Mutual Information Informed Novelty Estimation of Materials Along Chemical and Structural Axes

Andrew R. Falkowski, Taylor D. Sparks

December 2024

Abstract

This work presents a parameter-free method for estimating materials novelty along chemical and structural axes using mutual information informed density functions. The approach quantifies novelty by analyzing how MI changes with distance between materials, establishing objective criteria for determining meaningful neighborhoods without requiring predetermined parameters. We demonstrate the method's effectiveness using two case studies: a control dataset of materials with varying degrees of similarity and a practical application analyzing lithium-containing compounds from the GNOME dataset relative to known materials. The method successfully identifies meaningful patterns of novelty in both chemical and structural domains while providing interpretable results that align with materials science intuition. This framework offers researchers a quantitative tool for assessing candidate materials against existing knowledge bases and could support more informed selection of synthesis targets in materials discovery campaigns.

1 Introduction

The materials science field has witnessed an expansion of computational and experimental data, with significant resources devoted to developing and maintaining comprehensive data repositories [1–4]. These databases, which now contain several million materials, have enabled rapid computational screening for high performing materials using machine learning [5–7]. While claims of "new" materials frequently appear in the literature, the field lacks robust methods to quantify and assess the novelty of these new materials relative to what is known. It is likely that much of the low hanging fruit in the materials space has been picked and that future high-performing materials will need to be sought after in less explored regions. This necessitates the development of methods to assess and quantify relative novelty in materials databases.

Novelty in the materials science space can take on a variety of meanings depending on the subfield and the specific chemical and structural features that define differentiation therein. In thermoelectric materials, for example, the type, concentration, and spatial distribution of dopants serve as key differentiating features between compounds. At a general level, one can define material novelty along chemical and structural axes. Chemical differentiation is expressed in the use of different elements and formula templates. Structural distinctions are then drawn from the arrangement of these elements. Distinction can be quantified as a *distance* between materials along these axes. Two prominent approaches for computing chemical and structural distance are the element mover's distance (ElMD) [8] and local structure order parameters (LoStOP) [9], respectively. ElMD computes the Wasserstein distance between compounds discretized on a modified Pettifor scale [10], which aims to follow researcher intuition on element similarity. On the structural side, LoStOPs capture deviations from ideal coordination environments and compute the Euclidean distance between structural features. The reader is referred to the relevant publications for further information on these distance metrics. While ongoing research continues to advance materials distance representations [11, 12], this analysis employs the widely-adopted ElMD and LoStOP metrics.

Previous work in materials novelty estimation has explored various methodological approaches, each with distinct limitations. Baird et al. previously used ElMD with a density-based approach to quantify chemical novelty in active learning campaigns [13]. This approach, however, omitted structure and thus could not distinguish between polymorphs (same formula, different structure), which are an important axis of novelty. Additionally, their method computed material densities from multivariate Gaussian density functions over UMAP [14] projections, which makes assumptions of the local structure of the data and introduces stochasticity. This stochasticity leads to inconsistent density calculations that vary with the chosen random seed. Other approaches using variational autoencoders have shown promise in learning structural patterns from X-ray diffraction data and identifying materials outside the training distribution [15]. However, these methods require large training datasets that limit the method's applicability to small, specialized datasets. More recently, Gruver et al. used ElMD and LoStOP to define the chemical and structural novelty of materials generated with a large language model [16]. While differentiation along chemical and structural axes was assessed, their approach relied on fixed, arbitrary cutoff values that may not reflect the natural distance distributions in a materials dataset.

A variety of statistical approaches for novelty and outlier estimation methods exist within the literature [17–19]. While these offer convenient statistical interpretations, they are found to rely on user selected parameters that drastically influence novelty classification outcomes. Additionally, they often make distribution assumptions that are not guaranteed in a materials datasets and may not reflect the local structure of the data. The recent AUTOGLOSH [20] approach attempts to remedy this by providing a data-driven method for selecting optimal parameters. However, this method was found to perform poorly when sharp distinctions between points aren't observed within the data.

In this work, we present a simple, parameter-free method of assessing materials novelty along chemical and structural axes based on a mutual information (MI) informed density function. MI quantifies how much knowing the position of one material tells us about the positions of other materials, with higher values indicating stronger relationships between nearby points. By examining how MI changes with distance, one can establish objective criteria for determining meaningful neighborhoods and influence between materials without requiring predetermined parameters or assumptions about the underlying distribution. This approach preserves signal from the underlying distance metrics while adapting to the natural structure of the data. Through the use of density estimation, our method intuitively weighs closer neighbors more heavily than distant ones, reflecting how researchers typically assess data density. This method provides researchers with a tool for evaluating candidate materials along both chemical and structural axes, enabling more informed decisions in materials discovery campaigns. Furthermore, it offers a foundation for incorporating novelty considerations into active learning strategies.

