[Frontiers in Bioscience 16, 1276-1290, January 1, 2011]

Theoretical and computational approaches to ligand-based drug discovery

Angelo D. Favia

Istituto Italiano di Tecnologia (IIT), Drug Discovery and Development, via Morego 30, 16163 Genoa, Italy

TABLE OF CONTENTS

1. ABSTRACT

The basic idea behind ligand-based approaches is that the analysis of sets of molecules with experimentally determined activities can highlight those chemical features responsible for the activity changes. Historically, such approaches have been devised before structure-based methods. Nowadays, despite the ever increasing availability of experimentally determined structures, ligand-based approaches still play a major role in drug design either alone or in conjunction with structure-based efforts. This manuscript aims to provide a general overview of the main computational approaches in ligand-based drug discovery, particularly 3D QSAR methods, along with relevant references to the literature.

2. INTRODUCTION

2.1 Ligand-based drug design

Notwithstanding the complexity of its interdisciplinary nature, drug research is fundamentally based on simple concepts:

In living organisms, chemicals interact with macromolecules, triggering certain biological effects;

Similar molecules trigger similar effects (1).

Given a dataset of chemicals active towards a pharmaceutically relevant target, it is therefore common practice in drug research to look at similar molecules in order to conveniently modulate certain characteristics (potency, selectivity, bioavailability *etc.*). Here, however, the apparent simplicity stops and the first questions arise. Is there an unambiguous way to determine whether two molecules are similar, and hence likely to produce similar biological effects? What metric should be used to correctly classify molecular entities? Compelling evidence suggests that no satisfactory reply exists (2).

For decades, well before the advent of computational resources in medicinal chemistry, those questions were addressed using intuition bolstered by experience. Given a set of molecules sharing a common scaffold (known as 'congeneric molecules') with experimentally measured activities (a training set), researchers aimed to infer, mainly by visual inspection, the molecular features responsible for changes in biological activity. With varying degrees of success, researchers attempted to make predictions based on qualitative models (i.e. if A is more active than B, then C, whose activity has not yet been assayed, is likely to be more active than A). The reliability of these models then had to be established by experimental tests. This is known as a structure-activity relationship (SAR) study. It is a pivotal step in drug discovery. However, practitioners tend to work differently nowadays. This is because computers play a key role in almost every aspect of the drug discovery process. Model generation is now mostly left to complex mathematical algorithms, while human intervention is increasingly focused on model analysis and interpretation. In chemoinformatics (3, 4) computer and information technology are applied to a wide range of chemistry problems. Compounds thus cease to be ensembles of covalently bound atoms and instead are transformed, through the use of molecular descriptors, into numerical values that, taken together, say something about the biological activity and chemical reactivity of a given compound. According to the definition given by Todeschini and Consonni, a molecular descriptor is "the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment" (5). Mathematical applications are then used to find quantitative relationships between the descriptors and the measured activities. Their reliability is tested using sets of compounds, with known activity, omitted from the original model (test set). Data torturing rarely produces bad models. For this reason, the main duties of the final user are to interpret the outcomes of the model, to verify whether the model has any useful predictive capabilities, and, if it doesn't, to assess why. As in the days before computer-assisted drug design (CADD), new molecules with theoretical enhanced potencies are synthesized and bioassayed. The main difference now is that researchers attempt to make accurate quantitative predictions of the activities that will occur upon chemical modification (i.e. quantitative SAR, or QSAR). The final stage of the process (and perhaps the most delicate stage too) is the return from descriptor space to chemical space in pursuit of new molecules with hopefully improved features. Human abilities here still make the difference between a successful and a fruitless drug discovery campaign (6).

2.2. Historical perspective

Traces of elemental SAR studies can be found in the scientific literature as early as the end of the 19th century, when Richet reported on the relationship between solubility and toxicity for a series of compounds (7). Given the intuitive and reasonable logic behind SAR (*i.e.* similar things behave in similar ways), this early appearance is not surprising. A few years later, Meyer and Overton independently showed there was a close relationship between the olive oil solubility of molecules and their narcotic power in tadpoles (Figure 1) (8, 9).

However, it was Hammett who reported the very first case study in which a quantitative understanding of the relationships between structure and chemical properties was achieved. His linear free energy relationship (LFER) model was introduced in physical organic chemistry in 1937 (10). He discovered that equilibrium constants (*K*) or reaction rates (*k*) for many reactions involving benzoic acid derivatives with meta- and para-substituents can be related using just two parameters: a substituent constant (σ) and a reaction constant (ρ) (see Equation 1, K_0 is the reference value)

$$log\ (K/K_0) = \sigma\rho \quad \text{(Equation 1)}$$

In 1964, after a gap of almost three decades, Hansch and Fujita built on Hammett's findings to produce the first medicinal chemistry study capable of relating molecular properties to observed biological measures by means of numerical equations (11). In this milestone paper, Hansch and Fujita used π and σ, respectively, as descriptors of the electronic and lipophilic features of molecular structures to rationalize certain biological activities in several diverse test cases (Figure 2).

It is not by coincidence that these descriptors were successfully used and, moreover, seemed to have a general applicability. In a basic yet quite realistic representation of what goes on at cellular level, π accounts for hydrophobic interactions and for the molecular propensity to permeate biological barriers. It is, in practice, a very handy descriptor of the effective concentration of chemicals at the site of action, while σ roughly accounts for the molecular interactions occurring at the binding site. The study also highlighted the very important non-linear dependency between biological response and lipophilicity (this concept has been more properly addressed by Kubinyi) (12, 13). In the same year, Free and Wilson published a mathematical study about the estimation of biological response caused by structural changes (14): the QSAR era had officially begun. These papers had a great impact on the scientific community and profoundly influenced the way researchers handled molecular datasheets.

At the beginning of the 1980s, 3D structures of small ligands in complex with macromolecules were just becoming available to the scientific community (15). At the same time, researchers were starting to accept that drugs

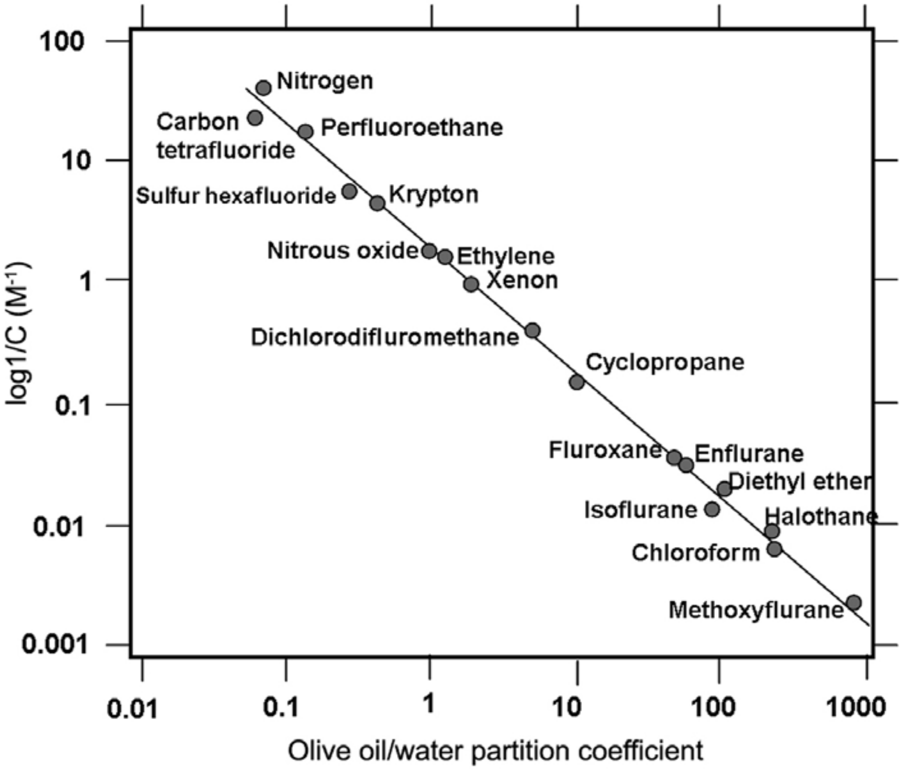

Figure 1. The Meyer-Overton correlation for anesthetics.

were more complex than bidimensional sketches on paper. The concept of bioactive conformation began to radically change the way scientists looked at molecules, at a time when computers were becoming increasingly powerful. As a consequence, the idea emerged that a target's chemical preferences could be inferred indirectly from studying the 3D arrangement at the site of action of its binders. The basic concept is that molecules are aligned onto each other in their putative bioactive conformation. Then, changes in experimentally determined activities are related to certain properties, which are calculated within a pre-defined volume around them using statistical techniques. In principle, one advantage of this approach is that comparisons are possible between non-congeneric molecules too. Based on the assumption of a shared binding mode, the exerted fields are what count, not the underlying molecular frames. The first example of 3D QSAR methodology dates back to 1988, when Cramer III and colleagues invented comparative molecular field analysis (CoMFA) (16). CoMFA immediately gained extreme popularity. It remains the approach of choice in many case studies today.

