A computational approach to the periodic system.

ANDRÉS MARULANDA-BRAN[*]

Universidad de Antioquia
acamilo.marulanda@udea.edu.co

February 18, 2021

Abstract

1. Introduction

Mendeleev's periodic system is, and for the last 150 years has been, regarded as one of the most important discoveries/inventions in the history of chemistry and in general the natural sciences. It was conceived after nearly one century of substantial efforts from the chemical community, remarkably from Döbereiner, Meyer and Mendeleev himself, the last of whom being the one that discovered and first used the periodic system as a predictive tool during the second half of the 18th century.

The periodic system (PS) was devised as a mean to systematize the generality of the chemical information of the time, which mainly consisted of chemical compositions of substances and their properties. To this time a few things were known about the elements and the way they gave rise to all the substances, the most relevant being that they can be sorted by weight, and that some of these show particular similarities such as the kinds of compounds they produce, and some chemical and physical properties of the substances they are present in. This allowed scientists to group some of the elements in small groups based on their similarity, giving rise to the question of whether the large set of known elements could be represented within a single system allowing to express and generalize this knowledge.

The main idea is to arrange the elements in a two-dimensional array, where the horizontal dimension expresses an order relationship, meaning an element is heavier than any element to its left, and strictly lighter than any element to its right (on the same row). Its vertical dimension expresses similarity in the sense that if two elements lie within the same column, they are -ought to be- chemically similar. Mendeleev's concept of similarity relies on the basis of the likeness of the chemical formulas the elements usually take part on, a concept very closely related to that of valence. One important note came with the first report of the PS, and is that not only vertical similarities occur throughout the table, such as for instance between K-Rb-Cs; but also horizontal ones, such as Pt-Ir-Os. It is important to note that this was a purely empirical approach 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 set of substances known at the time -the chemical space-; in fact, it has been reported that he used only very specific subsets of this chemical space, including hydrides, hydroxides, halogenides, oxydes, and some others. In comparison, it is now known that more than 11000 substances had been discovered to that time, 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 some clearly distinguishable regimes characterized by the different kinds of substances produced.

[*] Corresponding author

With this, the questions arise of whether the 1868's periodic system "fits" that time's data, and how it has been affected by the changes on the chemical space through time. All of this amounts to questioning the fitness of *a* periodic system, 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 periodic system, yet following Mendeleev's data-driven approach but in a big data setting.

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 preprocessing

The compositional formulas of all the available compounds were extracted from the database in such a way that the data ends up being a collection of text strings such as for instance C6H12O. The main idea of our preprocessing is to convert this corpus of strings into a more meaningful and easy-to-analyze format, over which mathematical operations and statistical measures can be performed. 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 for these, the latter formats are constructed and given to an algorithm, having the structure of the data being looked for *after* the preprocessing; 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.

For that end, sets of elements were constructed in such a way that two elements X,Y appear in the same set only if there exist 2 compounds $A = R - X_n$ and $B = R - Y_n$, where if n atoms of X were replaced by n atoms of Y in compound A, the result would be compound B. A set was constructed for every possible (R,n) pair found in the original dataset, and only those sets containing at least 2 elements were considered. This processing naturally expresses a similarity relationship between all the elements within a single set. 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 formula OH-X. For this particular example, R = OH and n = 1.

Note that this approach assumes that n is an integer, that is, there's a problem when treating 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. This was solved in the following way: if upon duplication of all subindices in the given compound this compound is stoichiometric, we work with the new, index-duplicated compound (this works for subindices $n = x/2$, $x \in \mathbb{N}$). Otherwise the compound is discarded. This first consideration should be studied in more detail as these compounds may correspond to, for instance, crystalline phases with non-stoichiometric amounts of water molecules. It is worthwhile noting that this approximation ignores every structural factor and relies only on compositional data.

For the purpose of visualization and analysis, such sets can then be conveniently represented in a 2-dimensional representation so as to include the information of some periodic system, that is, the relative positions of the elements in a given set, within the given PS. Examples of such a representation are shown in figure 1 where, for each element of a given set, its corresponding position in the given PS is pictured white.

Starting from 11356 substances known to the year 1868, a total of 6145 said tables were produced. The PT used is the standard long-version of the PT.

