



Applications of proteochemometrics (PCM) : from species extrapolation to cell-line sensitivity modelling

Isidro Cortes Ciriano

► To cite this version:

Isidro Cortes Ciriano. Applications of proteochemometrics (PCM) : from species extrapolation to cell-line sensitivity modelling. Other [cs.OH]. Université Pierre et Marie Curie - Paris VI, 2015. English. NNT : 2015PA066176 . tel-01212483

HAL Id: tel-01212483

<https://theses.hal.science/tel-01212483>

Submitted on 7 Oct 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ PIERRE ET MARIE CURIE - PARIS 6
ECOLE DOCTORALE COMPLEXITÉ DU VIVANT (ED515)

Thèse de Doctorat
Spécialité Bioinformatique

Applications of Proteochemometrics (PCM): From Species Extrapolation to Cell-Line Sensitivity Modelling

Isidro CORTÉS-CIRIANO

Thèse dirigée par Dr. Thérèse E. MALLIAVIN
Unité de Bioinformatique Structurale (BIS)
25 rue du Docteur Roux, 75015 Paris

Soutenue le 16 Juin 2015
à l'Institut Pasteur, Paris

Jury:

Prof. Alessandra CARBONE	Président du Jury
Prof. Véronique STOVEN	Rapporteur
Prof. Didier ROGNAN	Rapporteur
Dr. Raphael GUEROIS	Examineur
Dr. Andreas BENDER	Examineur
Dr. Thérèse E. MALLIAVIN	Directeur de thèse

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License (<http://creativecommons.org/licenses/by-nc-sa/3.0/>), with the following exception: chapter .1 which has been adapted from *I. Cortes-Ciriano et al.*, 2015 (DOI: 10.1039/C4MD00216D), and is Copyright *The Royal Society of Chemistry* 2014.

Writing began on January 8th and ended on February 21th 2015. Dr. Yannick Spill is thanked for his help on preparing the L^AT_EX template of this thesis.

I

En su grave rincón, los jugadores
rigen las lentas piezas. El tablero
los demora hasta el alba en su severo
ámbito en que se odian dos colores.

Adentro irradian mágicos rigores
las formas: torre homérica, ligero
caballo, armada reina, rey postrero,
oblicuo alfil y peones agresores.

Cuando los jugadores se hayan ido,
cuando el tiempo los haya consumido,
ciertamente no habrá cesado el rito.

En el Oriente se encendió esta guerra
cuyo anfiteatro es hoy toda la Tierra.
Como el otro, este juego es infinito.

II

Tenue rey, sesgo alfil, encarnizada
reina, torre directa y peón ladino
sobre lo negro y blanco del camino
buscan y libran su batalla armada.

No saben que la mano señalada
del jugador gobierna su destino,
no saben que un rigor adamantino
sujeta su albedrío y su jornada.

También el jugador es prisionero
(la sentencia es de Omar) de otro tablero
de negras noches y de blancos días.

Dios mueve al jugador, y éste, la pieza.
¿Qué Dios detrás de Dios la trama empieza
de polvo y tiempo y sueño y agonía?

El Ajedrez, Jorge Luis Borges

Acknowledgements

First of all, I would like to acknowledge Thérèse Malliavin for her always friendly mentorship and support throughout these years, which has been essential to my education. Likewise, I truly appreciate the freedom she endowed me to explore new methodologies and research topics. The amazing facilities and resources, both human and technical, provided by Michael Nilges and the BIS unit have also been a key factor to my education and to creating a working environment which I truly think is difficult to beat. I am very thankful to Thérèse and Michael for affording me the opportunity to attend many international conferences to present my work. On both the scientific and the personal side, I am very thankful to the BIS unit as a whole, and individually to every person therein. Beyond the warm welcome I received at my arrival, the friendly atmosphere in the lab has made me always happy, if not eager, to go to work.

I am very grateful to Andreas Bender for his mentorship and for creating a fruitful network of international collaborations, both within and beyond his group. Beyond these and many other things, I particularly appreciate his constant efforts to force me to use scientific terms in a precise way, thus letting no room for ambiguity. The members of the Bender group are very thanked for their collaboration in many joint projects and for always giving me the best welcome of all when visiting Cambridge. I would also like to express my gratitude to Gerard JP van Westen for his essential help at the beginning of my PhD, and for (among others) our nice discussions on theoretical aspects of PCM. Bart Lenselink and Ad IJzerman from Leiden University, and John Overington from the EBI, are acknowledged for their collaboration in several joint projects.

Last but not least, the Pasteur-Paris International PhD Programme is thanked for funding. The members of my Thesis Advisory Committee are thanked for their constructive suggestions.

Abstract

Proteochemometrics (PCM) is a predictive bioactivity modelling method to simultaneously model the bioactivity of multiple ligands against multiple targets. Therefore, PCM permits to explore the selectivity and promiscuity of ligands on biomolecular systems of different complexity, such proteins or even cell-line models. In practice, each ligand-target interaction is encoded by the concatenation of ligand and target descriptors. These descriptors are then used to train a single machine learning model. This simultaneous inclusion of both chemical and target information enables the extra- and interpolation to predict the bioactivity of compounds on targets, which can be not present in the training set.

In this thesis, a methodological advance in the field is firstly introduced, namely how Bayesian inference (Gaussian Processes) can be successfully applied in the context of PCM for (i) the prediction of compounds bioactivity along with the error estimation of the prediction; (ii) the determination of the applicability domain of a PCM model; and (iii) the inclusion of experimental uncertainty of the bioactivity measurements. Additionally, the influence of noise in bioactivity models is benchmarked across a panel of 12 machine learning algorithms, showing that the noise in the input data has a marked and different influence on the predictive power of the considered algorithms. Subsequently, two R packages are presented. The first one, Chemically Aware Model Builder (*camb*), constitutes an open source platform for the generation of predictive bioactivity models. The functionalities of *camb* include : (i) normalized chemical structure representation, (ii) calculation of 905 one- and two-dimensional physicochemical descriptors, and of 14 fingerprints for small molecules, (iii) 8 types of amino acid descriptors, (iv) 13 whole protein sequence descriptors, and (iv) training, validation and visualization of predictive models. The second package, *conformal*, permits the calculation of confidence intervals for individual predictions in the case of regression, and P values for classification settings.

The usefulness of PCM to concomitantly optimize compounds selectivity and potency is subsequently illustrated in the context of two application scenarios, which are: (a) modelling isoform-selective cyclooxygenase inhibition; and (b) large-scale cancer cell-line drug sensitivity prediction, where the predictive signal of several cell-line profiling data is benchmarked (among others): basal gene expression, gene copy-number variation, exome sequencing, and protein abundance data. Overall, the application of PCM in these two case scenarios let us conclude that PCM is a suitable

technique to model the activity of ligands exhibiting uncorrelated bioactivity profiles across a panel of targets, which can range from protein binding sites (a), to cancer cell-lines (b).

Correspondence to Previous Publications

The contents of the majority of this thesis have been previously published in refereed journals. The correspondence between these publications and the chapters of this thesis is as follows:

- Preface:
 - Some paragraphs were adapted from **I. Cortes-Ciriano**[†], A. Koutsoukas, O. Abian, R. C. Glen, A. Velázquez-Campoy and A. Bender. **Experimental validation of in silico target predictions on synergistic protein targets.** *Med. Chem. Comm.*, 2013, 4, 278-288.
- Chapter .1: Polypharmacology Modelling Using Proteochemometrics (PCM)
 - **I. Cortes-Ciriano**[†], Q. U. Ain, V. Subramanian, E. B. Lenselink, O. Mendez-Lucio, A. P. IJzerman, G. Wohlfahrt, P. Prusis, T. Malliavin, G. J. P. van Westen, and A. Bender. **Polypharmacology Modelling Using Proteochemometrics: Recent Developments and Future Prospects.** *Med. Chem. Comm.*, 2015, 6, 24-50.
 - S. Paricharak, **I. Cortes-Ciriano**[†], A. P. IJzerman, T. Malliavin, and A. Bender. **Proteochemometric modeling coupled to in silico target prediction: an integrated approach for the simultaneous prediction of polypharmacology and binding affinity of small molecules..** *J. Cheminf.*, 2015, 7, 15.
- Chapter .2: Predictive Bioactivity Modelling
 - D. S. Murrell, **I. Cortes-Ciriano**[†], G. J. P. van Westen, I. P. Stott, T. Malliavin, A. Bender, and R. C. Glen. **Chemistry Aware Model Builder (camb): an R Package for Predictive Bioactivity Modeling.** *In revision at J. Cheminf.*, <http://github.com/cambDI/camb>.
 - **I. Cortes-Ciriano**^{*}. **Conformal: an R package to calculate prediction errors in the conformal prediction framework.** <http://github.com/isidroconformal>, and since 21/01/2014 available at CRAN: <http://cran.r-project.org/>, 2014.

- **I. Cortes-Ciriano**[★], A. Bender, and T. Malliavin. **Prediction of PARP Inhibition with Proteochemometric Modelling and Conformal Prediction.** *Accepted at Mol. Inform.*, 2015.
- G. J. P. van Westen, R. F. Swier, **I. Cortes-Ciriano**, J. K. Wegner, J. P. Overington, A. P. IJzerman, H. W. T. van Vlijmen and A. Bender. **Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets.** *J. Cheminf.*, 2013, 5, 42.
- Chapter .3: Proteochemometric Modelling in a Bayesian Framework
 - **I. Cortes-Ciriano**[†], G. J. P. van Westen, E. B. Lenselink, D. S. Murrell, A. Bender, and T. Malliavin. **Proteochemometric Modelling in a Bayesian Framework.** *J. Cheminf.*, 2014, 6, 35.
- Chapter .4: Benchmarking the Influence of Simulated Experimental Errors in QSAR
 - **I. Cortes-Ciriano**[★], A. Bender, and T. Malliavin. **Comparing the Influence of Simulated Experimental Errors On 12 Machine Learning Algorithms in Bioactivity Modelling Using 7 Diverse Data Sets** *In revision at JCIIM.*
- Chapter .5: Isoform Selectivity Prediction
 - **I. Cortes-Ciriano**[†], D. S. Murrell, G. J. P. van Westen, A. Bender, and T. Malliavin. **Prediction of the Potency of Mammalian Cyclooxygenase Inhibitors with Ensemble Proteochemometric Modeling.** *J. Cheminf.*, 2015, 7, 1.
- Chapter .6 Large-scale Cancer Cell-Line Sensitivity Prediction
 - **I. Cortes-Ciriano**[†], G. J. P. van Westen, G. Bouvier, M. Nilges, J. P. Overington, A. Bender, T. Malliavin. **Large-scale prediction of growth inhibition patterns on the NCI60 cancer cell-line panel.** *In revision at Bioinformatics.*

[†]: first author; [★]: corresponding author.

Preface

Drug¹ discovery is the process of finding, improving and making suitable for human intake exogenous or endogenous (*e.g.* hormones) pharmacologically active substances, such as chemical formulations, antibodies or siRNAs, which will, *via* their administration, improve patients' impaired health, be the etiologic agent in an in-born disorder (*e.g.* autoimmune disease -*horror autotoxicus*-) or an acquired disease (*e.g.* infectious illnesses or mutations).

Historically, drug discovery has been performed phenotypically, which means that the addition of a substance, *e.g.* a plant extract or a fungal exudate, to a biological system, *e.g.* a bacterial colony or an entire organism, produces a visible and desirable effect to the system considered. Generally speaking, and in human terms, this means health recovery from a pathological state. Since immemorial time, the humankind has profited from the vast range of traditional remedies based upon varied substances, mainly extracted from medicinal plants, minerals, animals or secondary metabolites therefrom. However, there is no evidence that the underlying biological events for the observed phenomena were understood at a molecular level. How the humankind came across the *magic effects* thereof has yet to be clarified. In more recent times, some of the most revolutionary drugs, were discovered under the 'blind' guidance of serendipity, such as penicillin (quoting Sir Alexander Fleming: "One sometimes finds what one is not looking for"). Today, the process of screening (small) molecules on biological systems using the resulting phenotypes as biological readouts to find new medicines is termed *phenotypic-based screening* [Feng et al. (2009)].

The need for understanding the underlying mechanisms whereby poisons and drugs could exert their effect, not only against infectious diseases but also against other congenital or acquired disorders, led to the establishment of the first Pharmacology department by Buchheim in 1847. Since then, several models have been proposed aiming to shed light on how drugs and other substrates interact with their biomolecular targets, generally enzymes or structural proteins. These models have grown steadily in complexity, enabling the scientific community to introduce more explanatory parameters. By way of example, the consolidated lock-key model proposed by Fischer (1894) was amplified by Koshland (1958) to account for protein flexibility. The advancements in cell biology (and related areas such as biochemistry

¹from Proto-Germanic *draugijaz*: dry

and genetics) during the XIXth and XXth centuries, sometimes permitted the deconvolution of the observed phenotypes to the underlying molecular events. In some cases, the observed effect was ascribed to the modulation of a single biomolecular target, generally proteins. This opened new avenues for drug discovery, as it was hypothesized that the modulation of single biomolecular targets implicated in a given disease would be sufficient to overcome the pathological state. Therefore, large collections of molecules were screened on purified protein extracts, what is known as *target-based screening* [Brown (2007)].

Since the 1980s, while target-based approaches (together with developments such as combinatorial chemistry and high-throughput screening) have gained considerable attention, recent studies show that phenotypic screening is still one of the main sources of first in class drugs (*i.e.* those with a new mode of action; 27 out of 45 drugs approved by the FDA between 1999 and 2008) [Swinney and Anthony (2011)]. Both target-based and phenotypic-based screening are based on different but complementary principles. Target-based drug discovery makes it necessary to first identify biochemical factor(s) mediating disease biology [Hart (2005)], and in a second step high-affinity modulators of those proteins are identified in the hope that this modulation will reverse the diseased state back to the healthy state. While this approach is very amenable to high-throughput techniques, in this case no network information (neither of intra- nor inter-cellular networks) is taken into account, and in addition no information concerning ADME/Tox properties or the modulation of additional targets can be gathered. Hence, while a precisely defined area of chemical space is sampled in target-based screens (with the hope that this particular activity will be important in later stages), the information that can be gathered about a *e.g.* compound (in particular from the efficacy standpoint) is rather limited.

On the other hand, phenotypic screening evaluates the response of a biological system to the application of compounds at more complex biological levels, from cell cultures to whole organisms [Feng et al. (2009)]. Here, an efficacy assessment of chemical substances closer to disease biology is obtained, hence facilitating the collection of bioactivity information in a physiologically more relevant context. In turn it should be said that phenotypic screens generally deliver more noisy data than biochemical screens also due to the underlying biology, and that often high-dimensional readouts are obtained which are more difficult to interpret. However although the phenotypic readout is often thought to be more predictive for efficacy in man [Clemons (2004)], this approach does not permit the identification of the mode of action (MoA) of the compounds exhibiting bioactivity. Still, when combined with subsequent target elucidation techniques, phenotypic screening followed by identification of the target can be seen as a commonly applied and feasible screening strategy in early stage drug discovery.

Recent experimental and systems biology studies, *e.g.* [Cortes-Ciriano et al. (2013); Koutsoukas et al. (2013); Lounkine et al. (2012); Poroikov et al. (2007); Wahlberg et al. (2012); Westen et al. (2013)], permitted the characterization of compound-target bioactivity profiles in a more comprehensive manner than before, by testing and predicting broad bioactivity profiles against large numbers of targets, now also known as the *polypharmacology* that many compounds are thought to exhibit to achieve the observed clinical effects [Jalencas and Mestres (2013); Peters (2013); Reddy and Zhang (2013); Westen et al. (2011)]. Consistent with the notion that cross-reactivity and side-effects stem from the modulation of additional targets, which is an important cause of attrition in drug development [Hopkins (2007)], Lounkine et al. (2012) conducted the largest prospective evaluation of *in silico* target prediction to date for 656 known drugs approved for human use on 73 protein targets (focusing on undesired off-targets). Applying the similarity ensemble approach (SEA) [Keiser et al. (2007)], adverse drug reactions were linked to predicted off-target interactions. Biochemical assays confirmed about half of the predictions at in many cases pharmacologically significant concentrations (1 nM to 30 mM). Due to the large number of classes and relatively low random hit rates this represents a significant enrichment of active compounds at the anti-targets considered in the study.

In other words, the conventional *lock-and-key* or *one-compound-one-target* paradigm, which states that a compound exerts its activity via a unique protein target relevant for this particular disease, has been extended to a more complex scenario where small-molecules are acknowledged to interact with more than one biomolecular target at therapeutically relevant concentrations (*one-compound-multiple-targets* paradigm). In line with this thinking it was recently established that drugs modulate on average not only one, but around six targets (though of course the distribution between targets contribution to the desired action and those that are detrimental or toxic are often not known) [Mestres et al. (2009)]. This evidence, along with the fact that phenotypic-based screening appears more effective in the discovery of first-in-class small molecule drugs than target-based approaches, has revealed that the activity of compounds on multiple targets needs to be considered (i) for the anticipation of unwanted side-effects, which might arise from compound activity on functionally similar or dissimilar targets, (ii) to explore compound promiscuity, and (iii) to optimize their selectivity.

Predictive bioactivity modelling techniques follow the premise that structurally similar compounds exert similar activities on their biomolecular targets more often than dissimilar ones (*molecular similarity principle*). These techniques, of which [Quantitative Structure-Activity Relationship \(QSAR\)](#) is probably the most widely known, deal with information from one space, *i.e.* the chemical space, to predict compound activity on individual targets.

The main notion underlying this thesis is that compound and target information complement each other, and, thus, its integration permits to better understand and predict, with respect to models based on either chemical or biological information, complex interactions between compounds and their biomolecular targets. To substantiate this assertion, **Proteochemometric Modelling (PCM)** is applied on several case studies, where the complexity of the targets ranges from proteins to cell-line models. **PCM** is a predictive bioactivity modelling technique that combines chemical and target descriptors in single machine learning models. Therefore, the activity of multiple compounds against multiple (related) targets can be simultaneously modelled. This enables the prediction of the activity of (novel) compounds on (novel) targets, and has been found to be particularly useful when exploring compound promiscuity and selectivity.

The first chapter provides an overview of the current state of the field, paying special attention to: (i) novel machine learning developments, (ii) PCM applications on therapeutically relevant protein families, and (iii) novel PCM applications, such as the integration of bioactivity data from different species or the prediction of cell-line sensitivity. In **the second chapter**, both the theoretical and practical aspects of generating and validating PCM models are discussed, and illustrated with a tutorial on how to generate a full PCM study with the R package *camb*.

The next two chapters are devoted to the study of the influence of experimental errors in bioactivity modelling. **Chapter 3** proposes the application of Bayesian inference in PCM using **Gaussian Process (GP)**. This Bayesian framework provides two main benefits, illustrated on three PCM data sets, with respect to other machine learning techniques used in the field, namely: (i) the inclusion of the experimental errors of the bioactivity values as input to the model, and (ii) the generation of individual predictions as Gaussian distributions, which serve to assess their confidence. To further explore the tolerance of common machine learning algorithms to noisy data, **Chapter 4** benchmarks the influence of simulated experimental errors in the predictive power of **QSAR** models. The main conclusion is that the tolerance to noise is significantly different across the studied algorithms, indicating that some algorithms (and kernels) are better suited to model noisy data. Throughout chapters 3-6, special consideration is given to (i) the estimation of confidence intervals for individual predictions, and (ii) the assessment of model performance in the light of the uncertainty of the data. The main outcome is that the maximum model performance achievable is highly dependent on: (i) the range of bioactivity values modelled, (ii) their distribution, and (iii) the level of uncertainty thereof. For instance, in most of the data sets explored here the maximum achievable R^2 values on the test set are not higher than 0.7.

Next, **PCM** is applied on two cases studies. The first one illustrates the versa-

tility of PCM for the integration of multispecies bioactivity data from mammalian Cyclooxygenases (COX). In this chapter, ensemble modelling is introduced in PCM, showing that higher predictive power can be attained when combining the predictions of a diverse set of models into a meta-model. The complexity of the target space is notably amplified in Chapter 6, which is consecrated to the large-scale prediction of cancer cell-line sensitivity. This illustrates the flexibility of the *target* concept in PCM. The main purpose of the chapter is to posit that 'omics' data of cell-line panels, such as DNA Copy Number Variation (CNV) or gene transcript levels, can be used as descriptors in PCM. The results of modelling the growth inhibition 50% bioassay endpoint (pGI₅₀) values of 17,142 compounds screened against 59 cancer cell-lines from the NCI60 cell-line panel, lead to several conclusions of practical relevance: (i) gene transcript levels provide the highest predictive signal, (ii) no statistically significant differences are found between inter- and extrapolation of compound bioactivities to novel cell-lines and tissues, (iii) extrapolating compound bioactivities to structurally novel compounds is challenging, although the extrapolation power is not constant across chemical clusters. Overall, this is the first large-scale study benchmarking the predictive signal of various cell-line profiling data, which conclusions can help in the prediction of primary tumour sensitivity from the genomic data of cancer patients.

An epilogue closes the thesis with a discussion on the strengths and limitations of PCM.

Bibliography

- Brown, D (2007). "Unfinished business: target-based drug discovery". In: *Drug Discov. Today* 12.23-24, pp. 1007–1012 (cit. on p. viii).
- Clemons, PA (2004). "Complex phenotypic assays in high-throughput screening". In: *Curr. Opin. Chem. Biol.* 8.3, pp. 334–338 (cit. on p. viii).
- Cortes-Ciriano, I, A Koutsoukas, O Abian, RC Glen, A Velazquez-Campoy, and A Bender (2013). "Experimental validation of in silico target predictions on synergistic protein targets". In: *Med. Chem. Commun.* 4 (1), pp. 278–288 (cit. on p. ix).
- Feng, Y, TJ Mitchison, A Bender, DW Young, and JA Tallarico (July 2009). "Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds". In: *Nat. Rev. Drug. Discov* 8.7, pp. 567–578 (cit. on pp. vii, viii).
- Fischer, E (1894). "Einfluss der Configuration auf die Wirkung der Enzyme". In: *Ber. Dtsch. Chem. Ges.* 27.3, pp. 2985–2993 (cit. on p. vii).
- Hart, CP (2005). "Finding the target after screening the phenotype". In: *Drug Discov. Today* 10.7, pp. 513–519 (cit. on p. viii).
- Hopkins, AL (2007). In: *Nat. Biotechnol.* 25, pp. 1110–1111 (cit. on p. ix).
- Jalencas, X and J Mestres (2013). "On the origins of drug polypharmacology". In: *Med. Chem. Comm.* 4.1, p. 80 (cit. on p. ix).
- Keiser, MJ, BL Roth, BN Armbruster, P Ernsberger, JJ Irwin, and BK Shoichet (2007). "Relating protein pharmacology by ligand chemistry". In: *Nat. Biotechnol.* 25.2, pp. 197–206 (cit. on p. ix).
- Koshland, DE (1958). "Application of a Theory of Enzyme Specificity to Protein Synthesis". In: *Proc. Natl. Acad. Sci. U. S. A.* 44.2, pp. 98–104 (cit. on p. vii).
- Koutsoukas, A, R Lowe, Y KalantarMotamedi, HY Mussa, W Klaffke, JBO Mitchell, RC Glen, and A Bender (2013). "In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt Window". In: *J. Chem. Inf. Model.* 53.8, pp. 1957–1966 (cit. on p. ix).
- Lounkine, E, MJ Keiser, S Whitebread, D Mikhailov, J Hamon, JL Jenkins, P Lavan, E Weber, AK Doak, S Côté, BK Shoichet, and L Urban (2012). "Large-scale prediction and testing of drug activity on side-effect targets". In: *Nature* 486.7403, pp. 361–367 (cit. on p. ix).
- Mestres, J, E Gregori-Puigjané, S Valverde, and RV Solé (2009). "The topology of drug-target interaction networks: implicit dependence on drug properties and target families". In: *Mol. BioSyst.* 5.9, pp. 1051–1057 (cit. on p. ix).
- Peters, JU (2013). "Polypharmacology - foe or friend?" In: *J. Med. Chem.* 56.22, pp. 8955–8971 (cit. on p. ix).

