Mereology of Quantitative Structure-Activity Relationships Models

Guillermo Restrepo and Rom Harré

Abstract: In continuing with the research program initiated by Llored and Harré of exploring the part/whole (mereological) discourses of chemistry, we analyse Quantitative Structure-Activity Relationships (QSAR) studies, which are widespread approaches for modeling substances' properties. The study is carried out by analyzing a particular QSAR model, and it is found that different mereologies are needed: from those regarding bulk substances as wholes and molecular entities as parts and to mereologies where the wholes are molecules whose parts are atoms, structured subsets of atoms, nuclei and electronic densities. We suggest a relationship between successful QSAR models and a deep understanding of the mereologies used and the ways they are intertwined. We note that QSAR modelers prefer the mereology of substance-molecule and then discuss how that is related to simplicity and computational capacity. Historical questions are opened, *e.g.* how the mereologies of substances have changed over time? and why they are mostly oriented toward organic chemistry?

Keywords: *mereology*, *Quantitative Structure-Activity Relationships*, *substances*, *molecules*, *graphs*.

1. Introduction

Mereology is the theory of part-whole relations, *i.e.* the relations of part to whole and the relations of part to part within a whole (Harré 2015). Examples of mereological relationships are: carbon is part of methane, stirring is part of a chemical reaction, alkanes are part of organic substances, molecules are part of substances, electrons of molecules, to name but a few. Mereological discourses, as noted by Harré (2015), are at the core of chemistry and their study sheds light on the kind of reasoning that shapes the discipline. A major question for mereological studies concerns the distinction between predicates that can be ascribed both to a whole and to its parts, such as mass, and predicates which cannot, for example the color of a substance cannot be ascribed to its molecular constituents.

Just recently philosophers of chemistry have started to pay attention to such discourses, where the results by Needham (2005), Earley (2005) and Harré & Llored (2011, 2013), Harré (2015), Llored (2014) and Llored & Harré (2014) are worth mentioning.

2. Formal Rules for Mereological Reasoning

A usual description (Varzi 2014) of mereology is by taking parthood as a primitive, which is based on the following statements:

Everything is part of itself.
Two distinct things cannot be part of each other.
Any part of any part of a thing is itself part of that thing.

Which are formalized as a binary relation meeting reflexivity, antisymmetry, and transitivity, *i.e.* a partially ordered set. In the current work we use a relaxed parthood relation that may fulfil some, or all, of the above conditions. If we say that x is part of y and we write it as xPy, then the mentioned conditions are:

xPx (reflexivity)
xPy and yPx, then $x=y$ (antisymmetry)
xPy and yPz, then xPz (transitivity)

Some other definitions of importance for the ensuing discussion are:

$x=y := xPy$ and yPx (equality)
$xPPy := xPy$ and $\neg\, x=y$ (proper parthood)
$xOy := \exists\, z, zPx$ and zPy (overlap)
$xPPy$, then $\exists\, z, zPPy$ and $\neg\, z=y$ (company)
$xPPy$, then $\exists\, z, zPPy$ and $\neg\, zPx$ (strong company)
$xPPy$, then $\exists\, z, zPy$ and $\neg\, zOx$ (supplementation)
xPy and $\neg\, yPx$, then $\exists\, z, zPy$ and $\neg\, zOx$ (strict supplementation)
$Ax := \neg\, \exists\, y, yPPx$ (atom)

With this mereological background at hand, we now follow Llored & Harré (2014) suggestion that philosophers of science should be closer to the objects they study, which in fact they exemplified by analyzing the mereology of quantum chemistry studies by examining the actual procedures of quantum chemists. In the current paper, we analyze the mereological aspects of Quantitative Structure-Activity Relationships (QSAR) models, a widely used method for estimating the properties of substances by combining, mainly, chemical and mathematical insights.

3. Modeling in Chemistry

Models are created by establishing a linkage for the transfer of predicates between a source, which the model makers draw on, and a subject, which the construction is a model of (Harré 2004). Some examples of models in chemistry are molecular formulas (Schummer 1998) and chemical reactions (Fialkowski *et al.* 2005). Some others are found in three special issues (5.2, 6.1 and 6.2) of the journal *Hyle* devoted to the subject. In a molecular model the subject is a molecule and the source is a possible assembly of atoms along with their connectivities and sometimes with their spatial coordinates. One should note that these models are already mereological, consisting of parts assembled into wholes. Models of chemical reactions consider substances, whether or not they participate in chemical reactions, and sometimes concentrations as source; the subjects, at least initially, were chemical transformations occurring in natural processes.

Other models in chemistry use functional thinking (Restrepo & Villaveces 2012) and start by looking for the relevant variables characterizing the subject and by symbolizing them; in the final step variables are related to each other by functions. One of the most widespread cases of these models is the estimation of substances' properties on the basis of descriptors, where the subject is a property P of some substances and the source is a set of descriptors d. A descriptor is any experimental or theoretical feature characterizing the substance; some examples are melting and boiling points, or algebraic calculations on graphs associated with the substance. Finally, the model takes the form $P = f(d)$, f being a function. In modeling terms this modeling function maps the d predicates onto the space of possible predicates. That is it expresses a formal analogue representation.

A widespread framework for estimating substances' properties through descriptors is that of Quantitative Structure-Activity Relationships (QSAR), where a property (response variable) of a set of substances is modeled by representing substances as molecules, which are further represented by descriptors. These descriptors are worked out by different mathematical and statistical approaches to find a suitable function f for the relation $P = f(d)$. In the following we explore the details of these models from a mereological perspective.

