AstraZeneca
IMED Biotech Unit

A Million to One: Drug Discovery is a Numbers Game

Dr Graeme Robb
AstraZeneca, Cheshire, UK

The number of potential organic molecules that could exist is estimated to be more than 10^60, yet high-throughput screening (HTS) methods are restricted to 10^6 – 10^9 molecules, of which frequently less than 10^3 molecules will show any desired biological activity. For this approach to be successful, we must ensure our 10^6 subset of molecule is representative of the greater 10^60 set. However, the molecules in a pharma company's historical collection are typically unrepresentative of the greater 'chemical space'. Are there means of supplementing this set to be more representative? Following on from this the challenge of drug discovery is: which molecules should we design next in order to maximise information and minimise costly synthesis of new molecules? Normally a data scientist might consider Design of Experiments (DoE) to achieve this, but in the multi-dimensional world of chemical space, this is a challenging task. The unique combination of interactive visualisations, DoE capabilities and data manipulation tools within JMP enable us to incorporate chemically-aware methods to systematically explore and assess large, complex datasets. In this way we analyse the existing data in order to determine what to make, so as to maximise input for the next iteration, accelerating progress in drug discovery.

Introduction

Drug discovery is a long journey, from millions of potential start-points to a single drug candidate molecule that may one day become a marketed drug. An analysis across the pharmaceutical industry reveals that it takes between 11 and 14 years from project inception till a drug reaches market. The average cost per successful drug is $5 billion (Forbes, 2014). There are many reasons for these long timescales and huge costs. From a statistical point of view one of the reasons is that we are looking for an extremely unlikely outcome.

Given the variety of ways in which atoms can be put together to make molecule, and restrictions on what could exist in nature and what would be a 'drug-like' molecule, we can estimate the total number of drug molecule that can exist. Total enumeration has been performed up to 17 atoms, which results in 166 billion combinations. Drugs frequently contain 40 or more atoms, however and it has been estimate that there are 10^60 potential drug-like molecules.

The traditional method of finding 'hits' (molecules with some activity against the target protein) is to screen large collections of compounds against a protein assay, known as high-throughput screening (HTS). The numbers of compounds that can be screened in this way is around a million. Recently we have seen the creation of DNA-encoded libraries, where mixtures of several compounds can be tested at once. These methods can easily screen billions of compounds. The number of compounds that can be easily screened is however just a tiny fraction of the potential compounds available to us. How then can we hope to be successful with these methods?

We face a similar problem in the next stage of drug discovery. The identification of a hit compound (or series of similar compounds) allows drug designers to focus on a single 'scaffold'. The task is then to synthesise and test analogues, looking to explore

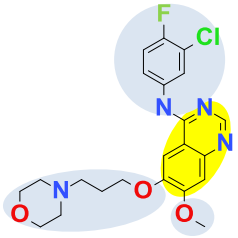


Fig. 1 – A typical drug molecule (Iressa) showing scaffold (yellow) and substituent groups (blue).

different groups off the core scaffold and in so doing investigate the local structure-activity relationships (SAR), see figure 1. The challenge here is that, even restricting ourselves to small numbers of atoms, there are hundreds of potential substructure groups to choose from and many more for larger numbers of atoms. As drug designers, how do we choose which compounds to make in order to allow rapid and efficient exploration of the SAR?

In order to simplify the problem we invoke the argument that similar chemical structures have similar properties. This implies the existence of a 'chemical space' with multiple dimensions in which all potential chemical structures exist. Therefore an ideal screening set for HTS would consist of a set of compounds with properties that space them evenly throughout chemical space. This works fine conceptually, but we have no workable definition of chemical space we can use. We do have a variety of methods of calculating properties of chemical structures but find that these will always give an incomplete description of chemical structure and the range of potential structures is simply too big. In reality, though drug companies supplement their screening collections with new compounds on a regular basis, we must acknowledge that screening sets cannot represent the wider chemical space, which is the reason that so many HTS's return only very weak hits or no hits at all.

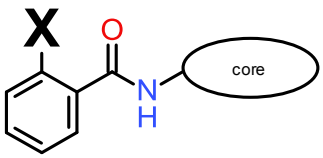


Fig. 2 – Representative scaffold of GPCR active compound, where X shows the position of the group under investigation.

