A computational approach to the periodic system.

Andrés Marulanda-Bran*
Universidad de Antioquia
acamilo.marulanda@udea.edu.co

May 2, 2021

Abstract

1. Introduction

Mendeleev's periodic system is, and for the last 150 years has been, recognised as one of the most important icons of chemistry and of all of the natural sciences, both by the general public and the scientific community itself. Such a system has been shown, in several recent works, to have appeared as a natural consequence of the increasing size of the set of known substances -the chemical space-, while having as well been affected by social factors. It has been shown indeed, that the amount of chemical data was rich enough for allowing a formulation of the periodic system (PS), since as early as 1840; about 30 years before its original publication [1].

The PS was devised as a means to organize and capture the generality of the chemical information of the time, which naturally consisted of chemical compositions of known substances, some of their chemical and physical properties, and some of the chemical reactions in which these were known to participate, among others. The elements were known to be interrelated to one another by two main relations: order and similarity. It is important to remark how heavily both of these concepts relied on the chemical space, for the time under study: the order relation was provided by atomic weights, which were in turn calculated by finding the lowest common weight among large sets of compounds containing a given element [2], while similarity was assessed by comparison of the compounds elements were present in [3].

As recently reported by Leal et. al. [4], the mathematical structure of a periodic system is, in general, that of a directed hypergraph in which elements belonging to a given hyperedge are related -say, by a similarity measure-, while order exists among *and* within such hyperedges, and is not limited to one single property -e.g. atomic weight- but can arise from different orderings simultaneously. The Mendeleevian periodic table (MPT) is a special case of such a general structure, with atomic weight being the order relation used. In MPT, hyperedges are in reality partitions of the set of elements, each of these being represented by a column of the table, conforming what we call the periodic table groups. In such a setting, elements belonging to the same partition (group) are -ought to be- chemically similar, or at least more similar to the elements within the same group than to elements outside of it.

Taking hyperedges as mere partitions of the set of elements on a 2-dimensional array is, however, not an accurate description in general, and this was noted as early as Mendeleev's original publication of his periodic table [5]. Paraphrasing the author: "chemically-analogous elements show either sequentially incremental atomic weights (Pt, Ir, Os), or

*Corresponding author

an equal increment of this quantity (K, Rb, Cs).". Such statement can be interpreted to be referring to what, from now on, we call "horizontal" and "vertical" similarities among the MPT, respectively.

It is important to note that Mendeleev's approach was a purely empirical one and it depended on the known substances and elements of the time and, even further, depended on the substances a scientist knew at the time or even *decided* to use for her/his analysis.

Mendeleev, however, didn't use the whole chemical space of his time; in fact, it has been reported that he used only very specific subsets of it, including hydrides, hydroxides, halides, oxides, and some others [3]. In comparison, it is now known that more than 11,000 substances had been discovered by 1868, suggesting that the known chemical space was very likely undersampled, only because of its unbearable extension. In addition to this, the last 150 years have brought with them a consistently exponential increase in the size of the chemical space, with an annual growth rate of 4.4% [6]. For comparison, and as a consequence of this exponential growth, the number of substances reported *only* in 2015, amount to the total reported between 1800 and 1950 [7].

With a much more complete -and probably less biased- picture of 1868's chemical space, enough computational power, and inspiration from the formal formulation of the MPT, Mendeleev's original ideas of similarity shall be brought back to use on this paper. Here, we ought to extract the relationships between elements and, as a natural extension of this, the most suitable 2-dimensional representation of the PS out of such space, which may then be compared against Mendeleev's or Meyer's (or any other) periodic table.

All of this amounts, in principle, to questioning the fitness of *a* PS, on the basis of the foundations well established 150 years ago and the available chemical space. Going even further, having obtained a way of measuring the fitness of a PS, the next step is finding an optimal configuration for this fitness measure, leading possibly to a different and more expressive PS, yet following Mendeleev's data-driven approach but in a big data setting. This approach might as well be applied yearly for the expanding set of substances, which would amount to generating a 2D representation of the periodic table for every year, which would in turn allow to explore questions such as, for instance, how stable PSs are with changes on the chemical space, and which is the first optimal date, if any, for the formulation of the PS.

