Are the chemical families still there?
exploration of similarity among elements[†]

Eugenio Llanos Ballestas[a,b], Wilmer Leal[a,c], Andrés Bernal[d], Jürgen Jost[c,g] and Peter F. Stadler [a,c,e,f,g]

October 5 2022

Abstract

The Periodic Table (PT) is perhaps the most famous and widespread icon of chemistry. It orders chemical elements by their nuclear charge and groups them into families according to their similarity. Such arrangement was built using data about formulae of few compounds available in 19th century. Since then, the number of compounds has grown exponentially during the 20th and 21st centuries, and new types of compounds have been obtained that were unknown to pioneers, rising the question about the validity and generality of the PT. Can these patterns be extracted from current data or are they constrained to a particular chemical domain? To answer this question we conducted a Big Data exploration of chemical similarity using formulae of compounds reported since around 1800. We found that the similarities between elements of the same family are resilient to attacks and are ubiquitous along chemical contexts. We also found that PT groups approach true equivalence classes, being the most symmetrical and transitive on our data. These features point to an underlying structure in the PT ruling the similarity between elements, which agrees with its fundamental nature. Time analysis revealed that since around 1980 all similarity relations are waning by an increasing production of unique formulae on almost all elements, leading to a *singularization* of elements. Nonetheless, PT families tend to be more frequently found, showing they prevail over any other similarity pattern.

1 Introduction

The Periodic System of Chemical Elements (PSCE) perhaps is the most famous and ubiquitous icon of chemistry among science students and the general public. In this system elements are ordered and grouped into families according to their similarity. The PSCE helped to organize chemistry during the 19th Century, allowing to summarize chemistry and introduce newcomers to the field[1]. Although several chemists during this time organized elements into some type of schema according to their atomic weight, the decisive moment came with the contributions of Mendeleev and Meyer[2], since their systems allowed to foresee properties of elements and compounds then unknown. Mendeleev ordered elements by their atomic weight and classified them according to the stoichiometry of their compounds[3]. Meyer also ordered elements by atomic weight[4], but the main emphasis was put on valence. Both authors considered the PSCE as the basis of inorganic chemistry in their textbooks[5,6], and introduced adjustments to their subsequent systems. Finally, the discovery of atomic structure lead to a redefinition of the notion of element and to a new order based on atomic numbers[7].

Nowadays, the PSCE is a subject of great interest in the scientific community. Topics of interest cover ontology[8,9], epistemology[10,11], history[12–14], discussions about its graphical representation[15–17], classification of super-heavy elements[18–20], and many others[21]. Some authors have discussed the quantum-mechanical foundations of the system[22–24], pointing out that several periodic trends may have a relativistic nature, and the quantum origin of some "irregularities" of the heavy elements[25].

The PSCE was built upon a relatively small repertoire of compounds available to pioneers of chemistry in the 19th Century (around 11 thousand compounds were reported by 1869[26]). Chemists have exponentially expanded the number of compounds since then[27,28] and nowadays they amount to tens of millions. Compounds played a key role on the similarity relations found by Mendeleev, whose famous classification of elements was based on their specific ratios of combination with Oxygen and Hydrogen. Since then, chemistry has grown to new domains out of the scope of the knowledge in the 19th Century, and new combinations could produce new patterns of similarity that overshadow those of

[a] *Bioinformatics Group, Department of Computer Science, Universität Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany; E-mail: ellanos@sciocorp.org*
[b] *Corporación SCIO, Calle 57b 50-50 bloque d22 of. 412, 111321 Bogotá, Colombia;.*
[c] *Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany;*
[d] *Departamento de Ciencias Básicas, Universidad Jorge Tadeo Lozano, Carrera 4 22-61, 110311 Bogotá, Colombia;*
[e] *Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany;*
[f] *Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Vienna, Austria;*
[g] *The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico 87501*

the PSCE. This poses questions on the validity of composition based similarity measures for the classification of elements, as the number of new relations among the elements is growing exponentially. By obtaining new compounds, are chemists creating new similarity relations that go beyond the PSCE? Are PSCE similarity relations restricted to a specific domain of compounds? In other words: Could the PSCE be derived from current data?

If the PSCE can be counted as a natural law, it should be visible through the analysis of contemporary chemical data; otherwise it would be a temporary regularity that has little impact in the long run. Several attempts to derive PSCE families are found in literature but they are based on reduced subsets of known compounds[29,30] or are based on a few properties of isolated elements[31–33].

We conducted an exploration of similarity among elements based on stoichiometrical patterns on a large dataset of 26,234,355 compounds from the Reaxys Database [1], which correspond to all compounds published in scientific literature since 1779 (current *Chemical Space*). We defined a similarity measure based on the comparison of molecular formulae. We found that the majority of similarity relationships do not match any family of the PSCE. Nevertheless, similarity between elements of the same family is frequent and its domain spans more chemical contexts than other relations. The maximal spread of PSCE families provides them with a structure very close to equivalence classes.

The patterns reported here reveal the underlying structure of similarity among elements when it is derived from chemical properties, particularly the way elements combine to form compounds. Our approach and results complement those that understand similarity among elements as patterns that emerge when physical properties of atoms are considered. The chemical and physical approaches are the sources of structure for the system of chemical elements. This knowledge is key for the devise of an standard periodic table[25].

2 Results and discussion

2.1 Chemistry is abundant and diverse, but biased

Elements are the alphabet of the language of chemistry and the source of its immense diversity of compounds (see Fig. SI-1). During the past 250 years, chemists have reported 123,006 *combination motifs*. By motif we mean the set of elements of a given compound e.g. {H,N,O} is the motif of HNO_3 and all hydrocarbons have motif {C,H}. Chemists have explored a rich set of motifs, going into atypical regions of the chemical space and pushing the boundaries of the possible. With the exception of actinoids and noble gases, chemists have combined almost every element with almost every other one in at least one compound, bringing to light the combinatorial nature of the combinations of elements and the wealthy diversity of compounds chemists have studied.

Despite this diversity, chemistry shows bias in its compositions. We found that 89.4% of all compounds are combinations of a core of twelve elements: H, C, N, O, F, Si, P, Cl, S, Br, I, B. Combinations of core elements with other elements account for an additional 10.3%, and only 0.3% do not include any element from the core. Compounds do not distribute uniformly along the set of motifs, but they form highly skewed distributions (Fig. 1). Carbon, for instance, forms 25,944,743 compounds, which is a fairly high number, but these are mostly restricted to a comparatively tiny variety of motifs: the most frequent motif, {C, H, N, O}, covers 25.7% of all carbon compounds, followed by {C, H, O } with 10.0%; the top ten motifs encompass 69,4% of C compounds. Similar biases are present for all elements (see Fig. SI-4), showing that synthetic chemists have focused on a few interesting motifs of each element, perhaps obeying technological limitations and pursuing specific goals.

Since about 1980, the chemistry of the bulk of elements has become more stable, as the number of new compounds of each element becomes more steady[27]. Chemists have focused on certain kinds of compounds for each element: after 1980 the spread of the number of compounds over motifs stabilizes on lower values as shown in Fig. 2. There are noisy regions, mostly in the early years, which result from temporary sudden drops in the production of new compounds. In the less noisy regions production of new compounds becomes more stable, and certain combinations gather the majority of compounds, leading to a more steady expected rank. This stable regions are found after about 1980 for many elements and could explain the lower standard deviation of the number of new compounds reported before[27]. Such stable behavior presumably is the consequence of a growing scientific community, which can sustain the production of new compounds on almost all elements.

2.2 Similarity is asymmetrical

Often similarity is conceived as a symmetric relation i.e. *A* is just as similar to *B* as *B* to *A*. However, for many applications, asymmetric similarity measures are useful. Synonymy, for instance, is an asymmetrical similarity relation. Words are used in different numbers of contexts and have different meanings attached to them. The word *pooch* can be substituted by *dog* in pretty much any phrase while preserving the meaning; but there are lots of phrases where *dog* cannot be replaced by *pooch*. The same happens with chemical similarity. Chemists consider elements to be similar inasmuch as they can be
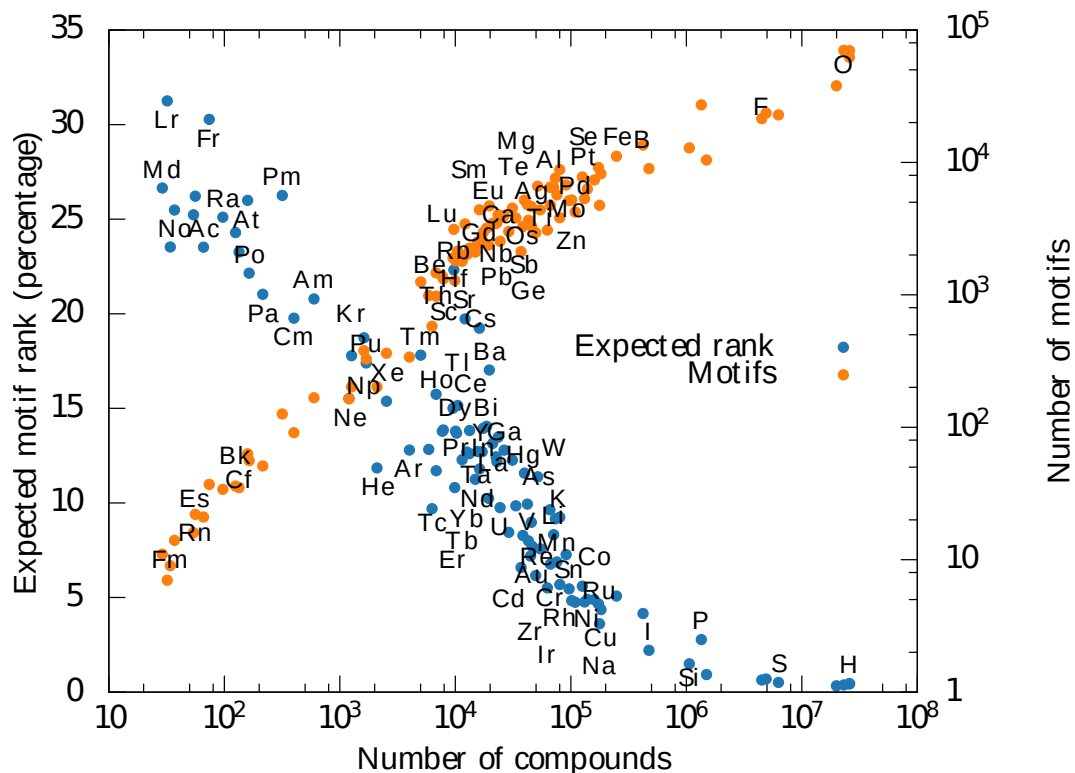

