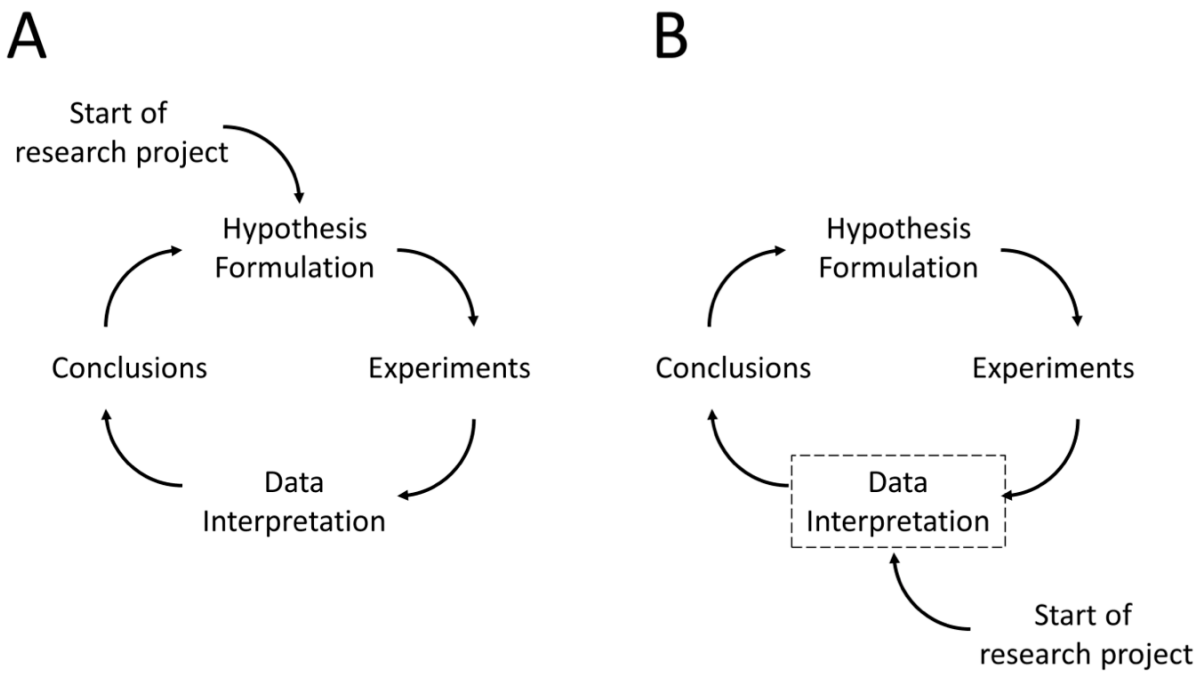
Déjà Vu - Réjà Vu : on knowledge-based approaches linking ligand and target information to bioactivity

Westen, G.J.P. van

Chapter 8

Conclusions and Future Perspectives



Contents

8.1 Personal observations in computational chemistry.

8.1.1 Primarily a Scientist. After four years of working as a PhD candidate I should like to start this chapter with several personal observations made during that time (these ought to be regarded as just that, personal observations).

Computational disciplines routinely handle datasets of immense size. Intuitive tools, slick graphical user interfaces (GUI) and standardized data formatting rest on advanced computational concepts. However, they also bring about two major pitfalls that should not be overlooked.

Firstly, in the predictable world of computational analysis it is sometimes difficult to remember concepts relevant in laboratory work like experimental error and reproducibility. It is tempting to treat K_i values that are presented as factual data points, however a 'data smear' is a more appropriate description. The trouble is that one doesn't know what the *true* value is for these 'data smears' while these points are often treated as if one does. Success is not always guaranteed when reproducing an experiment and no two experiments are identical. Whereas in computational approaches reproducing an experiment will usually lead to success and two experiments can be run on opposite sides of the world in different labs leading to identical results (down to three decimals or better).

Secondly, when confronted with a large data set it is often difficult to find a good starting location. However, the computational scientist should always remember to carefully curate the data to the best of his ability before embarking on any modeling approach. "Not all data points are created equal", and a model is as good as the error in the data. One should remember that the data feeding computational work is always a hand me down from other scientists. One is not always as lucky as to personally know this previous owner.

Hence, the most important conclusion from my thesis is that, despite the highly organized and reproducible fashion is which computational experiments are performed, computational chemistry remains a true scientific discipline with errors, uncertainty and limited capabilities. However, the location where this uncertainty resides is often not easily spotted in any experiment. Therefore no model or analysis should ever be treated as routine and getting to know one's data always pays off in the long run.

8.2 Observations from this thesis

8.2.1 Meeting the pre-set aims. This thesis focuses on knowledge-based computational approaches to combine data from different disciplines that are relevant in medicinal chemistry and drug discovery. The rationale was that these disciplines, chemistry, biology and bioactivity, are *complementary* and that there is much to be gained by combining them. After several research chapters we can conclude that this is indeed the case. Combining data, the addition of extra information improves models. However, when the matter in which this data is combined is not accurately represented by the descriptors (e.g., combining allosteric and orthosteric compounds without differentiating in the binding site) the effect is that the final combined models are inferior to individual models. Only a thorough validation can spot these problems and care should be taken to validate the performance of a model on each target individually rather than over all targets as this masks a single target that performs worse than the average.