Here, black means no element exist at this position, red means an element in this position exists in dataset, and white means there exist a compound $R - X_n$ as explained above. In this example, for instance, the data shows that both ZrF_4 and $ZrCl_4$ exist in dataset, but there exist no other compound with formula ZrX_4. It also shows that most lanthanides and actinides, as well as Tc, Re, Ga and Ge hadn't been discovered to this date. The vertical relationships on this side of

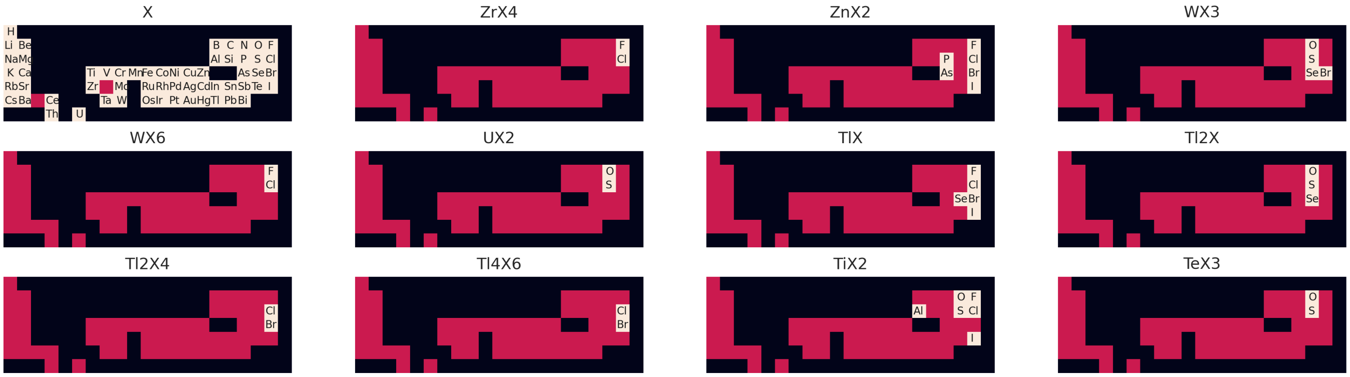

Figure 1: Periodic table representation of the constructed dataset.

the PS are clear from this sample.

3. Results and discussion

3.1. Mean Horizontal Neighbouring Distance

From figure 1 the typical vertical similarity can be spoted in the majority of tables from this sample (manifested as vertical blocks of white squares). We may ask, however, to what extent is this vertical similarity rule followed throughout the whole dataset? That is, is it more prevalent than, for instance, horizontal or diagonal (or any other) similarity and, in any case, how much deviation is there from the ideal arrangement?

To answer this, the horizontal neighboring distance (HND) was computed. For each element in each table, this is calculated as the averaged difference in PS groups there exist from the element to all its neighbors in the same table. This quantity is then averaged throughout all tables and the result is shown in figure 2. Note that this quantity is clearly a function of the available chemical space and the selected PS. By construction, an HND of 0 means the two elements belong in the same group, reinforcing the vertical similarity idea, while a larger distance implies a departure from this rule.

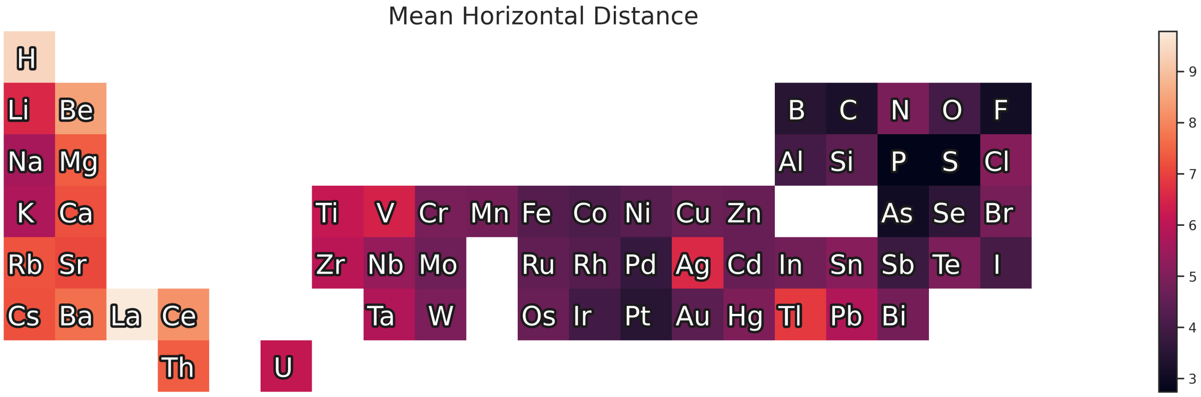

Figure 2: Mean horizontal neighboring distance (HND) for the elements in the PS. HND measures up to what degree elements are "misslocated" in the PS considering that similarity should be expressed as vertical relationships.

