

## A Primer on Molecular Similarity in QSAR and Virtual Screening

### Part I – Descriptor Choice

Andreas Bender, PhD

<http://www.andreasbender.de/>

---

#### 1. Introduction

Methods such as QSAR and various cheminformatics techniques have gained huge popularity in recent years. This can partly be attributed to increased productivity pressure in pharmaceutical industry and the assumption that computational models can replace some experiments, but also to the fact that more data more validated modeling methods and more computational power are readily available.

While methods abound, one should also take a step back from time to time to look at the bigger state of affairs to ask oneself what has been gained, and which expectations were simply hyped too much and failed to keep their promises.

In this article, the first of a series, we will discuss what progress has been made since the early works, such as those by Crum Brown and Frazer<sup>1</sup> and Hansch<sup>2</sup>, who were among the first to correlate chemical structure and biological (or physicochemical) properties and to postulate a causal relationship between the two.

Relating a molecular property to the underlying structure involves three broad steps<sup>3</sup>: Firstly, the *representation of a molecule* in a way suitable for computerized treatment, often referred to as the choice of a *descriptor*. Secondly, the *choice of the variable* one attempts to model, often called the *endpoint* – which can be any molecular property that can be experimentally measured. Frequently used endpoints (since they are relevant in practice) are solubility and logP as physicochemical properties or bioactivity as a biological measurement variable. Thirdly, descriptors (input variables) and endpoints (output variables) need to be connected, by means of one of a variety of available model generation methods.

Each of those steps deserves particular attention and it is of crucial importance that the descriptor chosen to represent the system, the mathematical method employed to generate the model, and the property or endpoint measured are a suitable combination that at least gives the *possibility* of a successful model generation. If the system is not adequately described by suitable descriptors as input variables, a model is based on effectively random variables and thus doomed from the beginning. If the modeling method is not able to handle the expected input-output relationships, such as the application of a linear method to a bilinear relationship, no suitable fitting of the function can be expected. If the endpoints measured are not purely a function of the molecular structure, but also *of the procedure used to obtain the measurements*, information is missing from the system and the model can't be more accurate than the available, error-riddled data. Large variability of measurements between laboratories, or also between experimental procedures, decreases the

quality of the model that will be obtained in the end, so a dataset as homogenous as possible is of crucial importance here.

## 2. Molecular Descriptors

A large series of molecular descriptors has been published and recently reviewed<sup>3, 4</sup> and here we are not attempting to give a comprehensive overview. Instead, we will focus on three distinct aspects of molecular descriptors which have been put forward in recent literature: Firstly, how to choose descriptors for establishing meaningful and “as-trustworthy-as-possible” structure-activity models; secondly, what the ability of 2D and 3D descriptors is to perform “scaffold-hopping”; and thirdly how to assess the information content of descriptors we currently use.

### a) Guidelines for Choosing Molecular Descriptors – Less is (Sometimes) More

A multitude of molecular descriptors exists which can be used to describe a molecular structure, certainly ranging into the hundreds if not thousands, and they range from one-dimensional descriptors and two-dimensional (fragment) representations, over three-dimensional, conformation-dependent descriptors, to those incorporating conformational flexibility (sometimes referred to as four-dimensional descriptors). All of them, in particular the geometric descriptors, which are often easier to back-project, have their advantages and disadvantages – but upon combination an analogy from a different area holds: beer and wine, taken separately, may be delicious drinks. But in a mixture, they should rather be avoided.

What does that mean in the world of QSAR or cheminformatics? The analogy is referring to a more and more common tendency in recent years to calculate a very large number of descriptors – since they are readily available – and, by means of feature selection methods, to retain only those variables, which are found to improve predictive performance. While this might be intuitively the right thing to do, consider the following example (more detailed described in the original publications such as those by Topliss and Costello<sup>5</sup> as well as Livingstone and Salt<sup>6, 7</sup>).