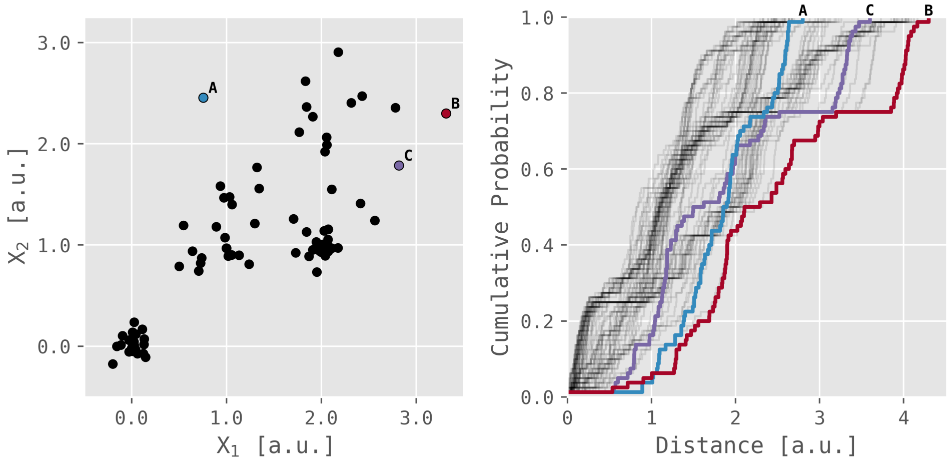
To demonstrate our methodology, we analyze a control dataset of materials with varying degrees of similarity and apply the approach to lithium-containing compounds from the GNOME [21] dataset to assess potential synthesis targets. Through these analyses, we show that our method provides explainable density estimates that conform to researcher intuition while offering new insights into the ways materials might be novel relative to existing compounds. This approach is expected to aid both experimental and computational materials scientists in selecting synthesis targets according to relative novelty in addition to predicted performance metrics.

2 Methods

Estimating the novelty of points in high-dimensional spaces requires balancing local and global characteristics of the data. Our approach addresses this challenge with a density-based novelty metric that reflects the underlying structure of the data. We employ MI analysis to determine both the extent of local neighborhoods and the degree of influence between points within these neighborhoods.

To demonstrate our methodology, we construct a synthetic two-dimensional dataset that exhibits several challenging features common in novelty estimation tasks: regions of varying density, global outliers, and local outliers. Figure 1a illustrates this dataset with three points of interest labeled A, B, and C, each representing different types of potential novelty.

A distance matrix, $D \in \mathbb{R}^{n \times n}$, is constructed from the pairwise Euclidean distances between points in the synthetic dataset. The cumulative distribution of distances for each point, as shown in Figure 1b, reveals distinct patterns of local and global relationships. Points with sparse local neighborhoods are characterized by initially flat profiles in their cumulative distributions. For example, point A exhibits a large gap between it and its nearest neighbors. Notably, point A's distribution span is relatively compact due to its central



(a)

(b)

Figure 1: Synthetic two variable dataset for method demonstration. (a) Data presented in real space with labeled points A, B, and C representing different outlier types of interest. (b) Cumulative distributions of pairwise Euclidean distances of points in dataset with points of interest highlighted.

position within the data space. By contrast, point B, situated at the dataset's edge, displays both local and global deviation patterns. Point C presents a more complex case, showing varying degrees of deviation at different distance scales.

Relying on a simple sum or average of distances would overemphasize edge points and neglect the inherent groupings that exist within the data. To address this limitation, we define a notion of influence that weighs neighbors according to a data-driven decay function. In the context of Figure 1, this is to say that points in the dense cluster in the bottom left corner, tell us little about point B and shouldn't influence its density. As such, we seek to find a cutoff beyond which points do not influence one another.

For a given distance matrix D, we determine an optimal cutoff distance τ^* by analyzing the mutual information between points at varying distance thresholds. We first discretize the distance domain into a set of potential thresholds τ that span the point distances. For each threshold, we construct a binary relationship matrix $R \in \{0,1\}^{n\times n}$ defined as:

$$R_{ij} = \begin{cases} 1 & \text{if } D_{ij} \leq \tau \\ 0 & \text{otherwise} \end{cases}$$

This binary matrix encodes whether each pair of materials are "neighbors" (within a distance τ of each other) or "non-neighbors." For example, if materials i and j are within distance τ of each other, then $R_{ij} = 1$, indicating they are neighbors. The mutual information $I(X;Y)$ between corresponding neighbors and non-neighbors in R is computed as:

$$I(X;Y) = \sum_{x\in\{0,1\}} \sum_{y\in\{0,1\}} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

where $p(x,y)$ represents the probability of observing a particular neighbor relationship ($x = 1$ for neighbors, $x = 0$ for non-neighbors) between one material and a pair of other materials y. The marginal probabilities $p(x)$ and $p(y)$ represent how often each type of relationship occurs independently. A high mutual information value indicates that knowing whether two materials are neighbors provides strong information about their relationships with other materials, suggesting a meaningful neighborhood structure at that distance threshold. The optimal cutoff τ^* is identified at the point where adding more distant neighbors no longer provides significant new information about the material relationships.