Figure 1 Diversity of combination motifs among elements. The number of motifs is directly proportional to the number of compounds (orange dots, right vertical axis). To determine how compounds are distributed over motifs, we ranked each element's motifs by the number of compounds they encompass, and calculated the expected motif rank (blue dots and left vertical axis, see **Methods**). A low value indicates that most compounds of the element belong to a small number of motifs. Note that this value is below 40% for all elements, indicating biased distributions of compounds over motifs. Also note that the expected rank decreases with the total number of compounds. For instance, elements such as C, N, O, conform tens of millions of compounds, but those compounds are concentrated in a very small percentage of motifs.
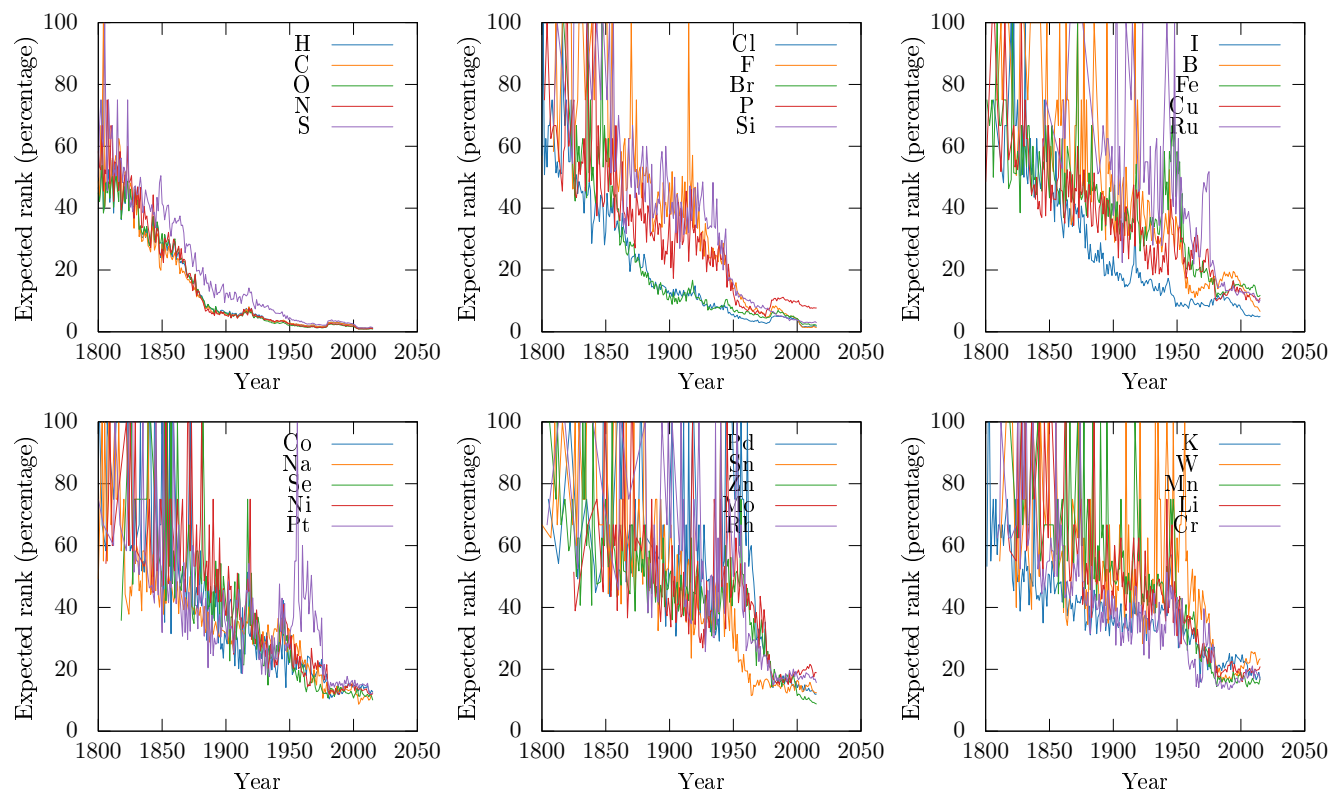
Figure 2 Growth of chemistry's bias towards favored combination motifs. We calculated the expected motif rank (see Methods) for each element in each year, taking into account only compounds reported on that year. The expected motif rank of the elements has become smaller and steadier as time passes: in early years the expected rank is high, and fluctuates widely; as time passes it becomes lower and the fluctuations diminish. Since around 1980 most of the elements occur in a relative small and steady set of motifs (see further details in Fig. SI-2 and Fig SI-3).

substituted for each other e.g. if they combine with the same other elements in the same ratio. For example, Na can be substituted by K in Na_2SO_4, yielding K_2SO_4. Because of the biases shown in the previous section, it is possible that an element A can substitute an element B in most of B's compounds, while B cannot substitute A in most of A's compounds. In that case B would be very similar to A but A would not be similar to B.

To measure similarity in this sense we use *Subsitution Formulae* (henceforth SF), e.g. X_2SO_4 is a SF of K, Na, and whatever other elements can take the place of X. We denote the set of all SFs of element e as $N(e)$. Note that an element has as many SFs as molecular formulae. SFs can be regarded as nodes in a bipartite graph where links exist between elements and their substitution formulae (Fig. 3). Since SFs conform the neighborhoods of elements in this graph, we also refer to them as *neighbors*. Common neighbors establish similarity relations between elements. When a neighbor is linked to just one element we say that it is *singular*. We measured the *singularity* of an element as the ratio of its singular neighbors to its total number of neighbors.
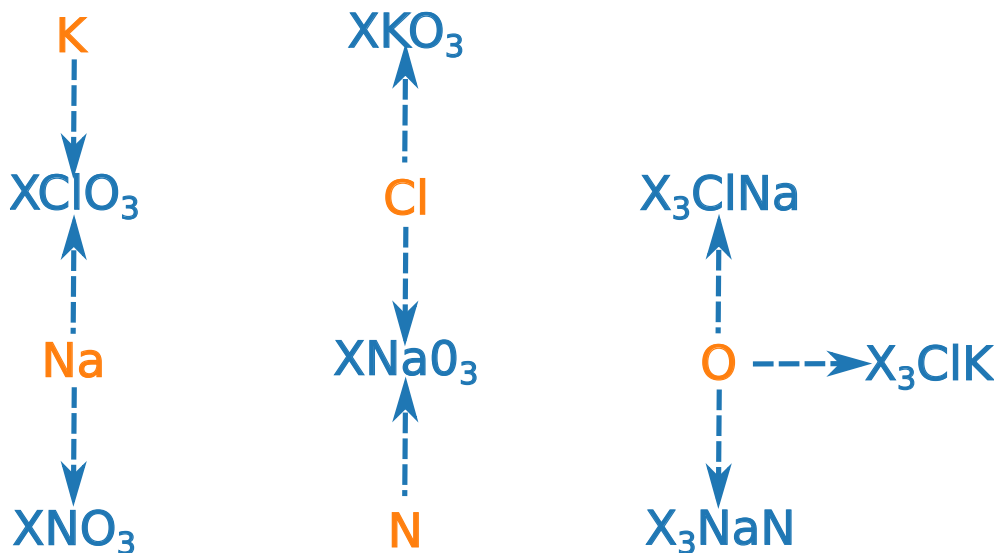

Figure 3 Illustrative example of Substitution Formulae (SFs) induced by the compounds $NaNO_3$, $NaClO_3$ and $KClO_3$. For each element a set of SFs is obtained as follows: Na $\to$ {XNO_3, $XClO_3$}, K $\to$ {$XClO_3$}, N $\to$ {$XNaO_3$}, Cl $\to$ { $XNaO_3$, XKO_3} and O $\to$ {X_3ClNa, X_3NNa, X_3ClK}. There is one common SF to Na and K ($XClO_3$), and one to Cl and N ($XNaO_3$). Oxygen is linked to three singular SFs (X_3ClNa, X_3NNa and X_3ClK) i.e. no other element is linked to them, while Na is linked to one singular SF (XNO_3).

Elements are more similar if they have more common neighbors; therefore, the size of the intersection of neighborhoods $|N(e_1) \cap N(e_2)|$ is a suitable measure of similarity between elements. Normalization with the size of the neighborhood of one of the elements results in the asymmetric similarity measure

$$S(e_1, e_2) = \frac{|N(e_1) \cap N(e_2)|}{|N(e_1)|} \tag{1}$$

By this definition $S(e_1, e_2) \neq S(e_2, e_1)$, so we are describing the similarity of one element *to* another rather than *between* elements. Note that $S(e_1, e_2) = 1$ if $N(e_1) \subset N(e_2)$, and $S(e_1, e_2) < 1$ otherwise; in particular, $S(e_1, e_2) < 1$ for $N(e_2) \subset N(e_1)$. Our measure captures asymmetry and allows to quantify differences of elements unevenly explored, even when they belong to the same family. We investigated deviations of chemical similarity from symmetry by plotting $S(e_1, e_2)$ vs $S(e_2, e_1)$. According to Fig. 4, some points lay on the diagonal (around the line $y = x$), which correspond to the small fraction of similarity relations that are approximately symmetric, e.g., Sn-Mo (0.035 and 0.037), Mn-W (0.031 and 0.031) and Zr-V (0.044 nd 0.044). Fig. 4 also shows that most similarity relations are asymmetric, e.g., Rn-Cl (0.92 and 2×10^2), Cd-Co (0.219 and 0.04) and Na-H (0.02 and 0.0006).

