

A Primer on Molecular Similarity in QSAR and Virtual Screening

Part III – Connecting descriptors and experimental measurements – model generation

By Andreas Bender, PhD

1. Introduction

The generation of models relating structure and some form of measurement of structural properties, be it bioactivity or any physicochemical property, consists of the description of the problem, the choice of a dataset of experimental endpoints, and the construction of the actual mathematical model. The descriptor generation was discussed in the first part of this primer, with the conclusion to choose descriptors relevant to the problem and of not overly complex nature. In the second part we were discussing experimental endpoints that can be used as “output variables” of models. Here the conclusion was that also experimental data are not without fault – data measured in a single laboratory may show significant differences (e.g. in high-throughput screenings), and combining results from different labs might be even more error-prone. Therefore, a clean dataset, measured by a single, well-defined experimental procedure should be used in the ideal case. In the current, concluding part of the primer I would like to discuss the basics of constructing a useful mathematical model that connects molecular descriptors as an independent variable, and outputs (predicted) properties as the dependent variable. The focus of this part of the primer will be on two steps commonly performed in the generation of QSAR/QSPR models, namely feature selection and model validation.

2. Feature Selection

One of the common first steps in QSAR/QSPR modeling (although not necessarily the best one) is to calculate a large number of features as a function of the molecular connectivity table (or 3D structure, or electronic wave function). Several thousand descriptors exist, with new ones published about weekly[1]. Next, often feature selection is employed to find which variables are beneficial to give a good regression or classification result – so a multitude of models is generated, validated against a set of internal and external validation sets, and the process of feature selection is repeated until the ‘best’ model according to user preferences is obtained. There is some reason for this ‘irrational’ process: Initially, when only a set of structures with their associated properties is known (but no knowledge whatsoever about the target protein, or the physical properties of a solute-solvent interaction) all the user is aware of is the structures – and no knowledge which properties of the structures may be responsible for the observed effect. If one knows that a target interaction in QSAR is dominated by hydrogen donor interactions, for example, one can tailor his descriptors – but this is often not known from the onset. While this trial and error approach seems ‘intuitive’ and hence can be a sensible approach in some cases, I would like to allude to an important point in this process: That models, generated *via* feature selection, are sometimes not as good as one might think from the statistics!

Firstly, imagine the situation that you have a small set of only ten compounds with measured activities against a target A. If you calculate hundred of descriptors for each compound, and combine this with a technique such as neural networks or support vector machines, then generate thousands of possible models ... then you *necessarily* will find models which are able to model your input data, simply by pure chance[2]! This is an important point, and it refers both to the number of features used as an input, as well as the modeling technique which might be more or less flexible. (‘Flexible’ here refers to the number of different models than can be fitted by a

given technique, also known as the ‘hypothesis space’ accessible to the model.) As a rule of thumb, you need to be aware of this issue in particular in case of a *larger number of descriptors, a flexible modeling technique (such as neural networks, support vector machines, and other nonlinear techniques)*, as well as *a small dataset to train and test with*. For neural networks in particular, a very good article by David Livingstone gives a discussion of the above points[3].

In the context of the “model significance” one of the often measured figures of merit of multiple linear regression models are the so-called “F measures”, which tell the user how “significant” (unlikely to occur by pure chance) a given model is. Now, imagine in the case of a small dataset, using lots of descriptors, from the millions of models you generate you find one that gives you a very high correlation coefficient with the test dataset. Is this model a “significant” model”? It might well be one – but it might also have occurred by chance, given that you firstly created *lots* of models to choose from. The F measure was initially derived if you have a single shot at creating a model, and not a large number of them. In a recent publication it was described how the significance of F measures varies if feature selection is performed[4] – and, as might be expected, the more features (models) to choose from, the less significant a given correlation becomes at identical dataset size.

3. Model Validation

As outlined above, a given correlation coefficient (or RMSE, for that matter) does not tell you a whole lot about the performance of your model. Of relevance are also the number of descriptors, the size of your dataset, the training/test/validation set split, the diversity of structures (which define the applicability domain for new compounds), the quality of your experimental data – and these are only the most important factors. But some of them are fixed and some are not: Let’s assume you have a given dataset, with data points of a given quality. Then how do you go about creating the best model possible with your data? How do you ensure applicability to future structures, how do you *validate* your model? We will now discuss some of the methods at our disposal to ensure quality of the models we create.