Within the identified similarity range, we expect the influence between points to decay with distance. Rather than imposing predetermined decay functions (such as linear, Gaussian, or exponential) that might not reflect the true data structure, we derive our decay function directly from the MI profile.

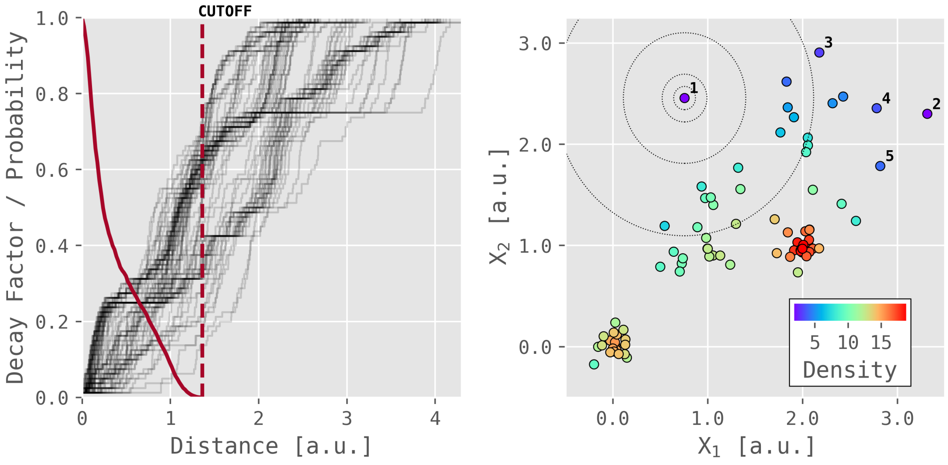
Using the cutoff value τ^* and the complete MI profile $I(\tau)$, we compute a density score ρ_i for each point i that reflects its local neighborhood density in the similarity space:

$$\rho_i = \sum_j f(D_{ij})$$

where $f(d)$ is a decay function derived from the normalized MI profile:

$$f(d) = \begin{cases} 1 - \frac{I(d)}{I_{\max}} & \text{if } d \leq \tau^* \\ 0 & \text{otherwise} \end{cases}$$

Figure 2a illustrates this MI-based decay profile overlaid on the cumulative Euclidean distance distribution of the synthetic dataset. The resulting density estimates and rankings are visualized as color and annotations in Figure 2b, with influence contours shown around point A to illustrate the spatial decay of influence. These contours demonstrate how the MI-based decay function naturally adapts to the local data structure, with appropriate falloff patterns both within dense clusters and in sparse regions.



(a)

(b)

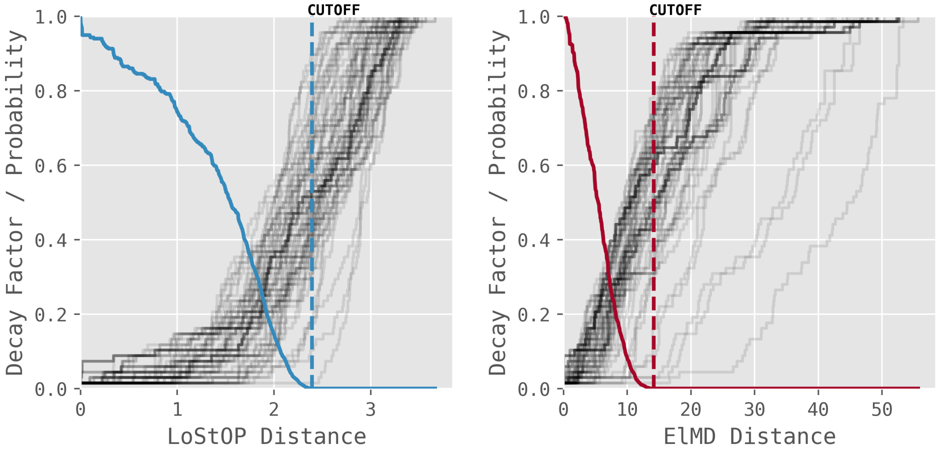
Figure 2: Density estimation approach and results for the synthetic dataset. (a) MI-informed decay function overlaid on Euclidean distance cumulative distributions. (b) Computed density scores with influence contours shown around point A. The five lowest density points are numbered in ascending order.