- Poroikov, V, D Filimonov, A Lagunin, T Gloriov, and A Zakharov (2007). "PASS: identification of probable targets and mechanisms of toxicity". In: *SAR QSAR. Env. Res.* 18.1-2, pp. 101–110 (cit. on p. ix).
- Reddy, AS and S Zhang (2013). "Polypharmacology: drug discovery for the future." In: *Expert. Rev. Clin. Pharmacol.* 6.1, pp. 41–47 (cit. on p. ix).
- Swinney, DC and J Anthony (2011). "How were new medicines discovered?" In: *Nat. Rev. Drug. Discov.* 10.7, pp. 507–519 (cit. on p. viii).
- Wahlberg, E, T Karlberg, E Kouznetsova, N Markova, A Macchiarulo, AG Thorsell, E Pol, A Frostell, T Ekblad, D Oncu, B Kull, GM Robertson, R Pellicciari, H Schuler, and J Weigelt (2012). "Family-wide chemical profiling and structural analysis of PARP and tankyrase inhibitors". In: *Nat. Biotechnol.* 30.3, pp. 283–288 (cit. on p. ix).
- Westen, GJP van, JK Wegner, AP IJzerman, HWT van Vlijmen, and A Bender (2011). "Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets". In: *Med. Chem. Commun.* 2, pp. 16–30 (cit. on p. ix).
- Westen, GJP van, A Hendriks, JK Wegner, AP IJzerman, HWT van Vlijmen, and A Bender (2013). "Significantly improved HIV inhibitor efficacy prediction employing proteochemometric models generated from antivirogram data". In: *PLoS Comput. Biol.* 9.2. Ed. by SL Kosakovsky Pond, e1002899 (cit. on p. ix).

Contents

Acknowledgements	i
Abstract	iii
Correspondence to Previous Publications	v
Preface	vii
Bibliography	xiii
List of Figures	xxi
List of Tables	xxxv
Introduction	1
.1 Polypharmacology Modelling Using Proteochemometrics (PCM)	3
.1.1 Introduction	3
.1.1.1 Available bioactivity data is growing: but can we make sense of it?	3
.1.1.2 Synergy between ligand and target space	4
.1.1.3 PCM as a practical approach to use chemogenomics data	9
.1.1.4 Practical relevance of PCM	11
.1.2 Machine Learning in PCM	12
.1.2.1 Support Vector Machines (SVM)	14
.1.2.2 Random Forests (RF)	15
.1.2.3 Gaussian Processes (GP)	16
.1.2.4 Collaborative Filtering (CF)	17
.1.3 PCM Applied to Protein Target Families	19
.1.3.1 G protein-coupled receptors	19
.1.3.2 Kinases	21
.1.3.3 Histone modification and DNA methylation	22
.1.3.4 Viral mutants	23
.1.4 Novel Techniques and Applications in PCM	25
.1.4.1 Novel target similarity measure	25

.1.4.2	Including 3D information of protein targets in PCM	25
.1.4.3	PCM in predicting ligand binding free energy	27
.1.4.4	PCM as an approach to extrapolate bioactivity data between species	28
.1.4.5	PCM applied to pharmacogenomics and toxicogenomics data .	30
.1.4.6	Other potential PCM applications	32
.1.5	PCM Limitations	32
.1.6	Conclusions	34
Bibliography		35
 Predictive Bioactivity Modelling		51
 .2 Predictive Bioactivity Modelling		53
.2.1	Compound standardization	54
.2.2	Descriptors	55
.2.2.1	Target descriptors	55
.2.2.2	Ligand descriptors	59
.2.2.3	Cross-term descriptors	60
.2.3	Statistical Preprocessing	62
.2.4	Generation of PCM Models	62
.2.5	Commonly used Algorithms	63
.2.6	Validation of PCM Models	69
.2.6.1	Statistical metrics	72
.2.7	Assessment of Maximum and Minimum Achievable Model Performance	75
.2.8	Conformal Prediction	75
.2.8.1	Regression	76
.2.8.2	Classification	77
 Bibliography		79
 Proteochemometric Modelling in a Bayesian Framework		87
 .3 Proteochemometric Modelling in a Bayesian Framework		89
.3.1	Introduction	89
.3.2	Materials and Methods	90
.3.2.1	Data sets	90
.3.2.2	Descriptors	91
.3.2.3	Modelling with Bayesian inference	93
.3.2.4	Computational details	95

.3.2.5	Assessment of maximum model performance	97
.3.2.6	Interpretation of ligand substructures	97
.3.3	Results	98
.3.3.1	Model validation	98
.3.3.2	Predicted confidence intervals follow the cumulative density function of the Gaussian distribution	104
.3.3.3	Analysis of GP performance <i>per target</i>	106
.3.3.4	Model interpretation of ligand descriptors	110
.3.4	Discussion	117
.3.5	Conclusion	119
Bibliography		121
Benchmarking the Influence of Simulated Experimental Errors in QSAR		127
.4	Benchmarking the Influence of Simulated Experimental Errors in QSAR	129
.4.1	Introduction	129
.4.2	Materials and Methods	133
.4.2.1	Data sets	133
.4.2.2	Data sets	133
.4.2.3	Molecular Representation	133
.4.2.4	Molecular Representation	134
.4.2.5	Compound Descriptors	134
.4.2.6	Model generation	136
.4.2.7	Machine Learning Implementation	136
.4.2.8	Simulation of Noisy Bioactivities	137
.4.2.9	Experimental Design	138
.4.3	Results	141
.4.4	Discussion	146
Bibliography		151
Prediction of the Potency of Mammalian Cyclooxygenase Inhibitors with En- semble Proteochemometric Modelling		157
.5	Isoform Selectivity Prediction: COX	159
.5.1	Introduction	159
.5.2	Materials and Methods	161
.5.2.1	Data set	161
.5.2.2	Descriptors	161

.5.2.3	Machine learning implementation	162
.5.2.4	Model generation	163
.5.2.5	Model validation	163
.5.2.6	Assessment of maximum model performance	164
.5.2.7	Ensemble modelling	164
.5.2.8	Estimation of the error of individual predictions	165
.5.2.9	Interpretation of compound substructures	167
.5.3	Results	169
.5.3.1	Analysis of the chemical and the target space	169
.5.3.2	PCM validation	172
.5.3.3	PCM models are in agreement with the maximum achievable performance	177
.5.3.4	PCM outperforms both Family QSAR and Family QSAM on this data set	178
.5.3.5	PCM outperforms individual QSAR models	179
.5.3.6	Model ensembles exhibit higher performance than single PCM models	179
.5.3.7	The ensemble standard deviation enables the definition of informative confidence intervals	180
.5.3.8	Ensemble modelling enables the prediction of uncorrelated human COX inhibitor bioactivity profiles	182
.5.3.9	Model performance per target is related to compound diversity	184
.5.3.10	Interpretation of compound substructures	187
.5.4	Discussion	192
.5.5	Conclusion	194
Bibliography		197
Large-scale Cancer Cell-Line Sensitivity Prediction		205
.6	Large-scale prediction of growth inhibition patterns on the NCI60 cancer cell-line panel	207
.6.1	Introduction	207
.6.2	Materials and Methods	208
.6.2.1	Data sets	208
.6.2.2	Compound descriptors	210
.6.2.3	Compound clustering	210
.6.2.4	Model generation	210
.6.2.5	Model validation	212
.6.2.6	Conformal prediction	214
.6.2.7	Pathway-drug associations	214

.6.2.8 Comparison to previous methods	215
.6.3 Results	217
.6.3.1 Summary of the cell-line profiling data set views	217
.6.4 Discussion	237
Bibliography	241
Epilogue	247
.7 Epilogue	249
Bibliography	253
Back Matter	255
Acronyms	257
Author Index	263
Subject Index	283

List of Figures

- .1.1 **Ligand-target interaction space.** The interaction between ligands (chemical compounds) and targets (biological macromolecules) can be envisioned as a matrix, where rows are indexed by target ids and columns by compound ids. Each matrix cell contains the binding affinity of a given compound on a given target, indicated by the following colors: blue means low affinity and yellow means high affinity. Traditional bioinformatics techniques have dealt with the similarity between targets, normally based upon sequence similarity. On the other hand, ligand based (QSAR) models have studied series of compounds acting on a given target. By contrast to both of them, PCM relates the chemical-target interaction space by describing targets and compounds with numerical descriptors permitting to predict activities of a given compound on a given target. 10
- .1.2 **Illustrative example of GP theory in a two-dimensional problem.** (A) The prior probability distribution embraces all possible functions which can potentially model the data set. A subset of six prototypical functions is depicted. Normally, the mean of the distribution is set to zero (black dashed line). (B) The inclusion of bioactivity information (red dots) accompanied by its experimental uncertainty (blue error bars) updates the prior distribution into the posterior probability distribution. In the posterior probability distribution, only those functions in agreement with the experimental data are kept. The uncertainty (pink area) notably increases in those areas with little experimental information available. The mean of the posterior distribution (black dashed line) is considered the best fit to the data. A prototypical function from the posterior is shown in blue. For a new compound-target combination, the bioactivity is predicted as a Gaussian distribution, in which the mean is the best prediction and its variance the uncertainty. A radial-kernelled GP with $\sigma = 1$ was employed to generate the figure. The python *infp*y package helped to produce the plots (John Reid, Version 0.4.13). 18

.2.1	Illustration of the descriptors used as input features in PCM. Compound-target pairs are encoded by the horizontal stacking of compound and target descriptors. The resulting matrix is used as input data to train a single machine learning model.	55
.2.2	Illustration of the computation of circular Morgan fingerprints. A. In the calculation of circular Morgan fingerprints, all substructures in a molecule, with a maximal user-defined bond radius (2 in the example), are assigned an unambiguous integer identifier. In the figure, 1 and 2 correspond to two arbitrary atoms. The atoms, namely Cl and N, are taken as root atoms to illustrate how the adjacent atom layers are considered in the definition of the substructures present in the molecule. B. The position in the array where a substructure will be mapped in the fingerprint is given by the modulo obtained by dividing the fingerprint integer identifier by the fingerprint size. The substructure IDs shown in the Figure are arbitrary.	61
.2.3	Toy example showing the influence of the range of the response variable (e.g. bioactivities) on R^2 values. A. R^2 and RMSE values of 0.91 and 0.90, respectively, are obtained when the response values range from 0 to 10 (arbitrary units). B. By contrast, the R^2 drops to 0.08 when the response value ranges from 0 to 1. Note that in both cases the RMSE values are the same, namely 0.90. To simulate y , random noise with mean 0 and standard deviation equal to 1 was added to x . The noise added was the exactly the same in both cases, namely (A) and (B).	74
.2.4	Calculation of conformal prediction errors (P values) in a binary classification example considering a confidence level of 0.80.	78
.3.1	Distribution of the maximum theoretical values of $RMSE_{test}$ (A, C and E) and $R^2_{0\ test}$ (B, D and F) for the adenosine receptors (A, B), G Protein-Coupled Receptor (GPCR)s (C, D) and dengue virus NS3 proteases data sets (E, F). These curves permit to estimate the reliability of $R^2_{0\ test}$ and $RMSE_{test}$ obtained for the GP models.	100
.3.2	Comparison between the performance of GP and Support Vector Machines (SVM) with either the radial or the Normalized Polynomial Kernel (NP) kernel. Ten models were calculated for each data set and for each combination of modelling technique and kernel, thus resulting in a total of 60 models. The performance of GP and SVM was assessed by kernel for the three data sets. Given that the distributions of $RMSE_{test}$ and $R^2_{0\ test}$ values were normally distributed, a two-tailed t -test of independent samples was applied to statistically evaluate their differences. These analyses let us conclude that SVM and GP perform <i>on par</i> for the modelling of the three data sets considered in this study.	103

- .3.3 **Noise influence in model performance.** $\text{RMSE}_{\text{test}}$ (red) and R_0^2 ext (black) values obtained when increasing the noise level (noise variance added to the diagonal of the covariance matrix) were calculated for: adenosine receptors (left figure), GPCRs (medium figure) and dengue virus NS3 proteases (right figure). Upper plots correspond to GP models calculated with the radial kernel while the bottom plots refer to GP models with the Normalized Polynomial (NP) kernel. In all cases, the radial kernel appears more sensitive to noise, while the NP kernel performs equally well when noise is added to the data. These data suggest that the NP kernel is more appropriate for the modelling of noisy PCM data sets. 104
- .3.4 **Analysis of the confidence intervals predicted on (left) the adenosine receptors and (right) aminergic GPCRs test sets.** The percentage of annotated values lying within the intervals of confidence of 68%, 80%, 95%, and 99% (ordinate axis) are depicted versus the size of the intervals. The blue line defines the theoretical proportionality between the size of confidence intervals and the number of points within the intervals, in the frame of the Gaussian cumulative function. The radial, Pearson VII Function-Based Universal Kernel (PUK), and Normalized Polynomial (NP) kernels are in close conformity with the cumulative Gaussian distribution in both data sets, while the Laplacian and Bessel exhibit a diverse behavior depending on the data set. Therefore, GP provide prediction errors in agreement with the Cumulative Gaussian distribution which can be reliably used to define intervals of confidence for the predictions. 105
- .3.5 **GP determine models applicability domain.** The differences between the true and predicted bioactivities (y axis) and the errors on predictions estimated by the GP model (x axis) are compared for the adenosine receptor data set with radial (A) and NP (B) kernel, and for the GPCRs data set with radial (C) and NP (D) kernels. The distribution of the differences between true and predicted bioactivities increases with the GP error on the prediction. This validates the GP error is a measurement of the Applicability Domain (AD) of the model. 107
- .3.6 **Model performance per target on the test set for the adenosine receptors data set.** The upper panel corresponds to R_0^2 test, while the lower panel to $\text{RMSE}_{\text{test}}$. These values were averaged for ten models trained on each subset corresponding to a given target. The best modeled target is the rat adenosine A_{2b} receptor (AA2BR RAT), while the worst is the rat A_3 receptor (AA3R RAT). In all cases, the mean $\text{RMSE}_{\text{test}}$ values are below 0.75 pK_i units, indicating that GP modelling can predict compound bioactivity on subsets corresponding to a given target. 111

- .3.7 **Distribution of pairwise compound Tanimoto similarity calculated on the target subsets extracted from the adenosine receptors data set.** The overall mean pairwise similarity is around 0.8. 112
- .3.8 **Evaluation of model performance per target on the GPCRs dataset.** RMSE_{test} values, averaged on ten models trained on different resamples of the dataset, are represented by bars, colored according to the number of datapoints per target. The standard deviations on RMSE_{test} are shown as error bars. Dark grey bars correspond to targets with more than two hundred annotated compounds. 113
- .3.9 **Evaluation of model performance per target for GPCRs dataset on the test set.** R_{0 test}² values, averaged on ten models trained on different resamples of the dataset, are represented by bars, colored according to the number of datapoints per target. Dark grey bars correspond to targets with more than two hundred annotated compounds. Both negative and infinite R_{0 test}² values were set to zero. 114
- .3.10 **Heatmap representing the contribution of each chemical substructure to compounds bioactivity on each adenosine receptor.** Columns are indexed by targets and rows by compound substructures. Depicted are some examples of compounds containing features beneficial (green) or deleterious (red) for bioactivity. Although a few substructures are predicted to have a beneficial or deleterious influence on the pK_i, there are others for which the effect depends on the target considered or on the rest of substructures present in a given compound. Therefore, over 90% of the substructures (black) are not implicated in compound bioactivity or their contribution depends on the other substructures present in a given compound. 115
- .3.11 **Descriptor importance for the dengue virus NS3 proteases data set.** Descriptor importance is calculated in the frame of Bayesian Automatic Relevance Determination (ARD) as the inverse of the value of the length scale of each descriptor. The descriptors of the first and second residues of the tetra-peptides (positions P1' and P2') are the most relevant for the model. This is in agreement with the higher influence of these two substrate positions for the cleavage rates of the proteases. 116

- .4.1 Illustration of two-way interactions between two-level factors, namely: Algorithm: RF and GP; and Noise: 0 and 100. There is interaction between two factors (**A**) when the difference of the mean RMSE values (response variable) across the levels of a factor (*e.g.* Algorithm) does not change across the levels of a second factor (*e.g.* Noise). In the example, the difference in performance between GP and RF is the same across all levels of factor Noise. This is illustrated by the presence of parallel lines. By contrast, the presence of non parallel lines (**B**) indicates that the performance of GP and RF changes depending on the noise level. Thus, GP outperforms RF at noise level 0%, whereas RF outperforms GP at noise level 100%. 140
- .4.2 **A,B.** Interaction plots. Median $RMSE_{test}$ values across all data sets for two-way combinations of factors. The data set-specific intercept was subtracted from the $RMSE_{test}$ values in order to make the results comparable across the seven data sets. 145
- .4.3 **A. $RMSE_{test}$ values across all data sets and noise levels for the 12 algorithms. B. $RMSE_{test}$ values across all data sets and models for the 11 noise levels studied.** The data set-specific intercept was subtracted from the $RMSE_{test}$ values in order to make the results comparable across the seven data sets. 147
- .4.4 $RMSE_{test}$ values for GBM models trained with increasingly higher bagfraction values across all Noise – Algorithm – dataset combinations. For low noise levels (up to 30%) the performance of all models is comparable irrespective of the bagfraction value. However, from noise level 30% upwards, the mean $RMSE_{test}$ difference between models trained with bag fraction values of 1 and 0.1-0.2 increases with the noise level. Overall, these data suggest that the noise sensitivity of GBM highly depends on the bagfraction value. 148
- .5.1 **Ensemble modelling with model stacking.** A. A set of models are trained with diverse machine learning algorithms (*Model₁ .. Model_n* in the Figure). The predictions of these models on each data-point in the training set calculated during cross validation, are used as descriptors to create a new training matrix, which rows are indexed by the data-points in the training set and columns by the models in the library. A machine learning model is trained on this matrix. The resulting model is the model ensemble. B. The model ensemble is then applied on the test set. 166

- .5.2 **Interpretation of compound substructures.** A. Predictive method. The average influence on bioactivity of a given substructure is calculated as the difference between the distributions corresponding to: (i) the predicted bioactivity for all compounds containing that substructure, and (ii) the predicted bioactivity using PCM for these compounds, from which that substructure was virtually removed by setting its count to zero. B. Student's method. In this case, the average substructure influence on bioactivity is evaluated as the difference between the pIC_{50} distributions for those compounds presenting and not presenting a given substructure. 168
- .5.3 **Principal Component Analysis (PCA) analysis of the target space.** A. Schematic overview of the COX binding pockets. B. PCA analysis was applied on the binding site descriptors used to train the models. The first two principal components explained more than 80% of the variance, thus indicating that there are mainly two sources of variability in the descriptor space, namely the isoenzyme type. This fact can be seen as COX-1 (triangles) and COX-2 (squares) define two distant clusters. Overall, the binding sites of orthologue cyclooxygenases are more similar than those of paralog sequences. These results also indicate that the amino acid descriptors account for structural differences between COX-1 and COX-2, which can be learnt by the models. Thus, it is expected that merging orthologues and paralogues will lead to more predictive models. 170
- .5.4 **Statistics of the repeated bioactivities for each compound-target combination.** A. The abscissa represents the mean value for the bioactivities repeated for each compound-target combination with more than one annotated bioactivity. The ordinate represents their standard deviations. Repeated bioactivities are equally distributed for low, moderate and high affinity COX inhibitors. B. Histogram of the standard deviation of the repeated bioactivities. The distribution is strongly skewed towards 0, thus indicating that the differences between repeated bioactivities are generally negligible. 171
- .5.5 **PCA of the compound descriptors used to train the PCM models.** The PCA was performed on the pairwise Pearson rank correlation matrix calculated with the compound descriptors used to train the models. The two first principal components (PC) explain 58.03% of the variance. COX-1 and COX-2 are represented with squares and triangles respectively. Overall, the overlap between the datapoints indicate that the compounds annotated on different targets cover the same regions of the chemical space. 173

- .5.6 **COX inhibitors selectivity on human COX-1 and COX-2.** A. Scatterplot corresponding to the comparison of bioactivities against human COX-1 and COX-2 for 1,288 compounds. A large proportion of the compounds present a COX-2/COX-1 selectivity ratio between 2 and 4 pIC₅₀ units. Therefore, the present data set includes COX inhibitors with highly divergent bioactivity profiles for COX-1 and COX-2 ($R_0^2 = -0.420$). B. Scatterplot of the observed against the predicted pIC₅₀ values for the compounds described in A. Blue squares correspond to the activity on COX-1, whereas orange squares correspond to the activity on COX-2. The PCM models explain more than 59% of the variance ($R_0^2 = 0.593$), thus highlighting the ability of the PCM models to predict the potency of compounds displaying uncorrelated bioactivity profiles on human cyclooxygenases. 174
- .5.7 **Model performance on the test set.** RMSE_{test} (A) and R_0^2 test (B) values for the following models: (group A) single PCM, (group B) Family QSAR and Family Quantitative Sequence-Activity Modelling (QSAM), (group C) individual QSAR, (group D) model ensembles comprising those single PCM models exhibiting the highest predictive power, and (group E) model ensembles comprising the whole model library. Bars are colored according to the groups defined in Table .5.2. Confidence intervals correspond to the mean value +/- one standard deviation calculated with bootstrapping [Efron and Tibshirani (1993)]. 175
- .5.8 **Y-scrambling.** Scatterplots corresponding to the percentage of bioactivities randomized, against (A) R_0^2 test and (B) RMSE_{test} values. The intercept in A becomes negative when 25-50% of the bioactivity variable is randomized. This finding indicates that PCM performance is not the consequence of spurious correlations in the descriptor space. 177
- .5.9 **Distribution of theoretical R_0^2 test (A) and RMSE_{test} (B) values.** The mean of the R_0^2 test distribution, 0.68, highlights that the uncertainty in public bioactivity data does not permit models with R_0^2 test values close to 1. Similar results were obtained for q_{test}^2 . The minimum RMSE_{test} value that a model can achieve without exhibiting overfitting is close to the experimental uncertainty. 178
- .5.10 **Pairwise Pearson correlation for the cross-validation predictions across the model library.** The predictive power across the model library is not uniformly distributed, as the predicted values for a large fraction of model pairs are uncorrelated (yellow areas). Therefore, the combination of these models in a model ensemble is expected to lead to higher predictive power than individual models ("wisdom of crowds"). 181

- .5.11 Confidence intervals calculated from the ensemble standard deviation of the models present in the model ensembles.** The percentage of data-points which predicted bioactivities lie within confidence intervals calculated with increasingly larger β values (Equation .5.1), is shown for: (i) the cross validated predictions calculated during model training (*Training* in the Figure), and (ii) for the predictions on the test set (*Test* in the Figure) calculated with the most predictive model ensemble, namely "Stacking SVM Radial Ensemble". The percentage of true values lying within the confidence interval derived for a given β value increases with the number of data-points available during model training. Overall, the confidence intervals derived from the ensemble standard deviation provide an estimation of the reliability of individual predictions, as in practice, this plot can be used to determine the β value corresponding to a given confidence level. 183
- .5.12 Jaccard pairwise similarity distributions for the compounds annotated on each target.** Compounds annotated on the human cyclooxygenases (annotated with a star in the plots) display compound similarity distributions with mean values skewed towards 1. By contrast, compounds annotated on targets with less than 30 annotated bioactivities display multimodal similarity distributions. A correlation between model performance and both the number of data-points and chemical diversity was established (see main text). Distributions were calculated with the same descriptors than the ones used to train the PCM models. 185
- .5.13 Target-averaged model performance.** The number of data-points is displayed through the size of the squares. A correlation can be established between the number of data-points and model performance, quantified by the standard deviation of the $\text{RMSE}_{\text{test}}$ values. Targets annotated with less than 30 compounds or with chemical structures displaying high structural diversity (*Oryctolagus cuniculus* COX-1, *Rattus norvegicus* COX-1, *Bos taurus* COX-1, and *Bos taurus*) are produced with high mean $\text{RMSE}_{\text{test}}$ values. These observations indicate that PCM models are not always able to extrapolate in the chemical or the target space if a given target or compound family is not sufficiently represented in the data set. 186

- .5.14 **Influence of compound substructures on potency and selectivity on human COX-1 and COX-2.** Rows in the heatmap are indexed by the isoenzyme type whereas columns correspond to compound substructures. Substructures are depicted in red within arbitrary molecules presenting it. The color represents the average influence (pIC₅₀ units) of each substructure on bioactivity. Red corresponds to an average increase in bioactivity, whereas blue indicates the a deleterious effect. Well-known chemical moieties, *e.g.* pyrrole rings (c), were singled out as selectivity determinants. For instance, substructure **d** is present in sulfonamides such as diflumidone, and substructure B in selective 1,2-diarylpyrroles COX-2 inhibitors. 188
- .5.15 **Volcano plots corresponding to the results of the Student's method applied on human COX-1 (A) and COX-2 (B).** The size of the points is proportional to the number of molecules in the data set containing a given substructure. Significant *P* values are shown in red (two-tailed *t*-test, $\alpha = 0.05$). 189
- .5.16 **Compound substructures predicted to increase the bioactivity on human COX-2.** The 20 substructures predicted to have the highest influence on bioactivity on human COX-2 (P35354) are plotted. Known chemical moieties such as pyrrole rings (1), aryl substituents (*e.g.* 4 and 5) or benzylsulfonamide (17) are represented. These substructures appear in diverse Non-Steroidal Anti-Inflammatory Drug (NSAID)s such as rofecoxib or etericoxib, as well as in chemical families of COX-2 inhibitors based on *e.g.* 1,5-diarylpurazoles or 3,4-diaryl-substituted furans [Blobaum and Marnett (2007); Dannhardt and Laufer (2000); Leval et al. (2000)] 190
- .5.17 **Compound substructures predicted to have the same influence on human COX-1 and COX-2.** Sub-structures predicted to decrease bioactivity are accompanied by a blue arrow, whereas that predicted to increase bioactivity are followed by a red arrow. Smaller substructures are found in this case, predominating substituents on the benzene ring. Therefore, substructure-activity relationships are difficult to be determined. 191