In the scenario where chemistry has been limited to a single scaffold we limit the scope of the available chemistry sufficiently to begin to apply statistical methods to maximise diversity in sets of compounds. Figure 1 shows a compound with three groups around a fixed scaffold. We tend to vary each of these in turn during the first round of exploration (before making best combinations in subsequent rounds), synthesising 'libraries' of similar compounds which differ only at one substituent position

The traditional method for designing a chemical library is to pick a tractable synthetic method and make as many compounds as possible from the available reagents. While this is efficient in terms of costs per compound, a single synthetic route will lead to a less diverse set of compounds.

Here we investigate if it is possible to accurately represent a constrained set of compounds with numerical descriptors and so apply the principles of Design of Experiment (DoE) to generate more diversity, with fewer compounds.

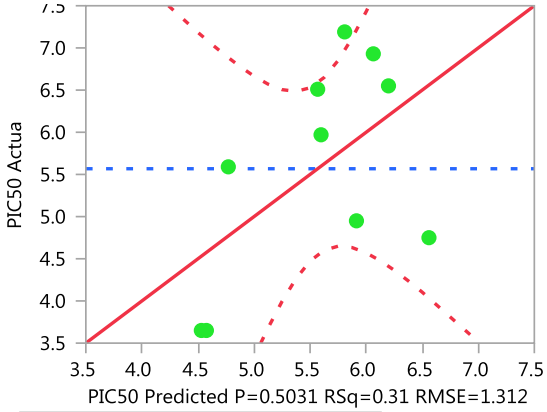
Methods

A hit chompund was identified from screening that was active against a class-A G-protein coupled receptor of interest. The identified molecular scaffold contained a substituted phenyl ring. We wished to explore the SAR of this substituent position. Based on experience in the field and an idea of the different underlying properties of molecular substructures we picked three properties to represent 94 small chemical substituents. The properties are:

- Pi (a measure of lipophilicity, a term describing partitioning between aqueous and organic solvent);
- Molecular Refractivity (MR, a measure of molecular size of the group)
- $Sigma_p$ (a measure of the electronic properties of the group, e.g. electron-withdrawing or electron-donating)

Judging from principal component analysis (PCA) the size and lipophilicity terms are moderately correlated, however it was judged to be important to leave both terms in the model so as to describe the cases where these terms diverge.

The problem with using these continuous values as inputs for DoE design is that the particular combination of variables that it selects may not exist within the compound set. Instead we converted each into a categorical variable by binning each into high, medium and low values ('high' and 'low' were the upper and lower quartiles, while 'medium' was the remaining 50%). A custom design based on these inputs selected a minimum of 9 compounds to represent the set, which were synthesised. A total of 10 compounds were available for modelling (including the original hit).



Summary of Fit

RSquare	0.305193
RSquare Adj	-0.04221
Root Mean Square Error	1.312043
Mean of Response	5.567264
Observations (or Sum Wgts	10

Fig. 3 – The output of a MLR model based on pi, MR and $sigma_p$ and built on the 10 initial compounds. There is no relationship to pIC50.

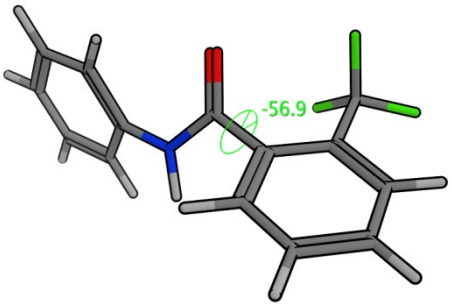


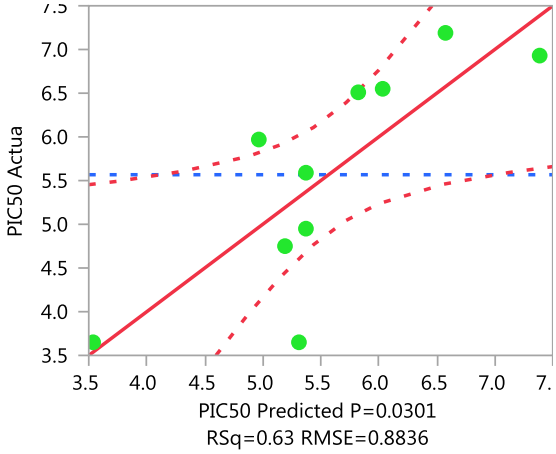
Fig. 4 – Representation of the 3D structure of one compound in the training set, showing how a bulky substituent (CF_3) can induce a twist in the molecule 57 degrees out of plane.