In retrospect, adding an extra dimension to the "classical" QSAR approach was an obvious evolution. Initially, however, scientists were skeptical about the usefulness of 3D interactions with respect to traditional monodimensional physicochemical descriptors. A number of new concepts had to be digested in order to persuade them. For instance, the number of variables (*e.g.* molecular fields) was far greater than the number of studied objects (*i.e.* molecules), hence *ad hoc* mathematical algorithms had to be designed or borrowed from the statistical mathematics community. Of these, partial least squares analysis (PLS) (17) probably contributed the most to the 3D QSAR legacy. At the time, of course, there was not widespread access to the computational power needed to deal with multivariate problems. This also slowed down the establishment of 3D QSAR.

At the beginning of the 1990s, there was a rapid increase in the number of applications of 3D QSAR techniques in drug design projects. The CoMFA-like use of different combinations of the Goodford's GRID (18) interaction points opened the door to countless practical applications (19, 20). In 1994, Klebe and colleagues published a methodology called CoMSIA (molecular similarity indices in a comparative analysis). It shared some similarities with CoMFA and aimed to overcome some of its pitfalls (21). Soon after, Richards and colleagues proposed a CoMFA variant, the self-organizing molecular field analysis (SOMFA). However, this was not as popular as CoMFA (22). Contextually, the number of available X-ray macromolecular structures was rapidly increasing. Researchers were producing new strategies for studying the interactions between small ligands and macromolecules in a more direct way (*e.g.* molecular docking). A bridge was needed between the structure-based and ligand-based worlds. Unfortunately, despite several cases in which 3D QSAR hypotheses were subsequently confirmed by X-ray-based studies (23-29), it became evident that, in ligand-based methodologies, certain general issues remained unsolved. Anyone who has dealt with classical 3D QSAR studies knows how much the final result depends upon a

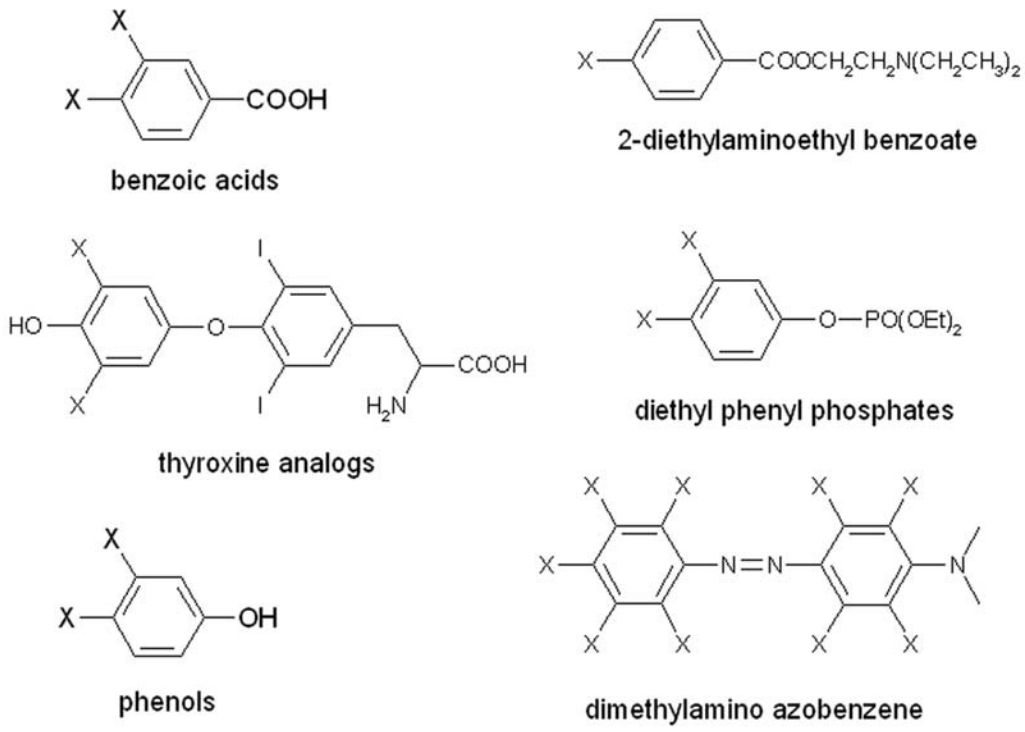

Figure 2. Classes of compounds reported in the Hansch-Fujita study.

number of user-dependent choices, such as the box size, the grid spacing, the field selection and, most importantly, the alignment procedure. Within the last 20 years, much effort has been made to improve 3D QSAR analyses. Eventually Dobler and colleagues extended the concept of traditional 3D analysis by adding three more dimensions. By taking into account multiple representations of the ligand molecules (4D) (30), various induced-fit hypotheses (5D) (31) and alternative solvation models (6D) (32) they could more accurately depict the biological process of binding. It is also worth mentioning the use of atomic property fields (APF) recently proposed by Totrov (33). Although its general applicability has not yet been proved, it is one of the most promising attempts to limit human bias across the procedure, from alignment to the final model.

When dealing with datasets of molecules active towards a given biological target, an alternative possibility is to derive a pharmacophore. The concept of pharmacophores was introduced by Kier in the late 1960s (34). According to the IUPAC definition, it represents "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response" (35). Commonly used features are hydrogen bond donor and acceptor groups, charged atoms, aromatic rings *etc.* This is in line with the concept of bioisosterism, which identifies distinct functional groups sharing similar biological, chemical, and physical properties (36). A pharmacophore can be thought of as a coarse representation of a prototypical molecule active towards a given biological target. Once developed, a pharmacophore model can be used to parse molecular databases in order to sort the probably inactive compounds from those likely to be active (*i.e.* those that match the pharmacophore hypothesis). At the end of the last century, this kind of analysis was mainly conducted by looking at the molecules. Unsurprisingly, the scientific community can now use complex algorithms to generate, handle, and elucidate pharmacophoric hypotheses.

None of the above-mentioned methodologies are error-free. Therefore, while older methodologies are still successfully applied (37), scientists are focused on creating improved protocols. Researchers are constantly proposing new descriptors (5), innovative ways of accounting for ligand flexibility, and more advanced protocols for analyzing the massive amount of data produced. The primary goal is still to infer, from a series of flexible small molecules with known activities, what drives the activity changes. Methodological advances should have an enormous impact on medicinal chemistry, allowing the prompt identification of candidates with improved features at sustainable costs.

2.3. Preamble

In the following section, the theoretical foundations and principles of ligand-based drug design will be briefly examined. In medicinal chemistry, the property under investigation is typically the affinity constants towards a given target (usually a protein) measured in terms of IC_{50} or $K_{i.}$ However, the investigative methods can

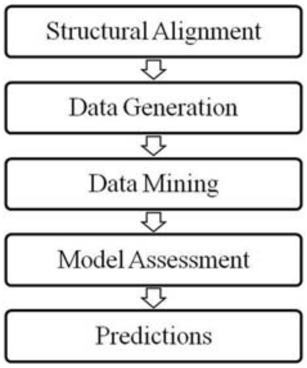

Figure 3. General framework of 3D QSAR protocols.

be used to study any observable that is dependent upon the geometric arrangement of molecules. In fact, in a broader context, the word 3D QSPR is sometimes used. This stands for 'quantitative structure-properties relationship', which is derived from knowledge of the three-dimensional arrangements of molecules. I note that, although most protocols share similarities to some extent (at least in the general framework), there can be remarkable differences from case to case. The following is thus intended to be a general scheme with a fairly broad applicability. A detailed mathematical description is purposely missing; however, the interested reader is encouraged to delve into the cited publications. This manuscript is not intended to be an exhaustive review of ligand-based approaches in drug discovery. I therefore apologize to those authors whose original work is not included in the references due to space limitations.