Imagine you have a small number of data points, say 10 data points, which represent your measurement (such as solubility), and you also have 2 variables which describe your system. If those 2 variables are sensibly chosen, for example as molecular weight and polar surface area, you will probably be able to correlate the output variable, solubility, to a reasonable degree with the 2 input variables in a linear model. The model won't be ideal, since the two variables insufficiently describe the system under consideration (and also the number of data points is probably not sufficient), but you will be able to achieve some kind of model, and it will be statistically significant.

Now, imagine you are not only using 2 input variables, but rather you employ all descriptors current software packages are able to calculate – this will be in the area of thousands of descriptors. What will be the case now if you plug *all* of those input variables into a model, and try to find a simple, linear model with 2 input variables, which describes the solubility of your dataset of 10 compounds? You will most certainly obtain a model with close-to-perfect fit, simply because there is a *huge* number of models to choose from – and some of them, *just by accident*, will be able to model your data. (One counterargument is that a suitable model validation routine can alleviate the problem. But this is only

true to a certain extent – if the number of models is large enough, always a random model can be found that, by pure chance, will be able to fit *both* the training and test set nearly perfectly.)

The more precise mathematical background for this phenomenon has been described in some recent publications – and if one is thinking about developing structure-activity models involving feature selection, a look at them would certainly be beneficial to avoid some common pitfalls<sup>5-7</sup>. Using few and interpretable features not only gives a neater model, which can be interpreted – it is also more likely to be of statistical significance. (And you will be in sync with the likes of Albert Einstein, who once said: “Everything should be made as simple as possible, but not simpler”. He was certainly referring to QSAR models here.)

#### b) Scaffold Hopping Capability of Molecular Descriptors

“Scaffold hopping” is a term that has been used extensively in recent literature on virtual screening, and it describes the ability of molecular descriptors to identify molecules with *similar properties, despite different underlying structures (and scaffolds)*. Often the opinion is stated, “3D descriptors are better at finding diverse scaffolds”. While this might be the intuitive answer, a look into recent literature doesn’t give as clear a picture.

Recently, while sub-structural keys were found to retrieve less scaffolds for diverse classes than 3D fingerprints, topological (2D) fingerprints were found to be at least *en par* with them<sup>8</sup>, and superior performance of 2D descriptors on other test databases was attributed to the large number of close analogues. On the other hand, circular fingerprints (which are a 2D representation) were in a large comparative study found to retrieve a large number of active compounds as well as a large number of different scaffolds: indeed, a similar percentage of scaffolds as of active compounds from the whole database – which would hint into the opposite direction, at least for fingerprints not based on sub-structural keys<sup>9, 10</sup>. In addition, it has been suggested that the question whether descriptors are able to identify novel scaffolds or not is heavily depending on the particular dataset under consideration as shown on four different targets (all from different target classes)<sup>11</sup>. Therefore, are 3D descriptors more likely than 2D descriptors to identify novel scaffolds in a virtual screening setting? Possibly this is true for some 2D/3D descriptor combinations, but it is still open to discussion whether this is due to an *inherent* property of the descriptor dimensionality, or each particular descriptor definition. Given that the global spatial arrangement of atoms is, by and large, already defined by the (local) connectivity information, *it might be possible that no too large intrinsic bias between 2D and 3D descriptors exists*. Clearly, further research is needed here.