To answer these, and other questions, this study makes use of the set of available substances discovered up to the year 2015, extracted from the Reaxys database. The article is structured as follows: Section II presents the general data-preprocessing procedure and the reasoning behind it. Section III, the calculations performed and an analysis and discussion of the results. Section IV presents an optimization setting for the PS with (probably, but not quite yet) possible candidate configurations and the analysis of these results.

2. Data description and preprocessing

The compositional formulas of all the available compounds were extracted from the database in such a way that the resulting data is a collection of text strings such as for instance C6H12O. The main goal at this point is to convert this corpus of strings into a data representation that directly expresses relationships between elements. For such a task, the underlying "grammar" behind each compound's composition is to be found, and for that matter the processing and analysis must be focused on interactions between compounds rather than the conversion of single compounds into machine-readable formats. Note that this approach differs from usual machine learning studies in that in these, the latter formats are constructed and given to an algorithm, with the structure of the data being looked for *after* training; while the approach presented here aims to directly extract structure from the data, allowing to make analyses and conclusions before any statistical learning is invoked.

Very much in the spirit of Mendeleev's concept of similarity: “The elements, which are most chemically analogous, are characterized by the fact of their giving compounds of similar form RXn", our preprocessing consists of decomposing all the available compositional formulas, into all possible rewritings of the form $R - X_n$. In this rewriting, X is any single element and n is a subindex, while R can be any combination of elements with subindices, possibly including element X too. As an example, the compound $SiCl_4$ can be rewritten as $SiCl_3{-}Cl$, $SiCl_2{-}Cl_2$, $SiCl{-}Cl_3$, $Si{-}Cl_4$, and $Cl_4{-}Si$. Note that, at this point, chemical compositions of substances of various sizes and elemental compositions have been converted into a "binary compound" representation, which allows already for an important increase in the size of the dataset, with which studies such as [8], where only binary compounds (nearly 4700) were used for a network study, may be reworked to yield important results with a naturaly much less biased picture of the chemical space.

For our purposes here, we regard a similarity relationship between two elements, as follows:

Definition 2.1 (Similarity relationship)**.** Two elements X and Y are connected through a similarity relationship (R,n), iff there exist two compounds $A = R - X_n$ and $B = R - Y_n$, such that substitution of n atoms of X in compound A, by n atoms of Y, would result in compound B.

In practice, sets of elements were constructed to represent such similarity relationships and, as such, a set of elements is constructed for each pair (R,n). An example of the computation is as follows: assuming the compounds KOH, NaOH and H_2O exist, then by this treatment a set K, Na, H will be formed corresponding to the similarity relationship (OH,1). Naturally, all sets of cardinality equal to one were removed as they contain no information about relationships between elements.

Note that this approach assumes that $n \in \mathbb{N}$, that is, a problem arises when considering non-stoichiometric compounds. In the given dataset there are actually examples of such non-stoichiometric compounds, and so this case needs to be properly handled. Although it was observed that some of these compounds correspond to crystalline phases with non-stoichiometric amounts of water molecules, implying that the dataset could be manually curated in order to include such substances, it was decided to ignore them for the sake of simplicity. It is worthwhile noting that this approximation ignores every structural factor and relies only on compositional data, which in turn disregards any information about isomers.

2.1. Results of preprocessing

Starting from more than 19 million compounds obtained after removing non-stoichiometric entries, further data reduction was obtained from the removal of isomers. As we are interested in historical analyses, and various isomers are discovered in different years, the effective year for each compositional formula was decided to be that of the earliest discovered isomer, which corresponds to the date of report of said composition formula. After such cleaning, nearly 3.5 million different formulas were obtained, upon which the previously discussed preprocessing procedure was applied.

A total of more than 280 million unique "binary representations" (rewrittings of the form RX_n) were obtained, from which more than 9.5 million different similarity relationships were extracted. Note that these "similarity relationships" relate, in general, more than two elements, so the actual number of binary similarity relationships within our dataset is much larger than this number. The actual number of such binary relations can be calculated as the sum of the number of 2-combinations of N_i, the cardinality of set i. Such quantity is equal then to $\sum_i \binom{N_i}{2}$.

3. Results and discussion

3.1. Similarity matrices

The establishment of binary similarity relations between elements, coming from the aforementioned treatment, leaves us with a large number of such relations. The question about which relations are stronger, and thus, which pairs of elements are more similar, can be explored through the computation of what will be referred to as a similarity matrix.