3. THEORETICAL BACKGROUND

3.1. Five steps to 3D QSAR

Ligand-based drug design is based on the theory that similar molecules cause similar biological effects. This intuitive concept does not always hold true (38, 39), but most of the time, it does. The reason is found at molecular level. Structurally similar molecules bind in a similar fashion (40) and are thus likely to be recognized by the same biological receptors, triggering similar responses (1). Everyday experience tells us that similarity means different things to different people, and similarity between ligands is no exception. This is because, in chemistry, molecules can only be compared indirectly by means of descriptors that capture defined aspects of their complex nature, one at a time (41). Moreover, there are several descriptors whose importance varies from case to case, depending on the nature of the ligand-receptor interactions involved. Additional complexity arises from the fact that molecules, with few exceptions, are flexible elements and, most of the time, their bioactive conformations are not known *a priori*. In the following sections, the five distinct phases of a general 3D QSAR framework (Figure 3) will be examined.

3.1.1. Structural alignment of molecules

Alignment is generally the first and most crucial phase in a 3D QSAR analysis (although it may be omitted from more recent methods, if a number of conformations per ligand are provided). In the alignment step, the ligands under investigation must be aligned onto each other in their putative bioactive poses (*i.e.* the conformations adopted upon receptor binding).

There are several ways to identify structural alignments for comparative purposes (42-45). One procedure considers the molecular framework and functional groups belonging to it. This usually involves choosing a reference structure and carefully selecting the tethering atoms. The reference structure can be either an active ligand whose spatial arrangement at the active cleft is known or, if that is unavailable, a known binder with a reduced number of rotatable bonds (to limit the uncertainty associated with its binding mode). In both cases, the reference can either belong to the studied dataset or be an active that shares the dataset's binding mode and that is already in the literature. The tethering atoms are chosen in order to guide the subsequent 3D ligands' superposition to the reference. By superimposing those parts of the ligands not directly involved in noncovalent interactions with the macromolecule, researchers can maximize the chances of getting a biologically meaningful alignment. Two objectives are thus achieved in one step:

The molecular degrees of freedom are sampled;

There is minimization of the distance function between identical pharmacophoric groups belonging to different molecules (one of which is always the reference structure).

In alternative methods, several chemotypes are assigned to each atom of the molecules in the dataset and then the optimal superposition between them is sought (46, 47). These protocols are easy to implement in computer algorithms. Their main advantage is that they converge quickly. On the downside, they do not properly capture what really goes on at a molecular level. This is because target-ligand recognition is exerted through molecular surfaces and atomic properties mapped onto them. Focusing on atomic superposition, therefore, might not lead to the expected results.

To overcome the above-mentioned limits, alignments based on molecular fields and surface properties have been proposed. In these protocols, properties are calculated at equally distributed points on the van der Waals surface or on a defined volume surrounding each conformer generated for each studied molecule. Then, the search algorithm tries to find the optimal superposition, which is the one that most minimizes the differences between calculated values in the confronted structures. At the end of the run, the generated alignments are ranked according to this fitness function. Hermann and Herron proposed an early implementation of this idea in 1991, using electrostatic potential values at the van der Waals surface to calculate the goodness of molecular superpositions (Figure 4) (48).

Since then, there have been a number of applications based mostly on the same principles or on the Lennard-Jones potentials, which are calculated at points within the buffer area around the molecules (49, 50). This

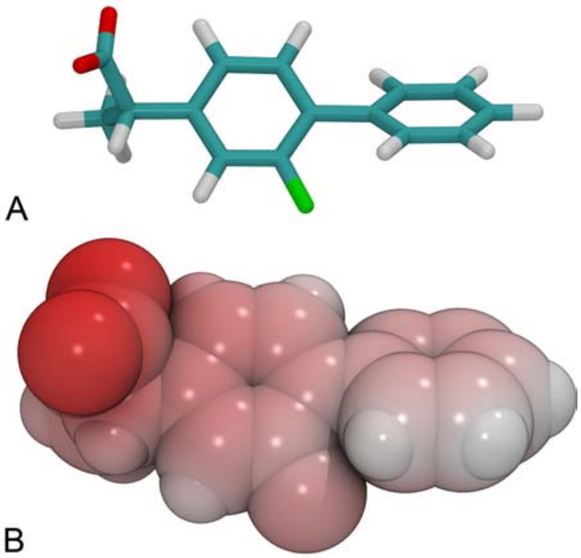

Figure 4. Comparison between two molecular representations. In A, (*S*)-Flurbiprofen is represented as a stick model, atom type colored. The same molecule is shown in B as a van der Waals surface, colored according to the electrostatic potential (from red to white for values from negative to positive, respectively).

approach yields more meaningful alignments from a biological standpoint, since molecules are treated as they are really "seen" by their natural counterparts. However, such accuracy comes at a cost. The calculation time increases considerably with respect to simpler punctual atomic superpositions. A recently developed way of superposing ligands takes into account the molecular topology by representing molecules as graphs of connecting atoms. It has produced results comparable to more time-demanding methods (51). Molecular superpositions can also be achieved without considering a reference molecule. However, the lack of reference to an established bioactive pose means there is a greater chance of deriving alignments that lack biological transferability.

3.1.2. Generation of data

Once the studied molecules have been pushed towards their putative bioactive conformations, the next steps are usually:

Creating a virtual box, big enough to contain all the molecules in the dataset plus a buffer area;

Calculating the physicochemical properties at regularly spaced points within the box.

The grid spacing should be set in a way that every field supposedly felt by the target is smoothly captured by the minimum amount of points. A value of between 1 and 2 Å is usually enough to deal with standard systems. In CoMFA (16), Coulombic and Lennard-Jones potentials are calculated to represent the steric and electrostatic effects accounting for the overall molecular reactivity. Similarly, the space around the molecules can be sampled by diverse combinations of chemical probes, which are accurately chosen to portray the intermolecular interactions acting upon target binding. Conversely in CoMSIA (21), the calculated properties at each grid point do not have a readily usable physical meaning. This is because their numerical value indicates the similarity between selected probes and the studied molecules in arbitrary units, in a distance-dependent functional form. In general, the enclosing box can be thought of as a virtual cage containing evenly distributed sensors (*i.e.* the grid nodes), which are receptive to pre-selected molecular features. Thus obtained, the signals are recorded and analyzed for correlations between the dependent variable (*i.e.* the property under investigation) and the independent variables (*i.e.* the values at each grid point). A general schematic representation of these steps is provided in Figure 5, where Donepezil is the prototypic molecule.

As stated above, the chemical determinants intervening upon target-ligand binding differ from case to case. This makes it difficult to choose *a priori* the molecular properties to consider in a 3D QSAR. Moreover, the best descriptive and predictive models could, in principle, be given by combinations of predominant properties, increasing the complexity of the analysis. Researchers are trying to find the single descriptor that, taken alone, could account for all molecular activity/reactivity. Quantum mechanical (QM) descriptors have been proposed, since they are directly connected with the intrinsic molecular reactivity. For example, deformations in the local atomic densities have not been proven to quantitatively predict the forces implicated in molecular interactions and in chemical processes such as van der Waals, Pauli, bonding and nuclear un-screening forces. However, their use in 3D QSAR has not yet been attempted (52). Although QM descriptors can give a more accurate picture of the molecular propensities, their use in 3D QSAR studies is greatly hampered by the associated computational burden. QM treatments are extremely time-demanding as compared to simpler, although less accurate, molecular-mechanics-based calculations. Recently, a simplified molecular description (named SAMFA) has been proposed by researchers at AstraZeneca Pharmaceuticals (53). Surprisingly, SAMFA seems to perform exceptionally well, compared to standard protocols that use more refined descriptors. Moreover, SAMFA models can be straightforwardly interpreted due to the simple nature of the descriptors used.