4. Quantitative Structure-Activity Relationships (QSAR)

In a QSAR model[1] the interest is on modeling a response variable, *e.g.* mutagenicity or octanol/water partition coefficient, that is creating a formal analogue of this variable. In general, every QSAR study involves the following steps (Todeschini & Consonni 2009):

1. Selecting the response variable.
2. Selecting substances to run the study.
3. Characterizing of substances through the characterization of molecules.
4. Splitting of the set of relevant molecules into training and test sets and the selection of molecules for external validation.
5. Model development.
6. Assessment of the model.[2]

Based upon the characterization of models through their subject and source, Harré (2004) has classified them as homeomorphic and paramorphic. The former have the same type of predicates in both subject and source, *e.g.* the scale model of an air-plane which has wings, fuselage and so on analogously to the real air-plane; or a mathematical model of a property based on abstractions from other (experimental) properties. Paramorphic models have other types of predicates in the subject than in the source, by virtue of the modeling link, as for example in the models of the behavior of gases through the physics of a number of interacting, moving billiard balls, or mathematical models whose subject is an experimental property, say boiling point, and whose source is a set of theoretical descriptors.[3]

Before continuing, we consider it important to discuss the mereological aspects of general approaches for modeling substances, which are a key point in QSAR studies.

5. Mereological Aspects of Modeling Substances

5.1 Mendeleevian mereology

A part/whole chemical relationship that is at the core of chemistry is that of element/substance. Mendeleev made the point that substances that display or are ascribed properties (simple substances) are made of basic substances (transcendental elements, devoid of properties) and which are uniquely characterized by a number – once the atomic weight, but currently the atomic number. As Mendeleev stated, "[T]he atomic weight does not belong to coal or to the diamond but rather to carbon" (Mendeleev 1869). This entails a mereology where the simple substances are the whole and the basic substances (chemical elements) the parts. This is a mereology meeting reflexivity (an element is part of itself) and atom conditions (oxygen has no parts – it is not made of other elements) but lacking antisymmetry and transitivity, given the atomic character of the parts, which also lacks company and supplementation. Although this mereology is of great importance for chemistry, it is too general and needs further refinements to cope with contemporary chemistry, rooted in a structural chemistry tradition, where molecules play a central role.

5.2 Substance-molecule mereology

In this mereology, quite widespread in chemistry textbooks and contemporary chemical discourses, substances (simple substances of Mendeleevean mereology) are the wholes and molecules the parts; but importantly, there is a one-to-one relationship between substances and molecules. This mereology is reflexive (a molecule is part of itself) and atomic (a molecule is not made of other molecules). It is neither antisymmetric nor transitive and lacks company and supplementation given the atomic character of its parts. However, as Schummer (1998) has pointed out, it is easy to see how rough the one-to-one assumption between molecule and substance is with the case of water, where the substance has several molecular species associated with, *e.g.* H_2O, H_3O^+, OH^- and clusters $(H_2O)_n$ with n taking different integers as values (Ludwig 2001). This leads to a more refined mereology of molecules and substances.

5.3 Substance-molecules mereology

Here substances are the wholes and molecular species the parts, but in a relation one-to-many, as explained before for the case of water. To explore the properties of this mereology, let us take x, y and z as molecular species. As xPx, reflexivity is met; xPy and yPz leading to xPz indicates transitivity, which is the case of, *e.g.*, $H_2OP(H_2O)_3$ and $(H_2O)_3P(H_2O)_{12}$, with $(H_2O)_3$ and $(H_2O)_{12}$ being molecular clusters. This also shows that the parts of this mereology are not atoms. In addition, if xPy and yPx, it is true that $x=y$; for overlapping is allowed, as seen in the case of H_2O, $(H_2O)_3$ and $(H_2O)_{12}$. It also meets the demand of company, *i.e.* $H_2OPP(H_2O)_{12}$, implying that there is, for example, a $(H_2O)_3$ such that $(H_2O)_3PP(H_2O)_{12}$ and $(H_2O)_3$ is not the same as $(H_2O)_{12}$. However, there is no supplementation, for $(H_2O)_3$ does overlap with H_2O.

Anyhow, most models in chemistry simplify this rich substance-molecules mereology to the substance-molecule one, and QSAR models are no exception.

The two latter mereologies and their strong reliance on molecules have led to the whole being replaced by the parts in chemical discourses. Such is the case of talking about the search of molecules with, *e.g.* antibreast cancer activity or of gaseous molecules. Perhaps authors really mean substances with antibreast cancer activity and molecules belonging to the gaseous phase of a substance, but we have observed that the confusion really exists at university chemistry courses and not surprisingly outside the university as well.[4] This interchange of wholes by parts is what Harré & Llored (2013), following Hacker & Bennett (2003), have called a mereological fallacy, *i.e.* assigning predicates the meaning that is determined by their role in describing wholes to parts of those wholes or vice-versa. Some such assignments are legitimate but some are not. There does not seem to be a general criterion for making the distinction, and so far this problem has been solved case by case.

The question arising is what kind of results and what kind of science would come about by using a different mereology, perhaps the substance-molecules one. This would entail more computational capacity, for now a substance is modeled on the basis not of atomic parts but of overlapping ones, as in the case of water modeled by their different overlapping molecular species. Perhaps the time has come with the current computational capacity and the hope of improving it to a large extent. However, one needs to take care of Borges' demon, *i.e.* developing models of the size of the modeled object, as Borges warns in his famous tale (Borges 2013). This latter risk is not taken with the substance-molecule mereology, but given its simplicity it loses more information than the substance-molecules one, for example lacking interpretation of liquid properties of substances, *e.g.* water (Ludwig 2001).

