J Math Chem (2013) 51:413–419
DOI 10.1007/s10910-012-0091-x

LETTER TO THE EDITOR

About the concept of Chemical Space: a *concerned* reflection on some trends of modern scientific thought within theoretical chemical lore

Ramon Carbó-Dorca

Received: 28 June 2012 / Accepted: 15 September 2012 / Published online: 27 September 2012

Sir,

In a recent paper [1], a footnote was written by the present author as co-author in order to comment about the concept of *Chemical Space*, which is a term appearing sporadically since old times in some available chemical literature. The mentioned footnote text was printed as follows, with the underlined part used to emphasize here some relevant parts:

> Often in the QSPR or QSAR common literature the term *Chemical Space* appears. This concept it is not currently used with the same sense as a possible *molecular space* notion or the equivalent term: *space of molecules* given here, but in a generic fuzzy manner, sometimes even using meaningless and preposterous literary ways. The authors would like to find within the QSPR literature a sharp definition of such *Chemical Space* term, but until the present time found none. So, in order to employ the terminology of the present work as far away as possible from the usual classical QSAR fuzzy one employed by some authors, the term "space of molecules" has been chosen and will be used here, although "molecular space" might be also synonymou-sly used. This notion has been and will be also employed here as a companion of the term: *space of parameters* or the synonymous concept: *space of descriptors*, which is also a well-defined definition with a precise meaning, commonly used in the task to build up a set of discrete linearly independent vectors, empirically describing a MPC. From such a vectorial descriptor space

This letter is written to the memory of Professor Odd Gropen, a faithful friend, a dedicated scientist and an occasional watcher of Star Trek.

R. Carbó-Dorca (✉)
Institut de Química Computacional, Universitat de Girona, 17071 Girona, Catalonia, Spain
e-mail: quantumqsar@hotmail.com

> construct it is normally obtained a QSPR linear equation intently employed for molecular unknown property estimation purposes. The statistical techniques, employed upon the vectors of the space of descriptors, within the empirical QSPR equation search algorithms, which drastically reduce the dimension of molecular descriptors' space, generate the *dimensionality paradox*.

The writer of this letter subscribes the ideas of the footnote as they are. Definitions about what Chemical Space must be thought nowadays are in essence within this small piece of work. But the present author thinks it is worthwhile to surf into some sources of information, which will lead to a fuzzy intellectual panorama. The subscriber of this letter is eager to say that it is necessary to describe such a way of thinking, as a warning in a travelling research track, which might lead to poor science background building in twenty-first century.

When talking about the difficulty of finding a sharp definition of Chemical Space, the authors of the previous footnote wanted to express the fact that no definition appearing in the literature seemed to be satisfactory. The basis of this opinion is tried to be explained here. However, an even more exhaustive analysis of the inquiry, in order to clarify the questions awakened in the quoted footnote, has aroused in the present author a deep concern about some contemporary scientific opinion trends, hopefully thinking they are just shared by few. The present letter is a consequence of this state of mind.

On the other hand, and to start this journey in the Chemical Space quest I have found that in the Wikipedia the search engine presented some site with the following text:

> **Chemical Space** is the space spanned by all possible (i.e. energetically stable) molecules and chemical compounds—that is, all stoichiometric combinations of electrons and atomic nuclei, in all possible topology isomers. Chemical reactions allow us to move in Chemical Space. The mapping between Chemical Space and molecular properties is often not unique, meaning that there can be multiple molecules which exhibit the same properties (?). Material design and drug discovery both involve the exploration of Chemical Space.

The question mark is mine. Fortunately, this not a so reliable source of information, provided as a real bonus with a complementary reference to the subject [2]. Within this reference it was found the following text:

> "Space", as Douglas Adams famously said "is big. You just won't believe how vastly, hugely, mind-bogglingly big it is". Change 'space' to 'Chemical Space', and his statement has similar resonance: the total number of possible small organic molecules that populate 'Chemical Space' has been estimated to exceed 10^{60}—an amount so vast when compared to the number of such molecules we have made, or indeed could ever hope to make, that it might as well be infinite. So, it is not surprising that our exploration of Chemical Space has so far been extremely limited. Taking the analogy further, just as much of astronomical space is a void, much of Chemical Space contains nothing of biological interest. But rarely, and often

