structure type groups represented with $N \in [500, 1000]$ different compositions. The original notations for structure type and crystal system in [?][?] are used. With the varying compactness of distribution in the two principal dimensions, all clusters are located in a distinct area of the map, correlated with the structure types and crystal systems they represent.

LEAFs' utility as elemental representation in composition-property ML models

Digital representation of chemical elements is essential for materials modelling. In the small data regimes that are prevalent in materials science, the choice of representation can have a significant impact on the model performance, especially for the models relying on composition-only input. We demonstrate the utility of LEAFs for such models, trained with CrabNet[7] in integration with the local environments matrix as described in Eq. 4 in the main text. The parameters of CrabNet for the training on the representative materials datasets[8] are unchanged from the original CrabNet study, where Mat2Vec[9] elemental features were employed instead of LEAFs. The two approaches are compared in terms of the average Mean Absolute Error (MAE) computed for the 5-fold cross validation (Table 2 in the main text), in which Mat2Vec and LEAF representations demonstrate overall comparable accuracy in six tests, with maximum MAE improvement of 4% for LEAFs in Dielectric test, and 6% for Mat2Vec in the JARVIS exfoliation energy test.

In contrast to engineered or machine learnt digital elemental representations, including Mat2Vec and random features, LEAFs enable analysis of the elemental local structural environments underpinning composition-property relationships, e.g., through feature selection (Figure 3a in the main text). Additionally, the structures of the Li-ion conducting materials[10] can be analysed in terms of the features, rendered important in Figure 3; in Figure S9, the distribution of such top 25 local structural environments for lithium atomic sites corresponding to high Li-ion conductivity in reported compounds[10] is illustrated.

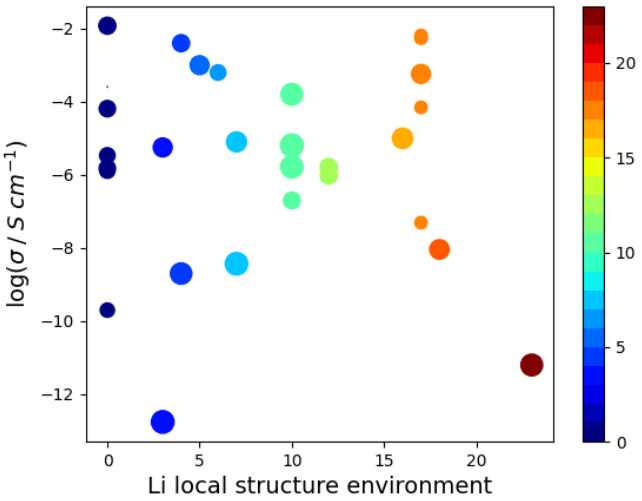

Figure S9. The most pronounced Li local structure environent in Li-ion conducting materials and their respective conductivity. The x-axis sequentially enumerates the considered Li local structure environments (also colour-coded) listed in full in Figure S2 , with the first 10 labelled in Figure 3 in the main text. The size of the markers corresponds to the degree of similarity of the Li coordination to the corresponding structural motif. The broad distribution of local structural environments for Li sites corresponding to high Li-ion conductivity in reported compounds[10] indicates the absence of a specific preferred Li coordination associated with high Li-ion conductivity.

References

1. R. Zimmermann, N. E. & Jain, A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC Adv.* **10**, 6063–6081 (2020).
2. Ward, L. *et al.* Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
3. Zagorac, D., Müller, H., Ruehl, S., Zagorac, J. & Rehme, S. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. *J. Appl. Crystallogr.* **52**, 918–925 (2019).
4. Jain, A. *et al.* A high-throughput infrastructure for density functional theory calculations. *Comput. Mater. Sci.* **50**, 2295–2310 (2011).

5. Onwuli, A., Hegde, A. V., Nguyen, K. V. T., Butler, K. T. & Walsh, A. Element similarity in high-dimensional materials representations. *Digit. Discov.* (2023) doi:10.1039/D3DD00121K.
6. Onwuli, A., Hegde, A. V., Nguyen, K., Butler, K. T. & Walsh, A. Element similarity in high-dimensional materials representations. Preprint at https://doi.org/10.48550/arXiv.2307.00784 (2023).
7. Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *Npj Comput. Mater.* **7**, 1–10 (2021).
8. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *Npj Comput. Mater.* **6**, 1–10 (2020).
9. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
10. Hargreaves, C. J. *et al.* A database of experimentally measured lithium solid electrolyte conductivities evaluated with machine learning. *Npj Comput. Mater.* **9**, 1–14 (2023).