3.1.3. Mining of data

Data mining represents the third step in a 3D QSAR study. As shown in Figure 5, for each point of the grid, property values are calculated, the accumulated data are analyzed, and a virtual model is generated. In this phase, the experimentally obtained measurements for each studied ligand in the training set are associated with the numerical values calculated at each grid node. Thus, the 3D information is folded in a 2D matrix. Statistical techniques are then used to spot, where present, relationships between the variance recorded at each grid node (indicated as x_1, x_2 *etc.* in Equation 2) and the variance of the known activity (indicated as biological response, *B.R.* in Equation 2). This

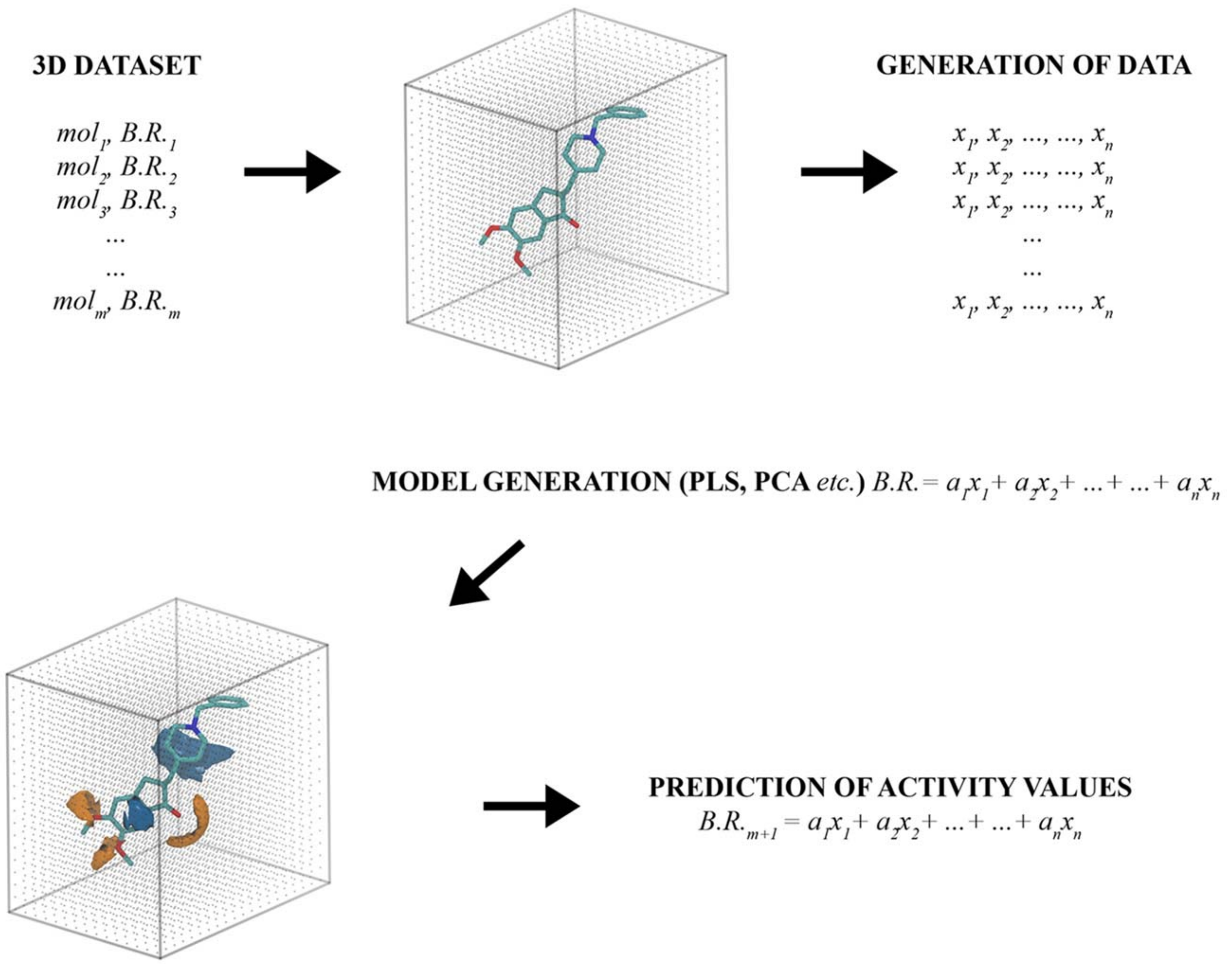
3D DATASET
$mol_1, B.R._1$
$mol_2, B.R._2$
$mol_3, B.R._3$
...
...
$mol_m, B.R._m$
GENERATION OF DATA
$x_1, x_2, ..., ..., x_n$
$x_1, x_2, ..., ..., x_n$
$x_1, x_2, ..., ..., x_n$
...
...
$x_1, x_2, ..., ..., x_n$
MODEL GENERATION (PLS, PCA *etc.*) $B.R. = a_1x_1 + a_2x_2 + ... + ... + a_nx_n$
PREDICTION OF ACTIVITY VALUES
$B.R._{m+1} = a_1x_1 + a_2x_2 + ... + ... + a_nx_n$

Figure 5. General steps of a 3D QSAR study. Three-dimensional structures of the molecular dataset, overlaid according to their putative binding poses, are immersed in a virtual box. For each molecule, the biological response (*B.R.*) is known along with the property values, calculated at each node of a predetermined 3D grid (the grey dots in figure). Statistical tools are then used to determine the coefficients that best correlate the dependent variable (*B.R.*) to the independent variables (field values at the grid nodes). The nodes that mostly correlate with the activity changes can be usefully highlighted in 3D (the colored isopotential surfaces in figure), indicating regions in space directly involved in the binding process where chemical modifications to enhance the observed property are desirable. Once assessed, the model can be used to predict the biological response of new molecules, given putative 3D arrangements at the binding cleft. To help interpretation, Donepezil is shown as stick model C-colored cyan, as a prototypical molecule surrounded by hypothetical interaction maps in an *ad hoc* built grid volume.

mathematical effort aims to find the optimal coefficients a_1, a_2 *etc.* in the equation below, so as to weight each node contribution to the biological response:

$$B.R. = a_1 x_1 + a_2 x_2 + a_3 x_3 + ... + a_n x_n \quad \text{(Equation 2)}$$

Historically, certain indexes have been used to monitor the wellness of the model from the earlier stages to its final version. For example, in the developing phase, it is important to check how well the model explains the activity of the dataset used to generate it. To do this, the squared correlation coefficient (r^2) can be conveniently calculated as:

$$r^2 = \frac{\sum_{i=1}^{N}(y_{calc,i} - y_m)^2}{\sum_{i=1}^{N}(y_i - y_m)^2} \quad \text{(Equation 3)}$$

Where y_{calc}, y and y_m indicate the estimation of the dependent variable, its actual value, and its mean value, respectively. The calculation is reiterated over the total number (N) of objects (*i.e.* molecules) considered. Values of r^2 can vary between 0 and 1, indicating, respectively, the total inability and the perfect ability of the model in explaining the variation in the dependent variable. In real life scenarios, values of r^2 between 0.8 and 0.9 indicate a good fitness of the model.

The most widespread way of generating robust linear equations in medicinal chemistry is the partial least squares regression analysis (PLS) (17). This is due to its strict association with the CoMFA protocol. However, scientists have also successfully used variable selections based on principal component analysis (PCA), design criteria (54), genetic algorithms (55), neural networks (56) and inductive logic programming (ILP) (57, 58). In recently developed 3D QSAR methodologies where the alignment is not required, a sufficiently large number of conformers per studied molecule are generated, each of which generates a data sheet. The sum of the generated data is then fed into a variable selector algorithm. This leads to the above equation, with the constraint that, in the end, each molecule can contribute just one conformer to the model. The advantage of this approach is that there is no need to guess the bioactive conformations *a priori*. These can be inferred directly from the final model. On the downside, this methodology cannot be readily applied to molecules with many degrees of freedom. This is because too many hypothetical conformers would be needed to represent the likely physiological poses, making the method unfeasible. Furthermore, the amount of data thus produced cannot be treated with traditional methods. More robust and time-consuming tools must be used, like machine learning methods. Moreover, due to the lack of a human supervised superposition, there is a real risk of producing a self-consistent model that lacks prompt chemical transferability. The latter issue will be addressed in the forthcoming sections.

3.1.4. Validation of the model

Once generated, the model must be validated to assess its descriptive and predictive capabilities. Descriptive and predictive powers refer to model's ability to estimate the known activities of the molecules in the training set, and the activities of the molecules in the test set, which can be either known or not. The descriptive power gives a measure of how well the model performs while retrieving the original data (upon which it was developed), a measure of which is the above-mentioned r^2 (Equation 3). However, the predictive power can be assessed either internally, by predicting the dependent variable of molecules arising from the original dataset and temporarily left out from it (*i.e.* cross-validation), or externally. In the latter and more challenging type of test, the model's robustness is tested in a real-life scenario by making predictions about molecules with known activity values that were never included in the model's development.

The cross-validated r^2, or q^2, is a useful index. It is used to measure the ability of the model to predict the activity values of molecules temporarily omitted from the model.

$$q^2 = 1 - \frac{\sum_{i=1}^{N}(y_{pred,i} - y_i)^2}{\sum_{i=1}^{N}(y_i - y_m)^2} \quad \text{(Equation 4)}$$

In this equation, y_{pred}, y and y_m are the predicted, the real, and the mean value of the observables under investigation. Intuitively, cross-validated r^2 are lower than pure r^2. Values of between 0.6 and 0.8 are generally considered reasonable, at least in 3D QSAR.

Extreme care should be taken when treating models with a low predictive power. This low predictive power could be due to the presence of molecules with unique features whose importance cannot be accounted for by the remaining members of the dataset (59).