Figure 2 clearly shows that, overall, the vertical similarities do not occur throughout the whole PS and, furthermore, this occurs more frequently to the right of the PT. Groups 1 and 2 show surprisingly high HNDs considering how chemically similar elements within these groups are expected to be (Na-K-Rb-Cs and Mg-Ca-Sr for instance). An initial explanation, for which more evidence will be shown further on, calls for the consideration of the metallic character of these elements which, in some aspects, resemble the chemistry of the transition metals (group 3-12) and thus are expected to appear in some of the tables, increasing the mean HND for the elements of these groups.

A number of other different distances can be calculated as well, but this has been selected as the most descriptive as the PS is constructed on the basis of order and similarity, with the similarity expresed as a strictly vertical one on this representation, as already stated above. Given that HND is parametrically a function of the selected PS, a new PS may now be devised by optimization of this quantity, which would lead to one where HND is minimized while preserving some given elemental ordering (be it atomic weight, Pettifor's scale, etc), which should naturally lead to a more expressive PS at least, and would allow a comparison between systems. Optimization and comparison between systems is, however, a topic for the next section.

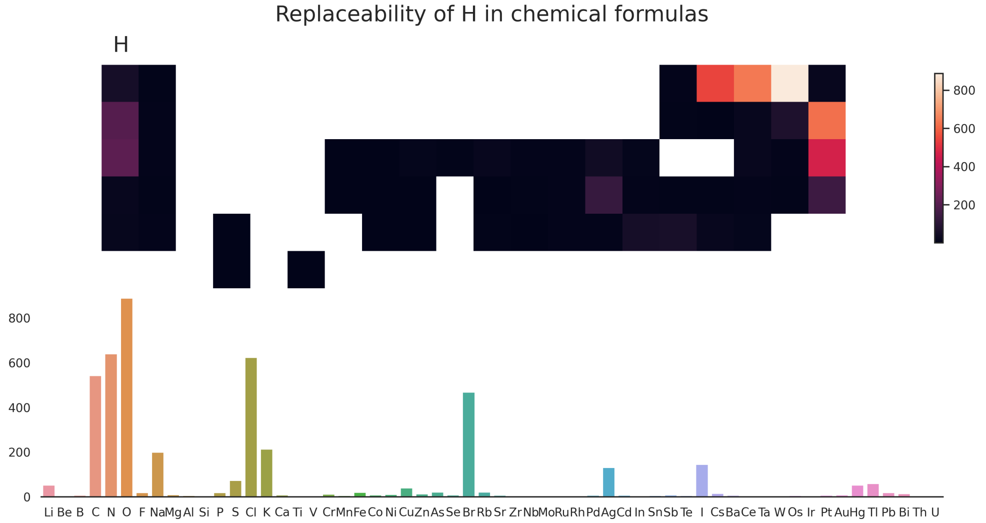

Figure 3: Replaceability of H in compounds in dataset. a) PS format showing periodic relationships and b) Barplot showing elements in atomic number order, bar height is number of times H and this element appear in the same table.

3.2. Replaceability of elements.

Not only can we calculate how far away elements are from their neighbors, but we might as well ask who their neighbors are. From the way the data was preprocessed, the answer to latter question can be interpreted as a measure of replaceability, as elements in the same constructed set can be interchanged within this set's corresponding formula. For a given element, the calculation consists of taking all the tables that contain this element, and then for each other element count the number of times these are present in this subset of tables. From this, some more detailed insights can be obtained about the relationships between elements, as shown in figure 3 for hydrogen.

It can clearly be seen that H is, under our approximation, more similar to C, N and O, and the halogens (except F). A similarity with the halogens are expected, as in most of organic chemistry these appear as substituents of H, but the resultant similarities to C, N and O must be further explored in order for it to chemically make sense.

Although the relationship between O, C and N with H may seem weird, 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. 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.

The other important relation found here is that expressed with the halogens. It can clearly be seen that the H-to-halogens relations are much stronger (more than twice) than that with the alkaline elements (group 1). This is clearly an important result as it suggests that H should be nearer to the halogens in the PS, as opposed to being in group 1 and treated as an alkaline element.

To show the power of these plots, another example is brought into play (figure 4).