- .6.1 **Modelling workflow and compound clustering.** A. pGI_{50} values for 17,142 compounds on 59 cancer cell-lines (941,831 data-points) were modeled with PCM Random Forests and conformal prediction. B. U-matrix for the SOM used to cluster the compounds. Black lines delimit the 31 clusters defined, whereas red labels indicate the cluster number. The similarity between each neuron and its 8 neighboring neurons defines the color code: blue corresponds to high similarity (homogeneous areas), and red corresponds to low similarity (heterogeneous areas). Therefore, clusters presenting blue and red neurons exhibit higher levels of intra-cluster chemical diversity. 209
- .6.2 **Distribution of respective maximum and minimum $RMSE_{test}$ (A,B) and R_0^2 test (C,D) values for the complete data set.** Average maximum and minimum values of 1.42/0.35 and 0.96/-0.96, were obtained respectively for $RMSE_{test}$ / R_0^2 test with the simulated data. The performance of the PCM models on the test set was in agreement with the uncertainty of the experimental measurements, as mean $RMSE_{test}$ and R_0^2 test values of 0.40 +/- 0.00 pGI_{50} unit and 0.83 +/- 0.00 (with $n = 10$ models) were obtained. These values are between the two extreme, maximum and minimum, theoretical $RMSE_{test}$ and R_0^2 test values. . . 219
- .6.3 **Y-scrambling validation.** Mean $RMSE_{test}$ (A) and R_0^2 test (B) values were calculated for the observed against the predicted bioactivities on the test set calculated with models trained on pGI_{50} values increasingly randomized ($n=3$). R_0^2 test values become negative when 75% of the bioactivity values are randomized. These data suggest that the relationships established by the PCM models between compound and cell-line descriptors, and the pGI_{50} values did not arise from chance correlations. 220
- .6.4 **Distribution of respective maximum and minimum $RMSE_{test}$ (A,B) and R_0^2 test (C,D) values for the variable bioactivity profile data set.** Average maximum and minimum values of 1.90/0.54 and 0.94/-0.90 were obtained respectively for $RMSE_{test}$ / R_0^2 test with the simulated data. The performance of PCM models was in agreement with the uncertainty of the experimental measurements, as mean $RMSE_{test}$ and R_0^2 test values of 0.580 pGI_{50} unit and 0.79 were obtained. These values are between the two extreme, maximum and minimum, theoretical $RMSE_{test}$ and R_0^2 test values. 221

- .6.5 Benchmarking of the cell-line profiling data set views for compound sensitivity prediction.** A. The predictive power of the 16 data set views (Table .6.1) was quantified by the RMSE values on the test set. For each data set view, we trained ten models on the variable bioactivity profile data set. We found significant differences among the data set views (Analysis of Variance (ANOVA), $P < 0.01$). Post-hoc analyses (Tukey's Honest Significant Difference Test (HSD), $\alpha 0.05$) were used to cluster the data set views according to their predictive power. 223
- .6.5 Figure .6.5 caption continuation** Data set views sharing a letter label performed at the same level of statistical significance and are depicted in the same color. We consistently found that gene transcript levels, and the abundance of proteins and miRNA led to the most predictive models (labeled with *a*). B. The evaluation of both interpolation and extrapolation power was evaluated on the complete data set. After finding significant differences among groups (ANOVA, $P < 0.01$), we found that PCM interpolates and extrapolates to new cell-lines and tissues at the same level of statistical significance (Tukey's HSD, $\alpha 0.05$). By contrast, we found statistically significant differences in performance between extrapolation and interpolation to new chemical clusters. 224
- .6.6 Interpolating compound bioactivities to novel cell-lines, tissues, and chemical clusters.** A. Cell-line-averaged $\text{RMSE}_{\text{test}}$ values ranged from 0.41 ± 0.01 (U251) to 0.86 ± 0.01 pGI₅₀ unit (HOP-92). We found significant differences for tissue-averaged performance (Tukey's HSD, $P < 1 \times 10^{-16}$), with $\text{RMSE}_{\text{test}}$ values ranging from 0.48 ± 0.01 (prostate) to 0.70 ± 0.01 (leukemia) pGI₅₀ unit. Cell-lines originated from the same tissue are depicted in the same color (breast: red, central nervous system: magenta, colon: yellow, lung cancer: grey, leukemia: green, melanoma: blue, ovarian: orange, prostate: cyan, renal: brown). We did not observe significant differences in tissue-averaged performance for tissues labeled with the same letter. B. Compound-cluster averaged performance for the 31 clusters defined with Self-Organizing Map (SOM)s. 225
- .6.6 Figure .6.6 caption continuation (B)** One-way ANOVA among the 31 chemical clusters ($P > 0.05$), with compound cluster-averaged $\text{RMSE}_{\text{test}}$ values in the 0.48 ± 0.01 and 0.65 ± 0.01 pGI₅₀ unit range. This analysis illustrates that the models do not constantly favor specific chemical clusters, thus making it possible to interpolate compound bioactivities across the chemical space covered by the data at the same level of statistical significance. By contrast, interpolating on the cell-line side depends significantly on the tissue source. 226

- .6.7 Learning curves.** Mean $\text{RMSE}_{\text{test}}$ (A) and R_0^2 test (B) values were calculated for the observed against the predicted bioactivity values on the test set calculated with $n=3$ models obtained using training sets covering an increasingly higher fraction of the complete data set. Models trained on 5% of the data set exhibited a mean $\text{RMSE}_{\text{test}}$ value of 0.52 pGI_{50} unit, which decreased till 0.39 pGI_{50} unit when 95% of the data-points were included in the training set. These data suggest that PCM models exhibit high interpolation capabilities. In practice, the compound-cell-line interaction matrix could be completed with in silico predictions, with a $\text{RMSE}_{\text{test}}$ values of 0.39 pGI_{50} unit, without requiring further experimental testing. 226
- .6.8 Correlation between observed and predicted pGI_{50} values.** Density correlation plot corresponding to the observed against predicted pGI_{50} values on the test set for: (A) the Leave-One-Target-Out (Leave-One-Tissue-Out in chapter .6) (LOTO) model for melanoma (Root Mean Squared Error ($\text{RMSE}_{\text{test}}$) and R_0^2 test values of 0.43 pGI_{50} unit and 0.80), and (B) the Leave-One-Cell-Line-Out (LOCO) model for the melanoma cell-line SK-MEL-5 ($\text{RMSE}_{\text{test}}$ and R_0^2 test values of 0.37 pGI_{50} unit and 0.87. The color bar indicates the density of points at each region of the plot. For the rest of LOCO and LOTO models comparable results were obtained, with bioactivity values correctly predicted along the whole bioactivity range. 227
- .6.9 Correlation between observed and predicted pGI_{50} values for the 81 drugs present in the complete data set for the following model validation scenarios: (A) LOCO, (B) LOTO, and (C) Leave-One-Compound-Cluster-Out (LOCCO).** The x-axis reports the drug NSC identifiers. Compounds discussed in the main text, namely NSC 630176 and NSC 707389, are marked with asterisks. Bars are colored according to drug mechanism of action (MoA). The abbreviations of the mechanisms of action are: A2: alkylating at N-2 position of guanine; A7: alkylating at N-7 position of guanine; AM: antimetabolite; Ang: angiogenesis; Apo: apoptosis inducer; Db: DNA binder; Df: antifolates; DNMT: DNA methyltransferase inhibitor; Dr: ribonucleotide reductase inhibitor; Ds: DNA synthesis inhibitor; HDAC: Histone deacetylase; Ho: hormone; P90: hsp90 binder; PI3K: PI3kinase; PKC: Protein kinase C; ROS: reactive oxygen species; RSTK: serine/threonine kinase inhibitor; T1: topoisomerase 1 inhibitor; T2 : topoisomerase 2 inhibitor; Tu: tubulin-active antimitotic; YK: tyrosine kinase inhibitor. 228

- .6.10 **Validation of conformal prediction.** For each confidence level (ϵ), represented in the x-axis, the number of data-points in the test set which true value lies within the predicted interval is calculated, y-axis. The high Spearman's r_s is likely due to the large size of the test set (188,366 data-points) and to the fact that the Confidence Interval (CI) produced by conformal prediction are always valid [Norinder et al. (2014)]. These data indicate that the modeling framework combining PCM models and conformal prediction is more information rich than what would be possible with only point prediction algorithms. 230
- .6.11 **Consistency between the pathway-drug associations calculated with the experimental and the predicted bioactivity values.** Box plots reporting the distribution of Spearman's r_s coefficients for pathway-drug associations calculated with the experimental and the predicted values over the 56 drugs present in the variable bioactivity profile data set, using all pathway-drug associations (FDR < 20%) (A), or only significant associations (C), as estimated in the variable bioactivity profile data set. Bar plots representing the drug-averaged Spearman's r_s coefficients calculated with all B. or with only significant (D) pathway-drug associations, averaged over the models labeled with *a* in (A). Missing bars in (D) correspond to drugs for which we did not find significant drug-pathway associations. 232
- .6.11 **Figure .6.11 caption continuation** E. Data view-averaged Spearman's r_s coefficients for patterns of growth inhibition calculated with the experimental and the predicted values. F. Bar plot reporting the drug-averaged Spearman's r_s coefficients for the patterns of growth inhibition calculated with the observed and the predicted bioactivities. Data views sharing a letter label and color in (A,C,E) perform at the same level of statistical significance. Significance for the Spearman's r_s in (B,D,F) is represented with an asterisk if two-sided P value < 0.05, for the Spearman's r_s coefficients calculated with the predictions generated with a model trained on the 'G.t.l. 1,000 genes' data view. Bars in (B,D,F) are colored according to compound MoA. Abbreviations of mechanisms of action: MoA: Mechanism of action; A2: alkylating at N-2 position of guanine; A7: alkylating at N-7 position of guanine; AM: antimetabolite; Ang: angiogenesis; Apo: apoptosis inducer; Db: DNA binder; Df: antifolates; DNMT: DNA methyltransferase inhibitor; Dr: ribonucleotide reductase inhibitor; Ds: DNA synthesis inhibitor; HDAC: Histone deacetylase; Ho: hormone; P90: hsp90 binder; PI3K: PI3kinase; PKC: Protein kinase C; ROS: reactive oxygen species; RSTK: serine/threonine kinase inhibitor; T1: topoisomerase 1 inhibitor; T2 : topoisomerase 2 inhibitor; Tu: tubulin-active antimitotic; YK: tyrosine kinase inhibitor. 233

- .6.12 Evaluation of the predicted growth inhibition patterns for methotrexate (MTX) on the NCI60 panel.** A. Relative growth inhibition pattern (z-scores) on the NCI60 panel calculated from the experimental pGI_{50} values. The experimental uncertainty of the measurements is also displayed. B. Predicted relative growth inhibition pattern of growth inhibition along with the 75% confidence interval calculated using con-formal prediction. We used the predicted values on the test calculated with 10 PCM models (interpolation). Complex inhibition patterns are reflected by the predictions. 234
- .6.12 Figure .6.12 caption continuation** For instance, renal cell-lines TK-10, RXF-393 and A498 (marked with an asterisk) were predicted to be highly resistant to MTX, whereas the effect of MTX on sensitive cell-lines, namely UO-31, SN12C, CAKI-1 and ACHN, was also correctly predicted. Cell-lines originated from the same tissue are in the same color (breast: red, central nervous system: orange, colon: olive green, lung cancer: dark green, leukemia: turquoise, melanoma: blue, ovarian: blue, prostate: purple, renal: magenta). 235
- .6.13 Correlation of gene expression profiles for the 44 cell-lines present in both the NCI60 panel and the Cancer Cell Line Encyclopedia (Cancer Cell-Line Encyclopedia (CCLE)).** A. Pairwise Spearman's r_s correlation of the 1,000 most varying genes between the DTP-NCI60 and the CCLE data sets. Both data sets share 44 cell-lines. The correlation between the gene expression profiles of identical cell-lines is higher than 0.8 in all cases (diagonal of the matrix), with a median Spearman's r_s value close to 0.875. B. The first box plot on the left reports the Spearman's r_s correlation, above 0.98, between the gene transcript levels calculated in triplicates for the NCI60 cell-lines. The box plot in the middle corresponds to the correlation between the gene expression profiles of the cell-lines found in both the CCLE and the NCI60 data set (diagonal of the matrix in (a)). The average Spearman's r_s correlation is close to 0.875. The third boxplot reports the Spearman's r_s correlation of different cell-lines (the non-diagonal elements of the matrix in (a)). The high correlation between gene expression profiles for the cell-lines present in both the CCLE and the NCI60 cell-line panel, indicates that the PCM models reported in this study could be extended to the CCLE. 239

List of Tables

.1.1	PCM studies published between 2010 and 2013. The wide applicability of PCM is evidenced by the increased coverage of drug targets in the studies of the last three years. Although traditional drug targets, such as GPCRs or kinases, are still widely represented, new applications (e.g. the modelling of viral genotypes or pharmacogenomics) are gaining ground steadily. BPN - Back Propagation Networks, BS - Bootstrapping Validation, CTD - Composition and transition of amino acid properties, CV - Cross-Validation, DCNB - Dual Component Naive Bayes, DCSVM - Dual Component Support Vector Machines, DT - Decision Trees, DTV - Decoy Test Validation, ENR - Elastic Net Regression, EV - External Validation, GP - Gaussian Processes, KNN - k-Nearest Neighbors, LCO - Leave-Cluster-Out Validation, LOTO - Leave-One-Target-Out Validation, NB - Naive-Bayes, NN - Neural Network, MLR - Multiple Linear Regression, OOB - Out-Of-Bag Validation, PCA - Principal Component Analysis, PLS - Partial Least Squares, Random Forest - RF, RS - Random Splitting, SVM - Support Vector Machines, SVR - Support Vector Regression, Y-Sc - Y-Scrambling.	8
.1.2	Selection of machine learning prediction methods used for PCM . . .	13
.2.1	Amino acid descriptor sets used in PCM (adapted from Westen et al. (2013a)). Those descriptor sets marked with * can be computed with the R package <i>camb</i>	57
.2.2	Covariance functions (kernels) formula.	64
.3.1	Overview of the proteochemometric data sets modeled in this work. Whereas the completeness of the compound-target interaction matrix of the dengue virus NS3 proteases data set is almost complete (88.84%), the adenosine receptors and GPCRs data set are more challenging to model given: (i) their sparsity (31.11 and 2.43% of matrix completeness respectively), and (ii) the consideration of information from human orthologues, being the respective number of different sequences 8 and 91.	92

.3.2	Internal and external validation metrics for the PCM models. For the three data sets, the best models are obtained with GP, being the lowest $RMSE_{test}$ and highest $R^2_{0\ test}$ values: (i) adenosine receptors: 0.58 and 0.75 with NP kernel, (ii) GPCRs: 0.66 and 0.72 with NP kernel, and (iii) Dengue virus NS3 proteases 0.44 and 0.92 with Bessel kernel. Overall, GP models for the three data sets agree with the validation criteria. .	99
.3.3	Number of datapoints per GPCR. Those receptors highlighted by a '*' symbol correspond to those present in a subset of human GPCRs which was first modeled with GP (see subsection GP performance per Target). GPCRs are named according to UniProtKB/ Swiss-Prot database [Magrane and Consortium (2011)].	109
.4.1	Algorithms benchmarked in this study. The third column indicates the parameters that were tuned using grid search and cross-validation (CV). The default values were used for those parameters not indicated therein.	133
.4.2	Data sets modelled in this study. These data sets can be found in the references indicated in the last two columns.	135
.4.3	Values for the slopes (coefficients) and P-values for the fitted linear model. The factor levels Random Forest (Algorithm), F7 (Data set), Morgan fingerprints (Descriptor type), Replicate 1 (Replicate) and Noise : 0 (Noise level) were used to calculate the intercept term of the linear model. Noise levels are reported as percentage points. The significance level was set to 5%, whereas all P-values are two-sided. .	144
.5.1	Composition of the COX data set. The total number of bioactivities, after duplicate removal and selected from ChEMBL as described in Materials and Methods, and of distinct compounds are 4,937 and 3,228 respectively. The last column indicates the percentage of the total number of distinct compounds (3,228) annotated on each target.	162
.5.2	Internal and external validation metrics (mean values +/- one standard deviation) for the PCM (A), Family QSAM (B), Family QSAR (B), Individual QSAR models (C), Ensemble PCM models combining the most predictive models (D), and Ensemble PCM models combining the whole model library (E). "Best" refers to the ensembles trained on only the three most predictive RF, GBM and SVM models. MS of models trained with different algorithms in a models ensemble allows to increase predictive ability, as the highest $R^2_{0\ test}$ and $RMSE_{test}$ values, 0.652 and 0.706 pIC50 units respectively, were obtained with the "MS SVM Radial Ensemble". The standard deviation for the metrics was calculated with the bootstrap method [Efron and Tibshirani (1993)]. . .	176

.6.1	Description of the data set views benchmarked for compound sensitivity prediction on the NCI60 panel. The abbreviated names used in Figure .6.5 are indicated in the second column. Prior biological knowledge, such as pathway information, was included in some data set views, whereas the gene transcript levels and mutational status for genes implicated in cancer, kinases and ABC transporters were gathered independently and combined in data set views to assess the redundancy of their predictive signal.	211
.6.2	Notes: 1. N/A refers to data not reported in the corresponding study. 2. Low R^2 values do not necessarily mean inaccurate predictions, as R^2 values decrease significantly, even if the predictions closely match the observations, when the range of values considered is small (see Figure .2.3). 3. The values reported for the NCI60 data set correspond to those obtained with the complete data set (see main text for details). Abbreviations: CCLE: Cancer Cell-Line Encyclopedia; CNV: Copy Number Variation; CV: Cross-validation; GDSC: Genomics of Drug Sensitivity in Cancer; LOCCO: Leave-One-Chemical-Cluster-Out; LOTO: Leave-One-Tissue-Out.	236

Introduction

.1 Polypharmacology Modelling Using Proteochemometrics (PCM)

.1.1 Introduction

.1.1.1 Available bioactivity data is growing: but can we make sense of it?

THE cost of developing new drugs has been continuously increasing in recent years and it is now estimated to be in the order of \$1.8 billion *per drug*. In addition, price pressure from health care providers has been increasing and there is a growing relevance of more targeted medicine. Hence, the blockbuster model of the pharmaceutical industry is being challenged [Akella and DeCaprio (2010); Paul et al. (2010)]. However, at the same time the amount of bioactivity data available both inside companies as well as in the public domain has significantly increased, for example with introduction of ChEMBL and PubChem Bioassay [Gaulton et al. (2012); Wang et al. (2012)]. This trend can be expected to only gain further speed in the future. The question now arises how this growing amount of bioactivity data can be used in real-world drug discovery and chemical biology projects, both to make drug discovery in commercial settings more efficient, but also to understand on a more fundamental level how we can use data in order to design a ligand with desired properties in a biological system.

Predictive bioactivity methods, such as Quantitative Structure-Activity Relationship (QSAR) models, are based upon the compound similarity principle [Bender and Glen (2005); Willett (2009)]. However, it has been shown that the activity of a compound against a single target is not sufficient to understand its actions in a biological system. In fact, promiscuity is intrinsic to chemical compounds [Mestres et al. (2008, 2009)], bioactivity against related targets frequently needs to be considered for efficacy of *e.g.* CNS-active drugs and anti-cancer drugs [Bianchi and Botzoulakis (2010); Shoshan and Linder (2004)], and promiscuity has been used to anticipate side-effects [Bender et al. (2007)]. Hence, only the simultaneous modelling of both the chemical and the target domain, across a series of protein targets, permits the meaningful mining of the compound-target interaction space [Bieler and Koeppen (2012)].

The term chemogenomics comprises techniques capable to capitalize on this huge amount of bioactivity data by considering compound and target information, in order to find unknown interactions between (new) compounds and their (new) targets [Bredel and Jacoby (2004); Jacoby (2013)]. Proteochemometrics (PCM) modelling describes methods where a computational description from the ligand side of the system is combined with a description of the biological side being studied and both are related to a particular readout of interest [Lapinsh, Prusis, and Gutcaits (2001); Westen et al. (2011a)].

In this context, ligands are typically small molecules although biologics also have been explored, whereas the biological parameters in the model can comprise protein binding sites, but also *e.g.* gene expression levels of particular cell-lines. The readout describes the biological effect of a particular ligand on the protein or cell-line of interest (such as an IC_{50} value of this particular combination of compound and biological system). Additionally, PCM relates to personalized medicine as it can predict the effect of a ligand on a complex biological system, *e.g.* cell-line, from genotypic information [Cortes-Ciriano, I et al. (2015)].

.1.1.2 Synergy between ligand and target space

An analysis of the drug-target interaction network demonstrated that a given ligand interacts with six protein targets on average at therapeutic concentrations [Mestres et al. (2009)]. Targets with correlated bioactivity profiles might be related or distant from a sequence similarity standpoint. It has been recently shown that the classification of class A GPCRs based on ligand activity differs considerably from that obtained when using a classic description of proteins based upon sequence alignments [Lin et al. (2013); Westen and Overington (2013)]. Hence, full sequence similarity from multiple sequence alignments would not generally correlate with similar ligand affinity. Nevertheless, kinases exhibiting a sequence identity higher than 60% tend to have similar ATP-binding sites and hence they tend to be inhibited by similar compounds [Vieth et al. (2005)]. Similarly, compound binding is more conserved between human and rat orthologous proteins with respect to paralogues [Kruger and Overington (2012); Westen et al. (2012)]. Thus, to better understand intra-family and inter-species selectivity both the target and the compound space need to be considered simultaneously.

In ligand space, chemogenomic approaches relying only on ligand data have shown that there is an unequal distribution of ligand data. This is due to the fact that some target classes (*e.g.* GPCRs or kinases) have been traditionally regarded as more interesting from a medicinal chemistry standpoint, and are thus overrepresented in bioactivity databases [Gregori-Puigjané and Mestres (2008b)]. Moreover, while

some chemogenomic methods implicitly consider target information using bioactivity profiles of groups of similar ligands, *i.e.* the interaction between these compounds and a panel of targets, they are outperformed by techniques that explicitly consider target information [Gregori-Puigjané and Mestres (2008a); Rognan (2007)]. In addition, bioactivity profiles for related compounds are not always available.

In target space, techniques were employed which benefit from the structural or sequence information available and rely on groups of related targets with the aim to identify possible off-target effects and drug specificity for a particular target of interest [Rognan (2007)]. Based on the inverse similarity principle, related proteins are likely to interact with similar compounds. As in the previous case, the unavailability of data also constitutes a limitation for target-based chemogenomics.

The combination of ligand and target data allows the creation of predictive models that can rationalize *e.g.* viral or cancer cell-line selectivity, whereas models exclusively based on ligands cannot explain the role of the target in selectivity [Westen et al. (2011b)]. Merging data from ligand and target sources into the frame of a single machine learning model allows the prediction of the most suitable pharmacological treatment for a given genotype (personalized medicine), which ligand-only and protein-only approaches are not able to perform. This is precisely the underlying principle in proteochemometrics (PCM), which employs both ligand and target features simultaneously, and which therefore enables the deconvolution of both the target and the chemical spaces in parallel [Lapinsh, Prusis, and Gutcaits (2001); Westen et al. (2011a)].