6. *Modus operandi* and mereological aspects of QSAR modelling

As the aim of the current paper is to explore QSAR *modus operandi* with mereological eyes, we analyze a recent paper (Luo *et al.* 2014) on the subject, which we think reflects and brings together several of the traditions of the QSAR community and which help us understand the mereologies implicit in developing the model. In the following we describe, step-by-step, the procedure followed in such a study.

1. The substance 5-hydroxytryptamine, commonly known as serotonin, is a neurotransmitter acting upon neurons involved in processes such as emotions and memory. Serotonin receptors are cell membrane proteins that detect substances outside the cell and activate cellular responses; the particular serotonin 1A receptor binds serotonin to it. As this receptor is found in brain regions with functions involved in mood and anxiety disorders, it has been studied as a target for antidepressant drug discovery (Roth *et al.* 2000, Luo *et*

al. 2014). In fact, several drugs like buspirone and tandospirone are agonists of this receptor (*ibid.*), *i.e.* these drugs bind to the receptor activating a biological response, which, depending on the drug, causes more or less efficacy of the receptor 'tuning' the biological response.

Hence, knowing which substances have affinity by the 5-hydroxytryptamine 1A receptor is of interest for medicinal chemistry, as novel drugs for treating mood and anxiety disorders such as schizophrenia can be developed. The authors of the discussed paper look for QSAR models of 5-hydroxytryptamine 1A receptor binding activity, using data retrieved from the PDSP Ki database.

The selection of the endpoint to model, in this case 5-hydroxytryptamine 1A receptor binding, depends on the experience of the researchers on the subject and on the access to information for developing the model, normally a reliable and well managed/curated database, if possible. In this case the information came from the National Institute of Mental Health Psychoactive Drug Screening Program Ki database, which contains information on the receptor binding of several substances. Hence, the raw material for the study comes from an external repository, built up with knowledge from different sources and from different experimental approaches; that is the reason why the database needs to be refined.

Here the whole is the binding between the receptor (R) and the molecules (ligands, L), *i.e.* R-L. This whole is understood as an assemble where L can be substituted by different molecules, then it is a mereology where one of the parts is constantly changing. This mereology meets reflexivity, for ligands and molecules are part of themselves; and it is atomic, with R and L as atoms. But it is not antisymmetric and transitive, for overlapping between R and L is not allowed, which also excludes company and supplementation.

2. With the source of information at hand, the next step is to decide which substances to retain for further modeling.[5] In the current case that is to look for a criterion to decide when a molecule is either active or inactive regarding its binding to the receptor, which is customarily done through dissociation constant arguments.

If the binding is understood as the relation $R + L \leftrightarrow RL$, the dissociation constant is given by $K_i = [R][L]/[RL]$, with $[x]$ indicating the molar concentration of x. Hence, the smaller the dissociation constant, the stronger the binding R-L.

The cut-off value to define active versus inactive molecules was set to 10 μM, a decision that depends on the experience of the researchers, which includes, *e.g.* knowledge of the property to model and number of molecules to treat, which is also related to computational capacity. In former QSAR studies, less than one hundred molecules sufficed for the model, but now, with more computational power and more information, it is a common practice to work with hundreds and thousands of compounds.

The cut-off value is also chosen taking into account the number of molecules selected for the study (actives), which implies dropping some others. Hence, there is no general principle or criterion to select the substances for the model and the decision on which substances to include depends on the context of the study and on the resources to carry it out.

The above procedure led to 180 active receptor compounds, to which 78 inactives were added for statistical purposes. Here the authors decide on the number of inactive compounds to use, a decision that depends on the number of active compounds that have been selected. The idea is to have instances of inactive compounds to be able to develop a model that differentiates between active and inactive compounds regarding the receptor.

As seen, the mereology is a substance-molecule one, which is context dependent, for both the whole and the parts are negotiated.

3. It is important to consider also external data and compare the results for validation purpose, 66 additional actives were extracted from the World of Molecular Bioactivity (WOMBAT) database, which is different to the one of the source data.

In former QSAR studies this step was omitted. Tropsha, one of the authors of the analyzed paper, has repeatedly shown that lacking this step may lead to weak models, in statistical terms (Tang *et al.* 2009, Shen *et al.* 2004). However, even with Tropsha's warnings, there are studies that did not consider external validation. In short, this step depends on the experience of the researchers and on their ability to negotiate their results with the community.

The number of external compounds is normally decided taking into account the number of compounds used in the model, and the selection of the alternative source of information is given by the knowledge of the researchers of other databases. Hence, social aspects of science turn out to be important, for database developers need to let researchers know about their resources such as researchers need to communicate their needs of information. Scientific meetings, sometimes sponsored by database developers, scientific publications, and informal scientific communication are examples of these forms of communication (Björk 2007).

The authors mention that all 66 WOMBAT compounds were structurally different from the compounds of the modeling set. That is important, for in current QSAR studies the trend is to have a diverse set of molecular structures to develop general models rather than one based on a particular family of closely related molecular structures (Basak 2014), which was a typical feature of early QSAR models.

However the question is how the authors determined structural differences. As they did not mention that, it is likely that personal experience was used. Even if they would have mentioned it, it would have been a decision of the researcher. Developing methods for assessing molecular similarity and molecular diversity is an own subfield of chemoinformatics (Leach & Gillet 2007). Hence, the subject of molecular diversity involves much know-how and many pragmatic compromises.

