

Context-Guided Tree Similarity for Compound-Derived Structural Comparison of Rare-Earth Elements

Guoyong Mao^{a,*}, Xingrui Wei^a, Junwen Wang^b

^a*School of Electrical and Information Engineering, Changzhou Institute
of Technology, Changzhou 213032, China*

^b*Orient Magnet Technology Co. Ltd, Changzhou 213100, China*

maogy@czust.edu.cn, weixingrui@czust.edu.cn, wjw0212@hotmail.com

(Received June 15, 2026)

Abstract

Quantifying structural similarity among chemical elements is important for understanding compositional relationships in chemical systems. In this work, we introduce a compound-derived growth-tree representation that characterizes each element through its compositional expansion patterns across reported compounds.

Based on this representation, we propose a Context-Guided Tree Similarity (CGTS) measure, which evaluates similarity by aggregating local overlaps under shared compositional contexts rather than relying solely on global comparisons.

Applying this framework to rare-earth elements, we obtain a distinctly non-uniform similarity structure, with high similarity concentrated among heavy rare-earth elements such as Tb, Dy, Ho, Er, Tm, and Lu. Analysis of top-ranked pairs reveals strong consistency in local expansion behavior within shared chemical environments.

These results suggest that compound-derived structural similarity provides a meaningful basis for analyzing relationships among chemical elements from a compositional perspective. The identified similarity patterns are consistent with known chemical affinities

*Corresponding author.

within the rare-earth series, indicating that the proposed framework captures structurally relevant chemical behavior. The approach therefore provides a compound-derived descriptor for comparing elemental similarity in chemical systems.

1 Introduction

Quantifying similarity among chemical elements is an important problem in chemical informatics and computational chemistry, with implications for understanding structural relationships and compositional behavior in chemical systems. In particular, rare-earth elements exhibit rich and diverse compound chemistry and play important roles in a wide range of technological applications, including permanent magnets, phosphors, and catalytic systems [2, 3, 9]. Because rare-earth elements often exhibit partially overlapping chemical characteristics, establishing reliable similarity measures is important for analyzing their relationships and potential substitutability.

Existing approaches to quantifying element similarity are typically based on physicochemical descriptors, statistical correlations, or computational representations. Descriptor-based methods rely on properties such as ionic radius, electronegativity, and electronic configuration, and have been widely used to construct feature representations of elements [12]. More recently, machine learning and graph-based predictive models have been developed to infer similarity relationships from large-scale chemical and materials datasets [1, 13, 14]. Structural similarity has also been investigated using chemical network topology and graph-theoretical representations [7, 8]. While these approaches have achieved considerable success, they are primarily based on intrinsic elemental properties, learned representations, or global network structures and often do not explicitly account for how elements participate in actual compound formation. Moreover, conventional graph similarity measures typically compare entire graph structures globally and do not explicitly distinguish the local compositional contexts in which expansion patterns occur.

To address this limitation, we adopt a compound-based perspective that explicitly considers how elements participate in observed composi-

tional systems. Instead of relying solely on intrinsic properties or abstract representations, each element is characterized through the compositions in which it appears. Specifically, we construct a growth-tree representation in which nodes correspond to compositional states and edges represent incremental expansion through the addition of new elements. Starting from simple binary combinations, the structure progressively expands to higher-order compositions, forming a hierarchical structure that reflects how an element participates in increasingly complex compositions. This representation captures not only which elements co-occur with a target element, but also how these relationships evolve through successive compositional expansions.

Based on the growth-tree representation, similarity between elements is defined in terms of their compositional expansion patterns under shared compositional contexts. To this end, we define a Context-Guided Tree Similarity (CGTS) measure that evaluates the consistency of local expansions conditioned on shared parent nodes. For each compositional state that appears in both growth trees, the corresponding sets of expansion outcomes are compared. In contrast to global overlap measures or graph-level similarity metrics that compare entire structures, CGTS focuses on local structural consistency under shared compositional contexts and therefore emphasizes chemically comparable expansion patterns.