8.2.2 PCM, a technique with many names. Arguably the most important subject of the thesis is proteochemometric (PCM) modeling, which plays a central role in the majority of the chapters and is reviewed in **chapter 2**. We find that PCM is a technique that is gaining ground in scientific literature as it is applied by diverse groups to diverse targets. At the same time the diverse user base also leads to fragmentation in literature as the same technique carries multiple names (PCM,[1, 2] chemogenomics,[3] Protein-Ligand fingerprint,[4] Multi-Assay-Based SAR [5]) all of which combine data from related targets. Yet, the difference between these papers resides in what type of data (e.g. chemical structures linked to assay activity, or information about what interactions between a compound and target are possible linked to a docking score) is combined and what are considered 'related targets'.

Historically, most novel approaches can actually be classified within existing parameters based on the type of description included for the target. PCM is no exception and should therefore be classified as a method to statistically derive a Structure-Activity-Relationship (SAR). The same goes for the chemogenomics work by Jacob *et al*. and the Multi-Assay-Based SAR by Ning *et al.*[3, 5] Other approaches, like Protein-Ligand fingerprint, should be classified as structural methods much like the work presented in **chapter 7**.

What to consider related targets (and hence which targets to dismiss when creating PCM models) is highly dependent on the eventual goal of the model. For example, in a receptor deorphanisation experiment, related targets should mean almost any protein of a superfamily, allowing a full sampling of the target space. Conversely, when viral resistance is modeled, the group of related targets should be much narrower, for instance limited to mutated versions of the main protein of interest. Likewise, the descriptor used in the PCM model should also be suitable for the target under consideration as we show in **chapter 3**.

When these limitations are acknowledged, PCM can be very flexible and applied to almost any group of targets as we have shown in this thesis. Furthermore the technique can be applied both pre-clinically and clinically as we will illustrate below.

8.2.3 Novel pre-clinical applications of PCM. In literature, PCM is mainly applied in a conventional way meaning the modeling of a series of ligands to a series of targets. However, there are new methods of application flowing directly from the potential combination of different target spaces. Some of these new application areas, like concurrent allosteric and orthosteric SAR modeling, are covered hypothetically in **chapter 2** and others are dealt with in the research chapters of this thesis.

For example, in **chapter 4** we use PCM to concurrently model orthologs and paralogs. This combination provides a way to incorporate historic data. The adenosine receptors provide a superb example as early work investigating this receptor family was actually performed on rat receptor orthologs.[6] In modern science it is particularly relevant not to reinvent the wheel. Using computational tools one has the freedom to incorporate high amounts of data to arrive at a preliminary understanding of the ligand – target space *before* any 'wet' experiment is performed. However, a lot of work has been published and forgotten, in particular if that work was done on other species orthologs. Yet, the information contained therein is still very relevant. Addition of historic compounds found to be active can help improve future work by making the model predictions more robust, but also by widening the applicability domain (in this case the chemical space wherein the model can make reliable predictions).

Likewise, we explored the limitations of target space. Theoretically infinite, we found that in practice there are boundaries. These limitations were explored in **chapters 5** and **6**. We found that these limitations can indeed be quantified based on target similarity. This allows a measure of reliability to be added to model predictions. Furthermore, preliminary experiments were started when we applied PCM to create a class A GPCR wide predictive model.

In essence such a model would have the potential to describe the full class A receptor ligand interaction space. However we found that the target definition is still cumbersome and limits the predictive capabilities of such a model. While we were able to train a model on almost all class A GPCRs (limited only by the availability of active compounds on certain subtypes), we found that the variation in binding pockets of class A GPCRs is still significant. Previous papers have been published that were successful in distinguishing between receptor subtypes based on just these binding pockets, more specifically the binding pockets located within the trans-membrane (TM) domains. We could repeat that distinction but this does not guarantee that these residues are the ones important for ligand binding. In other words, it is possible that the actual similarity measurement between the receptors based on these binding pockets was possibly not in line with the similarity as it is in bioactivity space.

Recently a paper was published by Cheng *et al.* where the authors compare a PCM based approach to an approach which relies on the consensus of 100 individual QSAR models (deemed multitarget-QSAR).[7] The authors find the PCM based approach to predict a large number of false positives. In my view the large number of false positives can be attributed to the usage of the binding pocket defined by Gloriam *et al.* In 2010 Wu *et al.* published the crystal structure of the CXCR4 receptor, showing that binding pockets among GPCRs can differ to a large degree.[8] In the case of the CXCR4 receptor the ligand binds much higher (closer to the extracellular side of the receptor) compared to the previously published structures. Hence other residues are involved in this interaction.

Therefore, a possible follow up for this work would be to distinguish GPCRs based on their ligand type (i.e. purine, peptide, etc) and to define a binding pocket for each of these subclasses. Work by Surgand *et al*. can be a good starting point to define binding pockets per receptor cluster.[9] Furthermore, an increase of available GPCR crystal structures can be instrumental herein. PCM can form the bridge from receptors that have an available crystal structure to receptors that have a similar ligand but lack the availability of a crystal structure.

Lastly, PCM can be of instrumental value in optimizing compounds that should have an effect on multiple isoforms of a protein target. An example application is given in **chapter 5**. The target is formed by a viral protein, HIV-1 Reverse Transcriptase, which has been shown to mutate quickly under selective pressure by treatment with anti-retroviral drugs. An ideal drug is active on the isoform that occurs most frequently, wild type, but also on mutants that arise during treatment. When information about frequently occurring mutations is available, the optimal drug candidate can already be selected in the preclinical phase.

8.2.4 Novel clinical applications of PCM. Finally, applications of PCM are not limited to preclinical research. In **chapter 6** we show how PCM can be used to select and optimize treatment regimens for patients infected with HIV. Hence PCM can be a tool to create personalized treatment protocols. In **chapter 6** the chemical space is formed by the drugs that are FDA approved and the target space is formed by the mutants that have been characterized in an assay based personalized medicine methodology that has been approved for clinical use.