At this stage, it is also important to pinpoint possible outliers (molecules whose activities are badly predicted) and to understand the reasons behind such behaviors. Detecting outliers is an important task that can sometimes highlight incorrect experimental measurements and theoretical assumptions. It can also shed light on important features of the process under investigation, like activity cliffs in the structure-activity landscape (38, 39, 60).

3.1.5. Back to chemistry

The final phase of a 3D QSAR study is by far the least demanding in terms of CPU time. It aims to translate the obtained outcomes into suggested chemical modifications, so that molecules with improved characteristics can be developed. The relevant variables in the data matrix are those that most correlate with property changes. Once identified, they can be conveniently projected into 3D, highlighting exploitable regions in the space within the box. According to the model, appropriate chemical modifications in these regions can positively or negatively influence the observable property, making the rational design process relatively straightforward (Figure 5). Newly designed molecules can be fed into the model via the calculation of properties on given poses. Predictions of activity are immediately acquired since the optimal coefficients in Equation 2 (see above) are preset. When successful, virtual models can predict a given property, in silico, for molecules that have not yet been synthesized or purchased. However, even very robust models do not guarantee accurate predictions. There may always be unforeseen factors. It is therefore advisable to reassess the model from scratch, starting from the initial binding hypothesis.

In summary, the ideal model must fulfill both mathematical requirements (*i.e.* it must be robust from a statistical point of view) and more pragmatic requirements (*i.e.* the outcomes should tell us something relevant about the underlying biological processes, and give us useful indications of how to tune them).

3.2. 3D pharmacophore mapping

Pharmacophore mapping involves the creation of a pharmacophore. Its most challenging aspect is molecular alignment (61). Indeed, the underlying common features can be captured easily following the superimposition of an acceptable number of molecules (all of which actively bind the same target). Subsequently, these common features can be used to screen large databases of potential binders. However, the inclusion of inactive molecules can provide valuable information by validating the formulated

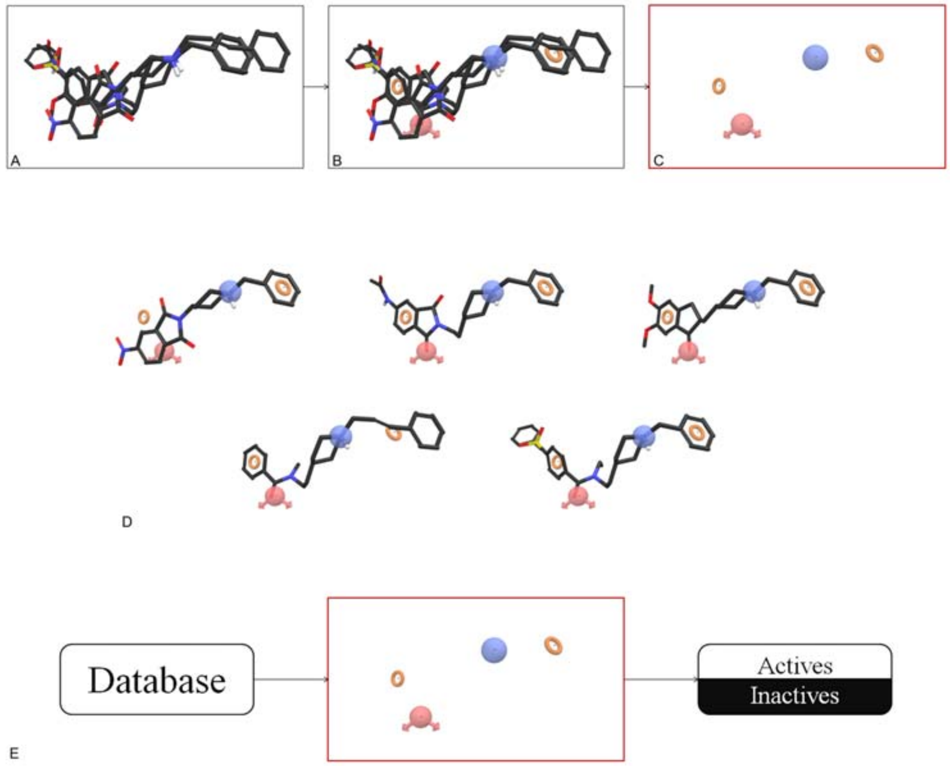

Figure 6. Graphic representation of pharmacophore-mapping and its application in drug discovery. From the analysis of the superimposed structures (A), some key features emerge (B). Eventually, a pharmacophore is derived (C). The orange ring represents an aromatic ring, the blue sphere a positively charged atom, while the red sphere with the two arrows represents a hydrogen bond acceptor. In D, the same molecules are mapped onto the pharmacophore. The active ones match all the extracted features (top middle, top right, and bottom right molecules), while the inactive ones do not. The pharmacophore can then be used to analyze databases in order to pinpoint the likely active molecules (E). The analysis depicted here was carried out with Phase (80) on the Acetylcholinesterase (AChE) dataset compiled by Sutherland and colleagues (76). The molecule on the top right is Donepezil, a well-known AChE inhibitor, approved for treatment of Alzheimer's disease.

hypotheses. There are a growing number of commercially available packages that allow researchers to develop pharmacophoric hypotheses (62, 63). The basic concepts behind them are similar. As with QSAR methodologies, the conformational space of fully flexible molecules must be sampled while trying to maximize the overlaps between the same features in different compounds. Eventually, active molecules will share features that are thought important for binding, while inactive molecules will not.

Feature assignment is the first step in pharmacophore development. Along with traditional readily usable properties such as hydrogen bond donor and acceptor, it is usually possible to customize attributes, thus shaping the model according to particular needs. This can be particularly useful when dealing with a well-known target whose binding preferences have already been extensively studied. Alternatively, molecules can be characterized by their exerting fields (37).

Once the minimum distance between features has been set, the second step involves the generation of a series of candidate alignments. Software packages usually have built-in conformational sampling engines, which run smoothly from input (2D or 3D models of the molecules) to final output (the pharmacophore). User-generated inputs can also be provided. Aside from the technical aspects of their generation, overlays are ranked according to a score that considers various terms, such as feature-matching or volume overlap. But the calculated score is just an indication, since it does not consider the relative relevance of each term (which is likely to vary from case to case). Hence, the only way to assess the alignments' ability to explain the difference between actives and non-actives is to validate them. This can be done internally (*i.e.* by visual inspection of the molecules used to derive them) or externally (*i.e.* by predicting the activity of new compounds with known activity that were omitted from the hypothesis generation).

Thus obtained, the pharmacophore can eventually be used to screen 3D databases in order to identify putative hits and leads.

Figure 6 provides a pictorial representation of pharmacophore-mapping and its application in screening databases.

Pharmacophore and 3D QSAR models are closely related. Both are based on binding hypotheses of related molecules to a target. For this reason, the two approaches can be integrated in a single workflow. Starting from a feature-based hypothesis capable of discriminating actives from inactives, this workflow can help develop a more refined mathematical model that attempts to relate the biological activity to certain 3D-calculated properties (63).

4. ADVANTAGES AND DISADVANTAGES OF LIGAND-BASED APPROACHES

Computational methods in drug design can be divided into two major classes: structure and ligand-based approaches. Structure-based approaches use knowledge of the 3D arrangement of the receptor of interest (either experimentally solved structures or comparative models). Usually, ligand-based approaches only consider molecules that reversibly bind using noncovalent interactions. Long before the first 3D receptor models became available, ligand-based methods were devised to make sense of biological assays. They continue to undergo improvement. However, after successfully gaining a wider following, their popularity greatly suffered from the appearance of more direct ways of simulating receptor ligand interactions (for example, molecular docking). Nowadays, the computational chemist who works in drug design must make a choice and decide which protocol best suits their needs.

Biological interactions are very involved processes and any simulation approach will have both advantages and disadvantages. It might seem straightforward to use the knowledge of the receptor's binding site when available. But certain considerations must be made. Experimentally obtained structures, such as those provided by X-ray determination, are single static impressions of molecular systems with thousands of degrees of freedom, brought, by the experimental conditions, towards one of the possible minima in the free energy landscape. For instance, the presence or absence of a ligand bound at the active site is likely to influence the receptor conformation, inducing severe structural rearrangements that can bias the output of subsequent structure-based procedures. Despite recent developments, incorporating protein flexibility into molecular docking calculations is still one of the major challenges of computational simulations (64-66). Ligand-based methodologies, in contrast, have an apparent advantage in not having to deal with the receptor. But this is counterbalanced by the fact that the negative image of the receptor, produced through the study of its ligands, is inevitably blurry and incomplete. The chemical characteristics of the dataset under investigation are what drive the learning process of the binding cleft. This automatically produces novel structural modification conjectures about the receptor's preferences. In other words, as long as the chemical boundaries dictated by the known active molecules are not trespassed, predictions are safer. But in drug discovery, researchers are generally interested in novel molecules that lie outside the known ground. To partly overcome these limits, it can be useful to incorporate, into ligand-based efforts, the information arising from the knowledge of the binding site. For instance, pharmacophoric hypotheses can be directly inferred from experimentally determined structures, making it possible to account for the presence of exclusion/inclusion regions (67).