2 Data and growth tree representation

2.1 Data source

The compound data used in this study are obtained from the ChemSpider database [11], which provides a comprehensive collection of reported chemical compounds. We extract compounds associated with selected rare-earth magnetic elements and use them as the basis for constructing growth trees. For each element, the collected dataset defines the set of compositions from which the growth tree is derived. The size of the dataset varies across elements, reflecting differences in reporting intensity and chemical research focus.

2.2 Growth tree construction

To characterize the chemical behavior of rare-earth elements, we construct a compound-derived growth tree for each element based on its reported compounds. The key idea is to represent the progressive expansion of chemical compositions as a hierarchical structure.

Prior to tree construction, all compositions are transformed into a canonical form to ensure consistency. Stoichiometric coefficients are removed so that each compound is represented only by its set of constituent elements. For instance, compounds such as E_2O_3 and EO are both reduced to $\{E, O\}$. The element symbols within each composition are then sorted alphabetically to obtain a unique representation (e.g., $\{E, Fe, O\}$), and duplicate entries are eliminated. As a result, multiple compounds with different stoichiometries may correspond to the same composition set. This preprocessing ensures that the representation captures combinational patterns rather than differences in notation or stoichiometry.

Starting from a target element E , the growth tree is constructed by progressively expanding compositions through the addition of new elements. Binary compositions involving E form the second layer of the tree, while higher-order compositions are generated by iteratively incorporating additional elements observed in known compounds. An edge is established between two nodes if one composition can be obtained from the other by adding a single element. For example, if both $\{E, O\}$ and $\{E, O, Fe\}$ are present in the dataset, an edge is created between them, and further expansions (e.g., $\{E, O, Fe, Al\}$) extend this hierarchical structure.

To ensure that the resulting structure forms a valid growth tree, only nodes that are reachable from the root element through such expansion paths are retained. Compositions that cannot be connected to the root via successive single-element additions are removed. This constraint guarantees that the tree reflects consistent compositional evolution pathways and avoids the inclusion of isolated or structurally inconsistent nodes.

Definition (Compound-derived growth tree). Let $\mathcal{C}_E$ be the set of canonical composition sets containing a target element E . The compound-derived growth tree of E is represented as a rooted directed

acyclic graph

$$T_E = (V_E, A_E, r_E),$$

where $r_E = \{E\}$ is the root, $V_E \subseteq \mathcal{C}_E$ is the set of composition nodes reachable from r_E , and

$$A_E = \{(u, v) : u, v \in V_E, u \subset v, |v| = |u| + 1\}.$$

Thus, each directed edge represents a single-element expansion from one composition set to another.

2.3 Coverage of growth trees

The constructed growth trees cover more than 90% of the collected compounds (see Table 1), indicating that the tree representation captures the vast majority of observed chemical combinations. This high coverage is essential, as it ensures that the growth trees provide a sufficiently complete structural representation of the underlying compound space.

Table 1. Coverage of compound data by growth tree representation

Element	Compounds	Sets	Included Sets	Coverage
La	767	191	180	0.94
Ce	783	180	165	0.92
Pr	308	102	98	0.96
Nd	540	132	125	0.95
Pm	8	8	5	0.63
Sm	631	141	138	0.98
Eu	678	147	142	0.97
Gd	1002	151	145	0.96
Tb	305	99	98	0.99
Dy	310	95	94	0.99
Ho	266	90	87	0.97
Er	371	105	104	0.99
Tm	209	82	80	0.98
Yb	492	130	129	0.99
Lu	223	87	84	0.97