In general, similarity between pairs of elements is low; more than half of the couples are less than 10% similar in both directions (2708 out of 5252). These results show that it would be very difficult, at least statistically unsound, to build a predictive tool based solely on data about the stoichiometry of chemical known chemical compounds. Trying to predict the compounds one element might form relying on composition of molecules seems hopeless, for elements tend to be very different from each other.
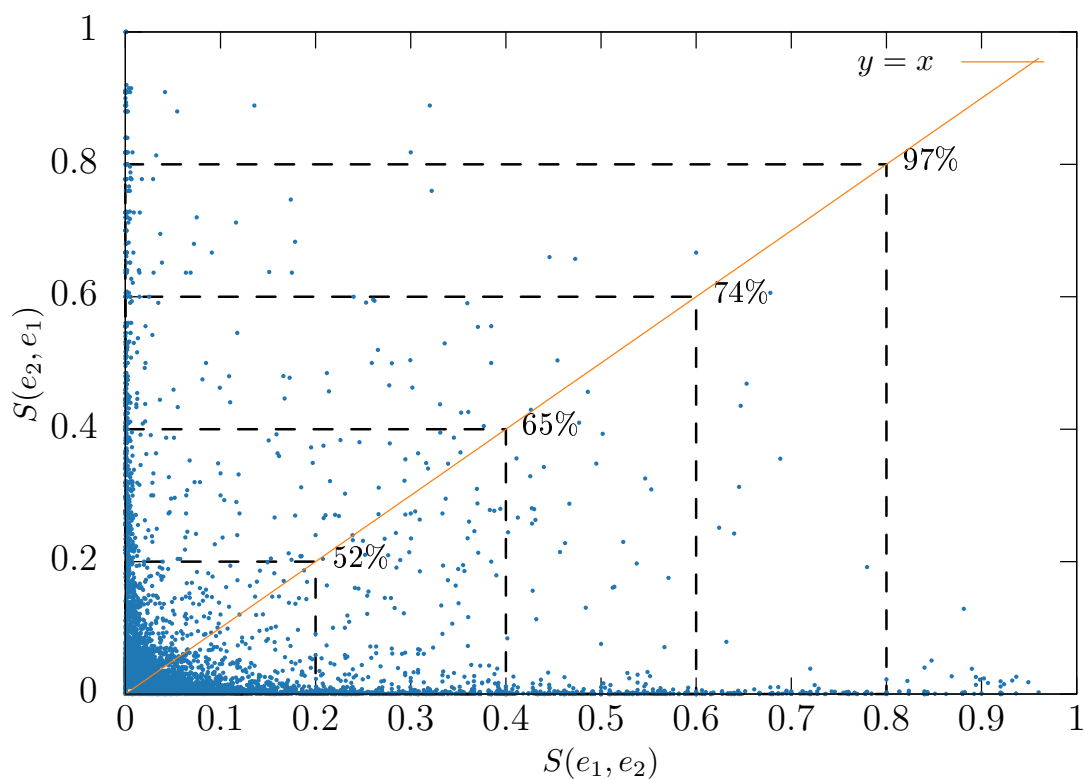
Figure 4 Comparison of similarity measures between elements in both directions $e_1 \rightarrow e_2$ and $e_2 \rightarrow e_1$. The asymmetry is clear from the plot, as there are few pairs close to the line $y = x$. The majority of close pairs to symmetry fall on low values of similarity, which is expected for elements moderately explored

2.3 Chemistry grows singular

Most of today's chemistry does not fit into any classification scheme based on stoichiometric patterns of molecular formulae, since unique combinations are the rule. 94.0% of the SFs are singular and encompass 77.9% of the compounds, and 42.2% of molecular formulae do not produce any similarity relation on any of its elements i.e. all their neighbors are singular.

However, singularity is unequally distributed among elements, so that very high singularity is rare: 34 elements have singularity > 0.5 and only the 4 main bio-elements (C, H, O, N) are above 0.8. These four elements alone are responsible for more than 75% of the singular SFs. This is expected in a compound based similarity measure, given the complexity of organic molecules. Typical singularity is much lower: the average is 0.394, the median 0.412. So there are, in fact, numerous similarity patterns on the SFs of most elements, but also numerous singularities. Elements having more formulae tend to be more singular (see Fig. SI-6), but there are many exceptions; even elements from the core such as I and Br may have a singularity around 0.5.

The past decades have seen an increase in the production of singular compounds, which has caused elements to grow more dissimilar among them (Fig. 5). In the early years of chemistry the bulk of elements had high singularity, as it is to be expected. Then, around 1860, their singularity started to decrease. This trend started to change after the 1950's, when the singularity of some elements (v.g. F, Mn) went into a steady increase for the first time in a century (Fig. 5). By the 1980's this phenomenon has spread to 75% of the elements, including almost all those that were discovered before 1860. Hence, in the last few decades chemists have been pushing the boundaries of the chemical space towards particular chemical contexts of each element that produce unique formulae. Around 75% of the elements are growing in singularity during this period. The resulting trend on the similarity of elements is that they are growing apart, all the similarity patterns are waning into a growing solitude.

These shifts in singularity coincide with previously reported shifts in chemistry's growth regime, arising from the analysis of the heteroscedasticity of the curve of production of new compounds[27]. The initial growth regime, named "proto-organic", is portrayed by its exploratory character, which is here again visible on the high frequency of formulae that are unique to elements during this period. The second regime, named "organic", covers the time period from 1860 to 1980; during this time, many new combinations were obtained and more similarity trends among the elements emerged,

which reflects in the drop of their singularity. The third regime, which is marked by the growth of compounds involving organic chains and metals, and hence was named "organometallic"[27], coincides with the increase in singularity for the majority of elements.
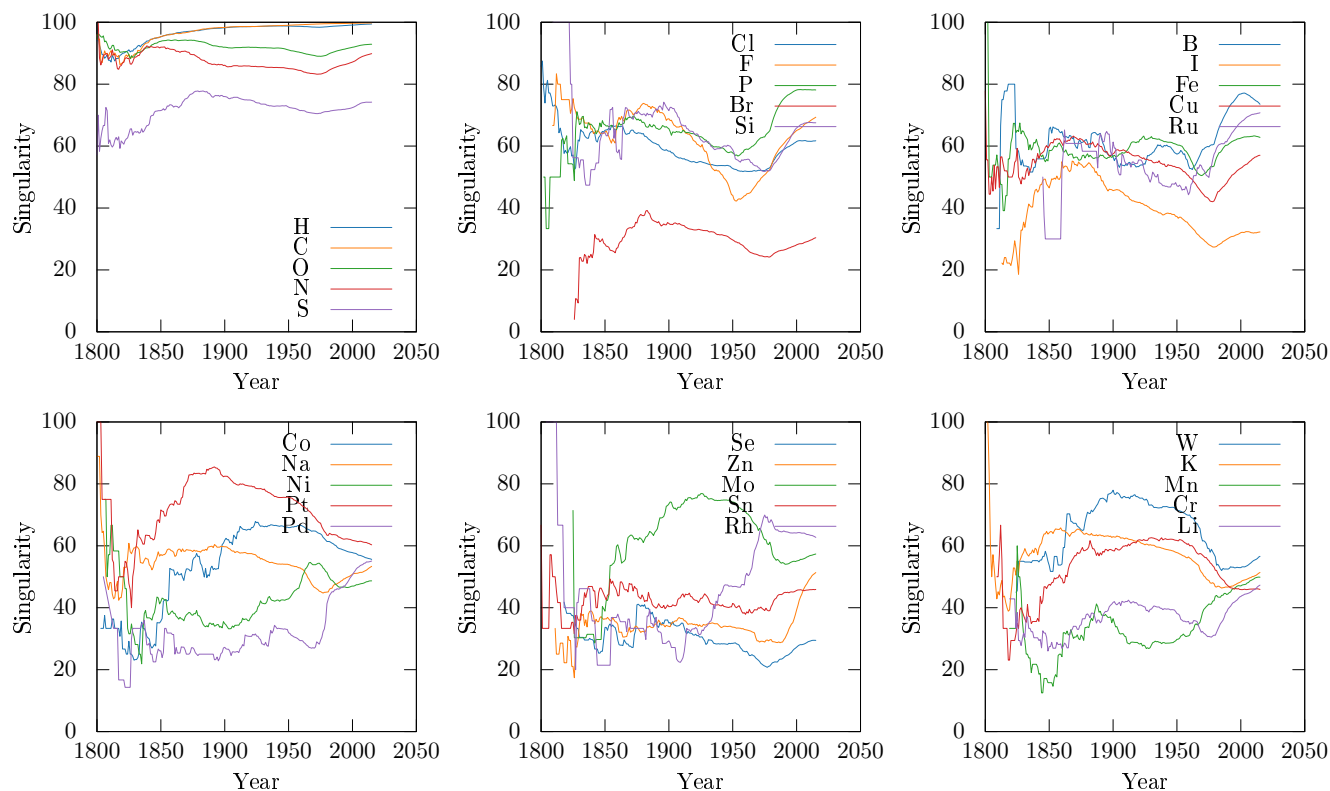

Figure 5 Singularity by year. Each element's singularity was computed considering all compounds published up to the corresponding year. Singularity decreases after 1860 on almost all elements and then started to grow after 1980 (see Fig. SI-7 and SI-8 for further details).

2.4 Similarity patterns match several PSCE families

Despite that elements tend to be different and singularity is growing with the discovery of new compounds, we found that couples of elements belonging to the same family fall among the most similar pairs. Members of the same chemical family (Intrafamily Similarity Relations – ISR) are counted among the most similar pairs ($e \to X$). In Fig. (6) each element is connected to its most similar one i.e. that element having the largest value in the similarity measure. 57% of the elements fall into ISR, and 71% have at least one member of the family among their four most similar elements. Nevertheless, several elements are connected to elements of different families. Many elements are most similar to Cl and S, therefore they appear as hubs in the graph.