Definition 3.1 (Similarity matrix)**.** Let L be a sorted list of the elements (say, by atomic number), and N be the length of such list. Then a similarity matrix M is an NxN matrix, where entry M_{ij} represents a net similarity measure between the ith and jth elements of list L. Entry M_{ii} is defined as the number of times element ith element is found in *any* similarity relation. Such net similarity measure will be computed as the total number of binary similarity relations found between ith and jth elements.

With this definition at hand, similarity matrices were calculated for each year between 1800 and 2015. As will be explored later on, these matrices can be analized globally in such a way that patterns are visible, and relations between pairs of elements can be understood as a colective.

3.1.1 Similarities: 1868

Let us now explore the similarities between elements for the year 1868. The numbers shown here represent a result of the analysis discussed earlier, performed over a dataset limited to the substances discovered before this year only. As such, only 60 elements are present in the matrix. This year is of primal relevance for this study, as this is the year of first publication of Mendeleev's PS and, as such, studying relations between elements for this particular moment of history allows for the exploration of the relations upon which Mendeleev and others based the construction of their periodic systems.

Figure 1 shows a normalized version of the formerly discussed matrix. The particular normalization procedure consists on the division of the entries M_{ij} by entry M_{ii}, and so the plot is of the matrix M', defined as $M'_{ij} = M_{ij}/M_{ii}$, $\forall j, 1 \leq j \leq N$. Such normalization allows to explore relations, while excluding statistical weights coming from the fact that some elements (e.g. H) appear in many more relations than others (e.g. La). It also is the reason why the matrix shown is not symmetric, while definition 3.1 guarantees M to be symmetric.

In this plot, the higher the color, the more similar the corresponding elements are. In general terms, some substructures can be seen within the matrix such as, for instance, vertical, horizontal and diagonal lines, which explicitly suggest the higher structure the PS is intended to capture. Particularly, some diagonal patterns can be observed, that are being repeated throughout the matrix. These diagonals, as will shortly be explained, are reminiscents of what will then be realised in the MPS as vertical, or group similarities.

To understand the meaning of these diagonals, let us explore one of the most prominent of these in figure 1; the diagonal starting from the relation As → P, followed by Se → S, and so on. Both of these are realised as sequential, pair-wise vertical similarities in MPT. What this means in general is that, provided entry M_{jk} is bright, the fact that entries $M_{j+n,k+n}$, n integer, are bright too, is reminiscent of the periodic behaviour expressed in the construction of groups in MPT.

Other patterns are prominent too, such as vertical and horizontal sequences in figure 1, as well as their combination to form somwhat consistent blocks. These are realised as horizontal and non-local relationships in MPT. In particular, we see a bright block formed by the elements from Mn to Zn, which represents the well known horizontal similarity among first row transition metals. Another important reminiscent of periodicity and non-locality, is a set of vertical lines formed to the sides of the aforementioned block. It represents a similarity between elements from Mn to Zn, and group 2 elements (Mg, Ca, Sr, Cd, etc). Very importantly, we note that this non-local relation is found in 1868, however MPT doesn't clearly express such relation, which may indicate a lack of dimensionality of this particular representation of the PS even for a chemical space as small as the one existent to this year.

One particularly important element for our discussion is hydrogen, as its position in the PT is still a controversial matter [cite]. Arguments from electronic structure arise for the placement of H on the first group, as its valence shell