> through serendipity rather than design, we have identified 'stars' in Chemical Space—molecules that can modulate biological processes. These molecules have formed much of the basis of our fight against disease and have greatly aided our understanding of biological systems. But such successful finds have been hard to come by, in part because of our lack of understanding of Chemical Space. Given that its enormous size makes a thorough exploration of Chemical Space impossible, a key question is how we should best direct our efforts towards regions of Chemical Space that are most likely to contain molecules with useful biological activity. This question is a central theme of the articles in this Insight, which were inspired by the Horizon Symposium on 'Charting Chemical Space: Finding New Tools to Explore Biology', the fourth in a series of unique scientific discussion meetings run by Nature Publishing Group and Aventis.

Kirkpatrick–Ellis introductory note printed above, quoting a literary description by Douglas Adams of fuzzy scientific content, talks about the Universe, where the solar system is and where are located an estimated number of some 10^{11} galaxies, each one provided with another estimated number of some 10^{11}stars, which is certainly a huge object collection. However, the authors obviously confuse Douglas Adam's description of a Galactic Universe (GU) say, with the Universe of Discourse, the Universal Set or simply Universe, a complete different subject, but a well-known definition of elementary set theory.

According to the Encyclopedia Britannica, one can read for instance: *when the admissible elements [of a set] are restricted to some fixed class of objects U, U is called the universal set (or universe)*, a description which any high school student of modern times might know.

Better, the Merriam Webster dictionary accepts the dictionary entry: *Universe of Discourse*, with two meanings, and the second one appears to be: *2) an inclusive class explicitly containing all the entities to be discussed in a given discourse or investigation or theory*.

That is, in chemistry one can contemplate the entity which must be associated to the set of all possible chemical structures, which is made of a huge quantity of elements too, but completely different in essence to the GU. Or in another way around: *Chemical Space*, accepting this term as a definition of the universe of discourse, which can be attached to chemistry, has *nothing* to do with the starry night universe, except in the large number of elements composing both. Even the estimation of the number of chemical structures provided by the introducers is conservatively made. Just taking all the structures with 100 atoms made of possibly 100 different elements it appears already as a nice quantity like: 10^{200}. That is, some number far greater than the estimated star number, which is contained into the GU and certainly far greater than the one provided by Kirkpatrick–Ellis.

However, not all in this introduction has to be blamed. Thanks to the Kirkpatrick–Ellis foreword, another reference was found [3]. Where a definition of Chemical Space could be gathered:

> Chemicals can be characterized by a wide range of 'descriptors', such as their molecular mass, lipophilicity (their affinity for a lipid environment) and

> topological features. ‘Chemical Space’ is a term often used in place of ‘multi-dimensional descriptor space’: it is a region defined by a particular choice of descriptors and the limits placed on them. In the context of this Insight, Chemical Space is defined as the total descriptor space that encompasses all the small carbon-based molecules that could in principle be created.

A definition which corresponds in many aspects to a better scientific proposal, than Kirkpatrick–Ellis one, but certainly naïve: as the author seems to ignore the possibility to describe molecular structures via quantum mechanical functions, thus making the possible Chemical Space definition infinite dimensional (therefore *really* huge, thinking by the intellectual standards of the previous references). Although the new definition provided by the reference [3] it is not free of curious logical bugs. For example: biological molecules are certainly carbon based molecules, but also some of them are far to be small, like some proteic structures of life science interest, which if the present author is not wrongly informed was the subject of the Insight. Also the sentence: ‘Chemical Space’ is a term often used in place of ‘multi-dimensional descriptor space’: it is a region defined by a particular choice of descriptors and the limits placed on them, induces some fuzziness, as the author do not seems to have a clear mind on what is a space and what a region of such space. May be the term subspace would have been more appropriate than region here. Such fuzziness demonstrates the confusion, which seems to be common to some users of Chemical Space concept, about everyday space where we live (including, of course, GU) and the abstract definition of vector space, certainly a vastly general conceptual idea, much better appropriate to contain the chemical structures, when described mathematically in discrete or continuous form, see for example references [4–7].