Many ISR form chains or other connected subgraphs. For instance, Alkali metals without Li, In-Ga-Al-B, Bi-As-Sb-P, Alkali earths, Hf-Zr-Ti, Ge-Si-Sn-Pb, F-Cl-Br-I. Fragments of transition metal families are found, such as W-Mo, Nb-Ta, Pd-Pt and Rh-Ir. These couples are reciprocal relations, being the most similar in both directions. Horizontal similarities were found among third period transition metals: Mn-Co-Ni-Cu and Zn-Cu. Cu, Fe and Cr are connected to S. Lanthanoids are divided into two connected groups: Ce-La-Pr-Pm-Nd-Sm and Tb-Eu-Cm-Gd-Dy-Yb-Er. The only missing element is Lu, which is connected to Y, which in turn is connected to Gd. Noble gases form a connected subgraph Kr-Xe-Rn-Ar, without He which is connected to O, one of the biggest elements so far. Rn has just a few compounds and is connected to several elements (I, Au, Cu and Se). The only group that is formed in Actinoids is Np-Th-Pa-Pu-U. The remaining elements are connected to several other elements different from their family.

The main deviations from the PSCE come from elements with exceptionally many compounds, elements with exceptionally few compounds, and the third period of Transition metals. This behavior shows that ISR are indeed frequent, but not necessarily the most frequent in the chemical space.
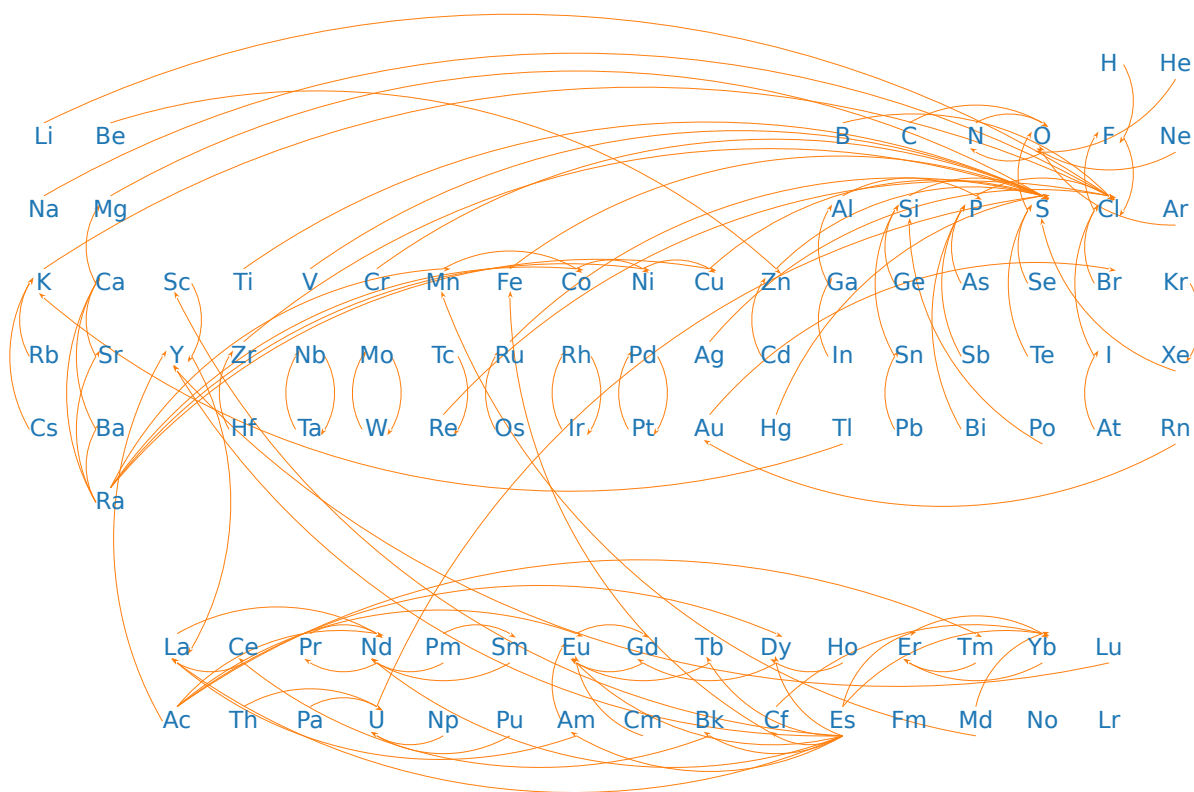
Figure 6 Network of the most similar element. Elements are nodes and the arrows bind each element with its most similar one $(e \to X)$. More than half (57%) of elements are similar to one element of the same family.

2.5 The ISR are ubiquitous in the chemical space

ISR are the most domain independent similarity patterns, occurring across the highest diversity of compounds and the largest range of molecular sizes. We found this through an exploration of how the similarity between couples of elements spreads over the chemical space, using three different measures:

1. The *Motif coverage* of a similarity relation $e_1 \to e_2$ is the ratio of the number of motifs of common neighbors of e_1 and e_2 to the total number of motifs of neighbors of e_1. In Fig. 3, for example, motifs NaClO and KClO add to the numerator of the motif coverage of K $\to$ Na, while NaClO and NaNO count towards its denominator.

2. The *Element coverage* of a similarity relation $e_1 \to e_2$, is the ratio of the number of elements in common neighbors of e_1 and e_2 to the total number of elements in neighbors of e_1. For example, in Fig. 3 Na and O are elements connecting Cl and N, while K is not.

3. The *Molecular size coverage* of a similarity relation $e_1 \to e_2$, was computed as follows: first we collected the set of sizes (number of atoms) of the neighbors of e_1, e.g. the neighbors XCl and XF are of size 2, while $XClO_3$ is of size 5, thus the set of sizes associated to the element X is $\{2,5\}$ and its cardinality is 2. Then, we count the molecular sizes of common neighbors of e_1 and e_2. Finally, we calculated the ratio of the number of molecular sizes of common neighbors to the number of molecular sizes of e_1. This ratio represents the number of molecular sizes of e_1 that are covered by the relation $e_1 \to e_2$ (see Methods).

ISR present higher values of these three measures than other similarity relations (see Fig. 7), showing that they are more scattered over the chemical space; that is, they tend to inhabit regions with more diverse compounds than other similarity relations. This is an interesting feature of ISR that can be understood as a manifestation of their fundamental nature: they are less tied to particular contexts than other similarity relations (see Fig. SI-9-SI-11, Fig. SI-12-SI-14 and Fig. SI-15-SI-17). If ISR were not fundamental, their occurrence would be contingent and would take place only on specific domains of compounds. But this is not the case for the majority of elements, revealing a special behavior of ISR that distinguishes them from other relations. Families such as Alkali metals, Halogens (excepting At), Ti-Zr-Hf, Cr-Mo-W and Lanthanoids clearly exhibit this behavior on all of their members. Many other families have some exceptions, but the
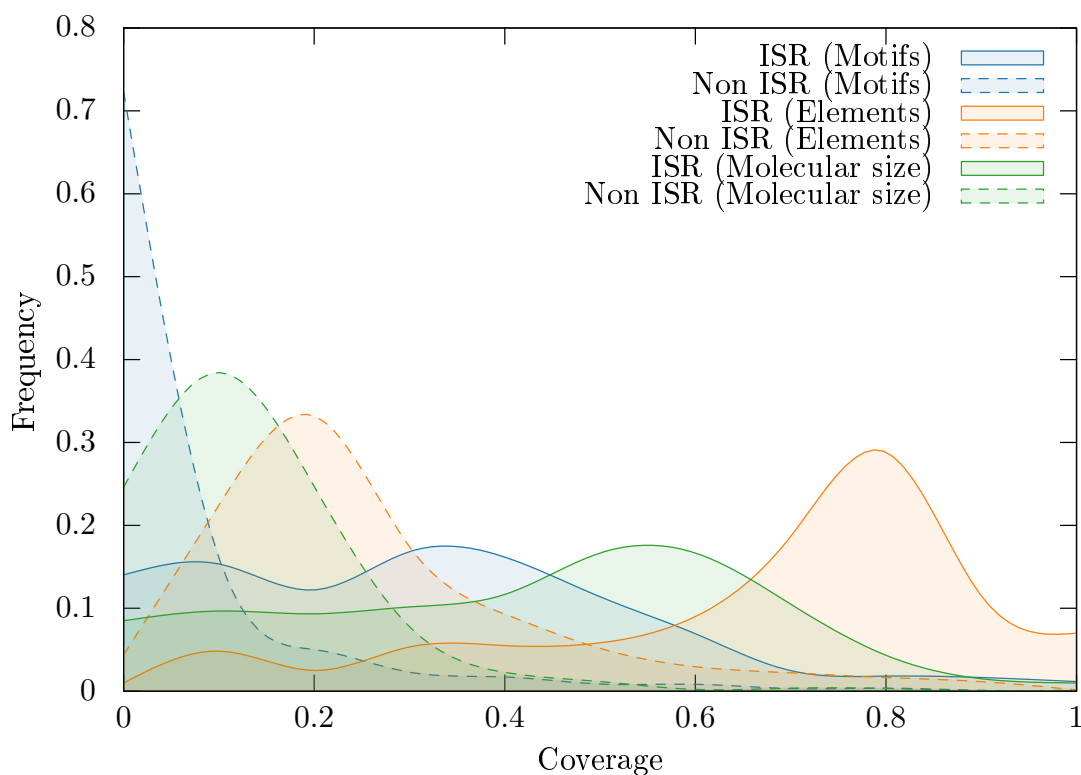
Figure 7 Distributions of coverage measures. ISR distributions correspond to continuous lines, while those of other relations are represented by dashed lines. The maximum frequency of ISR are found at higher values of the three coverage measures, relative to other relations.

trend is observed on almost all their members. Several elements have N (51), Cl (39) and S (38) among their four most scattered similarity relations, showing that ISR are not the only ones following this trend of independence on the chemical context. This behavior can be related to their high frequency among compounds and, hence, the high likelihood they have to form similarity relations with others elements outside of their families. Nevertheless, such relations are not reciprocal, for they have many neighbors that are not connected to other elements and, therefore, they are very dissimilar to other elements.