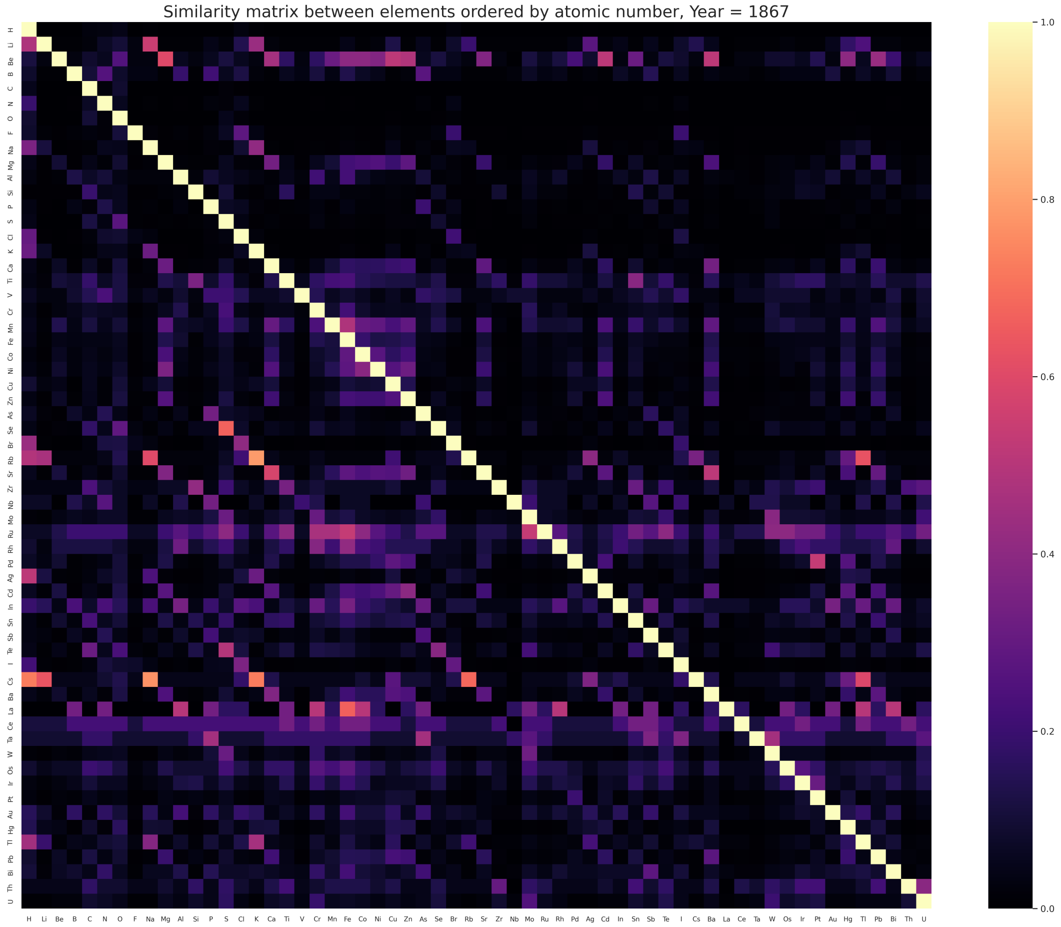

Figure 1: Similarity matrix for year 1868. Note that this matrix is not symmetric, meaning the relationship X → Y differs from Y → X.

contains one electron [cite], or on the 17th group (the halogens), as it lacks one electron for a full valence shell [cite]. There are even arguments stating that H should be on 14th group, above Carbon, as its valence shell is half full [cite]. In Figure 2, a plot is provided based on the current approach, that is intended to contribute to the discussion.

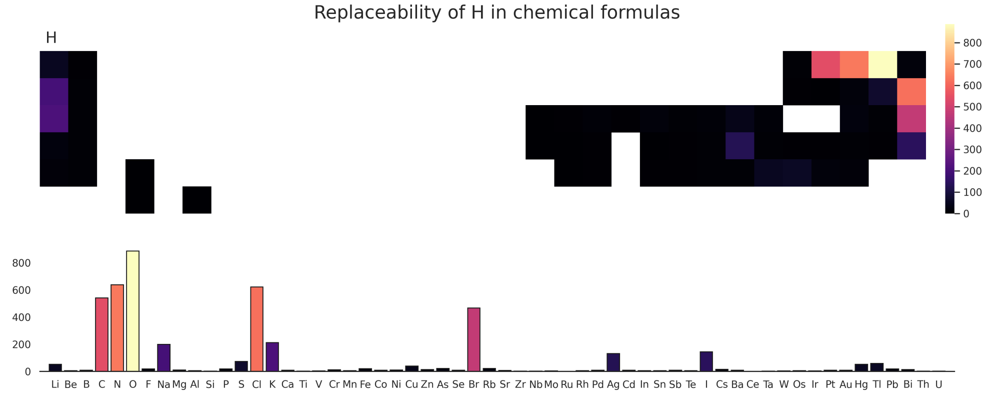

Figure 2: Replaceability of H in compounds in dataset. Year 1868.