4. Even with a refined database as source of information, the selected molecules were 'curated', as it has been found that even the most reliable databases contain errors (Fourches *et al.* 2010). Small structural errors within a data set may lead to significant loss of the predictability of QSAR models. The process of refinement started by removing inorganic or organometallic compounds, because most of the molecular descriptors can only be computed for organic molecules. This reflects the kind of problems QSAR researchers and descriptor developers want to address, *i.e.* those of organic chemistry. An open question is on the lack of interest for inorganic and organometallic chemistry. Perhaps part of the reason is that, today, QSAR practitioners are closer related to the organic chemical industry than to other fields.

In mereological terms this organic bias implies that the parts of the mereology are also biased, not homogeneously covering the space of compounds, but a reduced region, which is oriented to a big extent by the exploration made by the pharmaceutical industry. In addition, the bias has a pragmatic reason related to the mereology in which it is framed. A mereology of substance-molecule entails the definition of a molecule, which is understood as a connected finite set of atoms. As descriptors calculation requires computation, the number of atoms is always a constraint, which makes it easier to compute small assemblies, typical organic molecules, rather than lattices, typical of inorganic chemistry. On the other hand the connections among atoms also need to be defined, which are set up through ideas of chemical bonding. Traditionally, perhaps determined by the number of atoms, these connections have been understood as covalent bonds, disregarding other sorts of bonding, which also helps to disregard non-organic compounds.

The authors excluded inorganic substances by using an algorithm for detecting none-main group elements. Mixtures were also algorithmically removed because descriptors were calculated for single molecules, whereas the current chemoinformatics used for encoding molecules, *i.e.* SMILES (a text string), can also encode more than one molecule in a single representation. The SMILES data were then converted into 2D molecular graphs (where atoms are vertices and bonds are edges), as it has been found that calculation of descriptors from SMILES sometimes introduces errors (Luo *et al.* 2014). The graphs were normalized to attain a unique representation of same functional groups, which can also be made by home-made algorithms and commercial software. The authors, however, recommend manual working over of the original list, for there are reports of errors made by those algorithms.

At this point a further mereological refinement is made, for a molecule (now the whole) is made of several parts (graphs for the molecule[6]). This is the case, *e.g.*, of the tautomeric possibilities of a molecule where each tautomer may lead to different descriptors. This mereology is reflexive, given that if G_i, G_j and G_k are three graphs for the same molecule, it is found that G_iPG_i, which in graph theory is called a graph isomorphism,[7] *i.e.* G_iPG_i because G_i is isomorphic with G_i, which is called an automorphism. The mereology is antisymmetric and transitive, given that if G_iPG_j and G_jPG_i, then $G_i=G_j$; and if G_iPG_j and G_jPG_k, then G_iPG_k, as subgraphs,[8] *e.g.* backbones of tautomers, are common to graphs representing the molecule. The company requirement is also met, as G_iPPG_j implies the existence of G_k such that G_kPPG_j and it does not hold that $G_k=G_j$. Strong company, supplementation, and atoms are not part of this mereology, for there are common parts (overlapping) of graphs through common subgraphs.

What is sought in QSAR studies is to attain a similar substance-molecule mereology, where a unique graph (with its embedded subgraphs) is associated to a SMILES. The question arising is how to select such a unique graph. What is done is that the selection is driven by chemical experience and knowledge of the context of study, *e.g.* pH conditions of the modeled situation, in this case *R*-*L* interaction (Young *et al.* 2008).

Another step in the refinement and culling of the original list is the removal of duplicate structures, which can be done by using canonical SMILES and then by algorithmically comparing the strings. But this brings the question of which duplicate to keep, which implies, so far, a manual analysis of the properties of each duplicate to gather the information of all duplicates in a single molecule. Finally, the authors recommend a manual checking of the information. In the paper the structures retained after refinement of the list were 166 unique organic compounds (105 active and 61 inactive).

As seen refinement and culling is a process involving both automatic and manual procedures, which require several decisions, pragmatic compromises and further mereological refinements that mix mereologies of chemistry and mathematics.

5. The next step was the selection of training, test, and external validation sets, which was performed employing the external 5-fold cross-validation protocol, where the modeling set is randomly split into five subsets of approximately equal size. Each time, one subset is used as an external validation set, while the union of the other four subsets is used as the modeling set. The different modeling sets were further partitioned into multiple pairs of representative training and test sets of different sizes using the Sphere Exclusion algorithm developed by the authors.

As noted, the authors used particular methods to achieve their goals, but there are other methods, also derived from statistics, that are currently used in QSAR studies (Leach & Gillet 2007). All of them look for an unbiased way of grouping chemical substances. Hence, it is a decision of the authors which method to select.

6. The molecular descriptors for the curated molecules were calculated using Dragon (Tetko *et al.* 2005), a package of several commercially or freely available programs (Gugisch et al. 2014). There are more than one thousand molecular descriptors, *i.e.* more than thousand ways of numerically encoding particular features of the molecular graph (Todeschini & Consonni 2009). They range from constitutional ones, which count or assess the presence of particular functional groups or atoms or rings, to more elaborated descriptors, such as quantum chemical ones that depend on the level of theory used to compute them. There are also topological descriptors that mainly consider the connectivity of atoms in molecules. In the end, the decision which software is used to calculate descriptors depends on the number of descriptors available in the software, the computational cost of their calculation, whether the software is commercial or freeware, and on several other aspects that are part of the researchers' experience.