2.6 PSCE's families are resilient equivalence classes

In spite of the intrinsic asymmetry of the similarity relations between pairs of elements, chemical families form an interesting substructure: they approximate equivalence classes. Equivalence classes are a partition of a set by a symmetric and transitive relation, so elements that fall in the same class are considered equivalent through that relation. In this case, such structure is positional rather than metric, i.e. it is not about how similar two elements are (their similarity value), but if one is among the most similar elements of the other; so it becomes apparent when the similarity values are flattened into a ranking. Members of the same family of a any given element fall into the first positions of the most similar elements to it, so they form a symmetric and transitive relation. This unexpected result goes hand in hand with the traditional view of families where elements are divided into classes which are considered equivalent, and we found it emerges from an inherently asymmetric similarity relation. We first show how families approximate symmetric relations by calculating an asymmetry measure. Then we apply a hierarchical clustering algorithm that produce symmetric and transitive sets that match quite well families. For the sake of simplicity, in the following we use only the motif coverage measure to explore symmetry and transitivity of families, leaving out the other two coverage measures.

2.6.1 ISR ranking tends to be symmetrical

To analyze how asymmetrical these relations are, for each element e_1 we created a similarity ranking of the remaining elements ($r_{e_1}(e_2)$) and measured the asymmetry of the relation between e_1 and e_2 as the absolute value of the difference of the corresponding ranks ($\text{asym}(e_1, e_2) = |r_{e_1}(e_2) - r_{e_2}(e_1)|$). In Fig. 8 this asymmetry value is plotted against the maximum rank for every couple of elements. The values are scattered quite uniformly over the set of possibilities, although there is a small concentration at small values of asymmetry. ISR tend to concentrate at low values of asymmetry and low values of the ranking; around 65% of pairs of elements in the same family have asymmetry values below 8. These couples form

an orange spot at the bottom of the plot, showing these relations are relatively intense and reciprocal. So if one element is found very similar to another (the latter has a low value in the ranking of the former), it is highly likely the latter will also be very similar to the first one (the former also has a low value in the ranking of the latter).
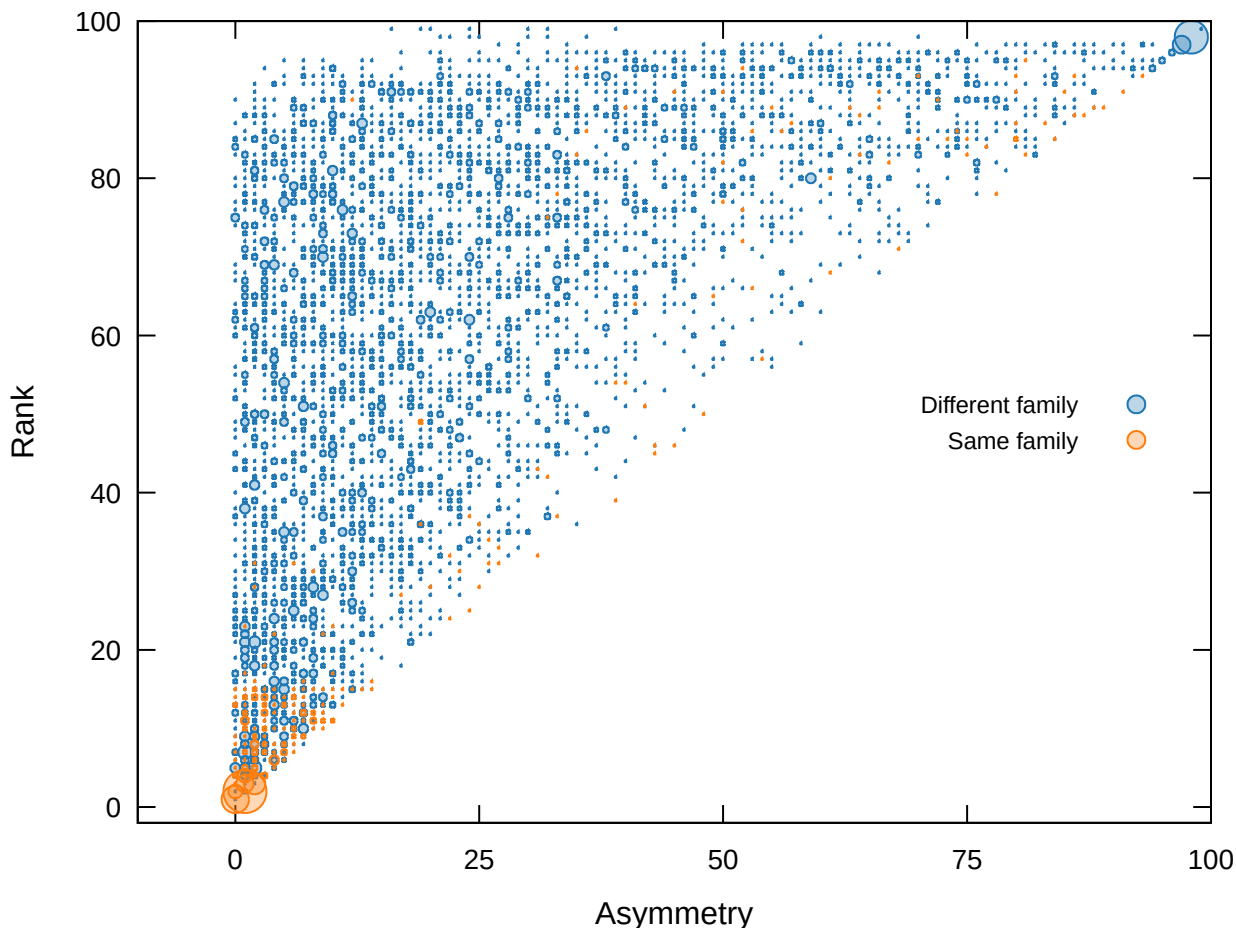

Figure 8 Rank asymmetry vs Rank value. Orange circles correspond to elements of the same family, while blue circles correspond to elements of different families. Circle radius represents the number of pairs that fall into the same value. Low values of asymmetry are related to couples on low ranking values (most similar). Asymmetry values are calculated through $\mathrm{asym}(e_1, e_2) = |r_{e_1}(e_2) - r_{e_2}(e_1)|$. The rank value on y-axis corresponds to $\max\{r_{e_1}(e_2), r_{e_2}(e_1)\}$.

2.7 PSCE's families form transitive clusters

Besides symmetry, we found the similarity relation between elements of the same family is also transitive. We obtained this by using a hierarchical clustering algorithm that captures both symmetry and transitivity into a simple method. The method is agglomerative, hence starts with the set of all elements and gradually forms clusters joining two nodes at minimum distance on each iteration. We defined the distance between two elements as the maximum rank value of each element on the ranking of the other one, $d(e_1, e_2) = \max\{r_{e_1}(e_2), r_{e_2}(e_1)\}$. The distance between two clusters is simply the maximum distance of every pair of elements in the two clusters $(d(C_1, C_2) = \max\{d(e_1, e_2) : e_1 \in C_1 \ \& \ C_2 \in C_2\})$. This clustering method is usually referred to as *complete linkage*. So, if one cluster C is formed by this method at the ranking position R, then the distance between any pair of elements in the cluster is at most R. We can define a relation between elements if they are separated by a distance less than a given value R. Clearly this relation is symmetrical. We denote this relation as $e_1 \sim_R e_2$. This relation is also transitive in the cluster C, that is $e_1 \sim_R e_2$ and $e_2 \sim_R e_3$, then $e_1 \sim_R e_3$, since $d(e_1, e_3)$ is also less than R.

Several clusters and the chemical families match quite well. Around 56% of the clusters obtained are formed by elements of the same family. Some other clusters are just the union of these ones. In Table 1 we report the biggest clusters that only contain elements of the same family (Family Clusters - FC). The majority of elements (78.4%) fall into some FC.

Five of the clusters obtained are complete chemical families: Alkali metals (excepting Fr), Ti-Hf-Zr, V-Ta-Nb, Cr-Mo-W, Cu-Ag-Au. The remaining families are partially found. In general the missing elements have either a high or a low

Table 1 Family Clusters. Left column contains the elements in the families and right column the Family Clusters formed by our agglomerative method. The majority of elements fall into FC (78.4%).

Family	Family Clusters
Li, Na, K, Rb, Cs	{Li, Na, K, Rb, Cs}
Be, Mg, Ca, Sr, Ba	{Be, Mg, Ba, Ca, Sr}
Ti, Hf, Zr	{Ti, Hf, Zr}
V, Ta, Nb	{V, Ta, Nb}
Cr, Mo, W	{Cr, Mo, W}
Mn, Tc, Re	{Tc, Re}
Fe, Os, Ru	{Os, Ru}
Co, Ir, Rh	{Ir, Rh}
Ni, Pd, Pt	{Pd, Pt}
Cu, Ag, Au	{Cu, Ag, Au}
Zn, Cd, Hg	{Cd, Hg}
B, Al In, Ga, Tl	{Al, Ga, In}
C, Si, Ge, Sn, Pb	{Si, Ge, Sn}
N, P, As, Sb, Bi	{P, As, Sb}
O, S, Se, Te, Po	{S, Se, Te}
H, F, Cl, Br, I, At	{I, Br, Cl}, {F H}
Rn, Xe, Kr, Ar, Ne, He	{He Ne}, {Ar Kr Xe}
Sc, Y, La, Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu	{Sc, Y, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu}
{Ac, Th, Pa, U, Np, Pu, Am, Cm, Bk, Cf, Es, Fm, Md, No, Lr}	{Th, U, Np, Pu}, {Fm, Md}, {Cf, Bk}, {Cm, Am}

number of compounds. Alkali earth metals are missing Ra, an element with just 24 compounds. The same happens on Lanthanoids, where Pm (which has only 317 compounds) is missing. These deviations are expected for elements having a handful of compounds, since they are more affected by noise and therefore prone to misclassifications. This is the case for Actinoids, where several elements have less than a hundred compounds and five elements are missing. Elements with many compounds also tend to separate from their families. This is true in many families on the transition metals, where 5 out of 9 have one missing element. These elements are form the lowest period in the Periodic Table, and is consistent with the singularity principle, which states the elements on first periods tend to separate from the others. A combination of rare elements and the singularity principle seems to be dividing the remaining groups, as in the case of Carbon group, where C and Pb are missing, as well as Nitrogen and Oxygen groups. Halogens are clustered together but H is closer to F than to any other element.