First of all, it is crucial to use multiple splits of your data – and this means, usually, three sets: One set to derive parameters from your model, one set to judge quality of your model (and perform model selection), and one set to assess performance of your model. An alternative is to use cross-validation: Your dataset is split into, for example, five different parts of equal size. In each run, four of the parts are used to train the model, and model performance is assessed on the fifth part in each run, and later averaged. Sometimes also “leave-one-out” cross-validation has been performed, where every compound is left out in turn and predicted, but this practice has been strongly advocated against and it should not be used anymore[5] – since the model performance one obtains this way is not predictive of the performance for new compounds at all. For details and on which method to use in which case, the reader is referred to the literature, since it all depends on the size of the dataset one uses[6]. Also of tremendous value is to use leave-multiple-out splits (so, to use a set of multiple compounds to judge model performance), but not to do so once and for all systematically (as in conventional cross-validation) but to repeat new splits over and over again[7]. From applications shown[8] this protocol gives better judgment of model performance, along with less complex and smaller models.

Several additional techniques exist to show that your model does not fit your output function by pure chance, one of them being Y-scrambling[9]. Here, the output variable is randomly permuted, and the model is fit to the new (random) variable. If similar model performance as that with the real output variable can be obtained, this indicates a high likelihood that the model

obtained its credentials purely by chance. (Typically, for example the correlation coefficients obtainable by Y scrambling should be much lower – if for the real model a correlation of for example 0.7 can be observed, the models obtained using random scrambling should not show correlation larger than maybe half that number. As always, this depends on the dataset size, the modeling algorithm, the number of random scramblings performed, etc., and for suitably large datasets correlations on the scrambles output variable should not differ significantly from zero.)

What also needs to be kept in mind, no matter what you generate a model for, is its applicability domain – the compounds for which you are confident to achieve good predictions. This relates much to the chemistry covered in the training set, but since models are based on descriptors instead of the structures themselves, one school of thought is to only make predictions for compounds whose descriptors fall into the ranges covered by the training set. Applicability domains have become more and more important in the recent cheminformatics literature, and for readers in the business of creating predictive models I would like to refer you to some recent articles in the area[10-12].

4. Summary and Conclusions

In this final primer of the series of generating structure-activity models we discussed some of the dangers we encounter when using descriptors, and experimental data, to generate a mathematical descriptor-property relationship. It should be clear from the paragraphs above that the generation of a reliable QSAR model is no trivial task, and that its quality depends on many different factors. To summarize, it is best to have sufficient and reliable experimental data (reproducible; usually from a single source); to use a small number of descriptors which are relevant to the problem (but not fewer of them); to apply a modeling technique that is as simply as possible (but no simpler), and to validate the model using repeated, appropriately-sized dataset splits. If all those points are considered, and you apply your final model to compounds you think are in its applicability domain, you can still not be sure that the model will make good predictions in the future, but at least you applied attention to all the details and did the best you could.

And all that this leaves me to do is wishing you good luck for your next QSAR study!

References

1. Todeschini, R.; Consonni, V., *Handbook of Molecular Descriptors*. Wiley-VCH: Weinheim, 2000.
2. Topliss, J. G.; Edwards, R. P., Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **1979**, 22, (10), 1238-44.
3. Livingstone, D. J.; Manallack, D. T.; Tetko, I. V., Data modelling with neural networks: Advantages and limitations. *J. Comput.-Aided Mol. Des.* **1997**, 11, (2), 135-142.
4. Livingstone, D. J.; Salt, D. W., Judging the significance of multiple linear regression models. *J. Med. Chem.* **2005**, 48, (3), 661-3.
5. Golbraikh, A.; Tropsha, A., Beware of q²! *J. Mol. Graph. Model.* **2002**, 20, (4), 269-76.
6. Hawkins, D. M.; Basak, S. C.; Mills, D., Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, 43, (2), 579-86.
7. Baumann, K.; Albert, H.; von Korff, M., A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I. Search algorithm, theory and simulations. *J. Chemometr.* **2002**, 16, (7), 339-350.
8. Baumann, K.; von Korff, M.; Albert, H., A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part II. Practical applications. *J. Chemometr.* **2002**, 16, (7), 351-360.
9. Papa, E.; Villa, F.; Gramatica, P., Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow). *J. Chem Inf. Model.* **2005**, 45, (5), 1256-1266.
10. Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O., A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *J. Chem. Inf. Model.* **2005**, 45, (4), 839 - 849.
11. Guha, R.; Jurs, P. C., Determining the validity of a QSAR model - A classification approach. *J. Chem. Inf. Model.* **2005**, 45, (1), 65-73.
12. Walker, J. D.; Carlsen, L.; Jaworska, J., Improving opportunities for regulatory acceptance of QSARs: The importance of model domain, uncertainty, validity and predictability. *QSAR Comb. Sci.* **2003**, 22, (3), 346-350.