1.1 Polypharmacology Modelling Using Proteochemometrics (PCM)

Dataset (data-points)	Receptor	Ligand descriptors	Target Descriptor	Bioactivity type	Machine Learning Technique	In silico Model Validation	Prospective Validation?	Remarks, Inferences	Reference
PDBbind [Wang et al. (2004)] (1,300)	1,300 protein-ligand complexes	Atom-type based	Atom-type based	K_d, K_i	RF	Y-sc, OoBV, EV	No	Increasing the training set size improves model predictability	Balaster and Mitchell (2010)
ProLINT database [Shandara] (3,595)	62 Kinases	Structural fragments and 2D autocorrelation vectors	Sequence-based structural fragments and amino acid sequence autocorrelation	IC_{50}	SVM	3-fold CV, EV	No	SVM based on autocorrelation descriptors perform better than fragment-based approaches	Fernandez, Ahmad, and Sarai (2010)
PDBbind [Wang et al. (2004)] (1,255)	Diverse Proteins	Property-encoded shape distributions	Property-encoded shape distributions	K_d, K_i	SVM	5-fold CV, EV	No	Training set enrichment and expansion enhances prediction accuracy	Das, Krein, and Brennan (2010)
Stanford HIV Drug Resistance database [Rhee et al. (2003)] (4,495)	728 Reverse transcriptases	Dragon descriptors [A Mauri (2006)]	z-scales [Sandberg et al. (1998)]	IC_{50}	PLS	7-fold CV, EV	No	Receptor-ligand and receptor-receptor cross-terms improved model performance	Junaid et al. (2010)
Immune Epitope Database [Vita et al. (2010)] (31,992)	12 HLA-DRB1 proteins	z-scales [Sandberg et al. (1998)]	z-scales [Sandberg et al. (1998)]	IC_{50}	PLS	7-fold CV, EV	No	Identified protein residues and peptide positions for binding predictions	Dimitrov and Garnev (2010)
Karaman et al. (2008) (12,046)	317 human Kinases	Dragon descriptors [A Mauri (2006)]	z-scales, [Sandberg et al. (1998)] Amino acid composition, sequence order and CTD	K_d	PLS, SVM, KNN, DT	Double CV	No	SVM outperforms all machine learning approaches	Lapinsh and Wikberg (2010)
CSAR-NRC HiQ [Kramer and Gedeck (2011b)]	346 protein-ligand complexes	Atom counts	Atom counts	K_d	MLR	RS	No	Distance dependent atom descriptors make the regression models more robust.	Kramer and Gedeck (ibid.)
Gold standard set (1,933)	313 diseases (OMIM) [Hamosh et al. (2002)]	Diverse Drug similarity measures	Disease-Disease similarity measure	Classifier score	Logistic regression classifier	10-fold CV, EV	No	Possibilities to include patient-specific gene expression profiles make the models suitable for pharmacogenomics studies	Gottlieb, Ruppin, and Sharan (2011)
Sc-PDB [Kellenberger et al. (2006)] (2,882)	581 targets	Hashed fingerprints	Protein sequence and 3-D structure based	Actives / inactives	SVM	5-fold CV, EV	No	Structure-based approaches perform better than sequence-based approaches	[Meslamani and Rognan (2011)]
GLIDA database [Okuno et al. (2006)] (5,207) and CVK Kinase database (15,616)	317 GPCRs and 143 Kinases	Dragon descriptors [A Mauri (2006)]	Protein sequence and feature-based	K_i, EC_{50}, EC_{50}	SVM	5-fold	9 compounds for ADRB2	Highly active compounds predicted by SVM not identified by	Yabuuchi et al. (2011)
Tibotec BVBA (4,024)	14 HIV RT	Circular fingerprints	Hashed fingerprints	EC_{50}	SVM	Y-Sc, CV, E	5 inhibitors for EGFR	ligand-based / structure-based approaches	Westen et al. (2013b)
Biointo-DB [Weill et al. (2011)] (336,678)	Oxytocin receptor	MACCS structural keys	Fingerprints based on the properties of amino acids in active site	Actives / inactives	RF	Losov 10-fold CV, EV	Biological evaluation of 37 compounds (2 hits)	PCM models yield better hits than the conventional virtual screening procedures.	Weill et al. (ibid.)

Dataset (data-points)	Receptor	Ligand Descriptors	Target Descriptor	Bioactivity type	Machine Learning Technique	In silico Model Validation	Prospective Validation?	Remarks, Inferences	Reference
PDBbind refined set (1,387)	23 protein families (1,387 proteins)	Atom-type based	Atom-type based, Distance-dependent protein ligand atom type pairs	K _d	MLR, PLS	5-fold CV, LCO	No	Inclusion of descriptors from PCM models predict free energies more accurately than docking programs	Kramer and Gedeck (2011a)
Stanford HIV Drug Resistance Database (4,794 protease and 4,495 RT sequence-inhibitor combinations)	828 HIV-1 protease variants	GRIND alignment independent descriptors [Pastor et al. (2000)]	z-scales [Sandberg et al. (1998)]	Inhibitor concentration	PLS	Double loop CV, Y-Sc and EV	No	Intra-protease cross-terms improve performance	Spiuth et al. (2011)
Kinase SARfari [Gaulton et al. (2012)] (85,908)	342 human Kinase domains	Extended Con-nectivity Finger-prints (ECFP-6) [Rogers and Hahn (2010)]	Fingerprints based on amino acid residues and physio-chemical properties	IC ₅₀ , K _i	DCSVM and DCNB	RS, EV	No	DCSVMs provide better activity prediction	Nijima, Shiraishi, and Okuno (2012)
BindingDB [Liu et al. (2007)] (1,275)	5 HDAC isoforms	Physical prop-erties and topo-logical indices of compounds	Sequence similarity, structure similarity, geometry descriptors	IC ₅₀	SVR	10-fold CV, EV	No	SVR models with PUK kernels have stronger mapping capabilities	Wu et al. (2012)
Docked complexes (2,335 PDB structures & 3,671 FDA drugs)	2,335 human targets	Ligand shape descriptors	Binding site shape descriptors	Ligand contact point score	PCA	DTV, EV	VEGFR2 inhibition by Meben- dazole and Cad-herin 11 inhibition by Cele-coxib were verified.	TFMS PCM approach can assist in drug reposi-tioning studies	Dakshanamurthy et al. (2012)
Literature (160 protein-ligand complexes)	47 HIV-1 proteases	Physical prop-erties, topological indices of com-pounds	z-scales [Sandberg et al. (1998)]	K _i	SVR	10-fold CV, EV	No	Protein-ligand interaction fingerprints improved models over cross terms	Huang et al. (2012a)
CHEMBL 2 [Gaulton et al. (2012)] (10,999)	8 Human and Rat adenosine receptors	Circular finger-prints	Hashed fingerprints	K _i	SVM	Y-Sc, EV, DTV	6 novel compounds were exper-imentally identified	Addition of orthologue information increased model quality	Westen et al. (2012)
CHEMBL 8 [Gaulton et al. (2012)] (81,689; 43,965)	136 GPCRs and 176 Kinases	MACCS keys	Sequence descriptors	K _i , IC ₅₀	SVM	5-fold CV, EV	No	Feature selection im-proved the predictive accuracy of the models	Cheng et al. (2012)
GVK Biosciences database (GVK Biosciences Private Limited, Hyderabad, India., 2007) (628,120)	238 Class A GPCRs	Chemical ker-nels based on ECFP-6 fingerprints and Dragon descriptors	Protein kernels based on full length, TM and loop sequences	Agonists / antago-nists	SVM	RS, DT, EV	No	Protein kernels based on TM sequences showed higher prediction accuracy	Shiraishi et al. (2013)
GDSC dataset [Yang et al. (2013)] (58,930)	639 cancer cell-lines	PaDEL descrip-tors (Vap, 2011)	CNV, sequence vari-ation and Microsatel-ite instability status	IC ₅₀	RF and NNs	8-fold CV, EV	No	PCM based on existing drugs allows drug reposi-tioning and pharmacoge-nomics studies	Menden et al. (2013)
Peptide library (180)	4 proteases	Binary and physiochemical descriptors	Binary descriptors	K _i	PLS	5-fold CV	No	Inclusion of intra-peptide cross-terms improved model performance	Prusis et al. (2013)
Kinase SARfari (54,012)	372 Kinases	Topological fin-gerprints	Amino-acid composi-tion and CTD	IC ₅₀ , K _i	RF and NB	OOB, 5-fold CV, EV	No	Random forests outper-form Naïve Bayes.	[Cao et al. (2013a)]

Dataset (data-points)	Receptor	Ligand descriptors	Target Descriptor	Bioactivity type	Machine Learning Technique	In silico Model Validation	Prospective Validation?	Remarks, Inferences	Reference
Virco (300,000)	HIV mutants (10,700 NNRTI, 10,500 NRTI, 27,000 PI)	Circular fingerprints	z-scales [Sandberg et al. (1998)]	IC ₅₀	SVM	Y-Sc, 5-fold CV, EV	No	Phenotypic resistance for novel mutants can be predicted via PCM	Westen et al. (2013a)
GPCRDB [Vrolijk et al. (2011)] (310)	9 human amine GPCRs	Physical properties and topological indices of compounds	z-scales [Sandberg et al. (1998)] and TM identity descriptors	K _i	SVR and GP	10-fold CV, EV	No	SVR is superior to GP.	Gao et al. (2013)
PubChem BioAssay dataset [Wang et al. (2012)] (63,391)	5 CYP 450 isoforms	Molecular signatures	CTD	AC ₅₀	KNN, SVM and RF	CV, EV	No	TM identity descriptors perform better than z-scales descriptors	Lapinsh et al. (2013)
Binding and PDSP database [Liu et al. (2007)] (13,079)	514 human targets	Topological fingerprints	Amino-acid composition and CTD	K _i	RF and NB	OOB, 5-fold CV, EV	No	Non-linear methods (SVM and RF) perform better.	Cao et al. (2013a)
In vitro OATP modulation data (2,000)	OATP1B1 and OATP1B3	Circular fingerprints	z-scales [Sandberg et al. (1998)] and feature-based ProfFP	K _i	RF	OOB, EV	Agreement between experiment and prediction	Random forests outperform KNN, SVM, NB and BPN	De Bruyn et al. (2013)
[Davis et al. (2011); Karaman et al. (2008); Metz et al. (2011)] datasets	50 Kinases	Mold2 [Hong et al. (2008)], Open Babel [O'Boyle et al. (2011)] and vol-surf [Cruciani et al. (2000)] descriptors	Knowledge-based fields ₁₂₃ and w-term _{ap124} derived fields	K _d / K _i	PLS	7-fold CV, EV, LOTO, Y-Sc	No	4class models are superior to z-class models and provide information about selectivity	Subramanian et al. (2013)

Table .1.1: PCM studies published between 2010 and 2013. The wide applicability of PCM is evidenced by the increased coverage of drug targets in the studies of the last three years. Although traditional drug targets, such as GPCRs or kinases, are still widely represented, new applications (e.g. the modelling of viral genotypes or pharmacogenomics) are gaining ground steadily. BPN - Back Propagation Networks, BS - Bootstrapping Validation, CTD - Composition and transition of amino acid properties, CV - Cross-Validation, DCNB - Dual Component Naive Bayes, DCSVM - Dual Component Support Vector Machines, DT - Decision Trees, DTV - Decoy Test Validation, ENR - Elastic Net Regression, EV - External Validation, GP - Gaussian Processes, KNN - k-Nearest Neighbors, LCO - Leave-Cluster-Out Validation, LOTO - Leave-One-Target-Out Validation, NB - Naive-Bayes, NN - Neural Network, MLR - Multiple Linear Regression, OOB - Out-Of-Bag Validation, PCA - Principal Component Analysis, PLS - Partial Least Squares, Random Forest - RF, RS - Random Splitting, SVM - Support Vector Machines, SVR - Support Vector Regression, Y-Sc - Y-Scrambling.

.1.1.3 PCM as a practical approach to use chemogenomics data

PCM modelling is a computational technique which combines both ligand information and target information within a single predictive model in order to predict an output variable of interest (usually the activity of a molecule in a particular biological assay) [Lapinsh, Prusis, and Gutcaits (2001); Westen et al. (2011a)]. It is this combination of orthogonous information that sets PCM apart from both QSAR and chemogenomics [Horst et al. (2011); Rognan (2007)]. Generally, the term *target* refers to proteins since the majority of PCM models in the literature have been devoted to the study of the activity of compounds on protein targets. Yet, target can also refer to a certain protein binding pocket (to allow distinction between binding modes, protein conformations, or allosteric/orthosteric binding) [Bahar, Chennubhotla, and Tobi (2007)], or even to a cell-line [Menden et al. (2013)]. Each binding site and each binding mode can be regarded (computationally) as a different *target*.

A PCM model is trained on a data set composed of a series of targets and compounds, where compounds have been measured on as many targets as possible (illustrated in Figure .1.1). The simultaneous modelling of the target and the ligand space permits to better understand complex drug-target interactions (*e.g.* selectivity) [Keiser et al. (2007); Ning, Rangwala, and Karypis (2009); Paolini et al. (2006); Wassermann, Geppert, and Bajorath (2009)] than it would be possible with chemogenomics. Indeed, the simultaneous modelling of compound and target data allows to assess the effect of target and chemical variability can be evaluated (*e.g.* protein mutations or the effect of chemical substructures on bioactivity). Thus, the aim of PCM is the complete modelling of the compound-target interaction space (Figure.1.1), including also the prediction of the bioactivity of novel compounds on yet untested targets.

Initial attempts to incorporate description of several proteins and their ligands in a single QSAR model involved modelling of the interaction between mutated glucocorticoid receptors and Deoxyribonucleic Acid (DNA) [Tomic, Nilsson, and Wade (2000); Zilliacus et al. (1992)]. The first full scale PCM study involving different proteins was devoted to the interaction of chimeric melanocortin receptors with chimeric peptides at Uppsala University [Prusis et al. (2001)]. The name *proteochemometrics* was coined later by the same research group [Lapinsh, Prusis, and Gutcaits (2001)]. Since then PCM has been applied on various diverse data sets (Table .1.1) [Bock and Gough (2005); Lapinsh et al. (2002)]. While the current chapter will focus on recent developments in the field between 2010 and 2013, a comprehensive discussion of PCM-related work before 2010 has been presented in a previous review by Westen et al. (2011a), to which we would like to refer the reader.

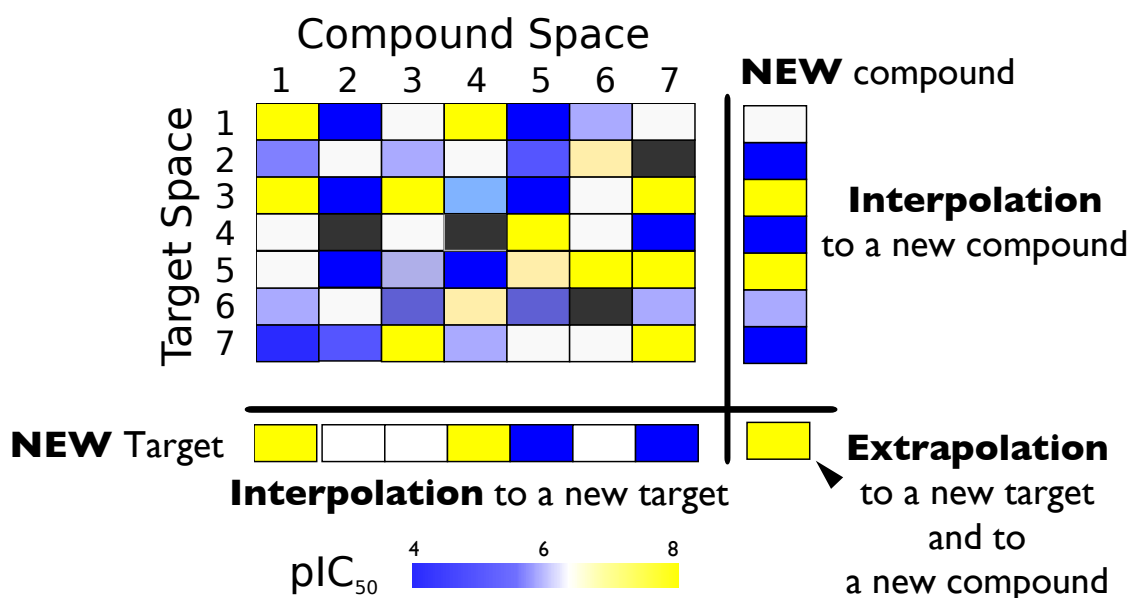


Figure .1.1: **Ligand-target interaction space.** The interaction between ligands (chemical compounds) and targets (biological macromolecules) can be envisioned as a matrix, where rows are indexed by target ids and columns by compound ids. Each matrix cell contains the binding affinity of a given compound on a given target, indicated by the following colors: blue means low affinity and yellow means high affinity. Traditional bioinformatics techniques have dealt with the similarity between targets, normally based upon sequence similarity. On the other hand, ligand based (QSAR) models have studied series of compounds acting on a given target. By contrast to both of them, PCM relates the chemical-target interaction space by describing targets and compounds with numerical descriptors permitting to predict activities of a given compound on a given target.

.1.1.4 Practical relevance of PCM

The novel way by which PCM brings together the chemical and the target spaces permits to better understand and predict the influence of target variability on compound activity. For instance, predicting compound activity on a cancer cell-line panel can identify selective compounds towards a particular cell-line [Cortes-Ciriano, I et al. (2015)]. Similarly, the influence of viral proteins mutations in compound activity can be quantified [Westen et al. (2011b)]. Therefore, PCM opens new avenues: (i) to mine drug affinity databases with the goal to create multi-target and multispecies models, (ii) to integrate toxicogenomics and phenotypic data in predictive models, (iii) to identify designed or natural ligands for orphan receptors (receptor deorphanization), (iv) and to design personalized medicine for viral infections or a defined cancer type based on genotypic information. The ability of PCM to model these data depends on the structure of the input matrix, as we will elaborate on below, and concrete examples referring to the above fields will be presented in the subsequent section.

Table .1.1 summarizes the main features of the PCM studies published between 2010 and 2013. In addition to traditional therapeutic targets (*e.g.* kinases or GPCRs), which continue to be well represented in recent PCM studies, other applications and techniques are gaining ground steadily, namely:

- The modelling of the selectivity of viral protein mutants, mainly Human Immunodeficiency Virus (HIV)
- The inclusion of bioactivity information from mammal orthologues
- The usage of 3D target information
- Toxicogenomics and pharmacogenomics

In the following sections, we will focus on:

- Section .1.2: (novel) machine learning techniques successfully applied in recent PCM studies (Table .1.2) and other predictive modelling contexts such as chemoinformatics
- Subsection .1.3: recent application of PCM on established protein target classes
- Subsection .1.4.1: Novel target similarity measure
- Subsection .1.4.2: including 3D information of protein targets in PCM
- Subsection .1.4.3: PCM applied to predict binding free energies (PCM-based scoring functions)

- Subsection .1.4.4: PCM as an approach to extrapolate bioactivity data between species
- Subsection .1.4.5: PCM applied to pharmacogenomics and toxicogenomics data
- Section .1.5: pitfalls of PCM
- Section .1.6: future perspectives and concluding remarks

.1.2 Machine Learning in PCM

Most of the currently used machine learning (Partial Least Squares (PLS), rough set modelling, neural net modelling, Naive Bayesian classifiers, and decision tree algorithms) as well as data preprocessing techniques in PCM have been described in recent reviews by Andersson, Gustafsson, and Strömbergsson (2011) and Westen et al. (2011a). Moreover, feature selection methods and common algorithms have been recently benchmarked, with the overall conclusion that kernel and tree methods, such as SVM or Random Forest (RF), do not benefit from feature selection, and that no particular algorithm-feature selection pair appears to be preferable [Bruce et al. (2007); Eklund et al. (2012, 2014)]. Therefore, only recent applications of novel techniques applied to PCM or chemoinformatic modelling will be discussed here, namely: Support Vector Machines (SVM), Random Forest (RF), Gaussian Processes (GP) and Collaborative Filtering (CF). A detailed description of the machine learning algorithms described in the following subsections is given in Table .1.2, whereas the mathematical formulation for these are given in Section .2.5.

Algorithm	Description	Advantages	Disadvantages	Reference
Support Vector Machines (SVM)	Maps the input space into a higher dimensional space where a hyperplane is at the interface between classes.	Medium training time. PUK kernel uses an approximation of linear, polynomial and RBF kernels	Optimize bandwidth hyper-parameter. No consideration of experimental error. No error bars for the predictions.	Ben-Hur and Ong (2008)
Dual-component SVM (DC-SVM)	Amino acid residues and compound fragments are treated as two components.	Accurate prediction of active versus inactive.	Huge kernel matrix. Reduced efficiency due to size.	Nijima, Shiraishi, and Okuno (2012)
Transductive SVM (TSVM)	Semi-supervised text mining technique.	Effective with unbalanced datasets. Smoothen the decision boundaries.	Difficult to implement without proper tuning.	Kondratovich, Baskin, and Varnek (2013)
Relevant Vector Machine (RVM)	Probabilistic counterpart of SVM.	Contains sparse descriptors. Smoothen the decision boundaries. Fast prediction. Easy retrieval of important descriptors.	Non informative predicted variance.	Tipping (2001)
Random Forest (RF)	Constructs multiple decision trees with random selection of variables	Computationally less expensive than SVM. Short training time (parallelization).	Requires relatively large amounts of memory	De Bruyn et al. (2013)
Gaussian Processes (GP)	Non-parametric Bayesian technique. Gives each prediction as Gaussian distribution.	Measurable interval of confidence (IC). Consideration of experimental uncertainty.	Long training time.	Schwaighofer et al. (2007)
Matrix Factorization	Calculates activities as dot product of compound and target features.	Missing values are predicted efficiently.	Interpretability.	Gao et al. (2012)
Collaborative Filtering	Multi-task learning.	Inferred features could be used as descriptors in the activity model. Estimates similarity between targets.	Performance on sparse data.	Erhan et al. (2006)

Table .1.2: Selection of machine learning prediction methods used for PCM

.1.2.1 Support Vector Machines (SVM)

Support Vector Machines (SVMs) are a group of non-linear machine learning techniques commonly used in computational biology [Scholkopf, Koji, and Vert (2004)], and in PCM in particular [Westen et al. (2011a, 2012)]. SVMs became popular in the last decade due to their performance and efficient capacity to deal with large data sets also in high-dimensional variable spaces, even though interpretability can be challenging [Ben-Hur and Ong (2008); Scholkopf and Smola (2001)] Furthermore SVMs require proper tuning of the so-called hyper parameters, usually determined by an exponential grid search.

In a recent study from Lapinsh et al. (2013) Random Forest (RF), k-Nearest Neighbours (KNN), and SVMs were applied to construct a PCM model of Cytochrome P450 (CYP) inhibition. The models were trained on 5 CYPs and 17,143 compounds. CYPs were described with transition and composition description of amino acids, while compounds were described with structural signature descriptors. These PCM models were shown to outperform single target models in terms of Area Under the Curve (Area Under the Curve (AUC): PCM: >0.90, QSAR: 0.79-0.89) that were constructed in parallel by Cheng et al. (2011). Of the methods used, RF and SVM were shown to be comparable in terms of accuracies and AUC. The high performance of the SVM model in the external validation (AUC: 0.94) evidences the suitability of this approach to correctly extrapolate in both the target and compound space.

SVMs can use different internal methods (kernels) to derive bioactivity predictions, the most dominant being the Radial Basis Function (RBF) kernel [M G Genton, N Cristianini, J Shawe-Taylor (2001)]. Radial basis function kernels have been shown to perform well on PCM data [Westen et al. (2011a, 2012)]. Recently the VII Pearson function-based Universal Kernel (PUK) [Üstün, Melssen, and Buydens (2006)] was also applied to PCM. Wu et al. (2012) showed that they were able to improve the mapping power of their PCM models for 11 Histone Deacetylase (HDAC)'s by using a PUK kernel. Nonetheless, the radial kernel still constitutes a common option when inducing bioactivity models given the necessity to tune only one kernel parameter, which in practice means shorter training times. Based on those results, the experienced user should keep in mind that although the radial kernel is a robust option with reliable results (in the experience of the authors), a proper kernel choice should be made on the basis of the data at hand [Duvenaud et al. (2013)].

Dual Component SVMs (DC-SVM) are an extension of the classical SVM and have been applied by Nijima, Shiraishi, and Okuno (2012) to a kinase data set spanning the whole kinome. They proposed a dual component Naive Bayesian model in which kinase-inhibitor pairs are represented by protein residues and ligand fragments that form dual components. Hence the probability of being active is simply estimated as

the ratio of bioactivity values between active and inactive pairs. This method was further extended to **SVMs** by modifying a Tanimoto kernel to include compound fragments. **PCM DC-SVMs** outperformed ligand based **SVMs (QSAR)** in internal validation, as accuracies of 90.90% and 86.20% were respectively obtained. However the same level of accuracy was not achieved when using external data sets, which produced accuracies of 73.90% and 81.30% for **DC-SVM** and ligand based **SVM**. Therefore, these results do not permit to conclude that **DC-SVM** outperform **SVM** although this might happen with other data sets.

A second type of **SVMs**, Transductive **SVMs (TSVMs)**, have been applied to model small (between 1,000 and 3,000 datapoints) and unbalanced **QSAR** data sets from the **Directory of Useful Decoys (DUD) HDAC** repository displaying a balanced accuracy higher than 30% on some data sets with respect to **SVM** [Kondratovich, Baskin, and Varnek (2013)]. The concept relies on transduction, allowing the modelling of partially labeled data which cannot be included using regular **SVM**. **TSVMs** could be potentially extended to **PCM** and have been shown to outperform **SVMs** in some cases [Collobert et al. (2006); Wang, Shen, and Pan (2007)].