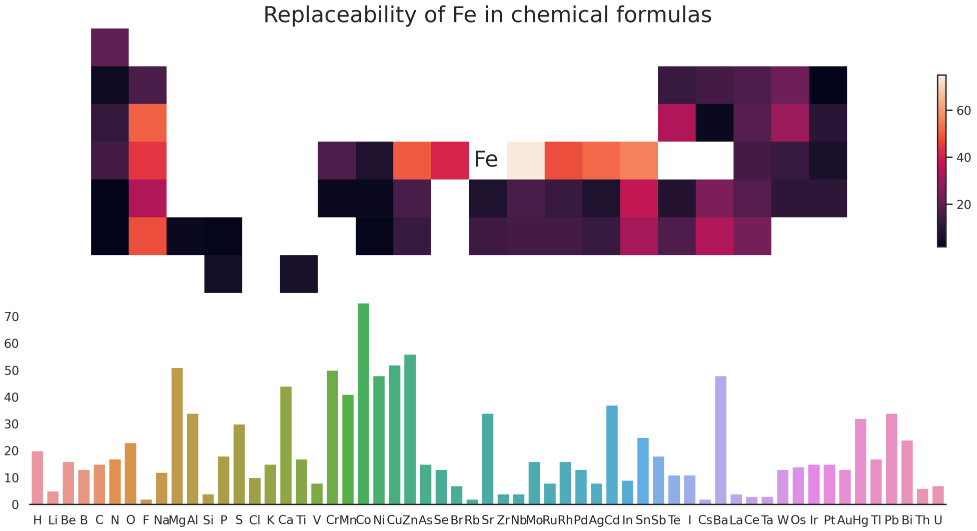

Figure 4: Replaceability of Fe in compounds in dataset.

The replaceability plot for Fe shows that, for this element, the similarity is mostly horizontal along the 4th period transition metals, as already expressed by Mendeleev on his original publication of the PS, but it also shows a comparable similarity with alkaline earth elements (group 2). Such a similarity is a non-trivial result and deserves a more in-depth exploration, as the PS expresses none of this information in it. Some questions arise as, for instance, can a new PS be constructed, in such a way that this information is more explicitly given? If not, is there some particular reason (probably a topological one) for this?. As an additional observation, the plots for alkaline elements (Supplementary Information) do not show such a marked similarity to transition metals and are instead more cluttered around group 1 of PS.

One of these plots could be drawn for each element, but this becomes quite tedious as it would imply the production and analysis of more than 90 of these individually. Instead, the PS representations can be unravelled into a 1-dimensional format and normalized, and finally concatenated into one large square matrix so as to be able to visualize all pairwise relative similarities between elements, as shown in figure 5. This ultimately looks much like a correlation matrix as it states element-wise relationships between pairs of elements in a more concise format. It looses, however, the possibility of studying the PS on the light of these results, showing that both representations give different important information and none should be disregarded in favor of the other. To facilitate the visualization of the first kind of representation (figures 3 and 4) an interactive webpage is to be constructed, where the similarities are calculated for an user-chosen element and shown in the PS representation.

The details of the normalization are important as it determines what is shown on the figure and the way it should be interpreted. In this case each row of the matrix was divided by the maximum value of said row, which naturally corresponds to the same element as the row, justifying the observation that the diagonal of the matrix is all ones. Due to this normalization the matrix is not symmetric, meaning the information on the upper triangle is different than that shown on the lower triangle, which is the same as stating that the similarity relationship X → Y is not the same as Y → X. As an example, take the couple La-Fe. As can be seen, the row corresponding to La has a high value (near 1) at the Fe column, meaning La can be replaced by Fe most of the time (within this dataset). When we look at the Fe row, the value corresponding to La is one of the lowest, meaning Fe can not be replaced by La most of the time.

The unidirectionality of these relationships is important as it may lead to several conclusions (drawn naively only from these results), namely:

- The chemistry of La wasn't as explored as that of Fe at the time. This can be thought of in sociological terms, as an exploration of the new element (La) limited to trying to copy the chemistry of a seemingly similar element (Fe).
- Or possibly the chemistry of La is more restricted than that of Fe, meaning possibly that the oxidation numbers for La are only a subset of those for Fe.