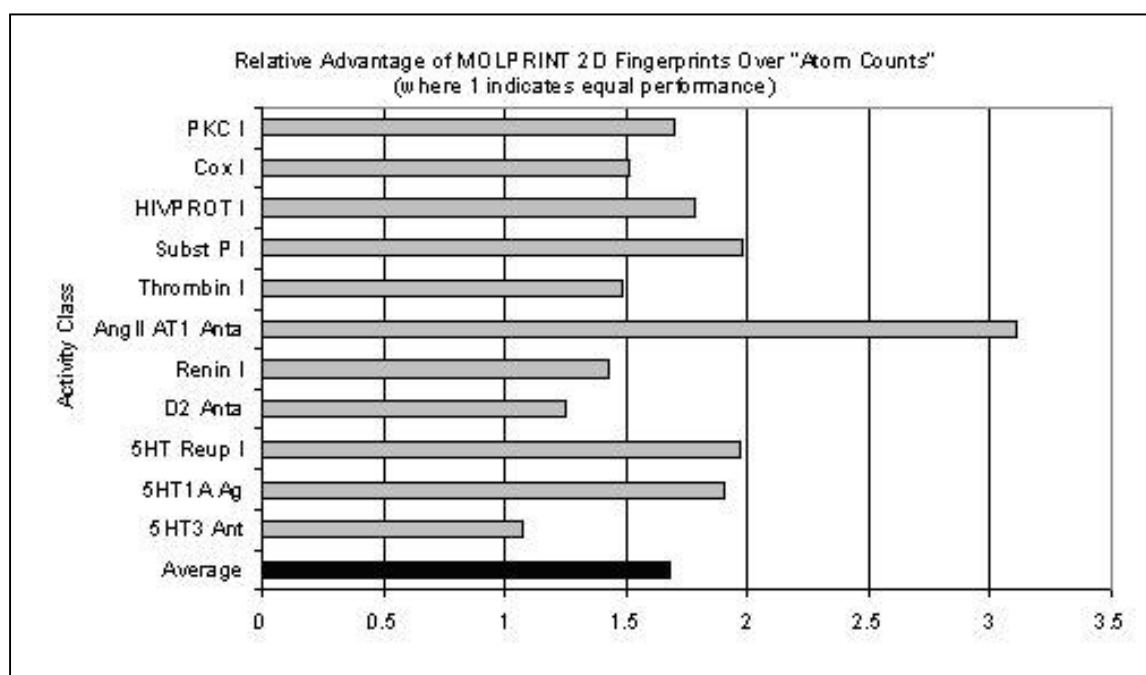
#### c) The Information Content of Current Descriptors

While circular fingerprints such as ECFP4<sup>12</sup> or MOLPRINT 2D fingerprints<sup>13</sup> have been shown to be information-rich, and currently the best-performing descriptors available as benchmarked on standard dataset<sup>9</sup>, the question arises how well those descriptors actually perform, *compared to not random, but a very dumb, basic classifier*.

This work has indeed recently been performed, with quite surprising results<sup>14</sup>. Namely, molecules in a virtual screening setting were described by simple descriptors, which didn’t include information about the connectivity of the molecule at all. Molecules were only assigned descriptors based on simple

“atom counts”, which contained the frequency of heavy atoms of different types (carbon, nitrogen, oxygen and so on) in the molecule – and nothing else. The similarity of molecules was assessed *via* the distance of those 12-dimensional count vectors, and the number of active compounds retrieved was compared to that *via* a standard circular fingerprint (MOLPRINT 2D) in combination with the Tanimoto Coefficient. Given enrichments for current virtual screening methods which are often in the range of 20-fold (20 times better than random) and higher, it could be expected that simply counting atoms was to perform much worse. But, to the contrary, the results obtained on a standard dataset<sup>15</sup> were rather surprising.

Shown in the figure below are the numbers of active compounds retrieved by the “dumb” atom counts, relative to the number of active compounds retrieved by conventional circular fingerprints. A number of 1 effectively means that one method performs as well as the other, while a factor of 2 for example indicates that the conventional fingerprints perform twice as well (retrieve twice as many actives) as the dumb atom counts.



Over the 11 classes of active compounds it can be seen that the difference between the methods varies between a factor of about 1 and a factor of about 3. This means, conventional fingerprints perform better than counting atoms overall, no questions. But overall, they *only perform not even twice as well* as simply counting atoms – where, on average, circular fingerprints are able to obtain enrichments of around 7, counting atoms also obtains enrichments of around 4. So – are we so much better than counting atoms right now? Overall, we certainly do perform better. But not as much better as one might expect, not even twice as well.