Nevertheless, from there it was found that another paper [8] was connected to the question of Chemical Space. Such study undertook a further “exploration” into the subject while talking now, perhaps in a better way than in the previous commented references, about a tenuously different subject: the Chemical Universe. Although, the previous confusion might be detected as present in this new study, because the paper title speaks of *exploration*, providing in this manner a plausible fuzzy connection of the Chemical Space again with the GU, the title appears written as:

> Virtual *Exploration* of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery

Still, reading the previous title it seems again that all the authors of this kind of works are impressed by the huge nature of the chemical compounds number possibility, or alternatively are trying to impress their readers, who still are not aware of the huge number of chemical structures possible, an obvious feature which any chemist might be aware since its scholar initiation. Reading the paper one realizes that, of course, generating molecular structures via heavy computing task is a feasible modern element. The paper describes how to generate the huge molecular database and the use one can perform of it, in order to test the generated structures for potential medicinal use. Certainly the universe of the paper title is the universe of discourse of all the structures with $\{C, N, O, F\}$ made of 11 atoms maximum.

Perhaps as a consequence of what has been said and read it will be wise adopting from now on within some appropriate context the term Chemical Universe (of discourse) instead of Chemical Space.

In order to peruse other possible sources of Chemical Space concept, a large amount of papers were retrieved in a search on the J. Med. Chem., among the most recent it has been pickup two, as anyone can retrieve the list, just typing Chemical Space in the journal web page search facility. In the first and modern reference [9], the concept of Chemical Space is employed in a plausible way as a vector space, even if the authors talk in the paper title about navigating it. Such kind of point of view makes one hopeful that the nonsensical feature attributed to Chemical Space will soon be corrected. While in a little bit older work [10], the authors remind of more of the same ideas as earlier quoted, as they write at the very beginning a copy of Kirkpatrick–Ellis foreword:

> "*Space is big. You just won't believe how vastly, hugely, mind-bogglingly big it is.*" Even though Douglas Adams' well known quotation relates to astronomy, these words are a striking description of Chemical Space. It is basically infinite, comprising all possible molecules, which has been estimated to exceed 10^{60} compounds even when only small…

These previous lines directly copied from the mentioned reference, constitute a nice example of repeated copies of the same information, which sometimes haunt scientific literature. So, it seems that even recently the same confusion about spaces is widespread in a great deal of printed papers. Yet another paper [11], two years older, taken as a sample from another medicinal chemistry journal, shows again a confusing point of view, as this novel author speaks now of Chemical Space *Travel*, which is a way to induce again the confusion of Chemical Space and GU or ordinary space to potential readers. Moreover, in this last paper another problem arises, as it uses docking among other items, to construct the research on his Chemical Space travel. However, in doing this, Chemical Space travel becomes kind of random walk, as it will be commented and explained next.

One can safely say this last sentence about this docking lead travel, because recently docking has been proven to be some kind of random result generator [12]. This last reference providing the readers with evidence of a digital chaotic behavior of docking procedures currently employed. In this recent study though, the authors try to salvage the docking procedures from the wreckage, which themselves have encountered in his endeavor to test different processes employed in this area within usual research.

In order to justify the randomly behaving docking procedures, the authors of reference [12] propose to rely on an intriguing question which arises in another reference [13], provided at the end of their study. Such a final statement appears as a final rescue boat or escape pod, trying to justify the obviously faulty docking programs behavior. Within this computational reference [13], the author apart of finding (rediscovering the obvious) the old knowledge about error accumulation in computational algorithms, which was known from the times of electromechanical computing machines, arrives at the astonishing conclusion that, due to error accumulation risk in modern computational procedures, a normal situation of *irreproducibility* might appear in scientific everyday life. A statement which certainly is contrary to scientific endeavor, as it is known from initial scientific thought up to date.

In digital machines there is no space for randomness, though. Even random number generation (see for example the report and references therein: [14]) becomes a hard task, due to the essential nature of digital computing, which today is fully deterministic. This situation will continue in this way, until someone demonstrates the contrary. See, for example, a manner on how to overcome the problem [15], with the unfaulty use of a new computer technology and behavior.