When dealing with drug optimization protocols, one tries to maximize the interactions occurring between a macromolecule and a small ligand, usually in terms of noncovalent interactions. However, depending on the nature of the assays, several other factors can influence the final measurements. These can be related to the probability that the ligand will reach the binding cleft, and can involve chemical stability, off-targets, membrane permeability, and so on. In fact, one aspect of deriving virtual models, which link a measurable biological response to certain chemical properties, is that the underlining causes of the observables are considered contemporaneously. In other words, each descriptor or feature included in a final ligand-based model must explain not only the interactions at the site of action, but also any preceding biological processes. This is an unsolvable shortcoming in any methodology that attempts to correlate an experimental outcome to the result of a model that only mimics the interactions at the site of action. Monodimensional descriptors in classical QSAR, such as those developed by Hansch, are probably more suited to capturing the foundations behind complex biological assays. This is because no assumption about 3D binding modes is made. Surprisingly enough, however, a recent study has shown that the addition of monodimensional descriptors (*i.e.* CLOGP and CMR) to a CoMFA analysis does not improve the quality of the final model. However, this could simply be due to the type of experimental assays considered (*i.e.* where receptor-ligand interactions play a predominant role) (59).

The structure-activity landscape is a high-dimensional representation of the property as a function of the descriptors used. For years, following the assumption that similar molecules share similar activities, such landscapes were thought to be so smooth that activity changes within confined regions happened gently. More realistically, depending on factors such as the nature of the assay, the distribution of the compounds, and, most importantly, the nature of the molecular representations used, structure-activity landscapes should be considered as rugged hyper planes instead (Figure 7).

Extreme care should be taken when dealing with molecules that do not fit the elaborated model, resulting in surprisingly poor predictions. This behavior could be caused by the presence of outliers (*i.e.* elements that appear to deviate markedly from other members of the sample in which they occur), or by true drastic changes in the activity landscape, called activity cliffs (39). The presence of activity cliffs can be revealed by additional experimental measurements in the area around the badly predicted molecules. Guha and Van Drie recently proposed an index to identify and quantify those behaviors (38, 60). Such phenomena are not entirely surprising since they lie at the core of the receptor-ligand recognition processes, where

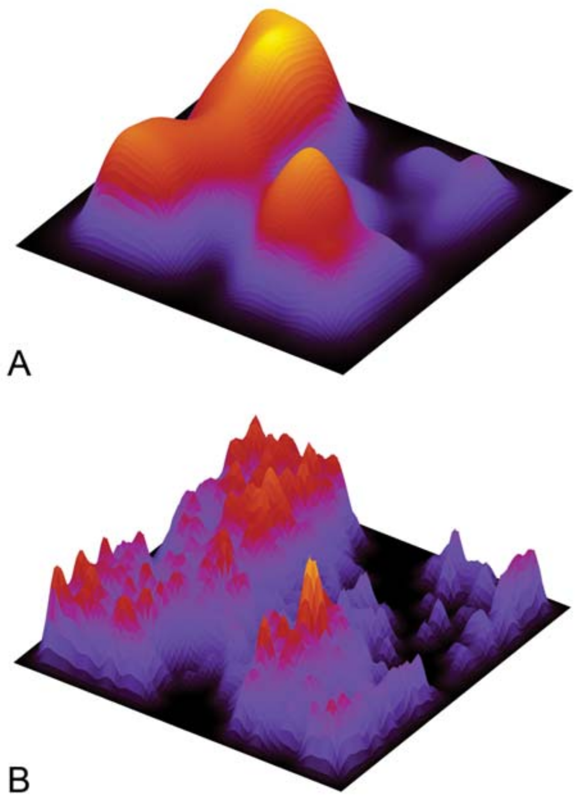

Figure 7. Bidimensional representation of a hypothetical structure-activity landscape. The x and y axes define a given descriptors' space where the molecules reside, while the z axis is the measured property. In A, the surface appears smooth. In B, the same surface is characterized by rough transitions and spikes.

very small chemical modifications can turn an inactive molecule into an active one. Here, specificity is achieved through almost perfect complementarity between interplaying parts (68).

5. CONCLUSIONS AND PERSPECTIVE

Ligand-based approaches establish a relationship between the measured responses exerted by a series of molecules and a set of parameters determined from their structures. Their central assumption is that the numerical outcomes obtained under experimental conditions for a set of molecules depend solely on their chemical nature.

Although 3D QSAR and pharmacophore methods have been around for some time, they are far from being perfect tools for medicinal chemistry. Each constituting step (*i.e.* structural alignment, data generation, data mining, model assessment, and predictions) helps determine the quality of the final outcome. Each step presents certain challenges. Novel protocols, tentatively free from these drawbacks, continue to be proposed. Due to the intrinsically modular nature of ligand-based methods, there is a general trend of combining expertise from different scientific branches to optimize each step.

In summary, the ideal methodology should:

- Be free from a user-dependent alignment;
- Use as few relevant descriptors and features as possible;
- Employ rapid and reliable data-mining methods to identify the significant properties;
- Return a mechanistic interpretation, where possible;
- Have high predictive capabilities.

One way to circumvent the molecular alignment is to provide all the possible conformations adopted under physiological conditions by the dataset molecules and then let the model select the relevant ones. Hence all the molecular degrees of freedom should be thoroughly sampled and a sensible number of minima in the free energy landscape should be retained for post-processing. To this end, because of their propensity to generate thermodynamics ensembles in a reasonable time, stochastic simulation such as Monte Carlo-like schemes have been extensively used. Alternatively, researchers could use molecular dynamics simulations conducted in explicit solvents, which more closely depict the true molecular behavior. However, even for small molecules, there is little possibility that all the relevant conformations can be explored in detail in a feasible CPU time. This is because of the system's propensity to remain trapped in local minima. Scientists have shown that this limitation can be conveniently overcome, boosting the dynamics by means of enhanced sampling methods. To date, however, practical applications seem confined to the study of very small sets of molecules of biological interest (69-71).

Usually, the establishment of noncovalent interactions with a ligand is what triggers a receptor's temporary activation/inhibition, leading to a pharmacological effect. Since electrons in the outer shells are responsible for the molecular reactivity profiles, QM calculations appear to be more adequate, with respect to force-field-based potentials, in providing an accurate description of the forces occurring during biological processes. However, such treatments are time-demanding, even for small molecular systems with a reduced number of atoms. This helps explains why the use of QM-based descriptors in 3D QSARs literature is hard to find (72, 73). In addition, such calculations should ideally be carried out for each and every conformer generated during the earlier step, exponentially increasing the computational burden. However, the development of more efficient algorithms and increasingly fast processors indicates promising future possibilities.

Part of the success of 3D QSARs is due to the use of algorithms such as PLS, which can separate the useful information in a given molecular description from the background noise, and then relate it to a dependent variable. However, when several conformers are provided for each studied molecule, more complex approaches are needed to analyze the increased number of data points. A promising approach is the use of advanced supervised machine-learning methods, such as support vector machines (SVMs), which are extensively used for pattern recognition and classification purposes involving a great number of variables (74).

One way to assess the strengths of a newly developed protocol is to use publicly available benchmarks. The steroid dataset used for the CoMFA validation (16) as well as the Selwood dataset (75) have been used for years; however they are currently too small or too confined to particular chemical classes to offer general applicability. In order to compare a number of QSAR methodologies, Sutherland and coworkers compiled eight datasets and made them freely accessible to the scientific community. These datasets ranged from 66 to 397 compounds, inhibitors of diverse proteins belonging to several EC classes (76). But Manchester and Czermiński showed that, apparently, no real benchmark can distinguish the merits of 3D QSAR methods. This is mostly because too few observations are used to describe the response. They advocate the use of simulated data instead (77).

Sadly, as Leach and colleagues have pointed out, validation standards for pharmacophore mappings are inadequate (61). Moreover, little effort has been made to develop novel features or more accurate ways of dealing with pharmacophore plasticity, which is a direct consequence of target flexibility.