A third flavor of **SVMs** are Relevance Vector Machines (**Relevance Vector Machine (RVM)**s) [Tipping (2001)]. The added value of **RVM** is the interpretability of the models, which is a consequence of their Bayesian nature. Each descriptor is associated to a coefficient, which determines its relevance for the model. Coefficients associated to low relevance descriptors are close to zero, hence the model becomes sparse and therefore permits shorter prediction times. Although the predicted variance is not informative in regression studies, class probabilities can be efficiently determined in classification [Lowe et al. (2011)]. **RVMs** have been demonstrated by binary classifiers trained on a subset of the **MDL Drug Data Report (MDDR)** database [Lowe et al. (ibid.)]. Therein, it was demonstrated that **RVMs** performed on par with **SVM**, encouraging the authors to conclude that **RVM** should be added to the current chemoinformatic tools and as such potentially applied to future **PCM** studies.

On the basis of the above, **SVM** constitutes a useful algorithm in which initial drawbacks such as interpretability (*e.g.* the determination of which chemical sub-structures most contribute to compound bioactivity) can be overcome with new developments (*e.g.* **RVM**).

.1.2.2 Random Forests (RF)

Random Forest (**RF**) models are often comparable in performance to **SVMs** [Westen et al. (2011a)], and are also non-linear. However, contrary to **SVMs** **RFs** tend to have relatively short training times and do not require extensive parameter tuning

[Svetnik et al. (2003)]. Furthermore, in addition to their comparable performance, RFs permit an evaluation of both feature contribution and feature importance in PCM models, as shown by Cortes-Ciriano et al. (2015); De Bruyn et al. (2013). An example of such evaluation is given in the identification of organic anion-transporting polypeptide (Organic Anion-Transporting Polypeptide (OATP)) inhibitors, where continuous descriptors, both z-scales (proteins) and physiochemical features (compounds), were binned into discrete classes. For each feature (protein and ligand) the correlation to activity and importance was calculated for each target class. In that way, compound inactivity was correlated with the presence of chemical substructures positively charged at pH 7.4, number of atoms < 20, and molecular weight < 300. Conversely, chemical substructures with a number of ring bonds between 18 and 32, without atoms with positive charge, and with a log D value between 3.4 and 7.5 were found to favour OATP inhibition.

Although RFs have a high interpretability it should be noted that they do not output error estimates (as is also the case with SVM), although recent papers suggest the usefulness of the variance along the trees of a random forest model to determine its applicability domain. Error estimates are of tremendous importance given the high levels of noise and error annotations in public bioactivity databases. Thus, fully informative predictions should be accompanied by individual uncertainties. This issue can be remediated by applying Quantile Regression Forests (QRF) which infer quantiles from the conditional distribution of the response variable [Meinshausen (2006)]. To our knowledge, QRFs have not been applied to QSAR or PCM yet. Gaussian Processes are a machine learning technique that has been used in PCM with inherent error estimation capabilities, as described below.

.1.2.3 Gaussian Processes (GP)

The determination of the applicability domain (AD) of a model (when are model predictions reliable or when can a model extrapolate) is one of the major concerns in bioactivity modelling (see previous studies [Bosnić and Kononenko (2009); Netzeva et al. (2005); Tetko et al. (2006)] for comprehensive reviews). Major obstacles to the AD determination are the errors and uncertainties contained in bioactivity databases [Kalliokoski et al. (2013); Kramer and L (2012); Kramer et al. (2012); Tiikkainen et al. (2013)], which are mainly due to data curation and experimental errors [Kramer et al. (2012)], as well as the accurate quantification of distances in the descriptor and the biological space, which would enable to anticipate prediction errors. Gaussian Processes (GP) aim to address these concerns by permitting to handle data uncertainty as input into a probabilistic model.

Figure .1.2 illustrates the basic idea underlying GP modelling. The prior prob-

ability distribution (Figure .1.2A) covers all possible functions candidate to model the data, each of which has a different weight determined by the kernel (covariance) parameters. Subsequently, only those functions from the prior distribution in agreement with the experimental data are kept (Figure .1.2B). The mean of these functions is considered as the best fit to the data. Given that each prediction is a Gaussian distribution, different confidence intervals can be defined from its variance (Figure .1.2B).

Gao et al. (2013) showed that SVMs performed, in general, slightly better than GPs when modelling a data set composed of 128 ligand and 9 human aminergic GPCRs, although the models trained on the best combination of descriptors exhibited equal Q^2 values of 0.74 for GP and SVM. Worth of mention, the difference in performance between GP and SVM was not assessed neither statistically nor by comparing the results of a series models trained on different resamples of the whole data set. Moreover, the predicted error bars by the GP PCM models were not considered. More recently, Cortes-Ciriano et al. (2014) showed the actual potential of GPs by applying both SVMs and GPs implemented with a panel of diverse kernels to multispecies PCM data sets, namely: human and rat adenosine receptors, mammal GPCRs and dengue virus proteases. GP and SVM performed comparably as absolute differences were statistically insignificant. However, GP provided notable added values via: (i) the determination of the model AD, (ii) the probabilistic nature of the predictions, and (iii) the inclusion of the experimental uncertainty in the model.

In the experience of the authors regarding the application of GP in PCM [Cortes-Ciriano et al. (ibid.)], and in agreement with [Schwaighofer et al. (2007)], the intervals of confidence (IC) calculated by GP are in accordance with the cumulative Gaussian distribution. Therefore, these intervals of confidence provide valuable information about individual prediction errors. In practice, knowing the error for each prediction can certainly guide decision-making about which compounds should be tested in prospective experimental validation of *in silico* PCM models. Overall, GP appear as an appealing approach for PCM in spite of the longer CPU time required for the training, as GP is an algorithm of $O(N^3)$ time complexity (*i.e.*, it scales with the third power of the size of the data set) [Rasmussen and Ws (2006)].

.1.2.4 Collaborative Filtering (CF)

One of the requirements for PCM is that target (protein) features need to be defined explicitly (usually by physicochemical characterization of amino acids). While this approach is effective, it nevertheless requires a certain level of information about target sequences and structures. An alternative approach would be to infer target features from an unsupervised approach and not use them as model input a priori. This was done quite recently in multi-target QSAR study for the hedhog signalling

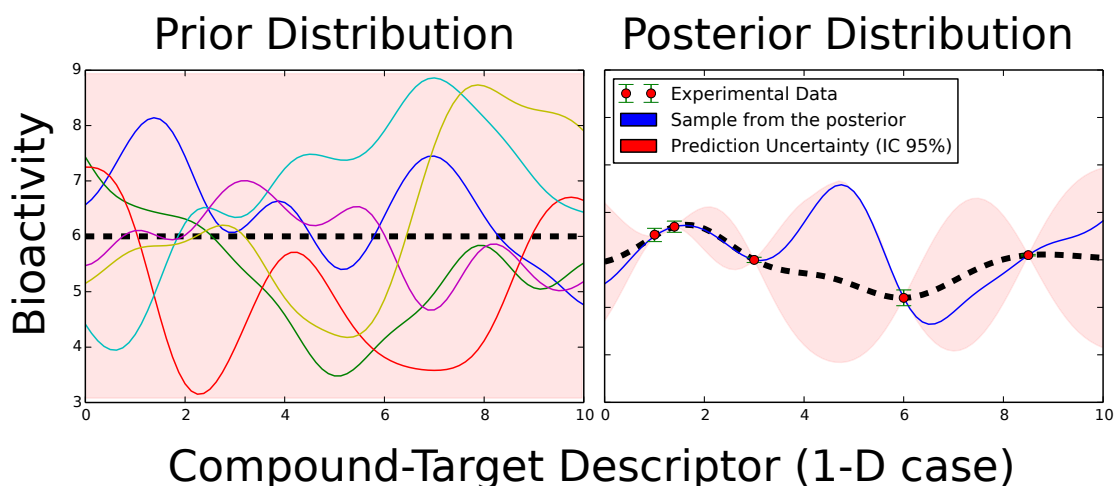


Figure 1.2: Illustrative example of GP theory in a two-dimensional problem. (A)

The prior probability distribution embraces all possible functions which can potentially model the data set. A subset of six prototypical functions is depicted. Normally, the mean of the distribution is set to zero (black dashed line). (B) The inclusion of bioactivity information (red dots) accompanied by its experimental uncertainty (blue error bars) updates the prior distribution into the posterior probability distribution. In the posterior probability distribution, only those functions in agreement with the experimental data are kept. The uncertainty (pink area) notably increases in those areas with little experimental information available. The mean of the posterior distribution (black dashed line) is considered the best fit to the data. A prototypical function from the posterior is shown in blue. For a new compound-target combination, the bioactivity is predicted as a Gaussian distribution, in which the mean is the best prediction and its variance the uncertainty. A radial-kernelled GP with $\sigma = 1$ was employed to generate the figure. The python *infp* package helped to produce the plots (John Reid, Version 0.4.13).

pathway across multiple cell-lines [Gao et al. (2012)].

Gao et al. (ibid.) incorporated a CF approach between 93 cyclopamine derivatives and four cell-lines (BxPC-3, NCI-H446, SW1990 and NCI-H157), and showed that Collaborative Filtering multi-target QSAR outperforms normal QSAR for their data set. The mean Root-Mean Squared Error (RMSE) for four cell-lines was 0.65 log units for CF while it increased to 0.85 log units for (single target) Support Vector Regression (SVR). The collaborative QSAR framework, combined with a feature selection methodology based on Collaborative Filtering and the content-based recommender systems (a system used by electronic retailers and content providers such as *amazon.com*) [Breese, Heckerman, and Kadie (1998)], enabled the definition of weights for the compound descriptors (drug-like index). When interpreting their models the authors could determine that molecular volume, polarity, and the cyclic degree are the most influential compound features for multi-cell-line inhibitors for this particular pathway. Erhan et al. (2006) also used CF with a large library of compounds against a family of 12 related targets screened in AstraZeneca's High-Throughput Screening (HTS) campaigns. The authors elegantly demonstrated how the principles of CF filtering can be used to derive a predictive model with the capability to extrapolate on the target side. However, better results were obtained when using target descriptors (binding pocket fingerprints of 14 bins in this case, where each bin accounts for a type of interaction -ionic, polar, or hydrophobic in the binding site). Another novelty of this work was the introduction of the kernel-based method Jrank (a kernel perceptron algorithm), which was able to outperform the multi-task neural network in most cases.

The overview presented above shows that PCM heavily draws on recent developments in the machine-learning field. In the following we will also summarize PCM applications in the medicinal chemistry and chemical biology fields, to different target classes as well as different types of biological readout.

.1.3 PCM Applied to Protein Target Families

As was touched upon above, PCM has been applied to a very diverse selection of protein targets. Here we will focus on a small selection of targets relevant for drug discovery, namely G Protein-Coupled Receptors (GPCRs), kinases, epigenetic markers, viral enzymes, and human cancer cell-lines.

.1.3.1 G protein-coupled receptors

Early PCM virtual screening studies by Bock and Gough (2005) to identify ligands of orphan GPCRs (oGPCRs) used physiochemical properties of the amino acids of the entire primary sequence of GPCRs, such as accessible surface area or surface tension,

rather than binding site residues. The authors screened 1.9 million ligand-oGPCRs combinations and were able to identify 4,357 highly active ligands of oGPCRs. The method, based on SVM, outputs a ranked list of putative oGPCRs ligands. In practice, the most relevant feature of their predictive pipeline is the description of GPCRs with only physicochemical descriptors, thus avoiding the usage of exact 3D information of the receptors [Bock and Gough (2005)]. Subsequently, Jacob et al. (2008) demonstrated that the usage of bioactivity data from 4,051 GPCR-ligand combinations (80 human GPCRs from classes A, B and C, and 2,446 ligands) extracted from the GLIDA GPCR ligand database [Okuno et al. (2006)] in PCM models improves the performance over single receptor models, leading to more reliable predictions. The authors used Tanimoto 2D and pharmacophore 3D kernels to describe the ligands, and kernels to describe the GPCRs, namely: Dirac, multitask, hierarchy, binding pocket and poly binding pocket. The best combination thereof was shown to be 2D Tanimoto for the compound side and the binding pocket kernel for the GPCRs, as the authors reported an accuracy of 78.1% when predicting ligands for orphan receptors.

These findings were further capitalized upon in the papers of Frimurer et al. (2005), and Weill and Rognan (2009). Both papers devised features for the 7Transmembrane (TM) core ligand-binding site and cavity fingerprints to improve the structure guided drug discovery approaches and provide a general class A GPCR similarity metric [Frimurer et al. (2005); Weill and Rognan (2009)]. The former approach introduced an *in silico* pipeline to relate 7TM GPCRs based upon the physicochemical properties of the ligand binding site, taken from the crystal structure of the bovine rhodopsin. The pipeline is composed of five steps, which are: (i) sequence alignment of the TM domain of the GPCRs of interest, (ii) selection of the residues in the core binding site important for ligand binding, (iii) definition of binding site signatures and generation of physicochemical descriptors for them, and (iv) use of these descriptors to rank, cluster or compare 7TM GPCRs. The authors applied this pipeline to identify ligands for the rhodopsin-like receptor, CRTH2, which by that time only had one annotated ligand besides prostaglandin D2, namely indomethacin. The screening of a library of 1.2 million compounds yielded 600 candidate hit compounds. 10% thereof were confirmed as ligands in a CRTH2 receptor-binding assay, using a IC₅₀ cut-off value of 10 mM to consider a compound as active.

On the other hand, Weill and Rognan (2009) introduced a new type of Protein-Ligand Fingerprints (PLFP), which encodes pharmacophoric properties of ligands and their binding cavities. These fingerprints were applied to two GPCRs data sets, namely: (i) 168,536 GPCR-ligand combinations (160,286 inactive and 8,250 active combinations), and (ii) 234,137 GPCR-ligand combinations (202,019 inactive and 32 118 active combinations). The total number of GPCRs considered was 160. The authors reported a cross-validated classification accuracy higher than 0.90 when using SVM, though the most predictive models on external data sets were not those

presenting the highest accuracy values in cross-validation [Kubinyi, Hamprecht, and Mietzner (1998)].

Overall, PCM models trained on GPCRs binding site amino acid descriptors have proven to be a powerful approach to identify the GPCRs targets for a given compound, and to predict ligands for orphan GPCRs. The increasing availability of bioactivity data on GPCRs of interest and orthologous sequences [Cortes-Ciriano et al. (2014)], as well as the development of novel methodologies to assess GPCRs similarity, is likely to increase the application of PCM on this target family.

.1.3.2 Kinases

Another important protein family in drug discovery subjected to PCM studies is the kinase superfamily which comprises more than 500 different human proteins [Manning et al. (2002)]. The role of kinases in cell signalling and their involvement in more than 400 human diseases have rendered this protein family an attractive target [Cohen (2002); Melnikova and Golden (2004)]. Each kinase generally contains a conserved kinase 1 domain that binds ATP in its active site, though some kinases contain more than one kinase domain. Inhibitors targeting this conserved binding site are known as Type I inhibitors. The activation loop of kinases, necessary for the transfer of a phosphate group, exhibits two different conformations, namely DFG-in and DFG-out (where DFG stands for the catalytic triad, Asp-Phe-Gly). Type II inhibitors bind to both the conserved ATP-binding site and to an adjacent pocket present in the DFG-out conformation. These compounds are more selective and thus attractive as drug candidates. Given the ability of PCM to model bioactivities against related targets, it is very well suited to model the affinity of small molecule inhibitors to the kinase family [Westen et al. (2011a)]. Different PCM models have been reported to analyze drug selectivity and predict bioactivity profiles against 15 kinases [Cao et al. (2013b); Subramanian et al. (2013)].

In a recent study by Cao et al. (2013b), the full kinase sequence space was described by alignment-independent -Composition, Transition and Distribution (CTD) features, 127 along with topological features of compounds. The data set comprised a total number of kinase-compound interactions of 54,012, with data from 22,229 compounds and 372 kinases. The best RF model exhibited a classification accuracy in five-fold cross-validation of 93.7%, and a sensitivity of 92.26%. Moreover, this high predictive power was maintained in the four validation levels suggested by [Park and Marcotte (2012)], as the following accuracies and sensitivities (respectively and in percentage units) were obtained: (i) L1: 93.15 and 91.23; (ii) L2: 89.53 and 88.24; (iii) L3: 90.71 and 89.48; and (iv) 87.30 and 85.82. Hence the statistical soundness of this PCM model enabled the classification of compound-kinase pairs as interacting, using

a 100 nM compound concentration as cut-off, or non-interacting. The high predictive ability of the models should be considered nevertheless with caution as the degree of completeness of the bioactivity matrix used in the training was only 0.65%. Therefore, to improve the predictive power of these PCM models they should be iteratively updated as more bioactivity values become available. Interestingly, kinases similar in the sequence space exhibited high dissimilarity when assessing their similarity with the inhibitors bioactivities. This was assessed using 120 kinases with more than 15 bioactivity annotations, 14,400 datapoints in total. Thus, these data highlights the adequacy of considering chemical and target space to optimize kinase inhibitors.

While high affinity is generally desired for drugs (except possibly in case of multicomponent therapeutics) [Borisy et al. (2003)], selectivity is equally important when targeting a protein family with highly similar binding sites, such as in this case kinases. Subramanian et al. (2013) applied PCM models to a kinase data set comprising 50 different proteins in the DFG-in conformation to better understand both the residue and compound features which determined whether the ATP-binding site of kinases are involved in compound binding. The resulting PCM models, using PLS including cross-terms (see Section 2.3), demonstrated the added value of PCM over ligand based approaches, as statistically satisfactory QSAR models were reported for only 44% of the targets. More importantly, the models could be visually interpreted, thus enhancing the practical usefulness of PCM for the optimization of compound selectivity. Further details on this study are given in Section 4.4, as models targets were encoded with 3D information.

As shown by these recent PCM studies on the kinase superfamily, PCM can support new concepts for kinase inhibition implicating the simultaneous interaction of kinase inhibitors with several targets leading to multi-target kinase chemotherapy [Gujral, Peshkin, and Kirschner (2014)].

.1.3.3 Histone modification and DNA methylation

Epigenetic markers have been identified as emerging therapeutic targets in various malignancies and diseases by correlating phenotypes and differential expression patterns [Prinjha, Witherington, and Lee (2012)]. Key protein families involved in these processes are readers (bromodomains), writers (DNA modifying enzymes, histone 1 acetylases, methyltransferases) and erasers (histone deacetylases) [Knapp and Weinmann (2013)]. Most of the bromodomain epigenetic targets have the ability to selectively modulate the gene expression pattern and contribute to post-translational modifications, chromatin binding, inflammation, oncogenesis [Prinjha, Witherington, and Lee (2012)], moreover there is a clear linkage to some diseases, *e.g.* multiple myeloma [Delmore et al. (2011); Floyd et al. (2013)]. Vidler et al. (2012) studied

the druggability of the different members of the bromodomain family focusing on amino acid signatures in the bromodomain acetyl-lysine binding site, which resulted in a bromodomain family classification more correlated with the binding of small molecules in comparison with a whole-sequence similarity classification. Numerous successful chemical probes like JQ1 have also been identified as bromodomain inhibitors by the Structural Genomics Consortium (SGC) [Gruetter (2012)]. However, the bromodomain family still has unexplored therapeutic potential. To date there are no PCM studies performed on this family.

Recently, Wu et al. (2012) utilized structural similarity between three classes of HDACs and generated a predictive model for a novel candidate anti-tumour drug. They implemented various descriptors (physicochemical properties) and similarity descriptors (sequence and structure) of compounds and targets in the PCM model and successfully identified class-selective inhibitors for class-I and class-II HDACs. The best model exhibited high predictive ability, as the authors reported a Q^2 value on the external set of 0.75. Overall, the increasing importance of epigenetic targets in drug discovery as well as the availability of large-scale resources of epigenetic targets and its modulators [Arrowsmith et al. (2012); Huang et al. (2012b)], will facilitate the application of PCM to this target family.

.1.3.4 Viral mutants

Previous sections highlighted the ability of PCM to model bioactivities of several human protein superfamilies, yet PCM based approaches are not bound to human protein targets. PCM has also been applied in a number of studies to predict activity profiles of ligands against different viral protein variants [Westen et al. (2011b)]. In the field of HIV, Westen et al. (ibid.) used 451 compounds tested against 14 HIV reverse transcriptase sequences to train a model that was able to predict the bioactivity of 317 new compound-mutant pairs. Interestingly, when the prediction was validated prospectively with *wet lab* experiments it was found that the prediction error (RMSE of 0.62 log units) was comparable to experimental uncertainty of the assay (0.50 log units). In a similar setting, Huang et al. (2012b) showed that the inclusion of Protein-Ligand Interaction Fingerprints (PLIFs) of viral residues and ligand structures as cross-terms improved model predictive power over models lacking them. PCM models were trained on 92 compounds and 47 HIV-1 protease variants with about 160 Ki values. The best PCM model exhibited a Q^2 value of 0.83 on the external set.

Next to these applications, PCM has been used to model the sensitivity of viral mutants to antiretroviral drugs, which could potentially guide HIV treatment [Westen et al. (2013b)]. Resistance testing and prediction using these models is achieved by incorporating genotypic (protein) and drug (chemical) data and subsequently

linking them to phenotypic data (resistance). PCM then allows the prediction of optimal treatment regimens. The advantage of PCM over established sequence-based approaches is that interpretation of a single model allows the combined elucidation of residues responsible for the change in efficacy and the complementary chemical features affected [Doherty et al. (2011); Junaid et al. (2010); Kontijevskis et al. (2009); Lapinsh et al. (2008)].

For instance, Westen et al. (2013b) trained PCM models based on a large clinical data set composed of circa 300,000 datapoints combining both phenotypic and genotypic data. The application of PCM enabled the integration of the similarity of marketed drugs together with protein sequence similarity. The best model exhibited a fold change error of 0.76 log units, which constitutes an improvement of 0.15 log units with respect to previously reported models trained on only protein sequence similarity (0.91 log fold change error). In addition, the authors identified novel mutations of both HIV reverse transcriptase and HIV protease conferring drug resistance, underlining the ability of PCM models not only to model bioactivity information, but to also learn about features relevant for activity from both the ligand and the protein target side.

Similarly, drug susceptibility profiles were predicted based on PCM. In that way, two models have been reported for the prediction of: (i) the susceptibility (bioactivity profile) of a given HIV protease genotype to seven commonly used protease inhibitors [Lapinsh et al. (2008)]; and (ii) the susceptibility of HIV reverse transcriptase to eight nucleoside/nucleotide reverse transcriptase inhibitors [Junaid et al. (2010)]. PCM models were trained on 4,792 HIV protease-inhibitor combinations, obtaining a Q^2 value on the external set for the best model of 0.87. These models have been made publically available via web-services available at <http://www.hivdrc.org/services>, allowing free use of these algorithms [Spjuth et al. (2011)].

While the ligands of most PCM studies discussed here were small molecules, protease peptide substrates are also amenable to PCM. This has been demonstrated recently by Prusis et al. (2008, 2013) to study the enzyme kinetics parameters for designed small peptide substrates on four dengue virus NS3 proteases using PCM modelling. It was found that the PCM models for K_m and K_{cat} were significantly different. Therefore, by optimizing peptide amino acid properties important for K_m activity it was possible to improve peptide affinity to protease, while preventing the catalytic activity of the proteases on the peptides.

These studies [Prusis et al. (2008, 2013); Westen et al. (2013b)] are some of the few reports in which predictions have been validated prospectively, demonstrating the predictive power of PCM in different scenarios.

.1.4 Novel Techniques and Applications in PCM

.1.4.1 Novel target similarity measure

In the context of GPCRs, studies developing better similarity metrics have helped to determine key binding residues within the GPCR trans-membrane (TM) helical bundle [Frimurer et al. (2005); Gloriam et al. (2009); Surgand et al. (2006)], have aided intra family similarity determination using cavity fingerprints [Andersson, Chen, and Linusson (2010)], and have fueled high-throughput homology models that supported cavity detection programs [Andersson, Chen, and Linusson (2010); Glinca and Klebe (2013); Liu et al. (2008); Weill and Rognan (2009)]. PCM approaches including these features have also helped in off-target predictions, and in target prediction for GPCR-focused combinatorial chemolibraries [Reutlinger et al. (2014); Weill (2011)].

The binding site focused techniques used allowed for the identification of orthosteric and allosteric sites on the same target for different ligand families. In this line, Gao et al. (2013) obtained the higher predictive ability with models trained on trans-membrane identity descriptors ($Q^2 = 0.74$) over z-scales ($Q^2 = 0.72$) when modelling the inhibition constant of 9 human aminergic GPCRs and 128 ligands, (310 ligand-target combinations). Similarly, Shiraishi et al. (2013) revealed specific chemical substructures binding to relevant TM pocket residues, which it is not only relevant to mutational analysis but also serves as a complementary approach to Structure-Based Drug Discovery (SBDD) [Shiraishi et al. (2013); Yabuuchi et al. (2011)]. TM identity descriptors and TM kernels behave more discriminatively than z-scales for GPCRs and allow identification and interpretation of GPCR residues associated with binding of ligands (of a particular chemotype). Therefore, the identification of chemical moieties and residues involved in ligand binding enables the development and optimization of GPCR inhibitors with respect to both potency and selectivity.