3 Results & Discussion

The developed method is applied to two materials datasets and used to assess novelty along chemical and structural domains. The first is a control dataset assembled from a variety of materials belonging to distinct materials sub-classes with varying degrees of similarity. The second analysis assesses the novelty of Li-containing compounds in the GNOME dataset relative to experimentally verified compounds in the Materials Project database.

3.1 Assessing Novelty in a Control Dataset

The cumulative distributions and the mutual information decay function are plotted in Figure 3 for both LoStOP and ElMD pairwise distances. The cumulative distributions and MI decay profile provide information on the underlying structure of the data. In Figure 3a the LoStOP MI profile has a sharp drop followed by an extended decay. This implies that there are some dense (high similarity) materials but generally the data is fairly spread out. This is further emphasized by the fact that the mutual information cutoff occurs at ∼65% of the maximum observed distance. The cumulative distribution in this plot is also relatively flat at the outset indicating that most points have relatively few near neighbors. In contrast, the ElMD decay profile shown in Figure 3b decays much quicker, indicating less distinction chemically. This is expected given that ∼70% of the control dataset are oxides. Several extreme outliers whose cumulative distribution deviate significantly are noted. Looking at these profiles ahead of the density analysis provides intuition for interpreting the computed densities.



(a) **(b)**

Figure 3: Distance distributions and MI-informed decay functions for the control materials dataset. (a) LoStOP distance cumulative distributions showing structural relationships. (b) ElMD cumulative distributions showing chemical relationships.

The normalized LoStOP and ElMD densities of the points in the control dataset are plotted in Figure 4. A pareto front showing the maximum tradeoff in structural and chemical density is plotted in red. It is worth noting that a linear relationship between structural and chemical density isn't observed, indicating that there is unique information derived from separating the analysis along the two axes. A wide gap in density is noted between the general data and the points on the Pareto front showing points with significant distinction and making the assessment of novelty easier.

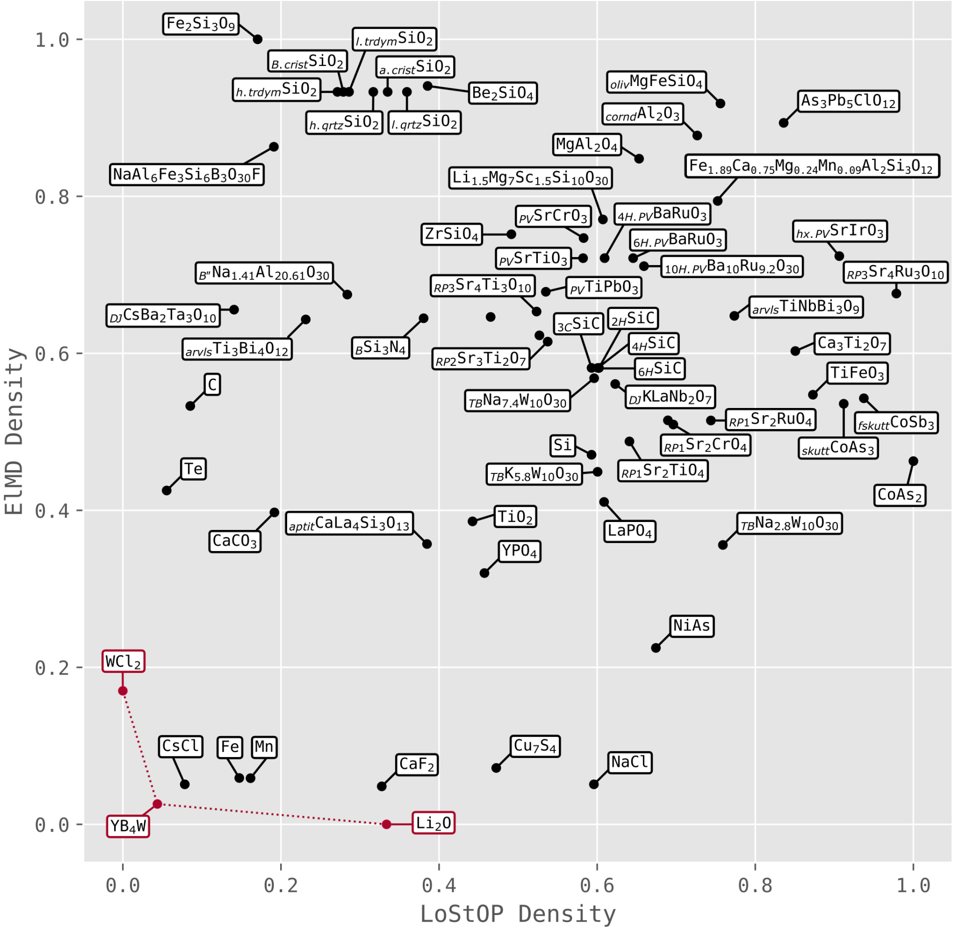


Figure 4: Normalized structural and chemical density plot for materials in the control dataset. A red color is assigned to points and labels of Pareto-optimal materials.