This step involves a further mereological change. Now the whole is the molecular graph, made by vertices (labeled atoms) and edges (bonds), with edges defined as couples of atoms that are connected. This mereology meets reflexivity, for a vertex is part of itself; but it does not display antisymmetry, because if i, j and k are vertices and $\{i,j\}$, $\{j,k\}$ and $\{i,k\}$ are edges, then $iP\{i,j\}$ holds but $\{i,j\}Pi$ does not; likewise transitivity is not met. It entails company, for if $iPP\{i,j\}$, it implies the existence of a j such that $jPP\{i,j\}$ and $\neg\ i=j$.[9] It also holds strong company as $iPP\{i,j\}$ implies the existence of a j such that $jPP\{i,j\}$ and $\neg jPi$. It meets supplementation: $iPP\{i,j\}$ implies the existence of a j such that $jP\{i,j\}$ and $\neg jOi$, *i.e.* there is no vertex k for which kPi and iPj. For a similar argument based on the absence of overlapping, the mereology is also strictly supplemented. Finally, it is a mereology without atoms because vertices are part of edges.

A further refinement of the above graph mereology, of importance for molecular graphs, is a mereology of subgraphs (see note 8) as parts of graphs, which is different from the already discussed mereology of molecule representations as made of graphs. If G is the graph and S_i, S_j and S_k are three of its subgraphs, the mereology fulfils reflexivity (S_iPS_i), antisymmetry (S_iPS_j and S_jPS_i implies $S_i=S_j$) and transitivity (if S_iPS_j and S_jPS_k, then S_iPS_k). It also meets company (if S_iPPS_j, then there is a S_k such that S_kPPS_j and $\neg S_k=S_j$), supplementation (if S_iPPS_j, then there is a S_k such that S_kPPS_j and $\neg S_kOS_i$) and strict supplementation (if S_iPPS_j, then there is a S_k such that S_kPS_j and

$\neg S_k O S_i$). There are no atoms in this mereology, given the overlapping among subgraphs.

Descriptors from molecular graphs are derived by mapping any of the graph mereologies onto real numbers. Once the descriptors are calculated, a decision is made on which ones to retain and which ones to drop. In some models, when the molecular features ruling a substance's property are roughly known, the descriptors are directly selected by the researcher, *e.g.* the number of rings and electronic properties in hepatotoxicity studies. In the paper, descriptors with zero values or zero variance were excluded and the retained ones were normalized. The way of selecting descriptors is also a matter of taste, here variance was used as the criterion, but other statistical measures may also be used, *e.g.* information content (Todeschini & Consonni 2009). In general, variable selection methods are used to select a subset of relevant descriptors for the studied property. The step of normalization is also a decision, because one may decide to run the study with raw descriptors, and even if one decides to avoid the different scales amongst descriptors, the question is whether to normalize or standardize and how to do it. In this procedure chemistry, computer science, and statistics are interacting; knowledge on data distributions and statistics to deal with them in a nice computationally smooth procedure, as well as time and cost are the most important factors. In fact, it is so important that it is customary to mention the specifications of the computers used in the study as well as the time for running the process, the algorithmic complexity is sometimes also mentioned. In the analyzed paper, computer features and computing times were published.

7. In the next step QSAR models are developed. There are different kinds of mathematical approaches to finding f in the sought for $P = f(d)$ relation (Todeschini & Consonni 2009). The authors used k Nearest Neighbours, Random Forest and Support Vector Machines as mathematical approaches. Once a model is generated it is important to assess whether the model is stable, *i.e.* if it yields similar results by perturbing it, for example by adding noise to the descriptor values or by introducing random values in place of them. The stability of the models was assessed through a Y-randomization test.

The choice of the method to analyze the validity of the models is based on the experience of the researcher, including statistical knowledge, and further decisions to be made while applying the validation method. Regarding the kinds of models to use, even if the authors used three, there are by far more possibilities (Todeschini & Consonni 2009), some simpler in mathematical terms and some more elaborated. The point here is that the researcher decides which model to use, which is basically rooted in experience. A decision that has to be made is if the QSAR model is intended to give insight into the relationship between molecular structure and activity or if the model is used only as a tool to estimate the endpoint. In the current case, and in many contemporary QSAR studies, the aim is estimation. In former QSAR studies there was more interest in interpreting models on chemical and physical grounds, an approach that has been gradually abandoned given the difficulty in interpreting mathematical combinations of descriptors (models) and even in the interpretation of descriptors. Currently, such interpretations are considered a plus of the models (see guidance document of the OECD 2007 for QSAR). It is worth noting that QSAR approaches relate structure to activity whereas the reverse relation is very difficult to establish by these methods.

8. The models with satisfactory results in the internal and external validation step were then used for virtual screening, which consists of exploring (screening) large databases of chemicals (libraries) that include information about one or several relevant properties. In this case the databases consisted of substances with potential 5-hydroxytryptamine 1A receptor binding activity. Hence, once the databases are selected, the QSAR models are run over the substances of the databases to estimate their 5-hydroxytryptamine 1A receptor binding activity and finally come up with candidate substances for experimental testing.

However, QSAR models are not of general applicability, *e.g.* models developed for alkanes are not suitable for heterocyclic compounds. Before applying a QSAR model to an external set of substances, it is important to determine whether it is able to yield reliable results for the external substances. Those substances for which the model is able to yield reliable results are part of the applicability domain of the model. Applicability domains are rooted in the similarity between the substances used to develop the model and other substances, where, once more, the different ways of assessing molecular similarity come into play.