.1.4.2 Including 3D information of protein targets in PCM

The binding of a ligand to a protein is a complex process, governed on the structural level by the 3D composition of the protein binding site, the 3D conformation of the ligands approaching, and the complementarity of their pharmacophoric features. Hence it is expected that inclusion of spatial information from the protein binding sites would improve the predictive power of PCM. Unfortunately, this approach is frequently limited by the lack of high quality 3D structures, poor understanding of ligand-induced conformational changes, and inaccurate superimposition of protein structures. The latter can be (partly) overcome by the use of alignment-free protein descriptors [Andersson, Gustafsson, and Strömbergsson (2011); Weill and Rognan (2009)], but usually at the cost of lower resolution, loss of target-related information and poor interpretability.

Jacob et al. (2008) found no improvement through the use of 3D information. In this study an analysis of 2,446 ligands interacting with 80 human GPCRs was performed using a linear vector representing conserved amino acids in the binding pockets. While the binding pocket kernel implicitly encodes 3D information, the spatial arrangements were derived from the comparison to only two template proteins. Overall, the 3D kernels (77% prediction accuracy) did not show improvements compared to lower dimensional protein descriptions (77% prediction accuracy with a protein similarity kernel). Likewise Wassermann, Geppert, and Bajorath (2009) found little improvement using 3D information in their analysis of interactions of 12 proteases with 1,359 ligands using the TopMatch similarity score [Sippl and Wiederstein (2008)], which used all amino acids within 8 Å around the catalytic residues to describe the target proteins. This 3D description did not perform better (61% recovery rate) than the sequence (57%) and protein class-based (62%) kernels used in this publication.

Conversely, early work by Strömbergsson et al. (2006) used local protein substructures, encoded as motifs of amino acid stretches, which are closer than 6.5 Å to each other. This local substructure method showed for a set of 104 enzymes an improvement over the use of global SCOP (Structural Classification of Proteins) folds and the RMSE values on the external validation set decreased from 2.06 to 1.44 pKi units. Additionally, it was found that local substructures close to the ligand binding sites were assigned more importance in the models than more distant ones, which is intuitively understandable. Similarly, Meslamani and Rognan (2011) did find an improvement by using 3D information. 581 diverse proteins were described by the 3D cavity descriptor FuzCav [Weill and Rognan (2010)], which is a vector of 4,834 integers reporting counts of pharmacophoric feature triplets mapped to C α -atoms of residues lining the binding site. The use of cavity 3D kernels showed a clear advantage (F-measure 0.66) over sequence-based descriptions (F-measure 0.54) in predicting target-ligand pairings for a large external test set (>14,000 ligands, 531 targets), especially in local models. This difference seems to be even more pronounced for data sets with limited ligand data (<50 ligands). Likewise, a recent study by [Subramanian et al. (2013)] described the superimposed binding sites of 50 (unique) kinases by molecular interaction fields derived from knowledge-based potentials and Schrodinger's Water-Maps [Hoppe, Steinbeck, and Wohlfahrt (2006); Robinson, SH, and Farid (2010)]. Also in this example a significant improvement for 3D methods ($r^2 = 0.66$, $q^2 = 0.44$) compared to sequence-based methods ($r^2 = 0.50$, $q^2 = 0.34$) was reported. Additionally, this combination of methods allows interpretation and easy visualization of PCM results within the context of ligands and binding pockets.

Earlier studies have not clearly shown the advantages of 3D PCM over solely sequence-based approaches, whereas more recent studies show that including 3D

information appears to improve performance. The particular data set used (e.g. number of ligands), and the quality of the data provided, likely determines if there is a possible gain in this type of description. However, the constantly increasing number of protein structures, more robust alignment-free methods (e.g. Nisius and Gohlke (2012) and Andersson, Gustafsson, and Strömbergsson (2011)), and introduction of protein descriptors with easier interpretability (e.g. Desaphy et al. (2013)), might help the interpretation and the visualization of PCM models in the future.

.1.4.3 PCM in predicting ligand binding free energy

The application of PCM to docking might not be directly obvious. Yet, the concepts used in PCM, quantitatively relating ligand and protein-side descriptors to affinity/activity, very much resemble empirical scoring functions. Molecular docking has led to the discovery of active compounds [Laine et al. (2010)], yet it suffers from several well described limitations, among which is the relatively low performance in prediction of interaction energies [Yuriev, Agostino, and Ramsland (2011); Yuriev and Ramsland (2013)]. In contrast, PCM models can predict the difference in Gibbs free energy ($\Delta G = -RT \ln K_d$) between the initial state, where the protein and the compound do not interact, and the final ligand-target complex. Therefore, the principles of PCM can be applied to develop PCM-based scoring functions.

Kramer and Gedeck (2011a) demonstrate this concept by building a structure-based PCM scoring function. Their method inducts a bagged stepwise multiple linear regression model with a subset of 1,387 protein-ligand complexes extracted from the PDBbind9-CN database [Wang et al. (2004)]. Subsequently a new compound-target interaction descriptor based upon distance-binned Crippen-like atom type pairs was introduced. The best model outperformed commercially available scoring functions assessed on the PDBbind9 database and was able to explain 48% of the variance of the external set, providing a RMSE equal to 1.44 K_d units. Although similar methods had been previously proposed [Artemenko (2008); Das, Krein, and Breneman (2010); Deng, Brenema, and Embrechts (2004); Sottriffer et al. (2008); Zhang, Golbraikh, and Tropsha (2006)], this was the first study where a sufficiently large validation was accomplished to ascertain the model's predictive power. Additionally, the implementation of bagged stepwise multiple linear regression (MLR) and PLS enabled the evaluation of the importance of ligand and target descriptors for the PCM model.

Similarly, a subsequent study reported the development of a scoring function based upon the CSAR-NRC HiQ benchmark data set (<http://csardock.org>) [Kramer and Gedeck (2011b)]. The best model exhibited acceptable statistics with a cross-validated $R^2 = 0.55$ and RMSE = 1.49 [Kramer and Gedeck (ibid.)]. Finally, Koppisetty et al. (2013) were able to predict for the first time ligand binding free energies where

the enthalpic and entropic contributions for a given binding event were deconvoluted. Therein, the authors demonstrated the importance of including ligand descriptors (QIKPROP and LIGPARSE calculated in Schrodinger suite [Schrödinger (2013)]) to the models in addition to 3D ligand-protein interaction descriptors.

As demonstrated above, PCM overlaps with methods that are originally coming from the structure-based field due to PCM describing in principle any method to relate ligand features and protein/target features on a large scale to an output variable of interest. Another source of complementary information is the information from divergent and convergent homologous sequences. This allows PCM models to extrapolate the bioactivity of ligands to the same protein target in different species as shown below.

.1.4.4 PCM as an approach to extrapolate bioactivity data between species

Given that PCM considers bioactivity data from related targets, these related targets can also include similar targets from different species. Given a group of related targets, a distinction can be made from an evolutionary standpoint between gene pairs originated from intra-species gene duplication events (paralogy, within species) or from speciation events (orthology, across species) [Koonin (2005)]. Since orthologous genes will tend to maintain the original function, binding modes will also tend to be more conserved than in paralogues, where the original protein function is less conserved.

This has also been shown to be true for affinities of ligands binding to these orthologues by analyzing bioactivity data in a recent study by Kruger and Overington (2012). The authors demonstrate that the same small molecule exhibits similar binding affinities when acting on orthologues (though some exceptions were found, e.g. Histamine H₃ receptor). Moreover, the authors verified that larger differences in binding affinity are observed for paralogues with respect to orthologues by analyzing the differences in binding for a total number of 20,309 compounds on 516 human targets, with 651 being the final number of orthologous pairs. These observations aid in optimizing ligands for their interaction with conserved residues across a given protein family, thus making them more desirable lead compounds (thus avoiding their interaction with unrelated targets) [Lounkine et al. (2012)].

In the field of PCM, Lapinsh et al. (2002) demonstrated for the first time the capability of PCM to successfully combine the pK_i values of 23 organic compounds on 17 human (paralogues) and 4 rat (orthologues) aminergic GPCRs. The authors were able to deconvolute the binding site interactions into two types, namely: those

involved in specificity and those involved in affinity. Therefore, compound design can be envisioned from the viewpoint of affinity or specificity. Similarly, the contribution to compound affinity of TM regions involved in the interactions of aminergic GPCRs and compounds was also quantified. For example, TM regions 2, 3, 4, 6 and 7 are responsible for low overall affinity in β_2 receptors; however, the same regions are positive contributors to overall high affinity in α_1 receptors. Westen et al. (2012) built on this by including in a PCM model bioactivity data from four human and rat adenosine receptors (A_1 , A_{2A} , A_{2B} and A_3). The authors screened a commercial chemolibrary composed of 791,162 compounds with the most predictive PCM model obtained, which exhibited Q^2 and RMSE values of 0.73 and 0.61 pKi units, respectively. Prospective experimental validation led to the discovery of new high-affinity inhibitors, among which a compound with a pKi value of 8.1 on the A_1 receptor.

Finally (chapter .5), the authors have applied PCM to model the pIC_{50} value of 3,228 distinct compounds on 11 mammalian cyclooxygenases (COX) using ensemble PCM [Cortes-Ciriano et al. (2015)]. The final ensemble PCM model, trained on the cross-validation predictions of a panel of 282 RF, SVM and Gradient Boosting Machine (Gradient Boosting Machine (GBM)) models, each one trained with different values of the hyperparameters, led to predictions on the test set with RMSE and R_0^2 values of 0.71 and 0.65, respectively. Additionally, the description of compounds with unhashed Morgan fingerprints permitted a chemically meaningful model interpretation, which highlighted chemical moieties responsible for selectivity towards COX-2 in agreement with the literature [Cortes-Ciriano et al. (ibid.)].

The ability of PCM to embrace multispecies information using only sequence descriptors allows the creation of models capable to predict compound activity on targets with little available data points on the human orthologue. The existing large body of bioactivity data collected on organisms other than human (*e.g.* rat and mouse) provides a good resource. This data was derived from the traditional usage of rodent tissues as a source of proteins for biochemical and pharmacological assays. Moreover, the difference in bioactivity between a compound acting on its human target with respect to its orthologue in another species (*e.g.* the CCR1 antagonist BX471) hampers the utilization of animal models to study human diseases at a molecular level [Horuk (2009)]. Thus, PCM can help not only to reduce the number of experiments required to complete the compound-target interaction matrix [Menden et al. (2013)], but also appears as a practical tool to understand complex diseases in scenarios where current experimental settings are insufficient (*e.g.* undeveloped enzymatic assays for a given protein). Similarly, PCM might be applied as a supporting tool in allometric scaling to predict the behavior of clinical candidate drugs in humans [Kagan et al. (2010); Zhang, Surapaneni, and Guan (2012)]. Nonetheless, the extrapolation capabilities of PCM models are subjected to the completeness of the bioactivity matrix (.1.1). In practice, even though high performance can be attained with a matrix completeness

level below 3%, the variability of the chemical space plays a key role in determining the extrapolation capability of a PCM model on the chemical side [Cortes-Ciriano et al. (2014)]. Therefore, a balance has to be found between the coverage of chemical and target space, and the degree of completeness of the bioactivity matrix.

.1.4.5 PCM applied to pharmacogenomics and toxicogenomics data

The biological space in a PCM model can be further extended from single proteins to whole cell-lines. A step forward in this regard is the inclusion of cell-line descriptors in a PCM model in order to model cell-line sensitivity to cancer drugs or toxic compounds. Given that individual cell-lines have been shown to demonstrate diverse profiles with respect to drug sensitivity, the variability on the cell-line side, which constitutes now the target side of PCM, can be exploited to concomitantly predict both drug potency and cell-line selectivity [Cortes-Ciriano, I et al. (2015)]. Additionally, PCM can also facilitate the interpretation of differential gene expression or mechanism of toxicity of compounds [Lapinsh et al. (2013)], as will be shown below.

The availability of pharmacogenomics and toxicogenomics data has enabled predictive modelling of cancer cell-line sensitivity. These models consider as the dependent variable the response of a whole cell to a given drug, such as in the form of EC₅₀ values, which determines the concentration at which a compound exerts half of its maximal effect. Therefore, the *target* component in the PCM model is no longer a single protein, described in terms of binding site properties, but by more complex (usually genomic) features such as oncogene mutations, cell karyotypes or gene expression levels.

In the context of human cell-lines, the work on the United States National Cancer Institute (NCI)-60 cell-line panel, which covers cells from 9 different cancer types, has helped to find novel molecular determinants of drugs sensitivity, as well as to develop drugs targeting concrete tumor types (disease-oriented); *e.g.* 9-Cl-2-methylellipticinium acetate for central nervous system tumours Shoemaker 2006. However, the number of cancer cell-lines with drug sensitivity data has vastly increased with the release in 2012 of two major cancer cell-line panels, namely: the CCLE consisting of 947 cancer cell-lines and the Genomics of Drug Sensitivity in Cancer (GDSC) consisting of 727 cancer cell-lines [Basu et al. (2013)]. The setup of both cell-line collections, sharing a total number of 471 cell-lines, enabled large scale pharmacological profiling thereof. In that way, Barretina et al. (2012) measured the chemotherapeutic effect of 24 drugs on the CCLE panel, while Garnett et al. (2012), tested 130 chemical compounds on the GDSC cell-line collection. In both cases, the cell-lines were further characterized genomically, by measuring gene expression data, chromosomal copy numbers, oncogene mutations, and microsatellite instability.

Recently, Basu et al. (2013) measured the sensitivity of 242 cell-lines from the CCLE panel to an Informer Set composed of 354 diverse molecules, including 54 clinical candidates and 35 United States Food and Drug Administration Agency (FDA)-approved drugs. The sensitivity data is publicly available at the Cancer Therapeutics Response Portal (CTRP, <http://www.broadinstitute.org/ctrp>).

The availability of public bioactivity profiles for compounds in combination with detailed genetic information of the cell-lines constitutes a scenario where machine learning can be applied for predictive cell-line sensitivity modelling. In this area, Menden et al. (2013) exploited cell-line drug sensitivity information from the GDSC and incorporated genomic features in combination with chemical descriptors in non parametric models, *i.e.* neural networks and Random Forests. These models allowed the authors to determine the missing drug response (IC_{50}) values in the original cell-line compound matrix. The best model predicted the sensitivity on the external (blind) test with a correlation between observed and predicted of 0.64, while a value of 0.61 was obtained when predicting the response on a tissue unseen by the model in the training phase. Recently, the authors have integrated PCM random forest models with conformal prediction for the large-scale prediction of cancer cell line sensitivity with error bars [Cortes-Ciriano, I et al. (2015); Norinder et al. (2014)]. Compounds were described with Morgan fingerprints, whereas a total of 16 cell-line profiling data sets were benchmarked for their predictive signal. Gene expression data constantly led to the highest predictive power. Interestingly, the authors found statistically significant differences in predictive power between PCM models trained on cell-line identity fingerprints (inductive transfer knowledge between cell-lines [Brown et al. (2014)]) and cell-line profiling data, suggesting that the explicit inclusion of cell-line information improves the prediction of cell-line sensitivity. Of practical relevance, the predicted bioactivities enabled the prediction of growth inhibition patterns on the NCI60 panel and the identification of genomic markers of drug sensitivity.

The cancer cell-line collections described above still remain to be fully exploited. While they constitute a great opportunity for PCM to integrate both drug sensitivity and genomics data in single models, this data integration still remains challenging due to the disagreement of drug sensitivity measurements between the CCLE and the GDSC [Haibe-Kains et al. (2013); Weinstein and Lorenzi (2013)]. Overall, the principles of PCM, namely the combination of chemical and cell-line (target) information in single machine learning models, are suited to integrate and exploit the increasing availability of drug sensitivity measurements on cancer cell-line panels. The application of PCM in pharmacogenomics is a recent sub-field of which the authors are certain it will grow in the near future. Moreover, *in silico* drug sensitivity prediction is a cost-efficient method capable to relate large-scale pharmacogenomics data, which is likely to foster the identification of chemotherapeutic lead compounds in both the academic and pharmaceutical cancer drug discovery pipeline.

.1.4.6 Other potential PCM applications

As reviewed above PCM has been applied in a wide range of drug discovery settings, yet more applications remain unexplored. The prediction of compound toxicity on cell-lines (toxicogenomics) [Heijne et al. (2005); McHale et al. (2010); Suter, Babiss, and Wheeldon (2004)], beyond the aforesaid cancer cell-line collections, is also amenable to PCM. Recently, Kaggle (<https://www.kaggle.com/competitions>), a crowd-sourcing platform, hosted two competitions in the field of chemoinformatic modelling. Two pharmaceutical companies, Boehringer Ingelheim and Merck, provided structure-activity relationship data sets to the community in order to find the most predictive machine learning algorithms. The Merck challenge consisted of 15 data sets, each of which containing the bioactivities of a series of molecules on a different target. The winners of the competition applied restricted Boltzmann machines (deep learning) [Hinton, Osindero, and Teh (2006)]. Interestingly, the winning team noted that the similarity between the data sets (targets) could be exploited by inducing a single neural network with all data sets, which output a layer with fifteen different units (neurons). On the other hand, Boehringer Ingelheim provided a data set with 1,776 compound descriptors. The response variable was binary, a value of 0 corresponded to a compound not eliciting the expected activity whereas a value of 1 corresponded to a compound showing activity. In this case, the highest predictive ability was obtained with model ensembles (Random Forests, Gradient Boosting Machines, and k-Nearest Neighbors). In a similar vein, the modelling challenge DREAM8 was proposed to the scientific community to model the toxicity of 106 compounds on 884 lymphoblastoid cell-lines, which were characterized by SNP genotypes and gene transcript levels quantified by RNA sequencing [*DREAM: Dialogue on Reverse Engineering Assessment and Methods project*; Norman et al. (2011); Stolovitzky, Monroe, and Califano (2007)].

As described in the previous sections, a large variety of protein targets have been modelled using PCM. Beyond the modelling of the activity of compounds on targets of diverse nature, the interaction between nucleic acids and proteins is also amenable to PCM modelling. In this context, Bellucci et al. (2011) predicted protein-RNA interaction based upon the physicochemical properties of both the polypeptide and the nucleotide chains. However, to best of our knowledge, few studies have been published in this area.

.1.5 PCM Limitations

The usefulness of PCM in computational drug design has been extensively proven *in silico* and in prospective experimental validation. Nevertheless, there are a number of limitations that should not be overlooked. Publicly available bioactivity databases contain a non-negligible degree of experimental uncertainty [Kalliokoski et al. (2013);

Kramer and L (2012); Kramer et al. (2012); Tiikkainen et al. (2013)], which should be certainly included in the modelling phase, as recently proposed by [Cortes-Ciriano et al. (2014)]. Similarly, intervals of confidence for individual predictions should be reported, which can be calculated with algorithm-dependent approaches, *e.g.* Gaussian Processes [Cortes-Ciriano, I et al. (2015)], or with algorithm-independent techniques, such as conformal prediction [Norinder et al. (2014)].

In addition to being informative for biologists, these confidence intervals constitute a valuable source of information about the applicability domain (AD) of a given model [Cortes-Ciriano et al. (2014)]. The AD is defined as the amount of ligand and target space to which a given model can be reliably applied. Thus, in addition to the model validation schemes presented above, an estimation of model AD should accompany any reported model in order to be of practical usefulness.

Another limitation which is often inherently related to bioactivity data is that of data skewness. Some data sets mostly report active [Morgan, Falcon, and Gentleman (*GSEABase: Gene set enrichment data structures and methods*)] or inactive molecules [Li, Wang, and Bryant (2009)], and thus compound-target combinations untested experimentally are normally considered as inactive or active interactions, respectively. Moreover, public data in general tend to favor a relatively small number of proteins classes that have been extensively explored (*e.g.* GPCRs and kinases). As such, for some targets the available data might not be sufficient for PCM projects given that imbalanced data sets can lead to models with high negative or false positive rates. Nevertheless, the modelling of cell-line sensitivity has shown that PCM displays high interpolation power, as the accuracy of prediction reached a plateau when 20% of the whole compound-cell-line matrix was included in the training set [Menden et al. (2013)].

Beyond the quality of the data, the descriptor choice still constitutes a field of active research, specially with respect to protein descriptors, which development will deeply influence the success of PCM in the coming years. A recent paper by Brown et al. (2014) suggested that PCM mostly relies on inductive transfer knowledge and that protein descriptors mostly act as labels and do not account for structural differences among them. However, we have recently shown that both amino acid descriptors and cell-line profiling data sets account for structural information of eukaryotic, mammal and bacterial DHFR, and cancer cell-lines, where the difference in performance on the test set between inductive transfer and PCM models was statistically significant [Cortes-Ciriano, I et al. (2015); Paricharak et al. (2015)].

PCM requires the concatenation of ligand and target descriptors, and sometimes also cross-terms, which substantially increases the dimensionality of the input space with respect to QSAR. Although this higher dimensionality might lead to overfitting

in PCM [Hawkins et al. (2006)], in practice, PCM has been shown to exhibit higher predictive power on the test set than QSAR [Cortes-Ciriano et al. (2014); Westen et al. (2012)].

.1.6 Conclusions

PCM is becoming a mature technique that allows the simultaneous use of both the chemical and the biological spaces in predictive bioactivity modelling. Both retrospective validation and prospective validation have underscored the advantages of PCM over ligand-based methods. However, it is the extensive expertise developed in the fields of QSAR and chemoinformatics on which PCM can build. Nowadays, a wide choice of properly benchmarked ligand and protein descriptors is available as well as different linear and nonlinear modelling algorithms. Nonetheless, conceptually diverse machine learning algorithms (e.g. GP), the inclusion of three-dimensional information of both ligands and targets, and the use of pharmacogenomics data are still under exploration.

Overall, the ability of PCM to become a customary technique in both the public and the private domain in the following years will certainly rest on its capability to capitalize on biological data of diverse nature, including personalized *omics* data (personalized medicine), in combination with structural data of ligands, be those small molecules, antibodies or peptides.

Bibliography

- A Mauri, VC (2006). "DRAGON software: An easy approach to molecular descriptor calculations". In: *MATCH Commun. Math. Comput. Chem.* 56, pp. 237–248 (cit. on p. 6).
- Akella, LB and D DeCaprio (2010). "Cheminformatics approaches to analyze diversity in compound screening libraries". In: *Curr. Opin. Chem. Biol.* 14.3, pp. 325–330 (cit. on p. 3).
- Andersson, CD, BY Chen, and A Linusson (2010). "Mapping of ligand-binding cavities in proteins". In: *Proteins* 78.6, pp. 1408–1422 (cit. on p. 25).
- Andersson, CR, MG Gustafsson, and H Strömbergsson (2011). "Quantitative chemogenomics: machine-learning models of protein-ligand interaction". In: *Curr. Top. Med. Chem.* 11.15, pp. 1978–1993 (cit. on pp. 12, 25, 27).
- Arrowsmith, CH, C Bountra, PV Fish, K Lee, and M Schapira (2012). "Epigenetic protein families: a new frontier for drug discovery". In: *Nat. Rev. Drug Discov.* 11.5, pp. 384–400 (cit. on p. 23).
- Artemenko, N (2008). "Distance dependent scoring function for describing protein-ligand intermolecular interactions". In: *J. Chem. Inf. Model.* 48.3, pp. 569–574 (cit. on p. 27).
- Bahar, I, C Chennubhotla, and D Tobi (2007). "Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation". In: *Curr. Opin. Struct. Biol.* 17.6, pp. 633–640 (cit. on p. 9).
- Ballester, PJ and JBO Mitchell (2010). "A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking". In: *Bioinformatics* 26.9, pp. 1169–1175 (cit. on p. 6).
- Barretina, J, G Caponigro, N Stransky, K Venkatesan, AA Margolin, S Kim, CJ Wilson, J Lehar, GV Kryukov, D Sonkin, A Reddy, M Liu, L Murray, MF Berger, JE Monahan, P Morais, J Meltzer, A Korejwa, J Jané-Valbuena, FA Mapa, J Thibault, E Bric-Furlong, P Raman, A Shipway, IH Engels, J Cheng, GK Yu, J Yu, P Aspesi, M de Silva, K Jagtap, MD Jones, L Wang, C Hatton, E Palescandolo, S Gupta, S Mahan, C Sougnez, RC Onofrio, T Liefeld, L MacConaill, W Winckler, M Reich, N Li, JP Mesirov, SB Gabriel, G Getz, K Ardlie, V Chan, VE Myer, BL Weber, J Porter, M Warmuth, P Finan, JL Harris, M Meyerson, TR Golub, MP Morrissey, WR Sellers, R Schlegel, and LA Garraway (2012). "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity". In: *Nature* 483, pp. 603–607 (cit. on p. 30).