Plotting the densities along each axis reveals several interesting features of the dataset. Although this is not a clustering algorithm, materials with high similarity along one or more axes tend to to be near one another. The SiC polymorphs are identical in formula with minor structural variation. As such, they are densely clustered within the scaled density space. The 3C structure differs from the nH polymorphs and is correctly found to be slightly more structurally novel. It's important to note, however, that neighboring points in the density space are not guaranteed to be neighbors in chemical or structural space, only that they have similar densities. In this case, the SiC polymorphs are neighbors in both spaces, and despite having close neighbors, their density isn't the lowest within the set. To understand this, we can consider the cumulative distributions of SiC polymorphs against a higher density material such as $As_3Pb_5ClO_{12}$ (top right in the plot). The cumulative distributions of the ElMD distances of the SiC polymorphs and $As_3Pb_5ClO_{12}$ are plotted in figure 5. Here, one can see that despite being immediate neighbors, the SiC polymorph cluster is generally distant from other compounds and hence has a lower density.

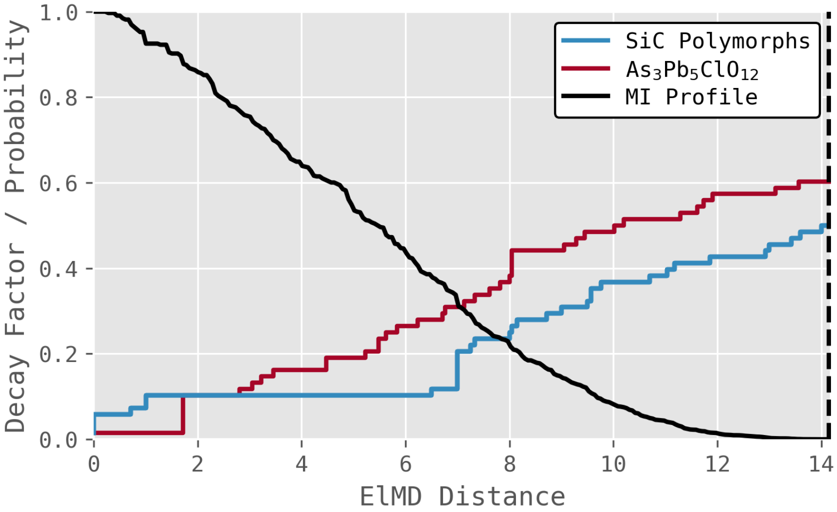
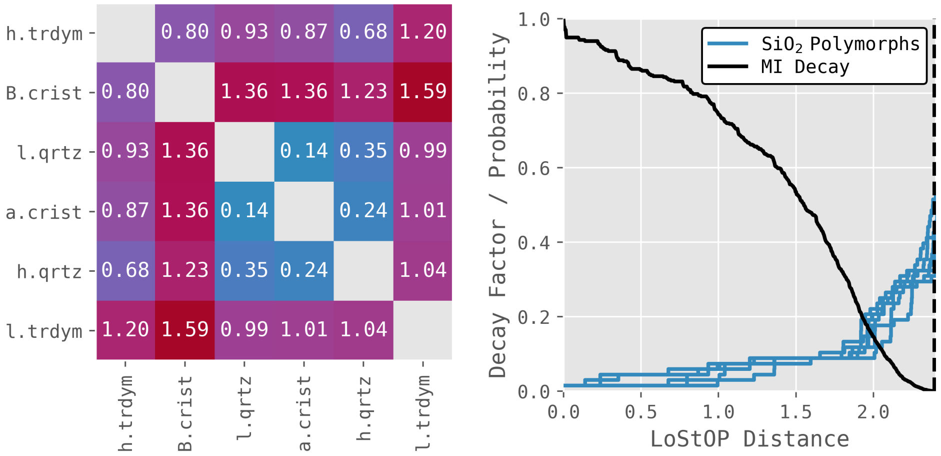


Figure 5: ElMD cumulative distribution of the SiC polymorphs and $As_3Pb_5ClO_{12}$ with The MI-informed decay function overlaid.