Table 1 summarizes the coverage of the growth tree representation for

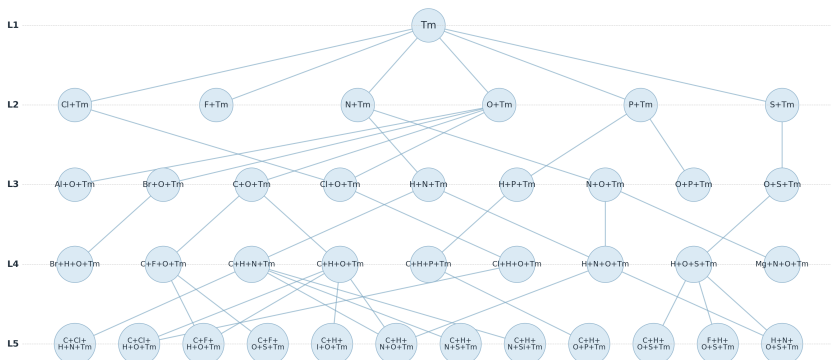


Figure 1. Representative portion of the Tm growth tree

each element. Most elements exhibit consistently high coverage, supporting the validity of using growth trees as the basis for structural representation. In contrast, Pm shows a significantly lower coverage (63%), which is attributed to its radioactive nature and the limited number of reported compounds. Due to insufficient data support, Pm is excluded from subsequent analysis.

2.4 Example of growth tree

Fig. 1 presents the growth tree of Tm as a representative example. The hierarchical structure illustrates how simple binary compositions expand into more complex multi-element combinations, highlighting the compositional pathways through which compounds are formed.

Although the completeness of compound data may vary across elements due to differences in experimental reporting, the analysis in this study is based on relative structural comparison between growth trees. Therefore, moderate variations in data coverage do not materially affect the comparability of the results.

3 Similarity measure

Based on the compound-derived growth tree representation, the similarity between rare-earth elements is defined by comparing their compositional expansion behavior under shared chemical environments.

Let

$$T_i = (V_i, A_i, r_i) \quad \text{and} \quad T_j = (V_j, A_j, r_j)$$

denote the growth trees of elements e_i and e_j , respectively.

For a node $p \in V_i$, the set of child nodes generated from p is defined as

$$\Gamma_i^+(p) = \{c \in V_i : (p, c) \in A_i\}.$$

Similarly, for $p \in V_j$,

$$\Gamma_j^+(p) = \{c \in V_j : (p, c) \in A_j\}.$$

These child-node sets represent the possible compositional expansions from a common chemical context.

The set of comparable parent nodes is defined as

$$P_{ij} = \{p \in V_i \cap V_j : \Gamma_i^+(p) \cup \Gamma_j^+(p) \neq \emptyset\}.$$

The Context-Guided Tree Similarity (CGTS) between elements e_i and e_j is then defined as

$$S_{ij}^{\text{CGTS}} = \frac{1}{|P_{ij}|} \sum_{p \in P_{ij}} \frac{|\Gamma_i^+(p) \cap \Gamma_j^+(p)|}{|\Gamma_i^+(p) \cup \Gamma_j^+(p)|}. \quad (1)$$

The local similarity term is computed using a Jaccard-type overlap [4–6], while the overall similarity is obtained by averaging across shared compositional contexts. Since each local similarity term lies in the interval $[0, 1]$, the resulting CGTS value is also bounded between 0 and 1. The maximum value is attained when the two trees exhibit identical local expansion behavior across all shared parent nodes. Thus, CGTS evaluates whether two elements exhibit consistent expansion behavior when placed

in the same chemical environment.

For a fixed parent node p , comparing child-node sets is equivalent to comparing the corresponding outgoing edge sets, since each child node uniquely defines an expansion step. Therefore, CGTS may also be interpreted as a similarity measure based on shared expansion edges under common structural contexts.

Compared with global overlap measures, graph edit distance, or tree edit distance approaches [10, 15], which evaluate similarity at the level of entire structures, CGTS performs comparisons only within shared compositional contexts. This design makes the measure less sensitive to peripheral structural differences and more directly aligned with the underlying chemical question.

The rationale behind CGTS is that chemically similar elements are expected to exhibit similar expansion patterns when placed in comparable compositional settings. If two elements repeatedly generate similar higher-order compositions from the same parent states, they are likely to play comparable roles in compound formation. Consequently, CGTS characterizes similarity through observed compositional expansion patterns rather than intrinsic elemental properties alone.