The results shown in Figure 2 show that H can be frequently replaced, expectedly, by Na and K, one of the groups it is attached to. Evidence is also provided, that H is even more similar to the halogens, as replaceability of H is higher for Cl, Br and I to a lesser extent. More prominently, however, strong relationships with O, C, and N are observed, which contribute to a much more interesting discussion: that of whether H should be a group 14 element.

The latter relations may seem weird at first, however, taking a look at the formulas responsible for this result, it is found that SX_2, OTlX, NX_3, MoOX, MnOX and similar substances offer an explanation for the relation H → O. These show that such a result is actually due to the presence of metallic oxides, hydroxides and hydrides, and happens in combination with elements with more than one oxidation state such as transition metals and other non-metals such as S and N, which allows them to combine in some compounds with a number x of Hs, while in other compounds with the same number of Os by doubling their oxidation state.

With the aim of exploring how these trends evolve with time, let us now explore the similarity matrix calculated for the year 2015.

3.1.2 Similarities: 2015

Figure 2 shows the results of the same calculation, as explained in the previous section, but with the whole dataset. This corresponds to the most complete map of similarities between elements that can be achieved with the current treatment and, as such, represents one of the most important products of this research.

In the light of the analyses earlier discussed, the results in figure 2 are quite interesting, as they seem to imply that the periodicity and the other relationships above discussed, are preserved and even *reinforced* with time.

A whole new group of elements is introduced to the study, namely, the lanthanides. These show remarkably high and consistent similarities within the series, which is the responsible for the bright large square in the middle of the matrix. Other manifestations of periodicity are observed too, this time more prominently than in the previous plot. Here, the diagonals go ver clearly and uninterruptedly, from Zn → Mg, all the way to La → Y, where a vertical relation is manifested (i.e. Y → La - Lu), and then the diagonal pattern is continued, starting from Zr → Hf and, arguably, going as far as Bi → Sb.

There are other less clear, but similar patterns. It is observed, for instance, another diagonal formed starting from Ag → Na, all the way down to Ba → Ca, where it is interrupted by a noisy vertical pattern correspondent to similarities between Sc and the lanthanides (La-Lu). The diagonal pattern is again observed starting from Hf → Ti, up to Bi → As. Although these are much more hidden, and probably less prominent patterns, they are, too, reminiscent of the periodic behaviour that is intended to be expressed in the PT groups.

Many things can be said just from these results. The lanthanides are observed to be a highly consistent group, meaning that all the elements here are very similar to all other elements in the same group. Additionally, the results in figure 2 and the discussion above provide evidence that Y and, to a lesser extent, Sc, behave much like lanthanides, a fact discussed in earlier works [cite]. The less intensive patterns discussed in the last paragraph, express a vertical, non-contiguous similarity between elements, such as that expressed by Ga → Ti. We see, however, that such observations are observed since Ag → Na and, less prominently, Mg → Cd, indicating the possibility that Na and Mg be placed above Cu and Zn, respectively. Such relations were noted by Mendeleev as early as his 1871 publication of the PS, to the extent that Ag was grouped with the alkaline elements and Cd with the alkaline-earths.

As can be seen, however, there are similarities for which no particular periodic pattern can be attained. In particular, columns for C, N and O show that the similarity between these elements, and many others, is in general high. However, when the rows correspondent to these same elements are analised, we see mostly dark bits. This highly antisymmetric behaviour can be understood considering that many elements can be *replaced by* C, N and O, while these elements can