A similar situation is found with the SiO_2 polymorphs, which are chemically identical but show structural variation and high novelty relative to other points in the dataset. Pulling the LoStOP distance matrix for these materials, shown in Figure 6a, the polymorphs are structurally dissimilar, which is in line with expectations. Excluding self similarity, the average pairwise structural distance is ~0.92, which is 40% of the average of all pairwise distances in the control dataset, ~2.39. Looking at the cumulative distribution of structural distances in Figure 6b shows similar but distinct cumulative densities that are generally distant. From this information, one can infer that the SiO_2 materials sit in a distant, low density region within the space.



(a)

(b)

Figure 6: SiO_2 materials distances in the control dataset. (a) LoStOP distance matrix between SiO_2 polymorphs showing relative structural similarity. (b) Cumulative structural distance distributions of SiO_2 materials.

In the interest of brevity, an exhaustive analysis of all groupings within the dataset will not be undertaken. However, a few interesting cases are worth noting. Te is structurally unique, but shows moderate density in chemical space despite being the only instance of Te in the dataset. This is odd relative to Fe which is in much lower density regions in chemical space despite there being other compounds containing Fe within the dataset. This is a consequence of the ElMD formulation which relies on the modified pettifor scale within which Te and O are near one another, with Si being relatively close. This results in Te being relatively close to the SiO_2 material cluster and many of the oxide containing compounds; hence why its chemical density is not as low as might be expected in this space. Fe and Mn are far from oxygen and are similar to one another chemically and so are in a globally more sparse region of chemical space.

The separation of two seemingly similar materials $Sr_4Ti_3O_{10}$ and $Sr_4Ru_3O_{10}$ is worth noting. Both materials have similar chemical formulas, but vastly different structural densities despite both of them being 3n Ruddlesden-Popper structures. For example, the Sr-O polyhedra are consistently 9-fold coordination in $Sr_4Ru_3O_{10}$, but in $Sr_4Ti_3O_{10}$ there are both 12-fold coordination as well as 9-fold coordination. $Sr_4Ti_3O_{10}$ incorporates Ti^{4+} ions, which typically adopt smaller ionic radii and favor more symmetric, less distorted octahedral environments. In contrast, $Sr_4Ru_3O_{10}$ contains Ru^{4+} ions, which can exhibit more diverse coordination geometries due to their participation in mixed-valence states and stronger spin-orbit coupling effects. These differences lead to variations in local distortions and overall structural arrangements, contributing to the observed divergence in structural densities.

This result highlights the importance of capturing subtle structural variations, even within the same structural family, as they can significantly impact materials' placement in structural density space. The LoStOP metric is particularly sensitive to these differences, allowing us to distinguish compounds that might otherwise be assumed to be structurally similar based on their chemical compositions alone.

Analysis of the materials on the convex hull is fairly straightforward. WCl_2 doesn't have any neighbors within the mutual information structural cutoff distances. Additionally tungsten and chlorine are relatively rare within the dataset. YB_4W is structurally unique with layers of yttrium and tungsten separated by a boron network. It is also space group 1 or in other words the least symmetric and is the only compound in the dataset with this space group number. Li_2O has an anti-fluorite structure and despite being an oxide is the only compound containing lithium in significant quantity.

Novelty is highly subjective and often includes domain specific nuance. As such, it is unlikely that any single, generalizable novelty estimator will be fully satisfying. That said, the novelty estimation method presented here provides a baseline for further investigations. The results of the control dataset analysis follow intuition based on underlying distance metrics and can be explained by materials intuition, demonstrating the method's utility in the materials space.

3.2 Li-Compounds in the GNOME Dataset

Next we apply this approach to selecting new, novel synthesis targets found in computational datasets. GNOME leveraged a deep learning framework to predict crystal stability. The approach resulted in the "discovery" of 2.2 million crystal structures, 380,000 of which are predicted to be thermodynamically stable [21]. A selection of these GNOME materials are available through the Materials Project. To assess the relative novelty of these materials are, we apply our approach to a small subset of the dataset and look at compounds containing lithium and at least one other element. To serve as an existing corpus, we pulled all experimentally verified lithium containing structures from the Materials Project, totaling 1834 materials. The mutual information cutoff and profile were computed over these to establish knowledge on the existing density data. Next the 44 lithium containing compounds in the contributed GNOME dataset were individually assessed against the existing corpus. This was performed so as isolate each GNOME materials density against the existing corpus and avoid the influence from other GNOME materials.