In conclusion, PCM can make robust models by using existing chemical, sequence and biological data. Compared to models created from only chemical information or models created with sequence information only, PCM is shown to create more robust models and hence improved predictions.

8.2.5 Linking crystal structures by consensus structures. While we have shown that the combination of multiple disciplines in the form of bioactivity and chemistry can improve predictability of models, more efficient combination of information from a single discipline can also lead to new insights. In **chapter 7** we introduce ‘consensus structures’. The consensus structures reinterpret existing information present in crystal structures, leading to new insights and visualization of hidden information. We demonstrate this by applying the method to a target that has been studied since 1995, HIV Reverse Transcriptase. Yet we were still able to identify novel features of the allosteric non-nucleoside reverse transcriptase binding pocket that had not been previously described.

This method relies on the presence of multiple crystal structures from a single target. This is the case for a number of enzyme targets, such as HIV-RT, but not for the GPCR superfamily yet. GPCRs have notoriously been very difficult to crystalize, moreover the first crystal structures have been published as recently as 2007 and 2008.[10, 11] Interestingly, now, in 2012, at least a dozen different adenosine A_{2A} receptor crystal structures have been reported, also with different ligands. Combined with the exponential growth rate of the number of structures in the PDB, the case of the adenosine receptors demonstrates that it is likely that the consensus structures can prove valuable on other targets in the near future, such as GPCRs.

8.3 General conclusions from the thesis

From the research chapters I conclude that novel approaches to better mine existing data are not definitive solutions. However, these approaches can be seen as complementary to existing methods as they provide answers that cannot be obtained with traditional methods.

Another important conclusion to be drawn is the utmost importance of thorough validation of computational models and predictions. Using a (larger) dataset with more descriptors (increasing feature space) to train models leads to a larger risk for chance correlations at the same time. With methods sensitive to predict a single drug – target interaction, presence of wrongly annotated drug – target interactions in the training set can severely reduce the quality of model predictions (as shown in **chapters 4** and **6**). However, these prediction errors are in the eye of the beholder as the model merely predicts something that is not expected by the scientist but still accurately reflects what is in the training set.

It is possible to estimate the total number of wrongly annotated values and the average resulting error from several samples, subsequently a confidence value can be estimated for each prediction. However, the cause of a model prediction uncertainty (error) can be one of many reasons, and as such it is virtually impossible to compensate for these errors before training a model. For instance a prediction uncertainty can be caused by an error in annotation, which resides in a small fraction of the total data points. Conversely, an uncertainty in predictions can also be due to large error on a single compound, leading to a small average uncertainty for the full set. Thirdly, an uncertainty can also be caused by a structural bias for a certain scaffold (see **chapter 4**). Moreover, a prediction error or 'wrong' can simply be that measured activity values are lower in a certain assay readout compared to another, therefore values obtained using that particular assay could be classified as inactive rather than active as they are lower than a certain threshold.

8.4 Future Perspectives for PCM.

8.4.1 Complementary tool. As concluded above, these novel approaches, like PCM, will never be a replacement for existing, e.g. Quantitative Structure-Activity Relationship (QSAR), models. The technique can rather be seen as complementary to QSAR. QSAR can be of high interest when optimizing for a single target, in this case PCM does not necessarily improve upon QSAR. However, in the early phase hit discovery, PCM can help locate hits for this particular target by making use of its similarities to nearest neighbors. The great strength comes from the fact that PCM is an expansion of current techniques, using ready available data.

As shown in this thesis, PCM is capable of creating a single model that predicts the activity of a single molecule on a large group of targets. Compounds that have a described bioactivity on not one but many targets are known as poly-pharmacologic compounds. Further expanding on methods like PCM might make predictive poly-pharmacology models a reality. I will illustrate this using several preclinical scenarios.

8.4.2 Compounds hitting a number of targets. Small molecules (or peptides) that should hit a number of targets are expected to become more relevant as novel drugs.[12] Single target hitting drugs have been pursued for many years; however it has been shown that many of the successful drugs actually display a poly-pharmacologic profile. Examples include: kinase inhibitors and anti-psychotics, drugs used in the treatment of cancer and depression.[13, 14] However, compounds that are active on multiple mutants of a virology target (HIV, Hepatitis, Influenza) are also polypharmacologic. We have shown in this thesis (**chapter 3** and **5**) that PCM is able to capture such a ligand – target interaction space. Likewise, these concepts can be applied in the discovery of new antibiotics. An optimal candidate should be a compound that is active on multiple mutants of a bacterial target. This is particularly relevant as there is a need for new antibiotics.[15]

8.4.3 Compounds with different functional poly-pharmacological effects. While most of this thesis focuses on activity which was defined as affinity, the effect compounds exert on targets differs. For instance in the field of GPCRs several effects are distinguished including: agonism (a compound activates a receptor and thereby one or more signaling pathways), antagonism (a compound blocks receptor activation and thereby signaling pathways), inverse agonism (a compound inactivates a constitutively active receptor), and these effects can also be partially depending on the level of (inverse) activation achieved in comparison to the natural ligand.[16]

In the field of central nervous system drug development it has been put forward that optimal drugs should be selectively non-selective rather than selective.[17] This selective non-selectivity has been described as 'magic shotgun' rather than 'magic bullet'. Roth *et al.* describe that D2 partial agonists which are full agonists on other receptors might be an interesting lead.[17] The discovery of these magic shotgun drugs is an ideal scenario for PCM where the ability to classify, model and predict these responses would be of great value in preclinical candidate selection.