The author, at this moment of the present letter, confesses his puzzlement about such pseudoscientific magical borderline trends, which seems to become widespread in some parts of modern chemical science and around heavy computational tools misuse and abuse as well.

As a final exercise, the author will try to resume his feelings on the Chemical Space definition and use, as follows. After reading all these fuzzy contributions to the shaping of a chemical universal set or vector space concept, and the literature embranchments found, he will try to outline this letter in a similar way as Kirkpatrick–Ellis, but entering the non-scientific quotation at the end of it.

Moreover, the author will try to do it, shifting from quoting the literary-motion picture work of Adams, but keeping the reference about the deep space (in the sense of GU) connection. Now, for this purpose he will use the appearance of futuristic reality proposed by the well-known Star Trek science fiction series; a production enterprise by far larger than Adam’s contribution to the subject of presenting a fanciful GU. In this science fiction context, one can wonder on what Mr. Spock (a legendary character of renowned logical brain structure) will opine, reading about such publications as these mentioned here, after analyzing them and finally giving his judgment about Chemical Space. Probably the Vulcanite will subscribe the basic ideas of the present letter against scientific fuzziness and nonsense, no doubt.

Apology

The author asks for pardon and oblivion to all: editor, staff and readers of this serious scientific journal, for the free use of a not so serious, but popular science fiction environment history and character.

Acknowledgments This piece of work has been written during the stay of the author in Riken (Kobe, Japan). The outstanding hospitality of Professor Kimihiko Hirao is warmly acknowledged.

References

1. R. Carbó-Dorca, E. Besalú, Centroid origin shift of quantum object sets and molecular point clouds: description and element comparisons. J. Math. Chem. **50**, 1161–1178 (2012)
2. P. Kirkpatrick, C. Ellis, Nature **432**, 823–865 (2004)
3. C.M. Dobson, Chemical space and biology. Nature **432**, 824–828 (2004)
4. R. Carbó-Dorca, A. Gallegos, A.J. Sánchez, Notes on quantitative structure-properties relationships (QSPR) (1): a discussion on a QSPR dimensionality paradox (QSPR DP) and its quantum resolution. J. Comput. Chem. **30**, 1146–1159 (2008)

5. R. Carbó-Dorca, A. Gallegos, in *Quantum similarity and quantum QSPR (QQSPR), Entry: 176*, ed. by R. Meyers Encyclopedia of Complexity and Systems Science, vol. 8 (Springer, New York, 2009), pp. 7422–7480
6. L.D. Mercado, R. Carbó-Dorca, Quantum similarity and discrete representation of molecular sets. J. Math. Chem. **49**, 1558–1572 (2011)
7. R. Carbó-Dorca, E. Besalú, Shells, point cloud huts, generalized scalar products, cosines and similarity tensor representations in vector semispaces. J. Math. Chem. **50**, 210–219 (2012)
8. T. Fink, J.-L. Reymond, Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. J. Chem. Inf. Model. **47**, 342 (2007)
9. H. Lachance, S. Wetzel, K. Kumar, H. Waldmann, Charting, navigating, and populating natural product chemical space for drug discovery. J. Med. Chem. (2012). doi:10.1021/jm300288g
10. J. Rose'n, J. Gottfries, S. Muresan, A. Backlund, T.I. Oprea, Novel chemical space exploration via natural products. J. Med. Chem. **52**, 1953–1962 (2009)
11. R. Van Deursen, J.-L. Reymond, Chemical space travel. ChemMedChem **2**, 636 (2007)
12. M. Feher, C.I. Williams, Numerical errors and chaotic behavior in docking simulations. J. Chem. Inf. Model. **52**, 724–738 (2012)
13. K. Diethelm, The limits of reproducibility in numerical simulation. Comp. Sci. Eng. **14**, 64–72 (2012)
14. R. Halprin, M. Naor, Games for Extracting Randomness. A study from the Dept. of Computer Science and Applied Mathematics. Weizmann Inst. of Science
15. I. Kanter, Y. Aviad, I. Reidler, E. Cohen, M. Rosenbluh, An optical ultrafast random bit generator. Nat. Photon. **4**, 58–61 (2010)