In the discussed paper the authors used three sorts of libraries to run the screening. The selection of these libraries depended on previous knowledge of the authors and on the access they had to them, because sometimes they belong to pharmaceutical companies that provide access only on the basis of a contract or a common research project. In these cases the researcher must negotiate the access to the information and needs close contact to industry. The applicability domain of the models was calculated using a fingerprint based similarity approach, which characterizes molecules as a string (fingerprint) over which a distance function is run to measure the nearness of two structures.

The similarity characterization of molecules implies an important decision; the authors selected fingerprints, but there are many other ways of doing it (Leach & Gillet 2007), *e.g.* by using the same molecular descriptors.

Even if fingerprints were the single form of characterization, the question is what kind of dictionary is needed to build up the fingerprint. These dictionaries are collections of molecular fragments that are searched in the molecule to build up the fingerprint, which in its simple version is a binary vector of presence/absence of information (Leach & Gillet 2007). But there are many dictionaries,[10] and the selection of the appropriate one depends on the molecules one is dealing with, and even on the access to the dictionary, for some of them are owned by companies. Now, supposing that the issue of the representation is solved, the next decision is on the distance function used to determine the nearness of the molecules. There are many of these functions and it is a decision of the researcher which one to choose, on the basis of the derived knowledge on the performance of those functions over particular types of data or for a particular kind of structures and several other factors (Todeschini & Consonni 2009, Todeschini *et al.* 2012, 2015).

Applicability domains are based on geometrical descriptions of the space of descriptors and on some other mathematical ideas, which in the end reduce subjectivities but do not make them vanish. In this particular case, the applicability domain was based on an Euclidean distance similarity threshold, but there are other methods to define it (Ellison *et al.* 2011).

In essence, all QSAR approaches imply, directly or indirectly, a simple similarity principle: compounds with similar structures are expected to have similar biological activities. A fundamental concept, derived from this principle, is the one of pharmacophore (Van Drie 2007), which is an abstraction of what is structurally needed for a molecule of a drug to cause a pharmacological effect (Wermuth *et al.* 1998). In mereological terms, the whole is the molecule and the part is the pharmacophore. When claiming that the molecule (whole) has a biological property and that a fragment (part) is responsible for that property, what is in play is a mereological discourse.

9. The final part of the study involved the experimental testing of the binding of the selected molecules from the libraries with the 5-hydroxytryptamine 1A receptor.

The authors finally suggest 15 substances from the three screened databases, which were experimentally tested on binding assays, obtaining nine actually active substances with binding affinity lower than 10 μM.

Here the binding tests involve also decisions on the kind of techniques used to run the competing binding assays. Beyond that, this step includes a novel communication, this time between QSAR modelers and experimentalists, in this particular case chemists and biochemists.

Conclusions

QSAR models of the properties of substances, exemplified by the case here studied, need different mereologies, for the wholes and their parts do not always agree. The whole, in principle, is a bulk substance, with experimental (wet-lab) measured properties. But this whole depends on the context, *e.g.* chemical, when the modeled property is a relational one amongst substances, as is the case of solubility in certain liquids; thermodynamic context, when the modeled property is boiling point, for example; biological context, when the property is a relational property between the substance and an organism, as the LD_{50} (Schummer 1998).

As a first approximation, the parts of the whole are single molecules associated to the bulk substance, in this case the whole is regarded as made of many equivalent parts. When this approximation does not enable the derivation of reliable models, a second level for the whole involves the further study of the associated molecular entities of the bulk substance, *e.g.* tautomeric forms or clusters. The selection of the molecule (the parts) to represent the whole is now based on a deep knowledge of the property to be modeled and on the bulk substances, *e.g.* pH, temperatures, cellular external conditions of the substances or even physical parameters associated to molecules such as volume in studies of interactions between molecules and pockets or particular regions on a protein.

Once the whole and the parts have been set up, the parts now become the new wholes and are studied following another mereological approach, namely the one of the molecular descriptors. Here a wealth of descriptors is at hand which consider the whole as constituted by atoms (parts), structured subset of atoms in the form of functional groups or ring systems, assemblies of nuclei and electronic density in quantum chemistry mereologies, where also molecular orbitals constitute parts of the whole in a sense. There are also mathematical mereologies where the molecules are regarded as composed of atoms (vertices) and pairs of atoms (edges), which both constitute a graph, in this case the whole is the molecular graph and the parts are the sets of atoms and atom pairs.

Different mereologies show up to treat a single target, *i.e.* modeling substances' properties, and how they meet different properties. Some of them are simple, *e.g.* substance-molecule mereology, which meets reflexivity and atom conditions as well as Mendeleevean mereology. Other mereologies are richer in their properties as the molecular graph-labeled atoms and bonds mereology that has reflexivity, company, strong company, supplementation and strict supplementation. These results show the complexity and richness of chemical discourses and how chemists move across different mereologies.

We also pointed out the different decisions researchers make when modeling substances' properties through QSAR methods, which are instances of the social and pragmatic aspects of chemistry.

There are a wealth of QSAR models for different properties, for different substances, with different approaches and many other variations, which in the end are tested by exploring their ability to estimate properties or to bring new knowledge and understanding of the processes modeled. Perhaps, part of the key for satisfactory models lays on the clear definition of the used mereologies, their limits, and the ways of intertwining them.

There is a reduction/emergence debate in philosophy of chemistry (Hettema 2013). The current paper contributes to the debate by showing how the reduction of substances, even to mathematical concepts, is currently used to come up with insight on the behavior of substances, in this case receptor-ligand interactions, which are experimentally quantified by measuring concentrations of bulk substances.