A similar idea is pursued in the search for partial agonists for HCA2, the nicotinic acid receptor.[18] Partial agonism at the HCA2 receptor might lead to favorable effects in the treatment of atherosclerosis while preventing the side effects caused by full agonists.[19] As this receptor is actually known to have two paralogs (the HCA1 and HCA3 nicotinic acid receptor) and no crystal structure is available, this is another area of interest where PCM could add value over existing approaches.[20]

8.4.4 Drug-Target residence time. Not only the effect compounds exert on a target can differ, the average time they remain bound to a target can also vary; this concept is known as drug-target residence time. It has previously been shown that differences in residence time can explain physiological effects not explained by affinity alone.[21, 22] Quantifying this residence time is known as a Structure-Residence-Time-Relationship (SRTR) and preliminary work has appeared in literature.[23] Like the addition of target information can be used to create better SAR models, the addition of target information might lead to better STRT models.

For example, the presence of certain functional groups on compounds might lead to better residence time on a certain targets but not another, through the addition of target information it might be elucidated what protein properties are responsible for this effect. This knowledge could then be used to optimize compounds active on another target that shares these properties. While this is speculation, there is no theoretical limitation.

8.4.5 Side effect screening of hit compounds. A final future prospect for PCM can be found in a completely different area of expertise. It is generally agreed upon that compounds should be selective for a certain target to reduce the chances of serious adverse effects arising from treatment with that compound. While it is currently impossible to accurately predict the complete pattern of interactions caused by a compound, it is possible to predict interaction with known anti-targets with a certain degree of reliability (e.g hERG, see below). PCM might be a tool in the early phase of drug discovery to expand the number of anti-targets for which interactions can be predicted. Examples include: GPCR-mediated side effects (via chemical similarity and binding pocket similarity) and Kinase inhibiting side effects (again via chemical similarity and binding pocket similarity).

Currently it is infeasible to predict the actual reliability of PCM when applied in side effect prediction. However the same principles apply that were used in **chapter 4**, therefore also here are no theoretical limitations to the application of PCM.

8.4.6 Novel developments in machine learning. PCM can also benefit greatly from the developments in the field of machine learning. Until recently, one was forced, when selecting a learning method, to choose between either high accuracy OR high interpretability (e.g. the choice between 'support vector machines' (SVM) or 'partial least squares' (PLS) in the case of PCM).[24, 25] Current developments allow for the combination of these two abilities in 'Random Forests', as we have shown in **chapter 3** and will further elaborate on below.[26]

8.4.7 Exponential growth of processing power. The exponential growth in computational power and available data that is currently taking place in computational chemistry is expected to continue over the coming decade. During my PhD project I have actually experienced an example of this growth first hand and I want to illustrate this with two examples.

The NNRTI data set used in chapters 3 and 5 was obtained already in November 2008. Early 2009 the first initial Leave-One-Sequence-Out (LOSO) experiments were performed on this data set. Using the Blosum protein descriptor (the largest and most training intensive, see chapter 3) training and validation of 14 LOSO models took approximately ***108*** hours using a standard workstation (Core 2 Duo E4600 2.4 GHz and 4GB memory). The models were built using the 'e1071' SVM package.[27]

In March 2012 the experiment was repeated with the same data set. At that point the models were trained using the 'forest' random forest package on an updated workstation (Core i7 860 2.8 GHz and 16 GB memory).[26] Total training time for the 14 LOSO models using the Blosum descriptor was ***2.5*** hours. The performance of the final models was in fact comparable and this represents a decrease in training time of *98 %*. Moreover, repeating this experiment for all fingerprints tested in **chapter 3** (13 times the dataset size) took a mere ***30*** hours. While this is caused by an increase in computational power and the advent of more efficient algorithms it illustrates how the progress in computational chemistry leads to more efficient data processing.

A similar example can be given for our structure-based approach. The consensus structures that are presented in **chapter 7** were originally trained in January 2007. At that point these density maps were calculated in approximately 15 minutes (0.25 hours) per protein structure binding pocket and 2 minutes (0.03 hours) per ligand structure. Hence, when calculating for 36 structures the total time was **10** hours. In March 2012 this experiment was repeated and applied to the structures available for the Adenosine A_{2A} receptor (then a total of 9 structures). The isolated binding pockets and co-crystalized ligands are comparable to those of the crystal structures in **chapter 7**. The total calculation time was now about 4 minutes for the 9 binding pockets combined (0.44 minutes per binding pocket) and less than 1 minute for the ligands combined. The total calculation time was thus 5 minutes (or **0.08** hours), representing a decrease of *97 %* in calculation time.

From these observations we can draw a number of conclusions other than just the experimental speed up. Firstly a dataset considered to be infeasible to model could be the subject of a standard procedure in a mere 3 years. Secondly, when optimizing a method and one is presented with a trade-off between speed and accuracy, it could very well pay off to optimize for accuracy rather than speed. Speed will catch up over the years, accuracy will not.