Results

A simple multiple-linear regression (MLR) model was fitted to the data for the 10 compound training set. Disappointingly, no model for the measured activity (pIC50) was possible, see figure 3.

We considered what other properties could be used to describe these compounds, particularly what property might describe the outlier compounds and explain their apparently anomalous behaviour. We realised that the outlier compounds would have a noticeably different shape to the majority of the set. We set about using molecular mechanics to calculate the minimum energy conformation of each molecule and selected the dihedral angle between the phenyl ring and adjacent carbonyl (see figure 4) to represent differences in the shape. Including this descriptor in the MLR model resulted in a greatly superior fit. Both pi and $sigma_p$ descriptors from out of the model as not significant (at 95% confidence). The final model is as shown in figure 5. The combination of MR (size) and dihedral angle adequately describe ~63% of the variance in pIC50.

Being a relatively simple model, we can also interpret it in order to understand the observed relationship. In Figure 5 the Parameter Estimates show that there is a

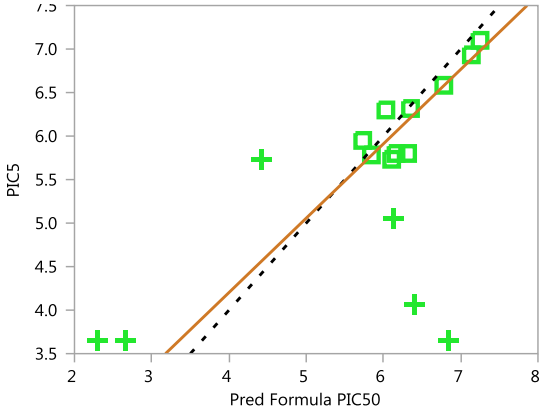


Summary of Fit	
RSquare	0.632318
RSquare Adj	0.527265
Root Mean Square Error	0.883647
Mean of Response	5.567264
Observations (or Sum Wgts	10

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.8427963	1.320405	2.15	0.0683
MR	-0.069176	0.023875	-2.90	0.0231 *
Dihedral Angle (corrected	0.083281	0.028355	2.94	0.0218 *

Fig. 5 – The output of a MLR model based on MR and dihedral angle and built on the 10 initial compounds. There is a reasonable relationship to pIC50.



Summary of Fit	
RSquare	0.775641
RSquare Adj	0.747596
Root Mean Square Error	0.253969
Mean of Response	6.2281
Observations (or Sum Wgts	10

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.7838726	1.03833	0.75	0.4719
Pred Formula PIC5	0.8545902	0.1625	5.26	0.0008 *

Fig. 6 – The predictions of the MLR model plotted against measured data for the 17 compound test-set. The dotted line is the 1:1 line while the red line ans summary of fir refer to the hollow squares only

negative correlation with MR and a positive correlation with dihedral angle, *i.e.* the most potent compounds are those with greatest twist out of plane but have the smallest size.

On the basis of this we made another 17 compounds, most of which had good predicted pIC50 (plus a few negative controls).

Figure 6 shows the measured pIC50 plotted against the predicted pIC50 for the 17 compounds of the training set. Though not every compound predicted well, 10 of the compounds (shown in hollow squares in figure 6) are predicted to be potent and confirmed to be so.

Satisfied that the model was both predictive (useful in a forward direction for telling us properties of unmade compounds) and interpretable (useful in a reverse direction for telling us which properties are important for potency) we determined that one substituent group from the existing set represented the optimum potency that could be achieved, even when considering the larger set of compound that was possible. We were also able to determine that it was not worth expanding our scope to larger substituents with more atoms as we had already observed a negative correlation with molecular size.

Conclusions

In this work we have shown that the principles of Design of Experiment (DoE) can be applied to drug design, but that careful consideration needs to be given to framing the SAR question and in limiting as much variability as possible so that it becomes possible to represent chemical diversity through a small number of descriptors.

We have also shown that descriptor selection requires a certain amount of trial and error, it being difficult to say which descriptors are most important *a priori*.

Successful application of DoE in this case permitted a full exploration of the structure activity landscape, making only a quarter of the potential compounds and ensuring the exploration was focussed on the most interesting property space.

Author biography

Dr Graeme Robb is a computational chemist at AstraZeneca, where he has worked for the last 12 years within oncology and diabetes disease settings. His role involves using 3D molecular structure and numerical modelling techniques to predict activity and properties of small, organic compounds in order to accelerate the process of finding potential new drugs. He holds a Ph.D. and a Masters degree in chemistry from the University of Edinburgh.