In Fig. 9 we visualize the ranking values of elements of the same family. To facilitate the comparison of different families, we normalized these values by number of elements in the family minus one. The elements of two families have all their normalized values below one: Alkalines (without Fr) and Wolframium group. Elements having lower numbers of compounds tend to have high ranks. This is the case of Ra, Po, At and Pm as well as almost all Actinoids. Lanthanoids are a very strong group, including Y and Sc, since almost every element of the family has normalized ranks below one. Halogens including Hydrogen but excluding At (having few compounds) are also a very strong group. There has been a small debate around the position of H on the PSCE ([34]). Our results favor Hydrogen at the head of Halogens, since it is positionally closer to them, and they are closer to it, than to any other element. However, Hydrogen is a very singular element, and there is no other element capable of binding to Carbon to produce such a wealthy variety of compounds in organic chemistry. In this sense, Hydrogen could be considered a group on its own, but positionally is very close to Halogens, hence we considered it as part of this family, even though it is usually counted as the head of Alkali metals.

Th falls into the same cluster as U, Np and Pu, forming a subgroup in the family of Actinoids. Quantum calculations have established the ground state for Th as $7s^{26d\ 2}$ [35], but it is considered an f element. Thorium coordination chemistry has applications on nuclear fuel industry, where it is used in sequestering Actinoids ions or as a surrogate of Pu [36], and it has applications on catalysis as an analogous of Pu and heavier Actinoids [37]. So, is not surprising that it shares several neighbors with these elements.

Symmetry and transitivity are characteristic of the FC. Other relations are either more asymmetrical or transitive at

greater values of the rank. Families are a sub-relation embedded into the asymmetrical similarity relation previously defined, they approximate an equivalence relation, breaking the inherent asymmetry of the original similarity measure. Since they tend to occupy the top positions in similarity rankings for each element, their symmetry and transitivity are easily met at low values of these rankings. So, the underlying structure of the similarity described by PSCE is an equivalence relation that partitions the set of elements into non intersecting classes. However, there are several elements that form clusters with elements of other families. There is a cluster formed by Ni, Co, Zn, Mn and Fe, all of them falling into the third period. This is the only region found where elements have such strong horizontal similarity relations instead of vertical, something that has been noted before and explained as the result of the stability for medium oxidation numbers (+3 and ±2) that leads to similar specific chemical properties and makes difficult to devise reliable group assignments[38]. Other interesting change from the usual Periodic Table is the position of Thallium, which falls into Alkali metals, a result that was recognized by Mendeleev himself, due to the stability of its oxidation state +1. This behavior of Thallium is related to inorganic compounds, since organic compounds tend to move it to its original family, as will be seen further below.
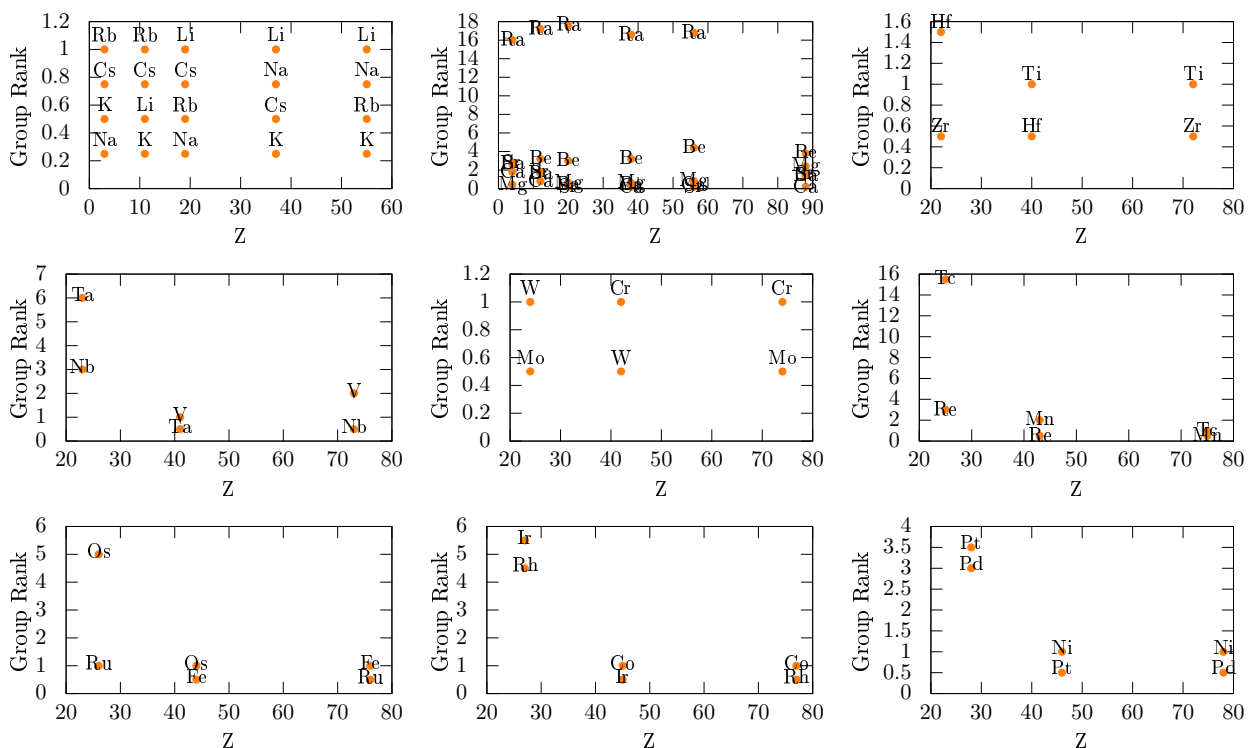

Figure 9 Family rankings. The abscissa represents the nuclear charge of each element. The ordinate axis gives the normalized rank of other elements. Values above 1 indicate dissimilarity (see further details in Fig. SI-18)

2.7.1 Family Clusters are highly resilient

FCs are also highly independent of the set of compounds used to calculate similarity relations. To test stability of FC, we carried out a series of experiments to obtain dendrograms using as input subsets of randomly selected compounds. We calculated dendrograms since it makes sense provided that the families match a relation that is both symmetrical and transitive. A set of 100 replicas of a fixed percentage of compounds were generated, on intervals from 90% to 10% of the total number of compounds. We also challenge the clusters found by removing compounds from the top motifs to test their stability against changes in the chemical context used to build the classification.

1. *FC are highly independent on the input set of compounds*. The similarity clusters formed by choosing randomly sets of input compounds match quite well the PSCE; even considering just 10% of the compounds the number of elements that fall into FC is above 60% (see Fig. 10 (a)), with a median above 72%. Such resilience of periodic patterns explains why they are frequently found even on small datasets, as the one available on the XIX Century. ISR are found on different kinds of compounds, they are ubiquitous and highly frequent. Periodic patterns are not constrained into specific sets of compounds, their fundamental nature is manifested on almost all domains. Mendeleev used these similarity patterns to predict elements yet undiscovered and the history of chemistry, as we see, has proven him right
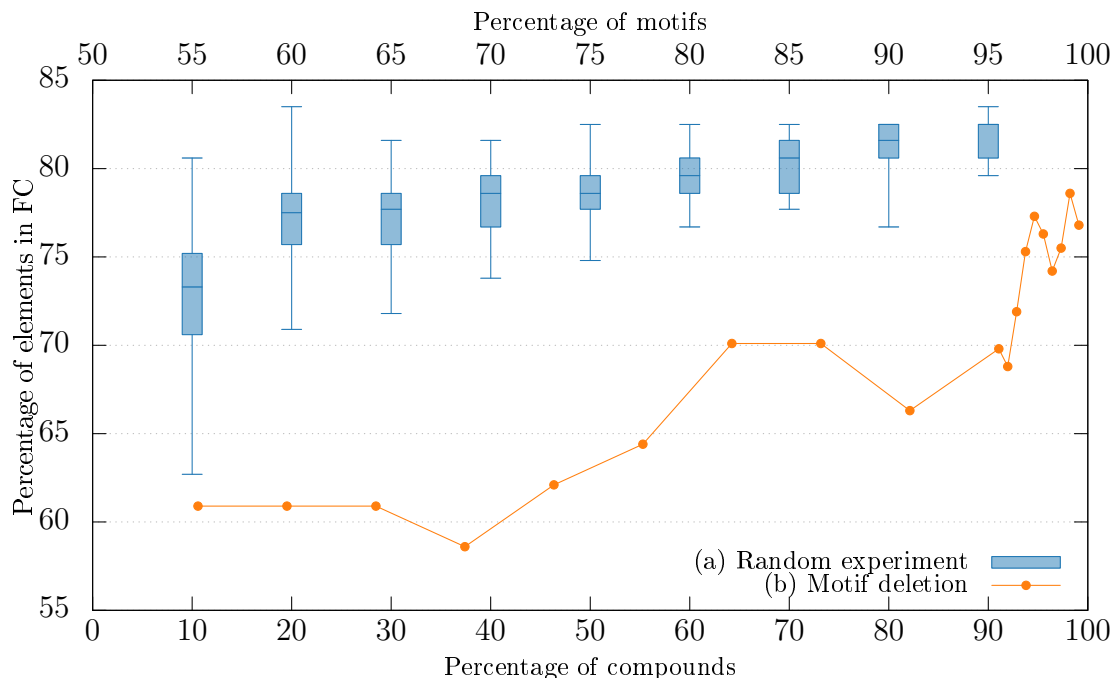
on domains beyond his reach.