The resulting chemical and structural densities are plotted in Figure 7 with data from the existing corpus in black and the GNOME data in blue. Pareto optimal materials are labeled in magenta if they are from the existing corpus and blue if they are from GNOME. Labels of chemical formulae are only provided for Pareto optimal materials and GNOME materials. In the interest of visibility the figure is cropped to the range of the GNOME data.

The density data shows that GNOME novelty is primarily in the chemical axis with mixed structural novelty. This is explained by the high presence of exotic elements within the bulk of the GNOME materials with many containing elements from the lanthanides and actinides. Against a large experimental corpus, chemical novelty is likely going to be more easily attained as many of these elements are expensive and difficult to work with experimentally. However, there remain a few high novelty compounds that have the potential for realistic synthesis including $Li_3Zr_3Co_8P_6$ and $LiBr_4O_{10}$. Although, being on the convex hull does not guarantee synthesizability. This view, however, provides a useful filter for selecting valuable materials for experimental synthesis based on their difference from an existing corpus and will hopefully enabled more diversified searches and quantification of novelty.

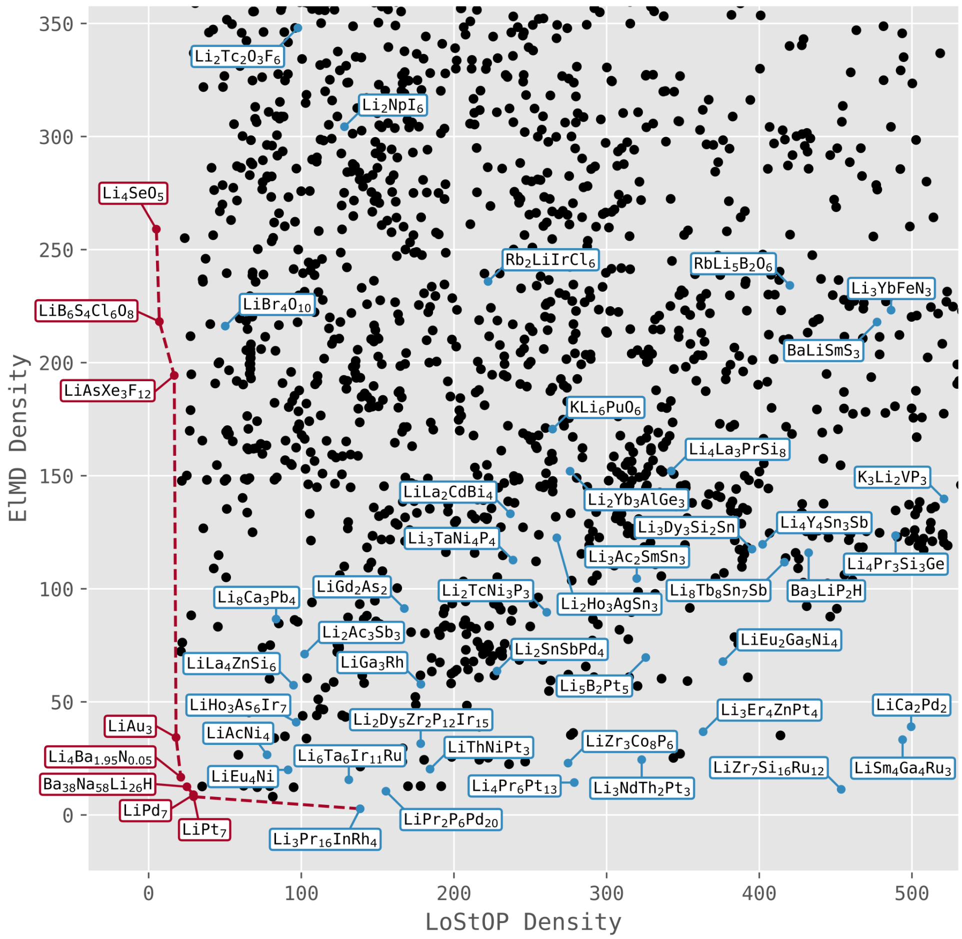


Figure 7: Chemical and structural density comparison between experimentally verified lithium-containing materials (black) in Materials Project and GNOME predicted lithium-containing materials (blue). Pareto-optimal materials are labeled in red if they are from the Materials Project and blue if they are from GNOME.

4 Conclusions

This work presents a parameter-free approach to materials novelty estimation that leverages mutual information to inform density calculations along both chemical and structural axes. By deriving influence functions directly from the mutual information profile of distance distributions, our method preserves the underlying signal from established distance metrics while adapting to the natural structure of the data. The approach successfully identifies meaningful novelty patterns in both a controlled dataset and characterizes novelty in a practical application to lithium-containing compounds from the GNOME dataset.