8.5 Future perspectives for structure-based methods

8.5.1 Millisecond molecular dynamics. It is in this light that we should consider the future perspectives for the structure-based methods. While the consensus structures provide a very valuable tool for the near future, their use might become obsolete by the advent of millisecond molecular dynamics (MD).[28] Molecular dynamics simulates the actual dynamic forces between individual atoms and is not novel. However these computational intensive calculations used to take days for simulations that only simulate nano- to microseconds of real time. The catalyst to make these simulations feasible on a millisecond scale is a specialized computer system named Anton (after Anthony van Leeuwenhoek).[28] Anton consists of 512 or more nodes (custom hardware subunits) designed to work together efficiently with custom software. Already this purpose-built approach has shown that it is capable of performing molecular dynamics calculations of GPCRs and has been used to retrospectively validate force fields currently in use.[29, 30]

The major advantage of MD over consensus structures is that MD is not bound by states of the protein which can be crystallized like consensus structures are. MD can also model transition states *between* states that can be crystalized. An obvious application area is modeling of extracellular loops (ELs) on GPCRs. These loops display an enormous degree of variation between the different GPCR crystal structures yet have also been shown to be important in ligand binding,[31, 32] and allosteric modulation.[33] Furthermore, MD allows the simulation of water molecules in the structure, the presence of which has previously already been shown to be very important.[34, 35]

The downside of MD is the fact that it is limited by the quality of force fields which are always an approximation whereas crystal structures are experimentally derived results. An extensive validation of these MD approaches is therefore key before their mainstream use and this validation has already begun to appear in literature.[30]

8.6 Drug discovery remains a challenging field

8.6.1 The drug discovery problem. When we compare drug discovery with other applied sciences like engineering or physics the most important difference is that in drug discovery we fail to completely understand the system we work on. An engineer designing a plane can accurately simulate the plane in flight *in silico* and the finally built prototype will behave near identical to those simulations. This omits a great deal of the experimental optimizations needed before the first test flight. While developments have progressed quickly in all fields mentioned in this thesis (chemistry, biology, bioactivity, computational approaches) we are not yet capable to fully understand our target organism 'homo sapiens'. Hence we cannot fully simulate a candidate drug in a human *in silico.*

It should however be noted that the first publication of a computational model simulating *a whole cell* (*Mycoplasma genitalium*) has appeared in literature July 20th 2012.[36] Still, one of the underlying causes preventing the simulation of 'homo sapiens' is the sheer dimensionality of the problem. Moreover, simply optimizing a compound to be a perfect binder on a single target is already a complex problem we do not fully understand.

Take the case of logP, a physicochemical property of compounds that cannot be perfectly predicted. Computational tools provide an estimate based on known observations and novel tools are still appearing in literature,[37-39] indicating that we do not fully understand our problem. Likewise, when presented with a crystal structure of a target and co-crystalized ligand, computational tools fail to present perfect predictions, again indicating a lack of complete understanding of the problem.[40, 41] While this is not novel information, it is relevant to be able to consider the drug discovery problem as outlined below.

8.6.2 Single solution for a complex problem. On the level of a drug being prescribed to a patient, the multi-dimensionality of the problem further explodes. The drug is the single solution to this very complex multi-dimensional problem. A very complex problem indeed as several of the parameters that need to be optimized have contradictory goals. A classic example case is the development of a compound with a target in the central nervous system. It might very well be that this compound requires several hydrogen bonds to have a high activity on a certain protein. However, from literature we known that an increase in hydrogen bond capacity is detrimental for blood-brain-barrier permeability.[42, 43]

Likewise the assumption that nanomolar affinity on a target leads to selectivity for that target is not always true. For instance the discovery of the hERG voltage gated potassium channel as the causative agent of cardiovascular sudden death has caused the withdrawal of several high affinity blockbuster drugs.[44, 45] Unwillingly, in the process of compound binding optimization, one might introduce hERG affinity by optimizing affinity to the actual target. Hence one is simultaneously creating a problem during the process of tackling another. The hERG channel itself is an example of the discovery of novel factors in the multi objective problem that change the drug discovery landscape continuously. Here I will classify these factors as the 'unknown problem'.

8.6.3 The unknown problem. Each drug is unique and each drug discovery track will encounter different hurdles. As we have outlined these hurdles might relate to the nature of the target or they might relate to the nature of the compound under development. Moreover, with the further unraveling of the target organism, possible *novel* problem areas are appearing in literature.[46] It might sound as an impossible task to create a drug, and the decrease in output of novel drugs supports this grim image.[47]

However, computational tools *can* be used to improve success rate, the key is to make use of the data available. Part of the solution to this problem is to be smart rather than to use the brute force method. Already universities have demonstrated incredible hitrates up to 70 % as shown by De Graaf *et al.* or high accuracy in crystal structures prediction Kufareva *et al.*, being universities their budget is limited and they cannot resort to brute force.[48, 49] Novel approaches are required in hit identification and the hit should preferably already meet several of the downstream requirements. [50]

8.6.4 Incorporating computational methods into existing research lines. Computational tools are cheap, quick and have become much more reliable. Can computational methods add value? I would argue that they can add value in any (drug) research project. But the question is how or when these methods add value. Naturally the nature of this added value depends on the specific research project, still there are some globally applicable ways in which to use computational tools.

- Helping to formulate a hypothesis that is partially based on previously available public data. Mining for this data is not necessarily limited to one's own field (for example problems one comes across in PK/PD may very well have been solved in the field of QSAR).
- Iterative data interpretation and comparison to known data might help in guiding a project while underway, and to prevent forgetting early lessons learned.

Both these two phases of a project consist of data interpretation. This is also what is known as e-Science. E-Science approaches can often obtain surprising results from unexpected sources. An example is work by Frijters *et al.*[51] Using only literature mining they identified novel associations between genes, drugs, pathways and diseases that have a high probability of being biologically valid.

In particular the fact that they did this automatically makes this a very useful tool. In this example the integration of computational approaches proved successful. Addition of computational methods into existing projects should not be cumbersome or difficult, it consists of merely shifting the starting point in a research project (**Figure 8.1**). The rationale here is that added data is added value as long as the data consists of information and not noise.