Quantitative structure-properties relationships have been successfully used for several years in a vast number of fields. Despite the advent of more readily interpretable structure-based approaches, they still play a major role in drug discovery. Regrettably, as Dearden and colleagues have noted, analyses are still sometimes carried out in a non-satisfactory manner (78). This is mostly because the practitioner has excessive freedom when using only measured responses for a series of molecules sketched on paper. Moreover, QSAR is not free from intrinsic weakness and limitations (79). Future improvements will likely focus on limiting human involvement as much as possible, while safeguarding chemical interpretability. Although the use of powerful tools (such as the ones mentioned above) could significantly improve the quality of each phase, there may be concerns about the creation of solid self-consistent models that lack chemical transferability. Having being used for over three decades, pharmacophore-based approaches are simple and adaptable. Although technically classified as a ligand-based approach, pharmacophore mapping can be conveniently tuned using the knowledge of the binders' molecular counterpart, constituting an ideal bridge between the ever-competing ligand- and structure-based followers. Notably 3D pharmacophores can also be used to pre-filter large datasets before virtual screening efforts in a docking framework, lightening the computational load. Despite some inevitable technical limitations, pharmacophore-based protocols are nowadays successfully applied to a number of different purposes spanning from database pre-filtering to hit enrichment. Given their robust nature, they are likely to play a pivotal role in the future.

No computational technique can be easily used as a black box for medicinal chemistry. They all reward users who have in-depth knowledge of the biological problem they are trying to solve. Finding quantitative relationships between chemical structures and properties should be seen as a round trip journey. During the outbound ride, molecules are transformed into numerical descriptors. At the destination, the variables correlated with the experimental observable are identified. The return journey should produce a mechanistic interpretation useful for designing molecules with improved features or detecting putative actives within a crowd.

6. ACKNOWLEDGMENTS

Special thanks are due to Prof. Angelo Carotti who read and commented on earlier versions of this manuscript.

7. REFERENCES

1. Y. C. Martin, J. L. Kofron and L. M. Traphagen: Do structurally similar molecules have similar biological activity? *J Med Chem*, 45(19), 4350-8 (2002)

2. T. I. Oprea and J. Gottfries: Chemography: the art of navigating in chemical space. *J Comb Chem*, 3(2), 157-66 (2001)

3. T. Engel: Basic overview of chemoinformatics. *J Chem Inf Model*, 46(6), 2267-77 (2006)

4. J. Gasteiger: Chemoinformatics: a new field with a long tradition. *Anal Bioanal Chem*, 384(1), 57-64 (2006)

5. Handbook of Molecular Descriptors. In: Ed R. Mannhold, H. Kubinyi&H. Timmerman. WILEY-VCH, Verlag GmbH, D-69469, Wemhem (Federal Republic of Germany) (2000)

6. H. Kubinyi: Chance favors the prepared mind--from serendipity to rational drug design. *J Recept Signal Transduct Res*, 19(1-4), 15-39 (1999)

7. M. C. Richet: Noté sur le rapport entre la toxicité et les proprietés physiques des corps. *Compt Rend Soc Biol*, 45, 775-776 (1893)

8. H. Meyer: Theorie der alkoholnarkose. *Arch Exp Pathol Pharmakol*, 42, 109-18 (1899)

9. E. Overton: Über die osmotischen eigenschaften der lebenden pflanzen und tierzelle. *Vierteljahrsshriften Naturforschungen Ges Zurich*, 40, 159-201 (1895)

10. L. P. Hammett: The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J Am Chem Soc*, 59(1), 96-103 (1937)

11. C. Hansch and T. Fujita: ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J Am Chem Soc*, 86(8), 1616-1626 (1964)

12. H. Kubinyi and O. H. Kehrhahn: Quantitative structure-activity relationships. VI. Non-linear dependence of

biological activity on hydrophobic character: calculation procedures for bilinear model. *Arzneimittelforschung*, 28(4), 598-601 (1978)

13. H. Kubinyi: Quantitative structure--activity relationships. 7. The bilinear model, a new model for nonlinear dependence of biological activity on hydrophobic character. *J Med Chem*, 20(5), 625-9 (1977)

14. S. M. Free, Jr. and J. W. Wilson: A Mathematical Contribution to Structure-Activity Studies. *J Med Chem*, 7, 395-9 (1964)

15. H. Berman, K. Henrick and H. Nakamura: Announcing the worldwide Protein Data Bank. *Nat Struct Biol*, 10(12), 980 (2003)

16. R. D. Cramer, D. E. Patterson and J. D. Bunce: Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc*, 110(18), 5959-5967 (1988)

17. S. Wold, A. Ruhe, H. Wold, W. J. Dunn and III: The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3), 735-743 (1984)

18. P. J. Goodford: A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem*, 28(7), 849-57 (1985)

19. J. Nilsson, H. Wikstrom, A. Smilde, S. Glase, T. Pugsley, G. Cruciani, M. Pastor and S. Clementi: GRID/GOLPE 3D quantitative structure-activity relationship study on a set of benzamides and naphthamides, with affinity for the dopamine D3 receptor subtype. *J Med Chem*, 40(6), 833-40 (1997)

20. G. Cruciani and K. A. Watson: Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b. *J Med Chem*, 37(16), 2589-601 (1994)

21. G. Klebe, U. Abraham and T. Mietzner: Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J Med Chem*, 37(24), 4130-46 (1994)

22. D. D. Robinson, P. J. Winn, P. D. Lyne and W. G. Richards: Self-organizing molecular field analysis: a tool for structure-activity studies. *J Med Chem*, 42(4), 573-83 (1999)

23. M. Catto, O. Nicolotti, F. Leonetti, A. Carotti, A. D. Favia, R. Soto-Otero, E. Mendez-Alvarez and A. Carotti: Structural insights into monoamine oxidase inhibitory potency and selectivity of 7-substituted coumarins from ligand- and target-based approaches. *J Med Chem*, 49(16), 4912-25 (2006)

24. C. Hansch, T. Klein, J. McClarin, R. Langridge and N. W. Cornell: A quantitative structure-activity relationship and molecular graphics analysis of hydrophobic effects in the interactions of inhibitors with alcohol dehydrogenase. J Med Chem*, 29(5), 615-20 (1986)*

25. A. Cavalli, G. Greco, E. Novellino and M. Recanatini: Linking CoMFA and protein homology models of enzyme-inhibitor interactions: an application to non-steroidal aromatase inhibitors. *Bioorg Med Chem*, 8(12), 2771-80 (2000)

26. A. M. Gamper, R. H. Winger, K. R. Liedl, C. A. Sotriffer, J. M. Varga, R. T. Kroemer and B. M. Rode: Comparative molecular field analysis of haptens docked to the multispecific antibody IgE(Lb4). *J Med Chem*, 39(20), 3882-8 (1996)

27. W. Sippl: Receptor-based 3D QSAR analysis of estrogen receptor ligands--merging the accuracy of receptor-based alignments with the computational efficiency of ligand-based methods. *J Comput Aided Mol Des*, 14(6), 559-72 (2000)

28. A. Carotti, R. N. Smith, S. Wong, C. Hansch, J. M. Blaney and R. Langridge: Papain hydrolysis of X-phenyl-N-methanesulfonyl glycinates: a quantitative structure-activity relationship and molecular graphics analysis. *Arch Biochem Biophys*, 229(1), 112-25 (1984)

29. A. Carotti, G. Casini and C. Hansch: Structure-activity relationship of the ficin hydrolysis of phenyl hippurates. Comparison with papain, actinidin, and bromelain. *J Med Chem*, 27(11), 1427-31 (1984)

30. A. Vedani, H. Briem, M. Dobler, H. Dollinger and D. R. McMasters: Multiple-conformation and protonation-state representation in 4D-QSAR: the neurokinin-1 receptor system. *J Med Chem*, 43(23), 4416-27 (2000)

31. A. Vedani and M. Dobler: 5D-QSAR: the key for simulating induced fit? *J Med Chem*, 45(11), 2139-49 (2002)

32. A. Vedani, M. Dobler and M. A. Lill: Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *J Med Chem*, 48(11), 3700-3 (2005)

33. M. Totrov: Atomic property fields: generalized 3D pharmacophoric potential for automated ligand superposition, pharmacophore elucidation and 3D QSAR. *Chem Biol Drug Des*, 71(1), 15-27 (2008)

34. L. B. Kier: Molecular orbital calculation of preferred conformations of acetylcholine, muscarine, and muscarone. *Mol Pharmacol*, 3(5), 487-94 (1967)

35. C. G. Wermuth, C. R. Ganellin, P. Lindberg and L. A. Mitscher: Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). *Pure Appl Chem*, 70(5), 1129-1143 (1998)

36. G. A. Patani and E. J. LaVoie: Bioisosterism: A Rational Approach in Drug Design. *Chem Rev*, 96(8), 3147-3176 (1996)