4 Results

4.1 Structural similarity between Tm and Tb

We first examine the pair Tm–Tb as a representative example to illustrate structural similarity among rare-earth elements. These two elements belong to the heavy rare-earth group and are expected to exhibit comparable chemical behavior, making them a suitable case for evaluating compositional similarity. Fig. 2 shows the overlap between the growth trees of the two elements, where the target rare-earth element is represented uniformly by X. A substantial portion of the two trees is shared, indicating that Tm and Tb participate in similar compositional environments and exhibit closely related expansion patterns.

Quantitatively, the Tm growth tree contains 80 nodes, whereas the Tb

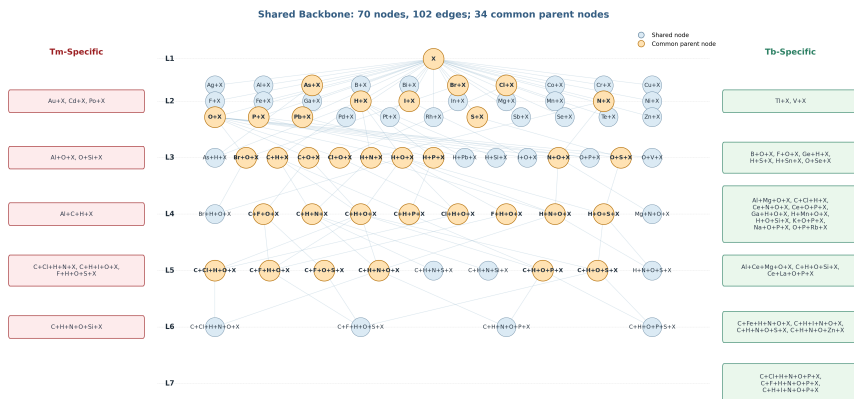


Figure 2. Structural overlap between the growth trees of Tm and Tb, with common parent nodes and their outgoing shared edges highlighted.

growth tree contains 98 nodes, among which 70 nodes are shared. This corresponds to a global node-level Jaccard similarity of 0.6481. In contrast, the CGTS value for Tm–Tb is 0.8594. This value is obtained by computing the Jaccard-type overlap of the child-node sets for each of the 34 common parent nodes and then averaging these local similarities. The sum of the 34 local similarity values is 29.2206, giving

$$S_{\text{Tm}, \text{Tb}}^{\text{CGTS}} = \frac{29.2206}{34} = 0.8594.$$

This indicates a much higher level of similarity when compositional expansions are compared under shared contexts.

This difference reflects an important structural feature visible in Fig. 2: although the overall coverage of compositional space differs between the two elements, their local expansion behavior within shared chemical environments remains highly consistent. Most structural discrepancies arise from peripheral regions of the trees, whereas the overlapping core exhibits strong agreement in compositional expansion patterns.

Overall, this comparison suggests that local structural consistency provides a more informative basis for similarity assessment than global overlap alone. Elements with different overall coverage may still display highly sim-

ilar combinational behavior when evaluated under shared compositional contexts.

4.2 Global similarity structure

To place the Tm–Tb result in a broader context, Fig. 3 presents the CGTS similarity matrix for all rare-earth elements, providing a system-level view of compositional similarity.