(More recently the “performance of the atom count descriptor” has also been evaluated on a another dataset, with similar results<sup>16</sup>. Thus, the above tendency seems to be general, hinting at the

possibility that descriptors which just count atoms already capture a surprisingly large part of the total information content of other descriptors employed in virtual screening.)

### 3. Summary and Conclusions

While a discussion of molecular descriptors for QSAR and virtual screening studies would fill several volumes, I here decided to focus on three distinct, but crucial aspects of how to select molecular descriptors for those tasks. Firstly, *a small number of meaningful descriptors* not only give models, which are *easier to interpret*, they are also *more likely to be of statistical significance*. If one is planning to develop QSAR models, it therefore saves disappointment later to think about which variables to include into your analysis beforehand. Secondly, it has often been claimed that 3D approaches are more able to discover novel scaffolds with similar properties than 2D approaches. While this is certainly true for some of the particular descriptors one can use, it is still open to discussion whether this is an *inherent* property of the dimensionality of the descriptor or the particular descriptor definition.

Since the (global) spatial arrangement of atoms is to a good extent defined by the (local) connectivity information though, *it might be possible that no too large an intrinsic bias between 2D and 3D descriptors exists*. Finally, while virtual screening and QSAR seem to work in many cases, they are no perfect descriptions of the system we are dealing with, but rather a statistical correlation that we are trying to interpret as causal relationships and make according modifications to structures. Many cases have shown that there is some validity to this approach – but current descriptors still deserve improvements (which are certainly to be delivered by bright current and future graduate students).

NOTE:

To be followed by

Part II – How reliable are experimental measurements (endpoints) in QSAR studies?

Part III – Connecting descriptors and experimental measurements – model generation.

## References

1. Crum Brown, A.; Frazer, T., *Trans. Roy. Soc. Edinburgh* **1868**, 24, 151.
2. Hansch, C.; Fujita, T., Rho-Sigma-Pi Analysis . Method for Correlation of Biological Activity + Chemical Structure. *J. Am. Chem. Soc.* **1964**, 86, (8), 1616-&.
3. Bender, A.; Glen, R. C., Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, 2, (22), 3204-3218.
4. Bender, A.; Jenkins, J. L.; Li, Q.; Adams, S. E.; Cannon, E. O.; Glen, R. C., Molecular Similarity: Advances in Methods, Applications, and Validations in Virtual Screening and QSAR. *Annual Reports In Computational Chemistry* **2006**, 2, 141 - 168.
5. Topliss, J. G.; Edwards, R. P., Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **1979**, 22, (10), 1238-44.
6. Livingstone, D. J.; Salt, D. W., Judging the significance of multiple linear regression models. *J Med Chem* **2005**, 48, (3), 661-3.
7. Salt, D. W.; Ajmani, S.; Crichton, R.; Livingstone, D. J., An improved approximation to the estimation of the critical f values in best subset regression. *J Chem Inf Model* **2007**, 47, (1), 143-9.
8. Renner, S.; Schneider, G., Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem* **2006**, 1, (2), 181-5.
9. Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J., Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, 9, (3), 199-204.
10. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A., Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, 2, (22), 3256-3266.
11. Good, A. C.; Hermsmeier, M. A.; Hindle, S. A., Measuring CAMD technique performance: A virtual screening case study in the design of validation experiments. *J. Comput.-Aided Mol. Des.* **2004**, 18, (7), 529-536.
12. SciTegic, I., San Diego, CA. - <http://www.scitegic.com>., SciTegic, Inc., San Diego, CA. - <http://www.scitegic.com>. In.
13. Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S., Molecular similarity searching using atom environments, information-based feature selection, and a naive bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, 44, (1), 170-178.
14. Bender, A.; Glen, R. C., A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Model.* **2005**, 45, (5), 1369-1375.
15. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A., Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, 44, (3), 1177-1185.
16. Melville, J. L.; Riley, J. F.; Hirst, J. D., Similarity by compression. *J Chem Inf Model* **2007**, 47, (1), 25-33.