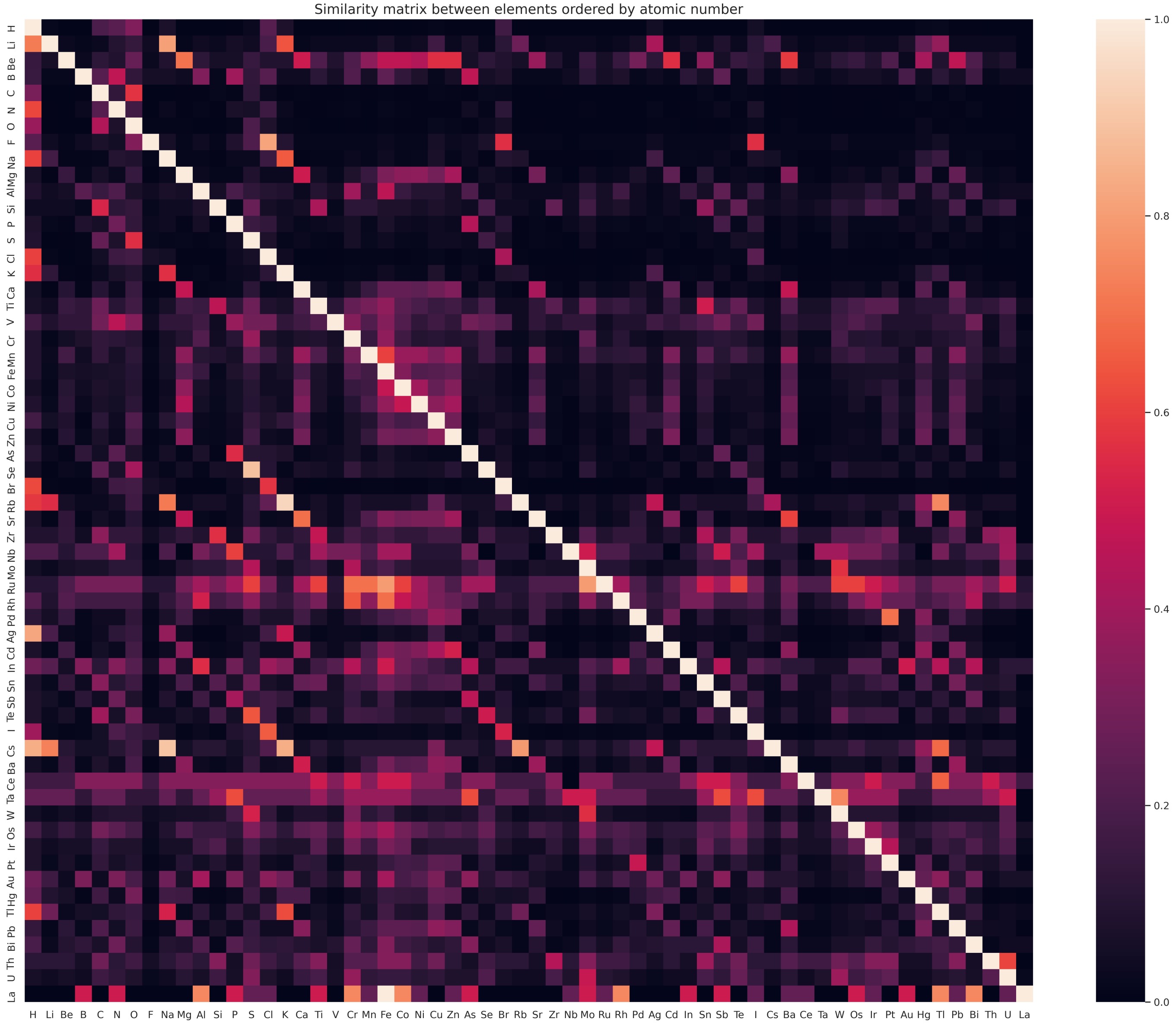

Figure 5: Replaceability matrix. Note that this matrix is not symmetric, meaning the relationship X → Y is not the same as Y → X.

In more 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. Namely, diagonal relationships on this matrix establish vertical pair-wise element similarities in the periodic table, for instance As → P followed by Se → S, Br → Cl, and so on. Vertical and horizontal relationships represent horizontal similarities in the PS, as well as other hidden ones such as the one shown in 4 between group 2 elements and first row transition metal elements. In this sense, this matrix (and others of this kind) allow for a much high level visualization of all the general relationships between elements without an embedding in any arbitrary PS. Again, some optimization work can be performed on this matrix similar to [1], 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. The present article could in fact be an extension of this, as it takes a very similar concept and extends it from inorganic solids to all known substances (up to some date).

References

[1] 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.