- Basu, A, NE Bodycombe, JH Cheah, EV Price, K Liu, GI Schaefer, RY Ebright, ML Stewart, D Ito, S Wang, AL Bracha, T Liefeld, M Wawer, JC Gilbert, AJ Wilson, N Stransky, GV Kryukov, V Dancik, J Barretina, LA Garraway, CSY Hon, B Munoz, JA Bittker, BR Stockwell, D Khabele, AM Stern, PA Clemons, AF Shamji, and SL Schreiber (2013). "An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules". In: *Cell* 154.5, pp. 1151–61 (cit. on pp. 30, 31).
- Bellucci, M, F Agostini, M Masin, and GG Tartaglia (2011). "Predicting protein associations with long noncoding RNAs". In: *Nat. Methods* 8.6, pp. 444–445 (cit. on p. 32).
- Ben-Hur, A and C Ong (2008). "Support vector machines and kernels for computational biology". In: *PLoS Comput. Biol.* 4.10, e1000173 (cit. on pp. 13, 14).
- Bender, A and RC Glen (2005). "A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication". In: *J. Chem. Inf. Model.* 45.5, pp. 1369–1375 (cit. on p. 3).
- Bender, A, J Scheiber, M Glick, JW Davies, K Azzaoui, J Hamon, L Urban, S Whitebread, and JL Jenkins (2007). "Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure". In: *ChemMedChem* 2.6, pp. 861–873 (cit. on p. 3).
- Bianchi, MT and EJ Botzakis (2010). "Targeting ligand-gated ion channels in neurology and psychiatry: is pharmacological promiscuity an obstacle or an opportunity?" In: *BMC Pharmacol* 10, p. 3 (cit. on p. 3).
- Bieler, M and H Koeppen (2012). "The role of chemogenomics in the pharmaceutical industry". In: *Drug Dev. Res.* 73.7, pp. 357–364 (cit. on p. 3).
- Bock, JR and DA Gough (2005). "Virtual screen for ligands of orphan G protein-coupled receptors". In: *J. Chem. Inf. Model.* 45.5, pp. 1114–1402 (cit. on pp. 9, 19, 20).
- Borisy, AA, PJ Elliott, NW Hurst, MS Lee, J Lehar, ER Price, G Serbedzija, GR Zimmermann, MA Foley, BR Stockwell, and CT Keith (2003). "Systematic discovery of multicomponent therapeutics". In: *Proc. Natl. Acad. Sci. U. S. A.* 100.13, pp. 7977–7982 (cit. on p. 22).
- Bosnić, Z and I Kononenko (2009). "An overview of advances in reliability estimation of individual predictions in machine learning". In: *Intell. Data Anal.* 13.2, pp. 385–401 (cit. on p. 16).
- Bredel, M and E Jacoby (2004). "Chemogenomics: an emerging strategy for rapid target and drug discovery". In: *Nat. Rev. Genet.* 5.4, pp. 262–275 (cit. on p. 4).
- Breese, JS, D Heckerman, and C Kadie (1998). "Empirical analysis of predictive algorithms for collaborative filtering". In: pp. 43–52 (cit. on p. 19).
- Brown, J, Y Okuno, G Marcou, A Varnek, and D Horvath (2014). "Computational chemogenomics: Is it more than inductive transfer?" In: *J. Comput. Aided Mol. Des.* Pp. 1–22 (cit. on pp. 31, 33).

- Bruce, CL, JL Melville, SD Pickett, and JD Hirst (2007). "Contemporary QSAR classifiers compared". In: *J. Chem. Inf. Model.* 47.1, pp. 219–227 (cit. on p. 12).
- Cao, D, Y Liang, Z Deng, Q Hu, M He, Q Xu, G Zhou, L Zhang, Z Deng, and S Liu (2013a). "Genome-scale screening of drug-target associations relevant to ki using a chemogenomics approach". In: *PLoS One* 8.4, e57680 (cit. on pp. 7, 8).
- Cao, D, G Zhou, S Liu, L Zhang, Q Xu, M He, and Y Liang (2013b). "Large-scale prediction of human kinase-inhibitor interactions using protein sequences and molecular topological structures". In: *Anal. Chim. Acta* 792, pp. 10–18 (cit. on p. 21).
- Cheng, F, Y Yu, J Shen, L Yang, W Li, G Liu, PW Lee, and Y Tang (2011). "Classification of Cytochrome P450 Inhibitors and Noninhibitors Using Combined Classifiers". In: *J. Chem. Inf. Model.* (Cit. on p. 14).
- Cheng, F, Y Zhou, J Li, W Li, G Liu, and Y Tang (2012). "Prediction of chemical-protein interactions: multitarget-QSAR versus computational chemogenomic methods". In: *Mol. BioSyst.* 8.9, pp. 2373–2384 (cit. on p. 7).
- Cohen, P (2002). "Protein kinases—the major drug targets of the twenty-first century?" In: *Nat. Rev. Drug Discov.* 1.4, pp. 309–315 (cit. on p. 21).
- Collobert, R, F Sinz, J Weston, and L Bottou (2006). "Large scale transductive SVMs". In: *JMLR* 7, pp. 1687–1712 (cit. on p. 15).
- Cortes-Ciriano, I, GJP van Westen, EB Lenselink, DS Murrell, A Bender, and TE Malliavin (2014). "Proteochemometric modeling in a Bayesian framework". In: *J. Cheminf.* 6.1, p. 35 (cit. on pp. 17, 21, 30, 33, 34).
- Cortes-Ciriano, I, DS Murrell, GJP van Westen, A Bender, and TE Malliavin (2015). "Prediction of the Potency of Mammalian Cyclooxygenase Inhibitors with Ensemble Proteochemometric Modeling". In: *J. Cheminf.* 7, p. 1 (cit. on pp. 16, 29).
- Cortes-Ciriano, I, van Westen, G J P, Bouvier, G, Nilges, M, Overington, J P, Bender, A, and TE Malliavin (2015). "Improved Large-scale prediction of growth inhibition patterns on the NCI60 cancer cell-line panel". In: *In revision Bioinformatics* (cit. on pp. 4, 11, 30, 31, 33).
- Cruciani, G, P Crivori, PA Carrupt, and B Testa (2000). "Molecular fields in quantitative structure-permeation relationships: the VolSurf approach". In: *J. Mol. Struct.* 503.1, pp. 17–30 (cit. on p. 8).
- Dakshanamurthy, S, NT Issa, S Assefnia, A Seshasayee, OJ Ps, S Madhavan, A Uren, ML Brown, and SW Byers (2012). "Predicting new indications for approved drugs using a proteochemometric method". In: *J. Med. Chem.* 55.15, pp. 6832–6848 (cit. on p. 7).
- Das, MP Krein, and CM Breneman (2010). "Binding affinity prediction with property-encoded shape distribution signatures". In: *J. Chem. Inf. Model.* 50.2, pp. 298–308 (cit. on pp. 6, 27).
- Davis, MI, JP Hunt, S Herrgard, P Ciceri, LM Wodicka, G Pallares, M Hocker, DK Treiber, and P Zarrinkar (2011). "Comprehensive analysis of kinase inhibitor selectivity". In: *Nat. Biotechnol.* 29.11, pp. 1046–1051 (cit. on p. 8).

- De Bruyn, T, GJP van Westen, AP IJzerman, B Stieger, P de Witte, PF Augustijns, and PP Annaert (2013). "Structure-Based Identification of OATP1B1/3 Inhibitors". In: *Mol. Pharmacol.* 83.6, pp. 1257–1267 (cit. on pp. 8, 13, 16).
- Delmore, JE, GC Issa, ME Lemieux, PB Rahl, J Shi, HM Jacobs, E Kastitis, T GilP, RM Paranal, J Qi, M Chesi, AC Schinzel, MR McKeown, TP Heffernan, CR Vakoc, PL Bergsagel, IM Ghobrial, PG Rson, RA Young, WC Hahn, KC Anderson, AL Kung, JE Bradner, and CS Mitsiades (2011). "BET bromodomain inhibition as a therapeutic strategy to target c-Myc". In: *Cell* 146.6, pp. 904–17 (cit. on p. 22).
- Deng, W, C Brenema, and M Embrechts (2004). "Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods". In: *J. Chem. Inf. Comput. Sci.* 44.2, pp. 699–703 (cit. on p. 27).
- Desaphy, J, E Raimbaud, P Ducrot, and D Rognan (2013). "Encoding Protein-Ligand interaction patterns in Fingerprints and Graphs". In: *J. Chem. Inf. Model.* (Cit. on p. 27).
- Dimitrov, I and P Garnev (2010). "Peptide binding to the HLA-DRB1 supertype: a proteochemometrics analysis". In: *Eur. J. Med. Chem.* 45.1, pp. 236–243 (cit. on p. 6).
- Doherty, KM, P Nakka, BM King, S Rhee, SP Holmes, RW Shafer, and ML Radhakrishnan (2011). "A multifaceted analysis of HIV-1 protease multidrug resistance phenotypes". In: *BMC Bioinformatics* 12.1, pp. 477–496 (cit. on p. 24).
- DREAM: Dialogue on Reverse Engineering Assessment and Methods project. URL: <http://dreamchallenges.org/> (cit. on p. 32).
- Duvenaud, D, JR Lloyd, R Grosse, JB Tenenbaum, and Z Ghahramani (2013). "Structure Discovery in Nonparametric Regression through Compositional Kernel Search". In: <http://arxiv.org/abs/1302.4922> (cit. on p. 14).
- Eklund, M, U Norinder, S Boyer, and L Carlsson (2012). "Benchmarking Variable Selection in QSAR". In: *Mol. Inf.* 31.2, pp. 173–179 (cit. on p. 12).
- (2014). "Choosing feature selection and learning algorithms in QSAR". In: *J. Chem. Inf. Model.* 54.3, pp. 837–843 (cit. on p. 12).
- Erhan, D, P L'heureux, SY Yue, and Y Bengio (2006). "Collaborative filtering on a family of biological targets". In: *J. Chem. Inf. Model.* 46.2, pp. 626–635 (cit. on pp. 13, 19).
- Fernandez, M, S Ahmad, and A Sarai (2010). "Proteochemometric recognition of stable kinase inhibition complexes using topological autocorrelation and support vector machines". In: *J. Chem. Inf. Model.* 50.6, pp. 1179–1188 (cit. on p. 6).
- Floyd, SR, ME Pacold, Q Huang, SM Clarke, FC Lam, IG Call, BD Bryson, J Rameseder, MJ Lee, EJ Blake, A Fydrych, R Ho, BA Greenberger, GC Chen, A Maffa, AM Del Rosario, DE Root, AE Carpenter, WC Hahn, DM Sabatini, CC Chen, FM White, JE Bradner, and MB Yaffe (2013). "The bromodomain protein Brd4 insulates chromatin from DNA damage signalling". In: *Nature* 498.7453, pp. 246–50 (cit. on p. 22).

- Frimurer, TM, T Ulven, CE Elling, L Gerlach, E Kostenis, and T Högberg (2005). "A physico-genetic method to assign ligand-binding relationships between 7TM receptors". In: *Bioorg. Med. Chem. Lett.* 15.16, pp. 3707–12 (cit. on pp. 20, 25).
- Gao, J, D Che, VW Zheng, R Zhu, and Q Liu (2012). "Integrated QSAR study for inhibitors of Hedgehog Signal Pathway against multiple cell lines: a collaborative filtering method". In: *BMC Bioinformatics* 13.1, p. 186 (cit. on pp. 13, 19).
- Gao, J, Q Huang, D Wu, Q Zhang, Y Zhang, T Chen, Q Liu, R Zhu, Z Cao, and Y He (2013). "Study on human GPCR-inhibitor interactions by proteochemometric modeling". In: *Gene* 518.1, pp. 124–131 (cit. on pp. 8, 17, 25).
- Garnett, MJ, EE Edelman, SJS Heidorn, CD Greenman, A Dastur, KW Lau, P Greninger, IR Thompson, X Luo, J Soares, Q Liu, F Iorio, L Surdez Dand Chen, RJ Milano, GR Bignell, AT Tam, H Davies, Ja Sson, S Barthorpe, SR Lutz, F Kogera, K Lawrence, A McLaren-Douglas, X Mitropoulos, T Mironenko, H Thi, L Rson, W Zhou, F Jewitt, T Zhang, P O'Brien, JL Boisvert, S Price, W Hur, W Yang, X Deng, A Butler, HG Choi, JW Chang, J Baselga, I Stamenkovic, Ja Engelman, SV Sharma, O Delattre, J Saez-Rodriguez, NS Gray, J Settleman, PA Futreal, DA Haber, MR Stratton, S Ramaswamy, U McDermott, and CH Benes (2012). "Systematic identification of genomic markers of drug sensitivity in cancer cells". In: *Nature* 483.7391, pp. 570–575 (cit. on p. 30).
- Gaulton, A, LJ Bellis, AP Bento, J Chambers, M Davies, A Hersey, Y Light, S McGlinchey, D Michalovich, B Al-Lazikani, and JP Overington (2012). "ChEMBL: a large-scale bioactivity database for drug discovery". In: *Nucleic Acids Res.* 40.Database issue, pp. D1100–1107 (cit. on pp. 3, 7).
- Glinca, S and G Klebe (2013). "Cavities Tell More than Sequences: Exploring Functional Relationships of Proteases via Binding Pockets". In: *J. Chem. Inf. Model.* 53.8, pp. 2082–2092 (cit. on p. 25).
- Gloriam, DE, SM Foord, FE Blaney, and SL Garland (2009). "Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design". In: *J. Med. Chem.* 52.14, pp. 4429–4442 (cit. on p. 25).
- Gottlieb, AS, E Rupp, and R Sharan (2011). "PREDICT: a method for inferring novel drug indications with application to personalized medicine". In: *Mol. Syst. Biol.* 7, p. 496 (cit. on p. 6).
- Gregori-Puigjané, E and J Mestres (2008a). "A ligand-based approach to mining the chemogenomic space of drugs". In: *Comb. Chem. High Throughput Screen.* 11.8, pp. 669–676 (cit. on p. 5).
- (2008b). "Coverage and bias in chemical library design". In: *Curr. Opin. Chem. Biol.* 12.3, pp. 359–365 (cit. on p. 4).
- Gruetter, M (2012). "Structural genomics: open collaboration is key to new drugs". In: *Nature* 491.7422, p. 40 (cit. on p. 23).

- Gujral, TS, L Peshkin, and MW Kirschner (2014). "Exploiting polypharmacology for drug target deconvolution". In: *Proc. Natl. Acad. Sci. U. S. A.* 111.13, pp. 5048–5053 (cit. on p. 22).
- Haibe-Kains, B, N El-Hachem, NJ Birkbak, AC Jin, AH Beck, HJWL Aerts, and J Quackenbush (2013). "Inconsistency in large pharmacogenomic studies". In: *Nature* 504.7480, pp. 389–93 (cit. on p. 31).
- Hamosh A S, A, J Amberger, C Bocchini, D Valle, and VA McKusick (2002). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders". In: *Nucleic Acids Res.* 30.1, pp. 52–55 (cit. on p. 6).
- Hawkins, DM, SC Basak, J Kraker, KT Geiss, and Fa Witzmann (2006). "Combining chemodescriptors and biodescriptors in quantitative structure-activity relationship modeling". In: *J. Chem. Inf. Model.* 46.1, pp. 9–16 (cit. on p. 34).
- Heijne, WHM, AS Kienhuis, B van Ommen, RH Stierum, and JP Groten (2005). "Systems toxicology: applications of toxicogenomics, transcriptomics, proteomics and metabolomics in toxicology". In: *Expert Rev. Proteomics* 2.5, pp. 767–780 (cit. on p. 32).
- Hinton, GEG, S Osindero, and YW Teh (2006). "A fast learning algorithm for deep belief nets". In: *Neural Comput.* 18.7, pp. 1527–1554 (cit. on p. 32).
- Hong, H, Q Xie, W Ge, F Qian, H Fang, L Shi, Z Su, R Perkins, and W Tong (2008). "Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics". In: *J. Chem. Inf. Model.* 48.7, pp. 1337–1344 (cit. on p. 8).
- Hoppe, C, C Steinbeck, and G Wohlfahrt (2006). "Classification and comparison of ligand-binding sites derived from grid-mapped knowledge-based potentials". In: *J. Chem. Inf. Model.* 24.5, pp. 328–340 (cit. on p. 26).
- Horst, E van der, JE Peironcelly, GJP van Westen, OO van den Hoven, WRJD Galloway, DR Spring, JK Wegner, HWT van Vlijmen, AP IJzerman, JP Overington, and A Bender (2011). "Chemogenomics approaches for receptor deorphanization and extensions of the chemogenomics concept to phenotypic space". In: *Curr. Top. Med. Chem.* 11.15, pp. 1964–1977 (cit. on p. 9).
- Horuk, R (2009). "Chemokine receptor antagonists: overcoming developmental hurdles". In: *Nat. Rev. Drug Discov.* 8.1, pp. 23–33 (cit. on p. 29).
- Huang, Q, H Jin, Q Liu, Q Wu, H Kang, Z Cao, and R Zhu (2012a). "Proteochemometric modeling of the bioactivity spectra of HIV-1 protease inhibitors by introducing protein-ligand interaction fingerprint". In: *PLoS One* 7.7, e41698 (cit. on p. 7).
- Huang, Z, H Jiang, X Liu, Y Chen, J Wong, Q Wang, W Huang, T Shi, and J Zhang (2012b). "HEMD: an integrated tool of human epigenetic enzymes and chemical modulators for therapeutics". In: *PLoS ONE* 7.6. Ed. by E Ballestar, e39917 (cit. on p. 23).
- Jacob, L, B Hoffmann, V Stoven, and JP Vert (2008). "Virtual screening of GPCRs: an in silico chemogenomics approach". In: *BMC Bioinformatics* 9.1, p. 363 (cit. on pp. 20, 26).

- Jacoby, E, ed. (2013). *Computational Chemogenomics*. Pan Stanford Publishing (cit. on p. 4).
- Junaid, M, M Lapinsh, M Eklund, O Spjuth, and JES Wikberg (2010). "Proteochemometric modeling of the susceptibility of mutated variants of the HIV-1 virus to reverse transcriptase inhibitors". In: *PloS One* 5.12, e14353 (cit. on pp. 6, 24).
- Kagan, L, AK Abraham, DE Mager, and JM Harrold (2010). "Interspecies scaling of receptor-mediated pharmacokinetics and pharmacodynamics of type I interferons". In: *Pharm. Res.* 27.5, pp. 920–932 (cit. on p. 29).
- Kalliokoski, T, C Kramer, A Vulpetti, and P Gedeck (2013). "Comparability of mixed IC₅₀ data - a statistical analysis". In: *PLoS One* 8.4, e61007 (cit. on pp. 16, 32).
- Karaman, MW, S Herrgard, DK Treiber, P Gallant, CE Atteridge, BT Campbell, KW Chan, P Ciceri, MI Davis, PT Edeen, R Faraoni, M Floyd, JP Hunt, DJ Lockhart, ZV Milanov, MJ Morrison, G Pallares, HK Patel, S Pritchard, LM Wodicka, and P Zarrinkar (2008). "A quantitative analysis of kinase inhibitor selectivity". In: *Nat. Biotechnol.* 26.1, pp. 127–132 (cit. on pp. 6, 8).
- Keiser, MJ, BL Roth, BN Armbruster, P Ernsberger, JJ Irwin, and BK Shoichet (2007). "Relating protein pharmacology by ligand chemistry". In: *Nat. Biotechnol.* 25.2, pp. 197–206 (cit. on p. 9).
- Kellenberger, E, P Muller, C Schalon, G Bret, N Foata, and D Rognan (2006). "sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank". In: *J. Chem. Inf. Model.* 46.2, pp. 717–727 (cit. on p. 6).
- Knapp, S and H Weinmann (2013). "Small-Molecule Modulators for Epigenetics Targets". In: *ChemMedChem*, pp. 1885–1891 (cit. on p. 22).
- Kondratovich, E, II Baskin, and A Varnek (2013). "Transductive Support Vector Machines: Promising Approach to Model Small and Unbalanced Datasets". In: *Mol. Inf.* 32.3, pp. 261–266 (cit. on pp. 13, 15).
- Kontijevskis, A, R Petrovska, S Yahorava, J Komorowski, and JES Wikberg (2009). "Proteochemometrics mapping of the interaction space for retroviral proteases and their substrates". In: *Bioorg. Med. Chem.* 17.14, pp. 5229–5237 (cit. on p. 24).
- Koonin, EV (2005). "Orthologs, paralogs, and evolutionary genomics". In: *Annu. Rev. Genet.* 39, pp. 309–338 (cit. on p. 28).
- Koppisetty, CAK, M Frank, GJL Kemp, and PG Nyholm (2013). "Computation of binding energies including their enthalpy and entropy components for protein-ligand complexes using support vector machines". In: *J. Chem. Inf. Model.* (Cit. on p. 27).
- Kramer, C and P Gedeck (2011a). "Global free energy scoring functions based on distance-dependent atom-type pair descriptors". In: *J. Chem. Inf. Model.* 51.3, pp. 707–720 (cit. on pp. 7, 27).
- (2011b). "Three descriptor model sets a high standard for the CSAR-NRC HiQ benchmark". In: *J. Chem. Inf. Model.* 51.9, pp. 2139–2145 (cit. on pp. 6, 27).
- Kramer, C and R L (2012). "QSARs, data and error in the modern age of drug discovery". In: *Curr. Top. Med. Chem.* 12.17, pp. 1896–1902 (cit. on pp. 16, 32).