Figure 10 (a) Percentage of elements in Family Clusters. Distribution of percentage of elements falling in Family Clusters, using 100 replicas of a fixed number of randomly selected input compounds. Even on small number of compounds (10% of the total Chemical Space) the number of elements that fall into Family Clusters is relatively high (above 60%). The spread decreases smoothly when more compounds are taken into account.

There are six patterns that occurred in all randomization experiments, and several others that occurred in the vast majority of experiments (see SI-1). Several couples of transition metals are frequently clustered together, as well as a subgroup of Halogens, and pairs of Alkalines and Alkali earths. Lanthanoids were a group frequently found (89.7%), including Sc and Y, which should be considered part of this family. Some subgroups of non metals are also frequent, Ge, Si and Sn (93.9%) as well as Al, Ga and In (90.9%). Nonetheless, there are several elements that clustered frequently with members of different families, such as Tl, which clustered with Alkalines (79.1%), Fe and Mn (86.7%), N and O (83.0%), and F and H (90.2%). These deviations to the expected periodic families show how intrincated are similarity relations among elements and the variability on intensity of the periodic law. Thalium is frequently clustered with Alkalines (79.1%), as previously found.

A few elements are not part of any frequent family cluster. Zinc, for example, is clustered several times with Mg, another element abscent from the table, but it is sometimes also grouped to {Co,Ni,Cu}. Zn and Mg share several formulae since both have extensive use on organic chemistry and several of their complexes have been compared as analogous[39–42]. The cluster {C, N, O} can be thought of as a group of outliers because of the high number of compounds these elements have, so there is a good chance they share neighbors and, therefore, they form a group separated from the remaining elements. Besides these clusters, no other clusters are as resilient as element families, for they occur on specific domains, on particular sets of chemical contexts, and hence their low frequency.

2. *FC are resilient to directed attacks*. We found that several families are resilient to deletion of compounds that have certain motifs. We performed a series of experiments by disregarding compounds from the top most frequent motifs, going from 1000th-100000th on steps of 1,000 or 10,000 motifs. The motifs are selectively disregarded, in such a way that the attack is directed to those chemical contexts having more compounds. The first 1000 motifs account for 73.9% of chemical formulae, so our experiments start with 26.1% of chemical formulae and end with 1.0% of them. In Fig. 10 (b) the percentage of elements that fall into FC is shown. Excepting one case, the percentage is above 60%, which is high when one considers the small number of compounds used to classify the elements (0.2% of all compounds after deleting 100,000 motifs). Several families prevail on the majority of experiments. Alkaline metals, many couples from Transition metals, some non metallic families and Halogens occur as clusters in the majority of experiments.

13

2.8 Inorganic *versus* organic compounds

Traditionally the PSCE is introduced in courses of inorganic chemistry, using several examples of compounds coming from the inorganic domain, such as oxides, salts, halides, and so forth. To test the influence on the similarity patterns found on both inorganic and organic chemistry, we obtained the FC using compounds with no Carbon on their formulae (inorganic domain) and compounds having Carbon and Hydrogen on their formulae (organic domain). In Table 2 we present the FC found on both sets of compounds.

Table 2 Family Clusters from organic and inorganic domains.

Inorganic	Organic
{Cs K Li Na Rb}	{Cs Rb} {K Li Na}
{Ba Ca Sr}	{Ba Be Ca Sr}
{Hf Ti Zr}	{Hf Ti Zr}
{Nb Ta}	{Nb Ta V}
{Mo W}	{Cr Mo W}
{Re Tc}	{Re Tc}
{Os Ru}	{Fe Ru}
{Ir Rh}	{Ir Rh}
{Pd Pt}	{Pd Pt}
{Ag Cu}	{Ag Au Cu}
-	{Cd Hg}
{Al Ga In}	{Al Ga In Tl}
{Ge Si Sn}	{Ge Si Sn}
{As N P} {Bi Sb}	{As Bi P Sb}
{O S Se Te}	{S Se Te}
{Br Cl F H I}	{Br Cl I}
{Ar He Kr Ne Xe}	{He Ne} {Kr Xe}
{Ce Dy Er Eu Gd Ho La Lu Nd Pr Sc Sm Tb Tm Y Yb}	{Ce Dy Er Eu Gd Ho La Lu Nd Pr Sc Sm Tb Tm Y Yb}
{Np Pu Th U} {Ac Es Fm Lr Md No} {Bk Cf Pa} {Am Cm}	{Np Pu Th U} {Cf Es} {Ac Lr} {Fm Md} {Am Cm}

79.4% of elements fall into FC in the inorganic domain, whilst 80% fall into FC in the organic domain, so PSCE is found on both domains. The change on the domain has little impact on the clusters formed. This is a consequence of the family resilience to the input set of compounds, so the FC do not differ to a large extent. Some changes occur, though. Tl on the inorganic domain tends to be closer to Alkalines than to its group, but on the organic domain this situation is changed. Oxygen forms a cluster with its own family on the inorganic domain, but moves apart on the organic domain.

2.9 Family Clusters are growing

The number of motifs where ISR are found is growing with time. We found a better match on the present than in the past. Elements are growing apart, becoming more dissimilar in the fraction of matching compounds, but our finding of a greater motif coverage of ISR is making more elements fall into FC. In Fig. 11 the percentage of elements that are part of ISR is shown. For each year, we calculated a dendrogram, using all compounds up to that year as input, and extracted FCs. Since elements grow on particularly frequent motifs, they are growing apart on the fraction of compounds that are common to every pair. But, since the number of motifs where ISR occur is growing everyday, the match between PSCE families and the clusters found is increasing. From almost 40% 1860 to almost 80% in 2015, the more compounds we explored, the better the match.

3 Conclusions

Our exploration of similarities among elements has shown that chemical families have several features that agree with their fundamental character. They have no particular domain to occur, they are ubiquitous over the chemical space, not being restricted to particular chemical contexts, nor constrained to specific domains of number of atoms, but they are highly independent of the nature of compounds. Moreover, they tend to be highly frequent on all chemical contexts, they are among the strongest similarity patterns. These features provide them with a distinctive behavior that makes them to stand out over any other pattern, so they prevail over others and are found even on small datasets. Such behavior
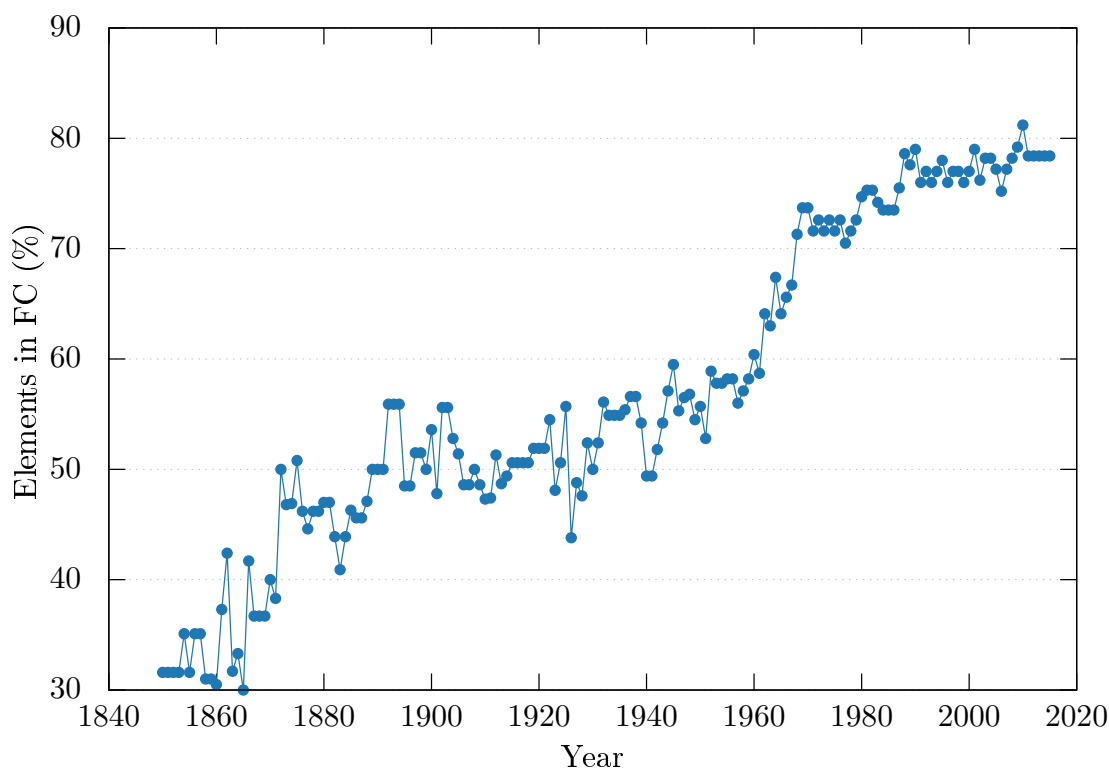
Figure 11 Percentage of elements falling in FC over the year. The match between the calculated clusters and the PSCE families is growing with time.

agrees with the fundamental character of the periodic table, since their similarity goes beyond any particular domain, and explains why they are found on many different studies. Our analysis shows that any study based on compounds will reach the same conclusions, provided the number of chemical contexts is wealthy enough on the nature of chemical compounds, whenever the starting set of compounds is unbiased, they will prevail over any other.

Moreover, we found that chemical families have a particular interesting structure: they are equivalence classes. Despite the similarity relation among elements is asymmetrical in general, and we captured this behavior into our similarity measure, families form a symmetric and transitive subrelation embedded into this asymmetric relation.

These features show that element families of the PT not only have remained recognizable despite the exponential growth of the number of compounds, but the match between data and families has improved over time. Our results also show that this match is going to be that way independently on the path chemistry can take in the future, as long as it remains unbiased. Nevertheless, we found that elements grow on particular domains of combinations, and their singularity is emphasized when more compounds are obtained. Such trend indicates elements are becoming more and more different from others, hence their similarity is waning. This behavior became much clearer since the 80's, a particular time where most of elements stabilized on their yearly growth. Despite this behavior, element families of the PT are becoming better defined.