37. S. Cross and G. Cruciani: Molecular fields in drug discovery: getting old or reaching maturity? *Drug Discov Today*, 15(1-2), 23-32 (2010)

38. R. Guha and J. H. Van Drie: Structure--activity landscape index: identifying and quantifying activity cliffs. *J Chem Inf Model*, 48(3), 646-58 (2008)

39. G. M. Maggiora: On outliers and activity cliffs--why QSAR often disappoints. *J Chem Inf Model*, 46(4), 1535 (2006)

40. J. Bostrom, A. Hogner and S. Schmitt: Do structurally similar ligands bind in a similar fashion? *J Med Chem*, 49(23), 6716-25 (2006)

41. J. Gasteiger: Of molecules and humans. *J Med Chem*, 49(22), 6429-34 (2006)

42. G. Klebe, T. Mietzner and F. Weber: Methodological developments and strategies for a fast flexible superposition of drug-size molecules. *J Comput Aided Mol Des*, 13(1), 35-49 (1999)

43. P. Labute, C. Williams, M. Feher, E. Sourial and J. M. Schmidt: Flexible alignment of small molecules. *J Med Chem*, 44(10), 1483-90 (2001)

44. C. Lemmen and T. Lengauer: Computational methods for the structural alignment of molecules. *J Comput Aided Mol Des*, 14(3), 215-32 (2000)

45. F. Melani, P. Gratteri, M. Adamo and C. Bonaccini: Field interaction and geometrical overlap: a new simplex and experimental design based computational procedure for superposing small ligand molecules. *J Med Chem*, 46(8), 1359-71 (2003)

46. M. Arakawa, K. Hasegawa and K. Funatsu: Application of the novel molecular alignment method using the Hopfield Neural Network to 3D-QSAR. *J Chem Inf Comput Sci*, 43(5), 1396-402 (2003)

47. M. Arakawa, K. Hasegawa and K. Funatsu: Novel alignment method of small molecules using the Hopfield Neural Network. *J Chem Inf Comput Sci*, 43(5), 1390-5 (2003)

48. R. B. Hermann and D. K. Herron: OVID and SUPER: two overlap programs for drug design. *J Comput Aided Mol Des*, 5(6), 511-24 (1991)

49. F. Manaut, F. Sanz, J. Jose and M. Milesi: Automatic search for maximum similarity between molecular electrostatic potential distributions. *J Comput Aided Mol Des*, 5(4), 371-80 (1991)

50. B. D. Hudson, D. C. Whitley, M. G. Ford, M. Swain and J. W. Essex: Pattern recognition based on color-coded quantum mechanical surfaces for molecular alignment. *J Mol Model*, 14(1), 49-57 (2008)

51. J. Marialke, R. Korner, S. Tietze and J. Apostolakis: Graph-based molecular alignment (GMA). *J Chem Inf Model*, 47(2), 591-601 (2007)

52. J. F. Rico, R. López, I. Ema and G. Ramírez: Chemical forces in terms of the electron density. *Theor Chem Account*, 118(3), 709-721 (2007)

53. J. Manchester and R. Czerminski: SAMFA: simplifying molecular description for 3D-QSAR. *J Chem Inf Model*, 48(6), 1167-73 (2008)

54. M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi and S. Clementi: Generating optimal linear PLS estimations (GOLPE): an advanced chemometric tool for handling 3D-QSAR problems. *Quant Struct-Act Relat*, 12(1), 9-20 (1993)

55. K. Hasegawa and K. Funatsu: Partial least squares modeling and genetic algorithm optimization in quantitative structure-activity relationships. *SAR QSAR Environ Res*, 11(3-4), 189-209 (2000)

56. T. A. Andrea and H. Kalayeh: Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. *J Med Chem*, 34(9), 2824-36 (1991)

57. R. D. King, S. H. Muggleton, A. Srinivasan and M. J. Sternberg: Structure-activity relationships derived by machine learning: the use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *Proc Natl Acad Sci U S A*, 93(1), 438-42 (1996)

58. R. D. King, S. Muggleton, R. A. Lewis and M. J. Sternberg: Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc Natl Acad Sci U S A*, 89(23), 11322-6 (1992)

59. R. D. Cramer and B. Wendt: Pushing the boundaries of 3D-QSAR. *J Comput Aided Mol Des*, 21(1-3), 23-32 (2007)

60. R. Guha and J. H. Van Drie: Assessing how well a modeling protocol captures a structure-activity landscape. *J Chem Inf Model*, 48(8), 1716-28 (2008)

61. A. R. Leach, V. J. Gillet, R. A. Lewis and R. Taylor: Three-dimensional pharmacophore methods in drug discovery. *J Med Chem*, 53(2), 539-58 (2010)

62. D. Barnum, J. Greene, A. Smellie and P. Sprague: Identification of common functional configurations among molecules. *J Chem Inf Comput Sci*, 36(3), 563-71 (1996)

63. S. L. Dixon, A. M. Smondyrev, E. H. Knoll, S. N. Rao, D. E. Shaw and R. A. Friesner: PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and

preliminary results. *J Comput Aided Mol Des*, 20(10-11), 647-71 (2006)

64. H. A. Carlson: Protein flexibility and drug design: how to hit a moving target. *Curr Opin Chem Biol*, 6(4), 447-52 (2002)

65. G. Bottegoni, I. Kufareva, M. Totrov and R. Abagyan: Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. *J Med Chem*, 52(2), 397-406 (2009)

66. G. Bottegoni, I. Kufareva, M. Totrov and R. Abagyan: A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE). *J Comput Aided Mol Des*, 22(5), 311-25 (2008)

67. M. Rella, C. A. Rushworth, J. L. Guy, A. J. Turner, T. Langer and R. M. Jackson: Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors. *J Chem Inf Model*, 46(2), 708-16 (2006)

68. I. Nobeli, A. D. Favia and J. M. Thornton: Protein promiscuity and its implications for biotechnology. *Nat Biotechnol*, 27(2), 157-67 (2009)

69. D. Bernard, A. Coop and A. D. MacKerell Jr: Conformationally sampled pharmacophore for peptidic delta opioid ligands. *J Med Chem*, 48(24), 7773-80 (2005)

70. J. Hritz and C. Oostenbrink: Efficient free energy calculations for compounds with multiple stable conformations separated by high energy barriers. *J Phys Chem B*, 113(38), 12711-20 (2009)

71. J. Hritz and C. Oostenbrink: Hamiltonian replica exchange molecular dynamics using soft-core interactions. *J Chem Phys*, 128(14), 144121 (2008)

72. A. J. Chalk, B. Beck and T. Clark: A quantum mechanical/neural net model for boiling points with error estimation. *J Chem Inf Comput Sci*, 41(2), 457-62 (2001)

73. B. Beck, A. Breindl and T. Clark: QM/NN QSPR models with error estimation: vapor pressure and logP. *J Chem Inf Comput Sci*, 40(4), 1046-51 (2000)

74. D. Meyer, F. Leisch and K. Hornik: The support vector machine under test. *Neurocomputing*, 55(1-2), 169-186 (2003)

75. D. L. Selwood, D. J. Livingstone, J. C. Comley, A. B. O'Dowd, A. T. Hudson, P. Jackson, K. S. Jandu, V. S. Rose and J. N. Stables: Structure-activity relationships of antifilarial antimycin analogues: a multivariate pattern recognition study. *J Med Chem*, 33(1), 136-42 (1990)

76. J. J. Sutherland, L. A. O'Brien and D. F. Weaver: A comparison of methods for modeling quantitative structure-activity relationships. *J Med Chem*, 47(22), 5541-54 (2004)

77. J. Manchester and R. Czerminski: CAUTION: popular "benchmark" data sets do not distinguish the merits of 3D QSAR methods. *J Chem Inf Model*, 49(6), 1449-54 (2009)

78. J. C. Dearden, M. T. Cronin and K. L. Kaiser: How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ Res*, 20(3-4), 241-66 (2009)

79. S. R. Johnson: The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J Chem Inf Model*, 48(1), 25-6 (2008)

80. Phase, version 3.1, Schrödinger, LLC, New York, NY, 2009

Key Words: 3D QSAR, 3D QSPR, ligand-based drug design, structure-based drug design, pharmacophore, CoMFA, CoMSIA, molecular descriptor, PLS, Review

Send correspondence to: Angelo D. Favia, Istituto Italiano di Tecnologia (IIT), Drug Discovery and Development, via Morego 30, 16163 Genoa, Italy, Tel: 0039-010-71781576, Fax: 0039-010-7170187, E-mail: angelo.favia@iit.it

http://www.bioscience.org/current/vol16.htm