From the same perspective of the reduction/emergence debate it is worth noting that QSAR modelers prefer a mereology of substance-molecule rather than the finer one of substance-molecules, a trend resulting from an offset between computational capacity and simplicity. It would be interesting to explore how alterations of these two factors favor the use of the substance-molecules mereology and how chemical discourses change thereby.

By discussing mereological discourses used in chemistry, a question arising is on how those mereologies have changed throughout the history of chemistry and which factors have influenced them.

We have pointed out the sociological aspects behind QSAR models; in this respect a question to be solved by the history and sociology of chemistry is why and how QSAR models began to focus on organic chemistry. Which conditions would have been needed to develop QSAR models for inorganic chemistry?

Initial QSAR modelers looked for mechanistic interpretations of models. However, since contemporary chemistry is oriented to structural chemistry, why have those mechanistic interpretations been practically abandoned?

Besides QSAR models, there are many other chemical discourses worth studying through mereology, which may help understand the kind of reasoning involved in chemistry. We hope the current work motivates other similar studies.

Acknowledgements

Guillermo Restrepo thanks the Universidad de Pamplona and the Alexander von Humboldt Foundation/Stiftung for supporting this research.

Notes

[1] More general models, of which QSAR models are part, are Quantitative Structure Property-Relationships, which are open to any substance's property and not only restricted to biological activities.

[2] These steps may be combined with others, *e.g.* virtual screening of databases to select substances for biological tests.

[3] QSAR models are more oriented toward paramorphic ones, for they are less expensive and even more environmentally friendly.

[4] In a rapid search on the Internet, we found that 'antibreast cancer molecules' has several entries, many of them in scientific journals. See for instance the abstract of Oliveros-Ferraros *et al.* (2011) or Zaki *et al.* (2012). It is also found that 'gaseous molecules' is a section of the scientific journal *Chemical Physics Letters*.

[5] The QSAR case aims at classifying substances (active/inactive) and the criterion for such classification is the cut-off value. However, for QSAR models where the aim is estimating the value of a property, where no classification is necessarily required, no cut-off value or any other threshold is needed.

[6] In QSAR studies, if the descriptors are calculated from molecular graphs or directly from SMILES, there are some descriptors that require to convert the graph into a 3-dimensional assembly, *i.e.* the traditional stick-and-ball depiction of molecules, were geometry is important. This entails further decisions, now on the way to convert the graph into the geometrical assembly, *e.g.* the selection of the force field to apply.

[7] If G and H are two graphs, they are isomorphic if there is a bijection between the vertex set of G and the one of H, such that any two vertices of G are adjacent (there is an edge for them) in G if and only if the bijection keeps the adjacency on the image of the vertices' bijection.

[8] A graph S is a subgraph of graph G if vertices of S are a subset of the vertices of G and if edges of S are a subset of the edges of G.

[9] However, company is not attained by graphs with loops, *i.e.* having $\{i,i\}$, which are not used for molecular graphs, as discussed here.

[10] It is also important to know the kinds of molecules to characterize, for the dictionaries may be created for particular subsets of molecules.

References

Basak, S.C.: 2014, 'Molecular similarity and hazard assessment of chemicals: a comparative study arbitrary and tailored similarity spaces', *Journal of Engineering, Science & Management Education*, 7, 178-184.

Björk, B.C.: 2007, 'A model of scientific communication as a global distributed information system', *Information Research*, **12**(2) paper 307 [Available at http://InformationR.net/ir/12-2/paper307.html, accessed 10 May 2015].

Borges, J.L.: 2013, 'Del rigor en la ciencia', in: *El hacedor*, Vintage, New York, p. 137.

Earley, J.E: 2005, 'Why there is no salt in the sea', *Foundations of Chemistry*, 7, 85-102.

Ellison, C.M.; Sherhod, R.; Cronin, M.T.; Enoch, S.J.; Madden, J.C. & Judson, P.N.: 2011, 'Assessment of methods to define the applicability domain of structural alert models', *Journal of Chemical Information and Modeling*, **51**, 975-985.

Fialkowski, M.; Bishop, K.J.M.; Chubukov, V.A.; Campbell, C.J. & Grzybowski, B.A.: 2005, 'Architecture and evolution of organic chemistry', *Angewandte Chemie International Edition*, **44**, 7263-7269.

Fourches, D.; Muratov, E. & Tropsha, A.: 2010, 'Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research', *Journal of Chemical Information and Modeling*, **50**, 1189-1204.

Gugisch, R.; Kerber, A.; Kohnert, A.; Laue, R.; Meringer, M.; Rücker, C. & Wassermann, A.: 2014, 'Molgen 5.0, a molecular structure generator', in: S.C. Basak, G. Restrepo & J.L. Villaveces (eds.), *Advances in mathematical chemistry and applications* (Volume 1), Sharjah: Bentham, pp. 113-138.

Guidance Document on the validation of (Quantitative)structure-activity relationships [(Q)SAR] models. OECD environment health and safety publications, 30 Mar 2007 [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono%282007%292, accessed 20 May 2015].

Hacker, P.M.S. & Bennett, M.R.: 2003, *Philosophical foundations of neuroscience*, Blackwell, Oxford.

Harré, R. *Modeling: gateway to the unknown*, Amsterdam: Elsevier, 2004.

Harré, R. & Llored, J.-P.: 2011, 'Mereologies as the grammars of chemical discourses', *Foundations of Chemistry*, **13**, 63-76.

Harré, R. & Llored, J.-P.: 2013, 'Molecules and mereology', *Foundations of Chemistry*, **15**, 127-144.