Analysis of the control dataset demonstrated the method's ability to distinguish between different types of novelty. The separate treatment of chemical and structural axes enables nuanced assessment of materials novelty, allowing researchers to identify materials that are novel in either or both domains. The interpretability of the method was also demonstrated and allowed unusual density rankings to be understood. Application to the GNOME dataset revealed that their computational predictions tend toward chemical novelty, particularly through the incorporation of exotic elements, while showing mixed levels of structural novelty relative to a known corpus of lithium-containing compounds. This approach also highlighted several potential experimental targets with high chemical and structural novelty relative to the existing corpus of data.

The developed framework provides materials scientists with a quantitative tool for assessing candidate materials against existing knowledge, enabling more informed selection of synthesis targets. This approach has potential applications in active learning strategies, where novelty metrics could guide exploration of undersampled regions of the materials space. Such application could enhance the diversity of training data and improve model robustness in previously unexplored domains. Future work might explore the integration of this novelty estimation approach with performance prediction models to optimize the balance between novelty and practical utility in materials discovery campaigns.

5 Code Availability

The code developed for this methodology and used to produce the figures in this manuscript is available at X.com

References

(1) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **2013**, *1*, 011002.

(2) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65*, 1501–1509.

(3) Curtarolo, S.; Setyawan, W.; Hart, G. L.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O., et al. AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science* **2012**, *58*, 218–226.

(4) Chanussot, L. et al. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catalysis* **2021**, *11*, 6059–6072.

(5) Abed, J. et al. Open Catalyst Experiments 2024 (OCx24): Bridging Experiments and Computational Models, 2024.

(6) Mansouri Tehrani, A.; Oliynyk, A. O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T. D.; Brgoch, J. Machine learning directed search for ultraincompressible, superhard materials. *Journal of the American Chemical Society* **2018**, *140*, 9844–9853.

(7) Chen, C.; Nguyen, D. T.; Lee, S. J.; Baker, N. A.; Karakoti, A. S.; Lauw, L.; Owen, C.; Mueller, K. T.; Bilodeau, B. A.; Murugesan, V., et al. Accelerating Computational Materials Discovery with Machine Learning and Cloud High-Performance Computing: from Large-Scale Screening to Experimental Validation. *Journal of the American Chemical Society* **2024**, *146*, 20009–20018.

(8) Hargreaves, C. J.; Dyer, M. S.; Gaultois, M. W.; Kurlin, V. A.; Rosseinsky, M. J. The Earth Mover's Distance as a Metric for the Space of Inorganic Compositions. *Chemistry of Materials* **2020**, *32*, 10610–10620.

(9) R. Zimmermann, N. E.; Jain, A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC Advances* **2020**, *10*, 6063–6081.

(10) Glawe, H.; Sanna, A.; Gross, E.; Marques, M. A. The optimal one dimensional periodic table: a modified Pettifor chemical scale from data mining. *New Journal of Physics* **2016**, *18*, 093011.

(11) Zhang, R.-Z.; Seth, S.; Cumby, J. Grouped representation of interatomic distances as a similarity measure for crystal structures. *Digital Discovery* **2023**, *2*, 81–90.

(12) Vaddi, K.; Li, K.; D. Pozzo, L. Metric geometry tools for automatic structure phase map generation. *Digital Discovery* **2023**, *2*, 1471–1483.

(13) Baird, S. G.; Diep, T. Q.; Sparks, T. D. DiSCoVeR: a materials discovery screening tool for high performance, unique chemical compositions. *Digital Discovery* **2022**, *1*, 226–240.

(14) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2020.

(15) Banko, L.; Maffettone, P. M.; Naujoks, D.; Olds, D.; Ludwig, A. Deep learning for visualization and novelty detection in large X-ray diffraction datasets. *npj Computational Materials* **2021**, *7*, 1–6.

(16) Gruver, N.; Sriram, A.; Madotto, A.; Wilson, A. G.; Zitnick, C. L.; Ulissi, Z. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379* **2024**.

(17) Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, Association for Computing Machinery: New York, NY, USA, 2000, pp 93–104.

(18) Papadimitriou, S.; Kitagawa, H.; Gibbons, P.; Faloutsos, C. In *Proceedings 19th International Conference on Data Engineering (Cat. No.03CH37405)*, 2003, pp 315–326.

(19) Wang, H.; Bah, M. J.; Hammad, M. Progress in outlier detection techniques: A survey. *Ieee Access* **2019**, *7*, 107964–108000.

(20) Ghosh, K.; Naldi, M. C.; Sander, J.; Choo, E. Unsupervised Parameter-free Outlier Detection using HDBSCAN* Outlier Profiles, 2024.

(21) Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. Scaling deep learning for materials discovery. *Nature* **2023**, *624*, 80–85.