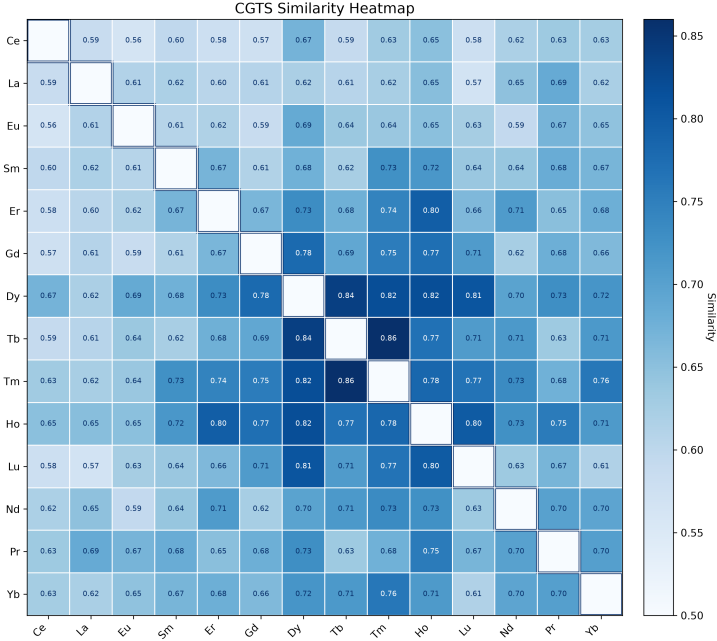


Figure 3. CGTS similarity heatmap for the rare-earth element system

The similarity structure is clearly non-uniform and exhibits a pronounced clustering pattern. High similarity values are concentrated among a subset of elements, particularly Tb, Dy, Ho, Er, Tm, and Lu, indicating that these elements share closely related compositional expansion behavior. In contrast, elements such as La and Ce show consistently lower similarity with the rest of the system, suggesting more distinct combinational patterns.

This pattern reveals a structured similarity landscape in which strong similarity is localized within specific groups rather than distributed uniformly across the system. Notably, the clustering of heavy rare-earth elements emerges directly from compound-derived structural information, indicating that their similarity is not only a consequence of periodic proximity but also reflected in their observed compositional behavior.

4.3 Top-ranked similarity pairs

To further identify the strongest structural relationships, Table 2 lists the ten element pairs with the highest CGTS values.

Table 2. Top 10 element pairs with the highest CGTS values

Element pair	CGTS
Tm–Tb	0.8594
Tb–Dy	0.8387
Dy–Tm	0.8186
Dy–Ho	0.8181
Dy–Lu	0.8093
Ho–Lu	0.7996
Ho–Er	0.7961
Ho–Tm	0.7828
Gd–Dy	0.7789
Gd–Ho	0.7717

The top-ranked pairs identify the strongest pairwise structural correspondences within the rare-earth system. Besides the highest-ranked Tm–Tb pair, combinations such as Tb–Dy, Dy–Tm, and Dy–Ho also exhibit high CGTS values, indicating highly consistent local expansion behavior under shared compositional contexts.

Although these results are broadly consistent with known chemical trends among heavy rare-earth elements, the ranking is not solely determined by periodic adjacency. For example, Tm–Tb obtains the highest CGTS value although the two elements are not adjacent in the periodic table. This indicates that CGTS captures compound-derived structural relationships beyond simple periodic position.

Table 3. Bottom 10 element pairs with the lowest CGTS values

Element pair	CGTS
Ce–Sm	0.5964
Nd–Eu	0.5940
Ce–Tb	0.5909
Eu–Gd	0.5863
La–Ce	0.5852
Ce–Er	0.5786
Ce–Lu	0.5776
Ce–Gd	0.5735
La–Lu	0.5673
Ce–Eu	0.5628

For completeness, Table 3 lists the ten element pairs with the lowest CGTS values. In contrast to the top-ranked pairs, the lowest-similarity combinations are dominated by Ce-related pairs, with La and Eu also appearing among several of the least similar combinations. This pattern is consistent with the broader similarity structure observed in Fig. 3, where Ce and La generally exhibit weaker structural correspondence with many of the heavier rare-earth elements. The bottom-ranked pairs therefore provide a complementary view of the CGTS distribution by highlighting element combinations with comparatively distinct compositional expansion behavior.

4.4 Implications for chemical substitutability

The similarity patterns identified above provide a structural perspective for evaluating substitutability among rare-earth elements. A high CGTS value indicates that two elements exhibit consistent compositional expansion behavior within shared chemical environments, suggesting comparable roles in compound formation. The clustering observed in the similarity matrix further indicates that such structural similarity is not uniformly distributed across the rare-earth series, but concentrated within specific subsets, particularly among heavy rare-earth elements such as Tb, Dy, Ho, Er, Tm, and Lu.