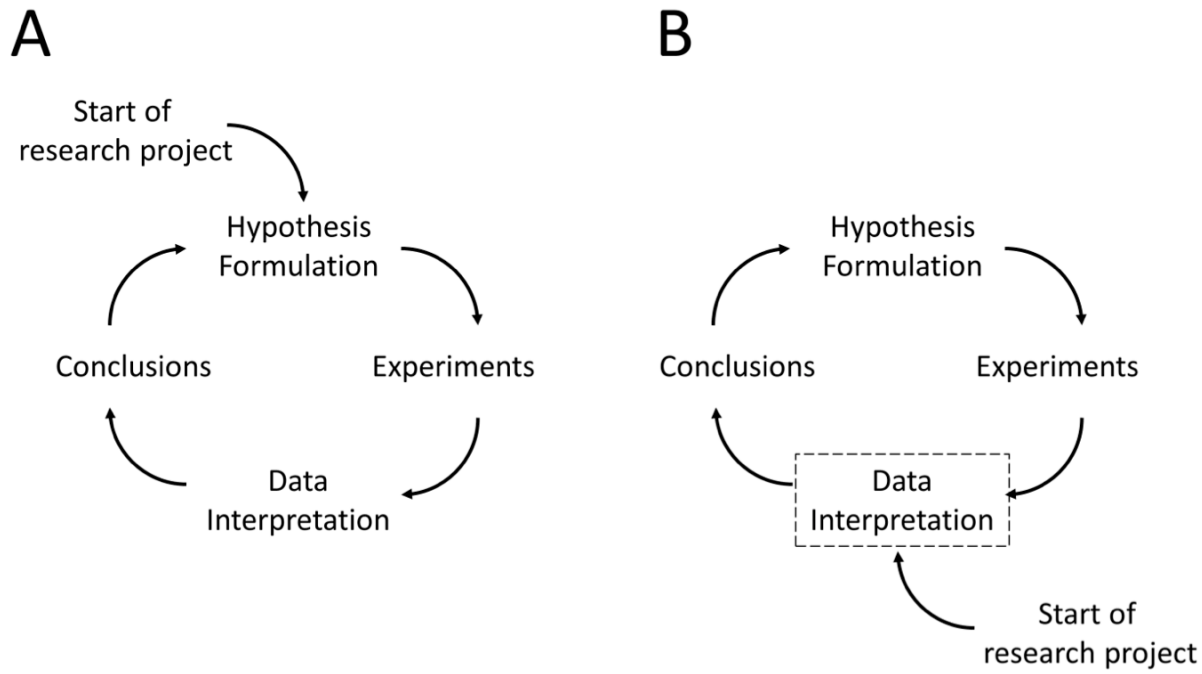


Figure 8.1: How computational methods can be integrated in existing research projects. (A) The classical way the scientific method guides research. (B) Addition of computational approaches. The fundamental set-up does not change, rather computational methods can be introduced early on in a research project (dashed box). By mining public data available these methods can help in constructing a solid hypothesis. Likewise, data gathered from experiments can be interpreted to arrive at a final conclusion.

8.6.5 Public data is not everything. Let us examine the public data we used for our model creation in **chapters 3** and **4** and which we propose to be used for hypothesis formulation in research. More specifically, let us compare the information used by companies to found their research on (proprietary data) with this data which is available in the public domain. While not much has been published, there seems to be a separation between public data and proprietary data (**Figure 8.2**). [52]

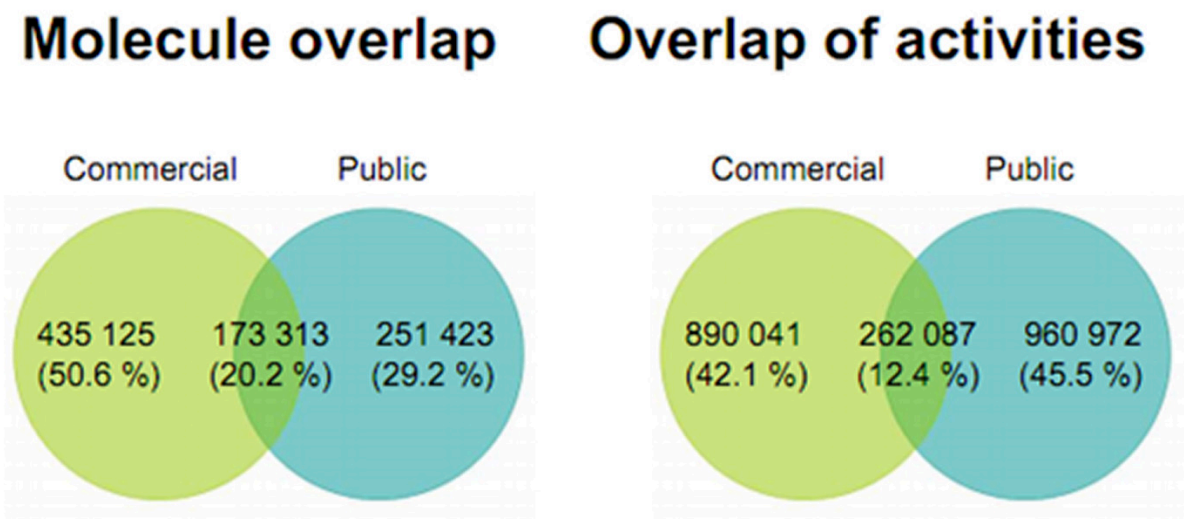


Figure 8.2: Overlap between public and proprietary databases. Numbers are for molecules associated with an activity. An activity is defined as a unique combination of Uniprot ID, small molecule, activity value, activity type and activity relation (Adapted from: Pekka Tiikkainen and Lutz Franke, Analysis of Commercial and Public Bioactivity Databases, 2011 [52]).