Harré, R.: 2015, 'Mereological principles and chemical affordances', in: E. Scerri & L. McIntyre (eds.), *Philosophy of chemistry: Growth of a new discipline*, Dordrecht: Springer, pp. 107-119.

Hettema, H.: 2013, 'Austere quantum mechanics as a reductive basis for chemistry' *Foundations of Chemistry*, **15**, 311-326.

Leach, A.R. & Gillet, V.J.: 2007, *An introduction to chemoinformatics*, Dordrecht: Springer, chap. 5, pp. 99-118.

Llored, J.-P.: 2014, 'Whole-parts strategies in quantum chemistry: some philosophical and mereological lessons', *Hyle: International Journal for Philosophy of Chemistry*, **20**, 141-163.

Llored, J.-P. & Harré, R.: 2014, 'Developing the mereology of chemistry', in: C. Calosi & P. Graziani (eds.), *Mereology and the sciences*, Dordrecht: Springer, pp. 189-212.

Ludwig, R.: 2001, 'Water: from clusters to the bulk', *Angewandte Chemie International Edition*, **40**, 1808-1827.

Luo, M.; Wang, X.S.; Roth, B.L.; Golbraikh, A. & Tropsha, A.: 2014, 'Application of quantitative structure–activity relationship models of 5-HT1A receptor binding to virtual screening identifies novel and potent 5-HT1A ligands', *Journal of Chemical Information and Modeling*, **54**, 634-647.

Mendeleev, D.: 1869, 'On the correlation between the properties of the elements and their atomic weights', *Zhurnal Russkoe Fiziko-Khimicheskoe Obshchestvo*, **1**, 60-77 (paper two in Jensen, W.B.: 2005, *Mendeleev on the periodic law, selected writings, 1869-1905*).

Needham, P.: 2005, 'Mixtures and modality', *Foundations of Chemistry*, **7**, 103-118.

Oliveras-Ferraros, C.; Fernández-Arroyo, S.; Vazquez-Martin, A.; Lozano-Sánchez, J.; Cufí, S.; Joven, J.; Micol, V.; Fernández-Gutiérrez, A.; Segura-Carretero, A. & Menendez, J.A.: 2011, 'Crude phenolic extracts from extra virgin olive oil circumvent de novo breast cancer resistance to HER1/HER2-targeting drugs by inducing GADD45-sensed cellular stress, G2/M arrest and hyperacetylation of Histone H3', *International Journal of Oncology*, **38**, 1533-1547.

Restrepo, G. & Villaveces, J.L.: 2012, 'Mathematical thinking in chemistry', *Hyle: International Journal for Philosophy of Chemistry*, **18**, 3-22.

Roth, B.L.; Lopez, E.; Patel, S. & Kroeze, W.K.: 2000, 'The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches?' *The Neuroscientist*, **6**, 252-262.

Shen, M.; Béguin, C.; Golbraikh, A.; Stables, J.P.; Kohn, H. & Tropsha, A.: 2004, 'Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds', *Journal of Medicinal Chemistry*, **47**, 2356-2364.

Schummer, J.: 1998, 'The chemical core of chemistry, I: a conceptual approach', *Hyle: International Journal for Philosophy of Chemistry*, **4**, 129-162.

Tang, H.; Wang, X.S.; Huang, X.-P.; Roth, B.L.; Butler, K.V.; Kozikowski, A.P.; Jung, M. & Tropsha, A.: 2009, 'Novel inhibitors of human histone deacetylase (HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation', *Journal of Chemical Information and Modeling*, **49**, 461-476.

Tetko, I.V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V.A.; Radchenko, E.V.; Zefirov, N.S.; Makarenko, A.S.; Tanchuk, V.Y. & Prokopenko, V.V.: 2005, 'Virtual computational chemistry laboratory design and description', *Journal of Computer-Aided Molecular Design*, **19**, 453-463.

Todeschini, R. & Consonni, V.: 2009, *Molecular descriptors for chemoinformatics*, Weinheim: Wiley-VCH.

Todeschini, R.; Ballabio, D. & Consonni, V.: 2015, 'Distances and other dissimilarity measures in chemometrics', in: R.A. Meyers (ed.), *Encyclopedia of Analitycal Chemistry*, New York: Wiley, pp. 60.

Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M. & Willett, P.: 2012, 'Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real datasets', *Journal of Chemical Information and Modeling*, **52**, 2884-2901.

Varzi, A. 2014, 'Mereology', in: *The Stanford Encyclopedia of Philosophy* [http://plato.stanford.edu/archives/spr2014/entries/mereology/, accessed 20 March 2014].

Wermuth, C.G.; Ganellin, C.R.; Lindberg, P. & Mitscher, L.A.: 1998, 'Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998)', *Pure and Applied Chemistry*, **70**, 1129-1143.

Young, D.; Martin, T.; Venkatapathy, R. & Harten, P.: 2008, 'Are the chemical structures in your QSAR correct?' *QSAR & Combinatorial Science*, **27**, 1337–1345.

Zaki, R.M.; Elossaily, Y.A. & El-Dean, A.M.: 2012, 'Synthesis and antimicrobial activity of novel benzo[f]coumarin compounds', *Russian Journal of Bioorganic Chemistry*, **38**, 639-646.

Guillermo Restrepo:
Bioinformatics Group, Department of Computer Science, Universität Leipzig, Leipzig, Germany;
and Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia; guillermorestrepo@gmail.com &
grestrepo@unipamplona.edu.co

Rom Harré:
Georgetown University, Washington, DC, USA; harre@georgetown.edu