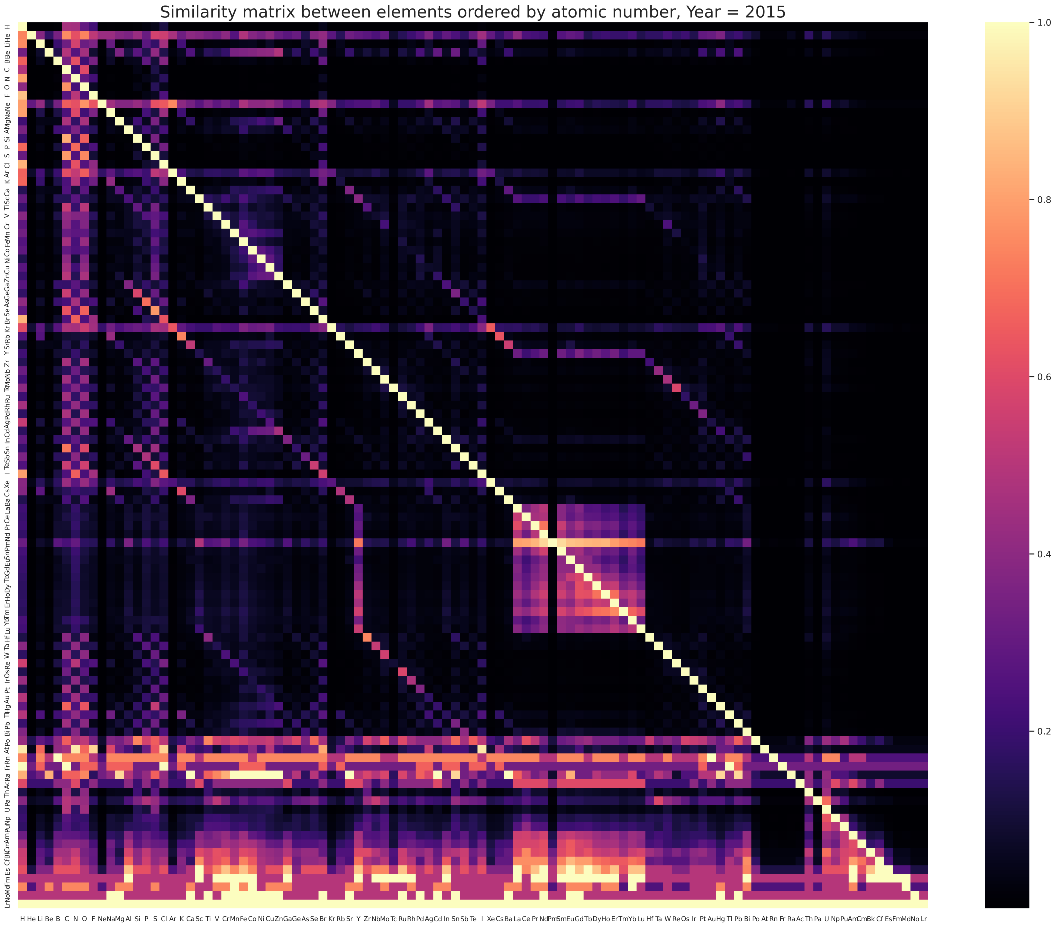

Figure 3: Similarity matrix for year 2015.

be replaced *only by a few* other elements. Such fact can be understood as a manifestation of the singularity principle, where second period elements show particularly different properties than the rest of the elements of their groups. This principle is also manifested in the fact that the diagonal patterns discussed above start from Na → Ag, or Mg → Ca, that is, third period elements and beyond.

As a final remark, we observe that very heavy elements (practically everything after Bi, where nuclei are highly unstable), show a similar singularity than that already discussed for C, N and O. This is due to the fact that very few compounds are known for such elements, and as such, the chemistry has not been explored enough, so as to make assertions on the similarity between these and other elements.

On the light of these results, many questions arise:

- How good does a selected PS express these patterns?
- How would an optimal tabular representation of the PS look, on the light of such metrics?
- How does such representation evolve with time?
- How *early* could we come up with a stable representation?

For the assessment of these, and many more questions, we formulate a *fitness* measure, general for any tabular representation of the PS.

3.2. Evaluation of a PS: Horizontal Neighbouring Distance

With the objective of comparing different periodic table representations of the PS (PTR), the aim of this chapter is to develop a measure for the adequateness of a particular PTR for describing the generalities of the relations between elements.

Although with well known exceptions, the general idea of a PTR is to encode both order and similarity between elements. In that sense, their construction requires an underlying, in principle sequential order (usually by atomic number), being "distorted" by similarity relations that are expressed in a second dimension through clasification of elements in a number of columns, which will be called groups. As the order relation is already fixed, our evaluation should be concerned with how well grouped are similar elements. Now, some degree of flexibility should be given to the proposition that similar elements be placed in exactly the same group; instead it may be proposed that similar elements should at least lie in neraby columns.