Figure 8.2 shows that, even with the quick growth of publicly available data, pharmaceutical companies still own a siginificant part of the structure-activity space that is unavailable in the public domain. Therefore it is elementary that pharmaceutical companies need to combine public and proprietary data and cannot merely rely on their in-house data.

8.7 Final conclusion

At the end of this thesis I should like to draw one final conclusion based on the research chapters covered in this thesis and the thesis itself. This conclusion is that universities and commercial companies should embark on collaborative research efforts. The academic and the commercial mindset are fundamentally different. None can be considered superior by any standard, but fact is that these mindsets are complementary to a large degree. Therefore it stands to reason that, like we show in this thesis, synergistic effects can result from combining these mindsets. With the advent of large academic-commercial cooperation platforms (so-called public-private partnerships, such as TIPharma, the Innovative Medicines Initiative of the EU, and the cooperation that was the source of this thesis) combining these mindsets is exactly what is happening in drug discovery…

8.8 References

1. M. Lapinsh, P. Prusis, et al.; *Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions.* Biochim. Biophys. Acta, Gen. Subj.; *2001*. **1525** (1-2): 180-190.
2. J. Wikberg, M. Lapinsh, and P. Prusis; *Proteochemometrics: A tool for modelling the molecular interaction space;* in *Chemogenomics in Drug Discovery - A Medicinal Chemistry Perspective*; H. Kubinyi and G. Müller; Editors. *2004*. p. 289 - 309.
3. L. Jacob, B. Hoffmann, et al.; *Virtual screening of GPCRs: An in silico chemogenomics approach.* BMC Bioinformatics; *2008*. **9** (1): 363-379.
4. N. Weill and D. Rognan; *Development and Validation of a Novel Protein−Ligand Fingerprint To Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands.* J. Chem. Inf. Model.; *2009*. **49** (4): 1049-1062.
5. X. Ning, H. Rangwala, and G. Karypis; *Multi-Assay-Based Structure-Activity Relationship Models: Improving Structure-Activity Relationship Models by Incorporating Activity Information from Related Targets.* J. Chem. Inf. Model.; *2009*. **49** (11): 2444-2456.
6. B.B. Fredholm, A.P. IJzerman, et al.; *International Union of Basic and Clinical Pharmacology. LXXXI. Nomenclature and Classification of Adenosine Receptors—An Update.* Pharmacol. Rev.; *2011*. **63** (1): 1-34.
7. F. Cheng, Y. Zhou, et al.; *Prediction of chemical-protein interactions: multitarget-QSAR versus computational chemogenomic methods.* Mol. BioSyst.; *2012*.
8. B. Wu, E.Y.T. Chien, et al.; *Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists.* Science; *2010*. **330** (6007): 1066-1071.

9. J.-S. Surgand, J. Rodrigo, et al.; *A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors.* Proteins: Struct., Funct., Bioinf.; *2006*. **62** (2): 509-538.
10. V.P. Jaakola, M.T. Griffith, et al.; *The 2.6 Angstrom Crystal Structure of a Human A2A Adenosine Receptor Bound to an Antagonist.* Science; *2008*. **322** (5905): 1211-1217.
11. S.G.F. Rasmussen, H.-J. Choi, et al.; *Crystal structure of the human [bgr]2 adrenergic G-protein-coupled receptor.* Nature; *2007*. **450** (7168): 383-387.
12. A.D. Boran and R. Iyengar; *Systems approaches to polypharmacology and drug discovery.* Curr. Opin. Drug Discovery Dev.; *2010*. **13** (3): 297-309.
13. A.L. Hopkins, J.S. Mason, and J.P. Overington; *Can we rationally design promiscuous drugs?* Curr. Opin. Struct. Biol.; *2006*. **16** (1): 127-136.
14. G.R. Zimmermann, J. Lehár, and C.T. Keith; *Multi-target therapeutics: when the whole is greater than the sum of the parts.* Drug Discov. Today; *2007*. **12** (1–2): 34-42.
15. World Health Organisation. *Antimicrobial resistance*. 2012 [cited 2012 March 16]; Available from: http://www.who.int/mediacentre/factsheets/fs194/en/.
16. K. Terry; *Principles: Receptor theory in pharmacology.* Trends Pharmacol. Sci.; *2004*. **25** (4): 186-192.
17. B.L. Roth, D.J. Sheffler, and W.K. Kroeze; *Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia.* Nat Rev Drug Discov; *2004*. **3** (4): 353-359.
18. W. Soudijn, I. van Wijngaarden, and A.P. IJzerman; *Nicotinic acid receptor subtypes and their ligands.* Medicinal Research Reviews; *2007*. **27** (3): 417-433.
19. T. van Herk, J. Brussee, et al.; *Pyrazole Derivatives as Partial Agonists for the Nicotinic Acid Receptor.* J. Med. Chem.; *2003*. **46** (18): 3945-3951.
20. S. Offermanns, S.L. Colletti, et al.; *International Union of Basic and Clinical Pharmacology. LXXXII: Nomenclature and Classification of Hydroxy-carboxylic Acid Receptors (GPR81, GPR109A, and GPR109B).* Pharmacol. Rev.; *2011*. **63** (2): 269-290.
21. R.A. Copeland, D.L. Pompliano, and T.D. Meek; *Drug-target residence time and its implications for lead optimization.* Nat. Rev. Drug Discovery; *2006*. **5** (9): 730-739.
22. D. Swinney; *The role of binding kinetics in therapeutically useful drug action.* Curr. Opin. Drug Discovery Dev.; *2009*. **12** (1): 31-39.
23. G. Tresadern, J.M. Bartolome, et al.; *Molecular properties affecting fast dissociation from the D2 receptor.* Bioorg. Med. Chem.; *2011*. **19** (7): 2231-2241.
24. C. Cortes and V. Vapnik; *Support-vector networks.* Machine Learning; *1995*. **20** (3): 273-297.