4 Data and Methods

4.1 Data

Data for this study comes from Reaxys database, downloaded on January 2017. We included all 26,234,107 stoichiometric compounds with recorded molecular formula. We excluded those compounds with non-definite molecular formula, such as several polymers containing an undetermined number of monomers. These compounds comprise 4,330.224 molecular formulae, which were used to calculate similarity relations. The calculations of neighbors, similarity relations and so on, were based on the field 'Element Counts'. For time analyses, we used a smaller set of compounds (20,649,870) associated with reactions with reported publication year. The earliest date of publication was taken as their date of discovery.

4.2 Expected rank

To measure how the distribution of compounds is spread over motifs we used the following definition of Normalized Motif Expected Rank. Let $M(e)$ be the set of motifs of the element e. Every motif $m \in M(e)$ has an associated fraction $f_e(m)$ of compounds having that combination motif. Sort $M(e)$ by f_e in descending order and let $r_e(m)$ be the rank of motif m in the ordered set. The NEMR of element e is:

$$\bar{r}_e = \sum_{m \in M(e)} \frac{f_e(m) \cdot r_e(m)}{|M(e)|} \tag{2}$$

4.3 Coverage measures

Coverage measures are used to analyze how the similarity relation $e_1 \rightarrow e_2$ is scattered over the chemical space of e_1. We defined three different coverage measures: 1) motif coverage ($C_M(e_1 \rightarrow e_2)$), 2) element coverage ($C_E(e_1 \rightarrow e_2)$) and 3) molecular size coverage ($C_S(e_1 \rightarrow e_2)$). The three measures are defined as a ratio of number of motifs, number of elements and number of molecular sizes where $e_1 \rightarrow e_2$ is found, and the total number of motifs, elements and molecular sizes of e_2, respectively.

Let be $M(e_1)$ the set of motifs of the element e_1, and $M(e_1 \rightarrow e_2)$ be the set of motifs of e_1 where $e_1 \rightarrow e_2$ is found. The motif coverage is simply the ratio:

$$C_M(e_1 \rightarrow e_2) = \frac{|M(e_1 \rightarrow e_2)|}{|M(e_1)|}$$

Let be $E(e_1)$ the set of elements in the neighbors of $e1$, and $E(e_1 \rightarrow e_2)$ the set of elements in the neighbors where $e_1 \rightarrow e_2$ is found. The element coverage is the ratio:

$$C_E(e_1 \rightarrow e_2) = \frac{|E(e_1 \rightarrow e_2)|}{|E(e_1)|}$$

Let be $S(e_1)$ the set of molecular sizes of neighbors of e_1 (number of atoms in compounds of e_1), and $S(e_1 \rightarrow e_2)$ the set of molecular sizes where the relation $e_1 \rightarrow e_2$ is found. The molecular size coverage is the ratio:

$$C_S(e_1 \rightarrow e_2) = \frac{|S(e_1 \rightarrow e_2)|}{|S(e_1)|}$$

4.4 Families

Family	Elements	Family	Elements
1	Li,Na,K,Rb,Cs	11	Zn,Cd,Hg
2	Be,Mg,Ca,Sr,Ba,Ra	12	B,Al,Ga,In,Tl
3	Ti,Zr,Hf	13	C,Si,Ge,Sn,Pb
4	V,Nb,Ta	14	N,P,As,Sb,Bi
5	Cr,Mo,W	15	O,S,Se,Te,Po
6	Mn,Tc,Re	16	F,Cl,Br,I,At,H
7	Fe,Ru,Os	17	He,Ne,Ar,Kr,Xe,Rn
8	Co,Rh,Ir	18	Sc, Y, La, Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu
9	Ni,Pd,Pt	19	Ac, Th, Pa, U, Np, Pu, Am, Cm, Bk, Cf, Es, Fm, Md, No, Lr
10	Cu,Ag,Au		

Table 3 Families in the Periodic Table. We consider H as part of Halogens instead of Alkali metals as usual.

4.5 Clustering method

Our clustering method is described in algorithm 1.

Author Contributions

Result: dendrogram
Input: $e_1, e_2, e_3, ..., e_n$
Output: Dendrogram
Function dendrogram**(clusterList)**

   if *only one cluster* **then**
      return clusterList;
   end
   else
      for $C_1, C_2 \in clusterList$ **do**
         $distance = \infty$;
         for $e_1, e_2 \in C_1 \cup C_2$ **do**
            if $d_{12} < \max\{r_{e_1}(e_2), r_{e_2}(e_1)\}$ **then**
               $d_{12} = \max\{r_{e_1}(e_2), r_{e_2}(e_1)\}$;
            end
         end
         if $d_{12} < distance$ **then**
            $distance = d_{12}$
         end
      end
      clusterList = merge clusters at minimum distance;
      clusterList = delete clusters from clusterList;
      return dendrogram(clusterList);
   end

Algorithm 1: Algorithm to cluster elements.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Notes and references

[1] B. Bensaude-Vincent, *The British Journal for the History of Science*, 1986, **19**, 3 – 17.

[2] E. Scerri, *The periodic table: its story and its significance*, Oxford University Press, 2019.

[3] D. Mendeleev, *On the relation of the properties to the atomic weights of the elements*, Dover, New York, 2002.

[4] L. Meyer, *Justus Liebigs Annalen der Chemie*, 1870, **Supplementary volume VII**, 354—-364.

[5] G. Boeck and A. J. Rocke, in *Lothar Meyer, Moderne Theorien der Chemie, vollständiger Text der ersten Auflage (1864)*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2022, pp. 49–155.

[6] D. Mendeleev, *Principles of Chemistry*, P. F. Collier and Son, 1901.

[7] R. G. Egdell and E. Bruton, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2020, **378**, 20190302.

[8] C. Boyce, *HYLE–International Journal for Philosophy of Chemistry*, 2019, **25**, 49 – 70.

[9] H. Kragh, *Substantia*, 2017, **1**, 7–17.

[10] S. G. M. Sereno, *Foundations of Chemistry*, 2020, **22**, 477–488.

[11] D. Reznick, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2020, **378**, 20190309.

[12] Y. Maeno, K. Hagino and T. Ishiguro, *Foundations of Chemistry*, 2020, **23**, 201–214.

[13] A. Brito, M. A. Rodríguez and M. Niaz, *Journal of Research in Science Teaching*, 2004, **42**, 84–111.

[14] E. Scerri, *Fundations of chemistry*, 2022, 285–300.

[15] N. S. Imyanitov, *Foundations of Chemistry*, 2016, **18**, 153–173.

[16] N. S. Imyanitov, *Russian Journal of Inorganic Chemistry*, 2011, **56**, 2183–2200.

[17] E. G. Marks and J. A. Marks, *Foundations of Chemistry*, 2010, **12**, 85–93.

[18] Y.-L. Wang, H.-S. Hu, W.-L. Li, F. Wei and J. Li, *Journal of the American Chemical Society*, 2016, **138**, 1126–1129.

[19] P. Pyykkö, *Physical Chemistry Chemical Physics*, 2011, **13**, 161–168.

[20] J. Mewes and P. Schwerdtfeger, *Angewandte Chemie International Edition*, 2021, **60**, 7703–7709.

[21] J. Vícha, S. Komorovsky, M. Repisky, R. Marek and M. Straka, *Journal of Chemical Theory and Computation*, 2018, **14**, 3025–3039.

[22] P. Pyykkö, *Chemical Reviews*, 2011, **112**, 371–384.

[23] N. C. Pyper, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2020, **378**, 20190305.

[24] E. Weber, M. Lefevere and K. G. Barman, *Foundations of Chemistry*, 2021, **23**, 407–429.

[25] W. H. E. Schwarz, U. Müller and F. Kraus, *Zeitschrift für anorganische und allgemeine Chemie*, 2022, **648**, e202200008.

[26] W. Leal, E. J. Llanos, A. Bernal, P. F. Stadler, J. Jost and G. Restrepo, *Proceedings of the National Academy of Sciences*, 2022, **119**, e2119083119.

[27] E. J. Llanos, W. Leal, D. H. Luu, J. Jost, P. F. Stadler and G. Restrepo, *Proceedings of the National Academy of Sciences*, 2019, **116**, 12660–12665.

[28] J. Schummer, *Scientometrics*, 1997, **39**, 107–123.

[29] W. Leal, G. Restrepo and A. Bernal, *MATCH Communications in Mathematical and in Computer Chemistry*, 2012, **68**, 417–442.

[30] R. Liu, G. Mao and N. Zhang, *Foundations of Chemistry*, 2018, **21**, 193–206.

[31] D. N. Georgiou and T. E. Karakasidis, *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 2004, **8**, 231–236.

[32] M. R. Lemes and A. D. Pino, *Journal of Chemical Education*, 2011, **88**, 1511–1514.

[33] J. T. Machado and A. M. Lopes, *Communications in Nonlinear Science and Numerical Simulation*, 2019, **78**, 104883.

[34] V. M. Petruševski and J. Cvetković, *Contributions, Section of Natural, Mathematical and Biotechnical Sciences*, 2017, **38**, 83.

[35] P. Pyykkö, *EPJ Web of Conferences*, 2016, **131**, 01001.

[36] C. D. Tutson and A. E. Gorden, *Coordination Chemistry Reviews*, 2017, **333**, 27–43.

[37] K. Krauskopf, *Chemical Geology*, 1986, **55**, 323–335.

[38] S.-G. Wang and W. H. E. Schwarz, *Angewandte Chemie International Edition*, 2009, **48**, 3404–3415.

[39] A. Friedrich, J. Eyselein, J. Langer and S. Harder, *Organometallics*, 2021, **40**, 448–457.

[40] M. H. Chisholm, J. Gallucci and K. Phomphrai, *Chemical Communications*, 2003, 48–49.

[41] N. Nimitsiriwat, V. C. Gibson, E. L. Marshall, P. Takolpuckdee, A. K. Tomov, A. J. P. White, D. J. Williams, M. R. J. Elsegood and S. H. Dale, *Inorganic Chemistry*, 2007, **46**, 9988–9997.

[42] A. Erxleben and D. Schumacher, *European Journal of Inorganic Chemistry*, 2001, **2001**, 3039–3046.