With such requirements in mind, the horizontal neighbouring distance (HND) is now proposed. Let us first understand what the calculation is about, and then a mathematical formula will be derived. First, let us assume elements X and Y are related by the similarity relationship (R,n). Assuming a PTR, such pair of elements have each an assigned group and as such, a *horizontal distance* can be calculated between them neighbors. In order to include the information about every single binary relationship, the mean is calculated as the sum of all these distances, divided by the number of binary relationships.

Naturally, the distances between every pair of elements is fixed, for a fixed PTR. As such, the above described mean can just be computed as a mean of all the pairwise distances, weighted by the number of similarity relations found for each pair of elements. This statement is formalized through the following definitions.

Definition 3.2 (Group vector)**.** A group vector G is a numeric array containing the group number of each element, in a particular PTR. The ith entry of G corresponds to the group to which the ith element belongs in the given PTR. As such, G is a functional of the PTR. $G = G[PTR]$.

Definition 3.3 (Mean Horizontal Neighbouring Distance)**.** mHND is defined as the weighted average of all binary similarity relations among elements. Considering that matrix elements M_{ij} (def. 2.1) contain the weight for similarity relation between ith and jth elements, then mHND$_i$, for element i, is defined as follows.

$$mHND_i = \frac{\sum_k M_{ik}|G_i - G_k|}{\sum_{k \neq i} M_{ik}} \tag{1}$$

Where the sumation in the numerator can be written as such, as $G_i - G_i = 0$.

By construction, a mHND$_i$ of 0 means that exactly all elements related to ith element by similarity relations, are in the same group in the given PTR as said element, reinforcing the vertical similarity idea, while a larger distance implies a departure from this rule. Figure 4 shows a plot of mHND for all elements, calculated using the long format of the periodic table and embedded in the same format.

Figure 4 clearly shows that, overall, the vertical similarities are not dominant the whole MPT and, furthermore, such patterns are observed more commonly to the right of such table. Groups 1 and 2 show surprisingly high mHNDs considering how chemically similar elements within these groups are expected to be (Na-K-Rb-Cs and Mg-Ca-Sr for instance).

Given that mHND$_i$ is parametrically a function of the selected PTR, a new PTR may now be devised by optimization of this quantity, which would lead to one where mHND is minimized for every element, while preserving some given elemental ordering (be it atomic weight, Pettifor's scale, etc), which should naturally lead to a more expressive representation, and would allow a comparison between systems. Optimization and comparison between systems is a topic for the next section.

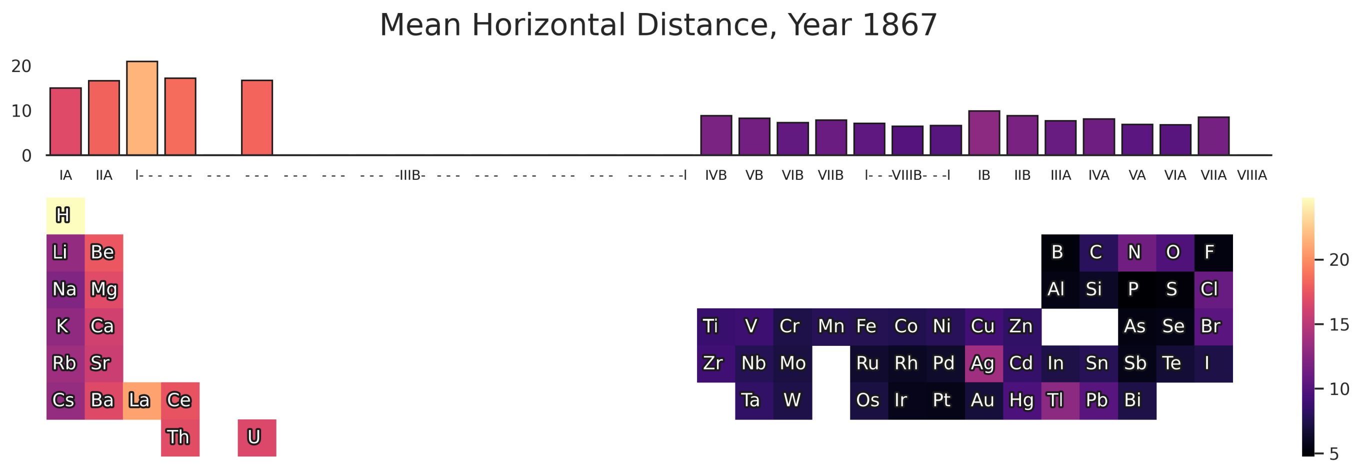

Figure 4: Mean horizontal neighboring distance (mHND) for the elements in the PS, year 1868.