25. S. Wold, A. Ruhe, et al.; *The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses.* SIAM journal on scientific and statistical computing; *1984*. **5** (3): 735.

26. A. Liaw and M. Wiener; *Classification and Regression by randomForest.* R News; *2002*. **2** (3): 18-22.

27. E. Dimitriadou, K. Hornik, et al. *Misc Functions of the Department of Statistics (e1071)* TU Wien 2006 1.5-15

28. D.E. Shaw, R.O. Dror, et al.; *Millisecond-scale molecular dynamics simulations on Anton*; in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis2009*; ACM: Portland, Oregon. 1-11.

29. A.C. Kruse, J. Hu, et al.; *Structure and dynamics of the M3 muscarinic acetylcholine receptor.* Nature; *2012*. **482** (7386): 552-556.

30. K. Lindorff-Larsen, P. Maragakis, et al.; *Systematic Validation of Protein Force Fields against Experimental Data.* PLoS One; *2012*. **7** (2): e32131.

31. M.C. Peeters, G.J.P. Van Westen, et al.; *Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation.* Trends Pharmacol. Sci.; *2011*. **32** (1): 35-42.

32. M.C. Peeters, G.J.P. Van Westen, et al.; *GPCR structure and activation: an essential role for the first extracellular loop in activating the adenosine A2B receptor.* The FASEB Journal; *2011*. **25** (2): 632-643.

33. M.C. Peeters, L.E. Wisse, et al.; *The role of the second and third extracellular loops of the adenosine A1 receptor in activation and allosteric modulation.* Biochemical Pharmacology; *2012*. **84** (1): 76-87.

34. V. Katritch, V.-P. Jaakola, et al.; *Structure-Based Discovery of Novel Chemotypes for Adenosine A2A Receptor Antagonists.* J. Med. Chem.; *2010*. **53** (4): 1799-1809.

35. W. Liu, E. Chun, et al.; *Structural Basis for Allosteric Regulation of GPCRs by Sodium Ions.* Science; *2012*. **337** (6091): 232-236.

36. Jonathan R. Karr, Jayodita C. Sanghvi, et al.; *A Whole-Cell Computational Model Predicts Phenotype from Genotype.* Cell; *2012*. **150** (2): 389-401.

37. A.K. Ghose, V.N. Viswanadhan, and J.J. Wendoloski; *Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods.* J. Phys. Chem.; *1998*. **102** (21): 3762-3772.

38. C. Kramer, B. Beck, and T. Clark; *A Surface-Integral Model for Log POW.* J. Chem. Inf. Model.; *2010*. **50** (3): 429-436.

39. L. Xing and R.C. Glen; *Novel Methods for the Prediction of logP, pKa, and logD.* J. Chem. Inf. Comput. Sci.; *2002*. **42** (4): 796-805.

40. E. Kellenberger, J. Rodrigo, et al.; *Comparative evaluation of eight docking tools for docking and virtual screening accuracy.* Proteins: Struct., Funct., Bioinf.; *2004*. **57** (2): 225-242.

41. J.B. Cross, D.C. Thompson, et al.; *Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy.* J. Chem. Inf. Model.; *2009*. **49** (6): 1455-1474.

42. F. Ooms, P. Weber, et al.; *A simple model to predict blood–brain barrier permeation from 3D molecular fields.* Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease; *2002*. **1587** (2–3): 118-125.

43. P. Crivori, G. Cruciani, et al.; *Predicting Blood−Brain Barrier Permeation from Three-Dimensional Molecular Structure.* J. Med. Chem.; *2000*. **43** (11): 2204-2216.

44. M.C. Sanguinetti and M. Tristani-Firouzi; *hERG potassium channels and cardiac arrhythmia.* Nature; *2006*. **440** (7083): 463-469.

45. G.-N. Tseng; *IKr: The hERG Channel.* Journal of Molecular and Cellular Cardiology; *2001*. **33** (5): 835-849.

46. J. Dudley, E. Schadt, et al.; *Drug Discovery in a Multidimensional World: Systems, Patterns, and Networks.* Journal of Cardiovascular Translational Research; *2010*. **3** (5): 438-447.

47. F. Pammolli, L. Magazzini, and M. Riccaboni; *The productivity crisis in pharmaceutical R&D.* Nat Rev Drug Discov; *2011*. **10** (6): 428-438.

48. C. de Graaf, A.J. Kooistra, et al.; *Crystal Structure-Based Virtual Screening for Fragment-like Ligands of the Human Histamine H1 Receptor.* J. Med. Chem.; *2011*. **54** (23): 8195-8206.

49. I. Kufareva, M. Rueda, et al.; *Status of GPCR Modeling and Docking as Reflected by Community-wide GPCR Dock 2010 Assessment.* Structure (London, England : 1993); *2011*. **19** (8): 1108-1126.

50. A. Bender, D. Bojanic, et al.; *Which aspects of HTS are empirically correlated with downstream success?* Curr. Opin. Drug Discovery Dev.; *2008*. **11** (3): 327-337.

51. R. Frijters, M. van Vugt, et al.; *Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases.* PLoS Comput. Biol.; *2010*. **6** (9): e1000943.

52. P. Tiikkainen and L. Franke; *Analysis of Commercial and Public Bioactivity Databases.* J. Chem. Inf. Model.; *2011*. **52** (2): 319-326.