These structural relationships are broadly consistent with practical substitution tendencies in rare-earth applications, where elements with similar chemical behavior are often considered interchangeable under specific performance requirements. Although substitution feasibility ultimately depends on the material system and application constraints, CGTS provides an interpretable approach for identifying chemically related candidate pairs.

5 Conclusion

This study presents a compound-derived structural framework for examining similarity among rare-earth elements through growth-tree representations of compositional expansion behavior. By defining the Context-Guided Tree Similarity (CGTS) measure, the proposed approach evaluates similarity at the level of local structural organization under shared chemical environments, rather than relying solely on global overlap or predefined elemental descriptors.

The results reveal that structural similarity across rare-earth elements is distinctly non-uniform. High similarity values are primarily concentrated among heavy rare-earth elements, including Tb, Dy, Ho, Er, Tm, and Lu, whereas elements such as La and Ce exhibit more differentiated structural patterns. The Tm–Tb case further illustrates that even when overall node coverage differs, elements may still display highly consistent local expansion behavior within shared compositional contexts. This contrast between global overlap and context-conditioned similarity highlights the importance of structural organization in characterizing chemical relationships among elements.

More broadly, the present framework provides an interpretable descriptor for analyzing compositional similarity from compound-derived structural information. Since the method is based on a general representation of compositional expansion, it may be extended to other chemical systems where entities can be characterized through hierarchical structural relationships. The resulting CGTS values are necessarily influenced by the coverage of the underlying compound database, while the composition-

set representation intentionally omits stoichiometric information. Future extensions may therefore incorporate additional data sources and stoichiometric information, together with thermodynamic stability or functional performance, to further examine structural similarity and practical substitutability.

Acknowledgment: The authors gratefully acknowledge the financial support from the National Natural Science Foundation of China under Grant no. 70971089.

Availability of data and materials: The data and code used in this paper are available at <https://github.com/czust-chem-teaching/CGTS>.

References

- [1] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, *Chem. Mater.* **31** (2019) 3564–3572.
- [2] G. B. Haxel, J. B. Hedrick, G. J. Orris, P. H. Stauffer, J. W. Hendley II, *Rare Earth Elements: Critical Resources for High Technology*, U.S. Geological Survey Fact Sheet 087-02, 2002.
- [3] M. Humphries, *Rare Earth Elements: The Global Supply Chain*, Congressional Research Service, Report R41347.
- [4] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et du Jura, *Bull. Soc. Vaudoise Sci. Nat.* **37** (1901) 547–579.
- [5] S. Kosub, A note on the triangle inequality for the Jaccard distance, *Patt. Recog. Lett.* **120** (2019) 36–38.
- [6] M. Levandowsky, D. Winter, Distance between sets, *Nature* **234** (1971) 34–35.
- [7] G. Mao, R. Liu, N. Zhang, Similarity transfer mechanisms of transition metals revealed by chemical network topology, *MATCH Commun. Math. Comput. Chem.* **95** (2026) 643–656.
- [8] G. Mao, R. Liu, N. Zhang, Similarity analysis of chemical elements based on compounds network, *J. Math. Chem.* **61** (2023) 1522–1531.

-
- [9] S. Massari, M. Ruberti, Rare earth elements as critical raw materials: Focus on international markets and future strategies, *Res. Policy* **38** (2013) 36–43.
 - [10] M. Pawlik, N. Augsten, Tree edit distance: Robust and memory-efficient, *Inf. Syst.* **56** (2016) 157–173.
 - [11] Royal Society of Chemistry, ChemSpider, <http://www.chemspider.com>, accessed September 14, 2025.
 - [12] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, *npj Comput. Mater.* **2** (2016) #16028.
 - [13] J. Xiao, L. Yang, S. Wang, Graph isomorphism network for materials property prediction along with explainability analysis, *Comput. Mater. Sci.* **233** (2024) #112619.
 - [14] T. Xie, J. C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, *Phys. Rev. Lett.* **120** (2018) #145301.
 - [15] K. Zhang, D. Shasha, Simple fast algorithms for the editing distance between trees and related problems, *SIAM J. Comput.* **18** (1989) 1245–1262.