3.3. Optimization

The present work could take many different directions from this point. In particular, the approach developed here could be directed towards the optimization of a one-dimensional ordering of elements, such as that reported in [9], where a reformulation of the Pettifor's scale was done with a similar approach to that presented here, but using only a set of inorganic substances obtained from the ICSD. Again, such work could highly benefit from the more general approach considered here with much more substances and relations, and the conclusions of such work could as easily be drawn.

Another direction is finding an optimal two-dimensional representation, that expresses a given order e.g. by atomic number), while expressing chemical similarity as well. In general, the investigation could be headed towards the optimization of a PS, given a dimensional constraint. As such, the question about the optimal dimensionality of a PS could be addressed, and all of this could be extended towards a historical analysis.

Equipped with def. 3.3, this work will take the second approach mentioned: the optimization of a two-dimensional representation of the PS (PTR). For that matter, the expanded description provided by individual mHND$_i$s is dumped into a single number S, which will be the sum of a modified expression for all individual mHND$_i$s.

$$S = \sum_i mHND_i = \sum_i \frac{\sum_k M_{ik}(G_i - G_k)^2}{\sum_{k \neq i} M_{ik}} \tag{2}$$

Where the absolute value is exchanged by the square operation. In that sense, the above equation represents a mean horizontal square distance, but for our purposes the focus will be on the quantity S only. Such quantity will be our optimization objective. To begin, as we have an (in principle) analytical expression, gradients can be calculated as follows.

$$\frac{\partial S}{\partial G_j} = \sum_i \frac{(G_i - G_j)N_{ij}}{\sum_k N_{ik} - N_{ii}} \tag{3}$$

As

$$\frac{\partial G_k}{\partial G_j} = 0, \forall k \neq j$$

The advantage of our modified version of mHND (eq. 2) is that the gradients of S contain explicit information about distances, while the gradient of absolute values would only contain information about the sign. With these quantities at hand, a variety of optimization techniques could be applied.

To facilitate the visualization of the first kind of representation (figures 4 and **??**) an interactive webpage is to be constructed, where the similarities are calculated for an user-chosen element and shown in the PS representation.

References

[1] W. Leal, E. J. Llanos, P. F. Stadler, J. Jost, and G. Restrepo, "The chemical space from which the periodic system arose." Submitted / accepted., 2019.

[2] I. Hargittai, "**The Periodic Table: Its Story and Its Significance**. By Eric R. Scerri. New York: Oxford University Press 2007. Pp. xxii + 346. Price (hardback) £19.99. ISBN-13: 978-0-19-530573-9.," *Acta Crystallographica Section B*, vol. 64, pp. 123–124, Feb 2008.

[3] *Mendeleev on the Periodic Law: Selected Writings, 1869 - 1905.* Dover Publications, 2002.

[4] W. Leal and G. Restrepo, "Formal structure of periodic system of elements," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 475, no. 2224, p. 20180581, 2019.

[5] D. Mendelejeff, "Ueber die beziehungen der eigenschaften zu den atomgewichten der elemente.," *Zeitschrift für Chemie*, vol. Band 5, pp. 405–406, Dec. 1868.

[6] "Exploration of the chemical space and its three historical regimes," *Proceedings of the National Academy of Sciences*, vol. 116, no. 29, pp. 14779–14779, 2019.

[7] G. Restrepo, "Compounds bring back chemistry to the system of chemical elements," *Substantia*, vol. 3, pp. 115 – 124, Dec. 2019.

[8] W. Leal, G. Restrepo, and A. Bernal, "A network study of chemical elements: From binary compounds to chemical trends," *Match*, vol. 68, no. 2, pp. 417–442, 2012.

[9] H. Glawe, A. Sanna, E. K. U. Gross, and M. A. L. Marques, "The optimal one dimensional periodic table: a modified pettifor chemical scale from data mining," *New Journal of Physics*, vol. 18, p. 093011, sep 2016.