- Kramer, C, T Kalliokoski, P Gedeck, and A Vulpetti (2012). "The experimental uncertainty of heterogeneous public K(i) data". In: *J. Med. Chem.* 55.11, pp. 5165–5173 (cit. on pp. 16, 33).
- Kruger, FA and JP Overington (2012). "Global Analysis of Small Molecule Binding to Related Protein Targets". In: *PLoS Comput. Biol.* 8.1, e1002333 (cit. on pp. 4, 28).
- Kubinyi, H, FA Hamprecht, and T Mietzner (1998). "Three-Dimensional Quantitative Similarity-Activity Relationships (3D QSiAR) from SEAL similarity matrices". In: *J. Med. Chem.* 41.14, pp. 2553–2564 (cit. on p. 21).
- Laine, E, C Goncalves, JC Karst, A Lesnard, S Rault, WJ Tang, TE Malliavin, D Ladant, and A Blondel (2010). "Use of allosterism to identify inhibitors of calmodulin-induced activation of Bacillus anthracis edema factor". In: *Proc. Natl. Acad. Sci. U. S. A.* 107.25, pp. 11277–11282 (cit. on p. 27).
- Lapinsh, M, P Prusis, and A Gutcaits (2001). "Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions". In: *Biochim. Biophys. Acta.* 1525.1-2, pp. 180–190 (cit. on pp. 4, 5, 9).
- Lapinsh, M and JES Wikberg (2010). "Kinome-wide interaction modelling using approaches for kinase description and linear and non-linear data analysis techniques". In: *BMC Bioinformatics* 11.1, p. 339 (cit. on p. 6).
- Lapinsh, M, P Prusis, T Lundstedt, and JES Wikberg (2002). "Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands". In: *Mol. Pharmacol.* 61.6, pp. 1465–1475 (cit. on pp. 9, 28).
- Lapinsh, M, M Eklund, O Spjuth, P Prusis, and JES Wikberg (2008). "Proteochemometric modeling of HIV protease susceptibility". In: *BMC Bioinformatics* 9.1, p. 181 (cit. on p. 24).
- Lapinsh, M, A Worachartcheewan, O Spjuth, V Georgiev, V Prachayasittikul, C Nantasenamat, and JES Wikberg (2013). "A unified proteochemometric model for prediction of inhibition of cytochrome P450 isoforms". In: *PloS One* 8.6. Ed. by DS Sem, e66566 (cit. on pp. 8, 14, 30).
- Li, Q, Y Wang, and SH Bryant (2009). "A novel method for mining highly imbalanced high-throughput screening data in PubChem". In: *Bioinformatics* 25.24, pp. 3310–6 (cit. on p. 33).
- Lin, H, MF Sassano, BL Roth, and BK Shoichet (2013). "A pharmacological organization of G protein-coupled receptors". In: *Nat. Methods* 10.2, pp. 140–146 (cit. on p. 4).
- Liu, T, X Lin Yand Wen, RN Jorissen, MK Gilson, and RN Jorissen (2007). "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities". In: *Nucleic Acids Res.* 35.Database issue, pp. D198–201 (cit. on pp. 7, 8).
- Liu, Z, L Wu, Y Wang, and X Zhang (2008). "Protein cavity clustering based on community structure of pocket similarity network". In: *Int. J. Bioinf. Res. Appl.* 4.4, pp. 445–460 (cit. on p. 25).
- Lounkine, E, MJ Keiser, S Whitebread, D Mikhailov, J Hamon, JL Jenkins, P Lavan, E Weber, AK Doak, S Côté, BK Shoichet, and L Urban (2012). "Large-scale prediction

- and testing of drug activity on side-effect targets". In: *Nature* 486.7403, pp. 361–367 (cit. on p. 28).
- Lowe, R, HY Mussa, JBO Mitchell, and RC Glen (2011). "Classifying molecules using a sparse probabilistic kernel binary classifier". In: *J. Chem. Inf. Model.* 51.7, pp. 1539–1544 (cit. on p. 15).
- M G Genton, N Cristianini, J Shawe-Taylor, RW (2001). "Classes of kernels for machine learning: a statistics perspective". In: *JMLR*, pp. 299–312 (cit. on p. 14).
- Manning, G, DB Whyte, R Mez, T Hunter, and S Sudarsanam (2002). "The Protein Kinase Complement of the Human Genome". In: *Science* 298.5600, pp. 1912–1934 (cit. on p. 21).
- McHale, CM, L Zhang, A Hubbard, and MT Smith (2010). "Toxicogenomic profiling of chemically exposed humans in risk assessment". In: *Mutat. Res.* 705.3, pp. 172–183 (cit. on p. 32).
- Meinshausen, N (2006). "Quantile Regression Forests". In: *JMLR* 7, pp. 983–999 (cit. on p. 16).
- Melnikova, I and J Golden (2004). "From the analyst's couch: targeting protein kinases". In: *Nat. Rev. Drug Discov.* 3.12, pp. 993–994 (cit. on p. 21).
- Menden, MP, F Iorio, MJ Garnett, U McDermott, CH Benes, PJ Ballester, and J Saez-Rodriguez (2013). "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties". In: *PLoS One* 8.4, e61318 (cit. on pp. 7, 9, 29, 31, 33).
- Meslamani, J and D Rognan (2011). "Enhancing the Accuracy of Chemogenomic Models with a Three-Dimensional Binding Site Kernel". In: *J. Chem. Inf. Model.* 51.7, pp. 1593–1603 (cit. on pp. 6, 26).
- Mestres, J, E Gregori-Puigjané, S Valverde, and RV Solé (2008). "Data completeness—the Achilles heel of drug-target networks". In: *Nat. Biotechnol.* 26.9, pp. 983–984 (cit. on p. 3).
- (2009). "The topology of drug-target interaction networks: implicit dependence on drug properties and target families". In: *Mol. BioSyst.* 5.9, pp. 1051–1057 (cit. on pp. 3, 4).
- Metz, JT, EF Json, NB Soni, PJ Merta, L Kifle, and PJ Hajduk (2011). "Navigating the kinome". In: *Nat. Chem. Biol.* 7.4, pp. 200–202 (cit. on p. 8).
- Morgan, M, S Falcon, and R Gentleman. *GSEABase: Gene set enrichment data structures and methods*. URL: <http://www.bioconductor.org/> (cit. on p. 33).
- Netzeva, TI, AP Worth, T Aldenberg, R Benigni, TD Mark, PO Gramatica, J Jaworska, S Kahn, G Klopman, A C, G Myatt, N Nikolova-Jeliazkova, GY Patlewicz, and R Perkins (2005). *Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships*. Tech. rep., pp. 1–19 (cit. on p. 16).
- Nijima, S, A Shiraishi, and Y Okuno (2012). "Dissecting Kinase Profiling Data to Predict Activity and Understand Cross-Reactivity of Kinase Inhibitors". In: *J. Chem. Inf. Model.* 52.4, pp. 901–912 (cit. on pp. 7, 13, 14).

- Ning, X, Rangwala, and G Karypis (2009). "Multi-assay-based structure-activity relationship models: improving structure-activity relationship models by incorporating activity information from related targets". In: *J. Chem. Inf. Model.* 49.11, pp. 2444–2456 (cit. on p. 9).
- Nisius, B and H Gohlke (2012). "Alignment-independent comparison of binding sites based on DrugScore potential fields encoded by 3D Zernike descriptors". In: *J. Chem. Inf. Model.* 52.9, pp. 2339–2347 (cit. on p. 27).
- Norinder, U, L Carlsson, S Boyer, and M Eklund (2014). "Introducing Conformal Prediction in Predictive Modeling A Transparent and Flexible Alternative To Applicability Domain Determination". In: *J. Chem. Inf. Model.* 54.6, pp. 1596–1603 (cit. on pp. 31, 33).
- Norman, TC, C Bountra, AM Es, KR Yamamoto, and SH Friend (2011). "Leveraging crowdsourcing to facilitate the discovery of new medicines". In: *Sci. Transl. Med.* 3.88, 88mr1 (cit. on p. 32).
- O'Boyle, NM, M Banck, CA J, C Morley, T Vandermeersch, and GR Hutchison (2011). "Open Babel: An open chemical toolbox". In: *J. Cheminf.* 3.1, p. 33 (cit. on p. 8).
- Okuno, Y, J Yang, K Taneishi, H Yabuuchi, and G Tsujimoto (2006). "GLIDA: GPCR-ligand database for chemical genomic drug discovery". In: *Nucleic Acids Res.* 34.Database issue, pp. D673–677 (cit. on pp. 6, 20).
- Paolini, GV, RHB Shapland, WP van Hoorn, JS Mason, and AL Hopkins (2006). "Global mapping of pharmacological space". In: *Nat. Biotechnol.* 24.7, pp. 805–815 (cit. on p. 9).
- Paricharak, S, I Cortes-Ciriano, AP IJzerman, TE Malliavin, and A Bender (2015). "Proteochemometric modeling coupled to in silico target prediction: an integrated approach for the simultaneous prediction of polypharmacology and binding affinity of small molecules". In: *J. Cheminf.* 7, p. 15 (cit. on p. 33).
- Park, Y and EM Marcotte (2012). "Flaws in evaluation schemes for pair-input computational predictions". In: *Nat. Methods* 9.12, pp. 1134–1136 (cit. on p. 21).
- Pastor, M, G Cruciani, I McLay, S Pickett, and S Clementi (2000). "GRIND-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors". In: *J. Med. Chem.* 43.17, pp. 3233–3243 (cit. on p. 7).
- Paul, SM, DS Mytelka, CT Dunwiddie, CC Persinger, BH Munos, SR Lindborg, and AL Schacht (2010). "How to improve R&D productivity: the pharmaceutical industry's grand challenge". In: *Nat. Rev. Drug Discov.* 9.3, pp. 203–214 (cit. on p. 3).
- Prinjsa, RK, J Witherington, and K Lee (2012). "Place your BETs: the therapeutic potential of bromodomains". In: *Trends. Pharmacol. Sci.* 33.3, pp. 146–53 (cit. on p. 22).
- Prusis, P, R Muceniece, P Andersson, C Post, T Lundstedt, and JES Wikberg (2001). "PLS modeling of chimeric MSo4/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand-receptor interactions". In: *Biochim. Biophys. Acta.* 1544.1-2, pp. 350–357 (cit. on p. 9).

- Prusis, P, M Lapinsh, S Yahorava, R Petrovska, P Niyomrattanakit, G Katzenmeier, and JES Wikberg (2008). "Proteochemometrics analysis of substrate interactions with dengue virus NS₃ proteases". In: *Bioorg. Med. Chem.* 16.20, pp. 9369–9377 (cit. on p. 24).
- Prusis, P, M Junaaid, R Petrovska, S Yahorava, A Yahorau, G Katzenmeier, M Lapinsh, and JES Wikberg (2013). "Design and evaluation of substrate-based octapeptide and non substrate-based tetrapeptide inhibitors of dengue virus NS_{2B}-NS₃ proteases". In: *Biochem. Biophys. Res. Commun.* 434.4, pp. 767–772 (cit. on pp. 7, 24).
- Rasmussen, CE and CKI Ws (2006). *Gaussian Processes for machine learning*. MIT Press (cit. on p. 17).
- Reutlinger, M, T Rodrigues, P Schneider, and G Schneider (2014). "Combining On-Chip Synthesis of a Focused Combinatorial Library with Computational Target Prediction Reveals Imidazopyridine GPCR Ligands". In: *Angew. Chem. Int. Ed.* 53.2, pp. 582–585 (cit. on p. 25).
- Rhee, S, MJ Gonzales, R Kantor, BJ Betts, J Ravela, and RW Shafer (2003). "Human immunodeficiency virus reverse transcriptase and protease sequence database". In: *Nucleic Acids Res.* 31.1, pp. 298–303 (cit. on p. 6).
- Robinson, DD, W SH, and R Farid (2010). "Understanding kinase selectivity through energetic analysis of binding site waters". In: *ChemMedChem* 5.4, pp. 618–627 (cit. on p. 26).
- Rogers, D and M Hahn (2010). "Extended-connectivity fingerprints". In: *J. Chem. Inf. Model.* 50.5, pp. 742–754 (cit. on p. 7).
- Rognan, D (2007). "Chemogenomic approaches to rational drug design". In: *Br. J. Pharmacol.* 152.1, pp. 38–52 (cit. on pp. 5, 9).
- Sandberg, M, L Eriksson, J Jonsson, M Sjöström, and S Wold (1998). "New chemical descriptors relevant for the design of biologically active peptides A multivariate characterization of 87 amino acids". In: *J. Med. Chem.* 41.14, pp. 2481–2491 (cit. on pp. 6–8).
- Scholkopf, B and AJ Smola (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. MIT Press, p. 644 (cit. on p. 14).
- Scholkopf, B, T Koji, and JP Vert (2004). *Kernel Methods in Computational Biology*. MIT Press, p. 400 (cit. on p. 14).
- Schrödinger, L (2013). *Small-Molecule Drug Discovery Suite 2013-3: QikProp*. URL: <http://www.schrodinger.com/> (cit. on p. 28).
- Schwaighofer, A, T Schroeter, S Mika, J Laub, A ter Laak, D Sulzle, U Ganzer, N H, and KR Müller (2007). "Accurate solubility prediction with error bars for electrolytes: a machine learning approach". In: *J. Chem. Inf. Model.* 47.2, pp. 407–424 (cit. on pp. 13, 17).
- Shiraishi, AG, S Niiijima, JB Brown, M Nakatsui, and Y Okuno (2013). "A chemical genomics approach for GPCR-ligand interaction prediction and extraction of ligand binding determinants". In: *J. Chem. Inf. Model.* 53.6, pp. 1253–1262 (cit. on pp. 7, 25).

- Shoemaker, RH (2006). "The NCI60 human tumour cell line anticancer drug screen". In: *Nat. Rev. Cancer*. 6.10, pp. 813–823 (cit. on p. 30).
- Shoshan, MC and S Linder (2004). "Promiscuous and specific anticancer drugs: combatting biological complexity with complex therapy". In: *Cancer Ther.* 2, pp. 297–304 (cit. on p. 3).
- Sippl, MJ and M Wiederstein (2008). "A note on difficult structure alignment problems". In: *Bioinformatics* 24.3, pp. 426–427 (cit. on p. 26).
- Sottriffer, C, P Sanschagrin, H Mer, and G Klebe (2008). "SFCscore: scoring functions for affinity prediction of protein-ligand complexes". In: *Proteins* 73.2, pp. 395–419 (cit. on p. 27).
- Spjuth, O, M Eklund, M Lapinsh, M Junaid, and JES Wikberg (2011). "Services for prediction of drug susceptibility for HIV proteases and reverse transcriptases at the HIV drug research centre". In: *Bioinformatics* 27.12, pp. 1719–1720 (cit. on pp. 7, 24).
- Stolovitzky, G, D Monroe, and A Califano (2007). "Dialogue on Reverse-Engineering Assessment and Methods: The DREAM of high-Throughput pathway inference". In: *Ann. N. Y. Acad. Sci.* 1115.11-22 (cit. on p. 32).
- Strömbergsson, H, A Kryshtafovych, P Prusis, K Fidelis, JES Wikberg, J Komorowski, and TR Hvidsten (2006). "Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures". In: *Proteins* 65.3, pp. 568–579 (cit. on p. 26).
- Subramanian, V, P Prusis, L Pietila, H Xhaard, and G Wohlfahrt (2013). "Visually Interpretable Models of Kinase Selectivity Related Features Derived from Field-Based Proteochemometrics". In: *J. Chem. Inf. Model.* 53.11, pp. 3021–3030 (cit. on pp. 8, 21, 22, 26).
- Surgand, JS, J Rodrigo, E Kellenberger, and D Rognan (2006). "A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors". In: *Proteins* 62.2, pp. 509–538 (cit. on p. 25).
- Suter, L, LE Babiss, and EB Wheeldon (2004). "Toxicogenomics in Predictive Toxicology in Drug Development". In: *Chem. Biol.* 11.2, pp. 161–171 (cit. on p. 32).
- Svetnik, V, A Liaw, C Tong, JC Culberson, RP Sheridan, and BP Feuston (2003). "Random Forest: a classification and regression tool for compound classification and QSAR modeling". In: *J. Chem. Inf. Comp. Sci.* 43.6, pp. 1947–1958 (cit. on p. 16).
- Tetko, IV, P Bruneau, H Mewes, DC Rohrer, and GI Poda (2006). "Can we estimate the accuracy of ADME-Tox predictions?" In: *Drug Discov. Today* 11.15-16, pp. 700–707 (cit. on p. 16).
- Tiikkainen, P, L Bellis, Y Light, and L Franke (2013). "Estimating Error Rates in Bioactivity Databases". In: *J. Chem. Inf. Model.* 53.10, pp. 2499–2505 (cit. on pp. 16, 33).
- Tipping, ME (2001). "Sparse Bayesian learning and the relevance vector machine". In: *JMLR* 1, pp. 211–245 (cit. on pp. 13, 15).

- Tomic, S, L Nilsson, and RC Wade (2000). "Nuclear receptor-DNA binding specificity: A COMBINE and Free-Wilson QSAR analysis". In: *J. Med. Chem.* 43.9, pp. 1780–1792 (cit. on p. 9).
- Üstün, B, WJ Melssen, and LMC Buydens (2006). "Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel". In: *Chemom. Intell. Lab. Syst.* 81.1, pp. 29–40 (cit. on p. 14).
- Vidler, LR, N Brown, S Knapp, and S Hoelder (2012). "Druggability Analysis and Structural Classification of Bromodomain Acetyl-lysine Binding Sites". In: *J. Med. Chem.* Figure 2 (cit. on p. 22).
- Vieth, M, JJ Sutherland, DH Rson, and RM Campbell (2005). "Kinomics: characterizing the therapeutically validated kinase space". In: *Drug Discov. Today* 10.12, pp. 839–846 (cit. on p. 4).
- Vita, R, L Zarebski, JA Greenbaum, H Emami, I Hoof, N Salimi, R Damle, A Sette, and B Peters (2010). "The immune epitope database 2.0". In: *Nucleic Acids Res.* 38.Database issue, pp. D854–862 (cit. on p. 6).
- Vroling, B, M Sanders, C Baakman, A Borrmann, S Verhoeven, J Klomp, L Oliveira, J de Vlieg, and G Vriend (2011). "GPCRDB: information system for G protein-coupled receptors". In: *Nucleic Acids Res.* 39.Database issue, pp. D309–319 (cit. on p. 8).
- Wang, J, X Shen, and W Pan (2007). "On transductive support vector machines". In: *J. Contemp. Mat.* 1998 (cit. on p. 15).
- Wang, R, X Fang, Y Lu, and S Wang (2004). "The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures". In: *J. Med. Chem.* 47.12, pp. 2977–2980 (cit. on pp. 6, 27).
- Wang, Y, J Xiao, TO Suzek, J Zhang, J Wang, Z Zhou, L Han, K Karapetyan, S Dracheva, BA Shoemaker, E Bolton, A Gindulyte, and SH Bryant (2012). "PubChem's BioAssay Database". In: *Nucleic Acids Res.* 40.Database issue, pp. 400–412 (cit. on pp. 3, 8).
- Wassermann, AM, H Geppert, and J Bajorath (2009). "Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects". In: *J. Chem. Inf. Model.* 49.10, pp. 2155–2167 (cit. on pp. 9, 26).
- Weill, N and D Rognan (2009). "Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands". In: *J. Chem. Inf. Model.* 49.4, pp. 1049–1062 (cit. on pp. 20, 25).
- Weill, N, C Valencia, S Gioria, P Villa, M Hibert, and D Rognan (2011). "Identification of Nonpeptide Oxytocin Receptor Ligands by Receptor-Ligand Fingerprint Similarity Search". In: *Mol. Inf.* 30.6-7, pp. 521–526 (cit. on p. 6).
- Weill, N (2011). "Chemogenomic approaches for the exploration of GPCR space". In: *Curr. Top. Med. Chem.* 11.15, pp. 1944–1955 (cit. on p. 25).

- Weill, N and D Rognan (2010). "Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites". In: *J. Chem. Inf. Model.* 50.1, pp. 123–135 (cit. on p. 26).
- Weinstein, JN and PL Lorenzi (2013). "Cancer: Discrepancies in drug sensitivity". In: *Nature* 504.7480, pp. 381–3 (cit. on p. 31).
- Westen, GJP van and JP Overington (2013). "A ligand's-eye view of protein similarity". In: *Nat. Methods* 10.2, pp. 116–117 (cit. on p. 4).
- Westen, GJP van, JJ Wegner, AP IJzerman, HWT van Vlijmen, and A Bender (2011a). "Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets". In: *Med. Chem. Comm.* 2.1, pp. 16–30 (cit. on pp. 4, 5, 9, 12, 14, 15, 21).
- Westen, GJP van, JK Wegner, P Geluykens, L Kwanten, I Vereycken, A Peeters, AP IJzerman, HWT van Vlijmen, and A Bender (2011b). "Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development". In: *PLoS ONE* 6.11, e27518 (cit. on pp. 5, 11, 23).
- Westen, GJP van, OOvd Hoven, Rvd Pijl, T Mulder-Krieger, and A Bender (2012). "Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data". In: *J. Med. Chem.* 55.16, pp. 7010–7020 (cit. on pp. 4, 7, 14, 29, 34).
- Westen, GJP van, R Swier, JK Wegner, AP IJzerman, HWT van Vlijmen, and A Bender (2013a). "Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets". In: *J. Cheminf.* 5.1, p. 41 (cit. on p. 8).
- Westen, GJP van, A Hendriks, JK Wegner, AP IJzerman, HWT van Vlijmen, and A Bender (2013b). "Significantly improved HIV inhibitor efficacy prediction employing proteochemometric models generated from antivirogram data". In: *PLoS Comput. Biol.* 9.2. Ed. by SL Kosakovsky Pond, e1002899 (cit. on pp. 6, 23, 24).
- Willett, P (2009). "Similarity methods in chemoinformatics". In: *Annu. Rev. Inform. Sci.* 43.1, pp. 3–71 (cit. on p. 3).
- Wu, D, Q Huang, Y Zhang, Q Zhang, Q Liu, J Gao, Z Cao, and R Zhu (2012). "Screening of selective histone deacetylase inhibitors by proteochemometric modeling". In: *BMC Bioinformatics* 13.1, p. 212 (cit. on pp. 7, 14, 23).
- Yabuuchi, H, S Nijima, H Takematsu, T Ida, T Hirokawa, T Hara, T Ogawa, Y Minowa, G Tsujimoto, and Y Okuno (2011). "Analysis of multiple compound-protein interactions reveals novel bioactive molecules". In: *Mol. Syst. Biol.* 7, pp. 472–484 (cit. on pp. 6, 25).
- Yang, W, J Soares, P Greninger, EJ Edelman, H Lightfoot, S Forbes, N Bindal, D Beare, JA Smith, IR Thompson, S Ramaswamy, PA Futreal, DA Haber, MR Stratton, C Benes, U McDermott, and MJ Garnett (2013). "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells". In: *Nucleic Acids Res.* 41.Database issue, pp. D955–961 (cit. on p. 7).

- Yuriev, E, M Agostino, and PA Ramsland (2011). "Challenges and advances in computational docking: 2009 in review". In: *J. Mol. Recognit.* 24.2, pp. 149–164 (cit. on p. 27).
- Yuriev, E and PA Ramsland (2013). "Latest developments in molecular docking: 2010-2011 in review". In: *J. Mol. Recognit.* 26.5, pp. 215–239 (cit. on p. 27).
- Zhang, D, S Surapaneni, and L Guan (2012). "Clinical Dose Estimation Using Pharmacokinetic/Pharmacodynamic Modeling and Simulation". In: *ADME-Enabling Technologies in Drug Design and Development*. Ed. by D Zhang and S Surapaneni. Hoboken: J Wiley & Sons, Inc. Chap. 6 (cit. on p. 29).
- Zhang, S, A Golbraikh, and A Tropsha (2006). "Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces". In: *J. Med. Chem.* 49.9, pp. 2713–2724 (cit. on p. 27).
- Zilliaccus, J, AP Wright, U Norinder, JA Gustafsson, and J Carlstedt-Duke (1992). "Determinants for DNA-binding site recognition by the glucocorticoid receptor". In: *J. Biol. Chem.* 267.35, pp. 24941–24947 (cit. on p. 9).

Predictive Bioactivity Modelling

.2 Predictive Bioactivity Modelling

IN general, predictive bioactivity modelling studies (among which [PCM](#)) share a common algorithmic structure, which can be summarised in 4 model generation steps:

- Compound structure normalization (section [.2.1](#))
- Compound and target descriptor calculation (section [.2.2](#))
- Preprocessing, feature selection, model training and validation (sections [.2.3](#), [.2.4](#), [.2.6](#), [.2.7](#) and [.2.8](#))
- Bioactivity prediction for new molecules

Although many programming languages provide libraries for machine learning and statistical modelling, the R programming language provides a flexible platform for statistical analyses, and its applicability in medicinal chemistry has been reviewed elsewhere [Mente and Kuhn ([2012](#))]. Although R is extensively used in diverse biological domains, *e.g.* genomics, the availability of R packages for cheminformatics and medicinal chemistry is limited. Moreover, it does not exist a unified and open-source framework in R for the generation and validation of predictive bioactivity models comprising all four steps mentioned above. Nonetheless, R still constitutes the most frequent choice in the medicinal chemistry literature for compound bioactivity and property modelling [Mente and Kuhn ([ibid.](#))].

In order to fulfill this shortage, we have created the R package *camb*: Chemically Aware Model Builder [Murrell et al. ([2014](#))]. *camb* provides a complete and open framework in R to:

- Manipulate compound structures
- Generate compound, amino acid, and protein descriptors
- Train and validate [QSAR](#), [Quantitative Structure-Property Relationship \(QSPR\)](#), [QSAM](#), [PCM](#) and chemogenomic models
- Process and make predictions for external molecules
- Visualize chemical structures and the output of varied statistical analyses

Thus, *camb* enables the generation of predictive models (QSAR, QSPR, QSAM and PCM) starting with: chemical structure files, protein sequences (if required), and the associated properties or bioactivities.

The present chapter provides an overview of the aforementioned 4 modelling steps, both theoretically and in practice, as each step is related (when possible) to the corresponding *camb* functions. A tutorial on how to use *camb* for the generation of a PCM model is provided with the package documentation.

2.1 Compound standardization

Chemical structure representations are highly ambiguous, even if, *e.g.* canonical SMILES are used for representation -for example when considering aromaticity of ring systems, protonation states, and tautomers present in a particular environment. Hence, standardization (also termed as normalization) is a step of crucial importance when either storing structures, which should later be usable *e.g.* for structural searches, but also before descriptor calculation, such as in the current case, since many molecular properties are dependent on a consistent assignment of the above criteria in the first place. If one looks into large chemical databases one can see how important this step is -a rather good explanation for PubChem [Wang et al. (2012)], one of the largest public databases around, can be found at this address: <http://pubchemblog.ncbi.nlm.nih.gov/2014/06/19/what-is-the-difference-between-a-substance-and-a-compound-in-pubchem/>. Hence, standardizing chemical structures is crucial in order to provide consistent data for later modelling steps.

This step can be achieved by the *StandardiseMolecule* function of the R package *camb* [Murrell et al. (2014)]. The function *StandardiseMolecules* enables the depiction of molecular structures in the same (*i.e.* standardized or normalized) form. *camb* makes use of Indigo's InChI [InChI (2013); Indigo (2013)] plugin to represent all tautomers in canonical SMILES by converting molecules to InChI, discarding tautomeric information, and converting back to SMILES. The different arguments of this function allow control over the maximum number of fluorines, chlorines, bromines and iodines the molecules can contain in order to be retained for training. Inorganic molecules (those containing atoms not in the following set: {H, C, N, O, P, S, F, Cl, Br, I}) are removed if the argument *remove.inorganic* is set to "TRUE", which is the default value. Additionally, upper and lower limits for the molecular mass can be set with the arguments *min.mass.limit* and *max.mass.limit*.

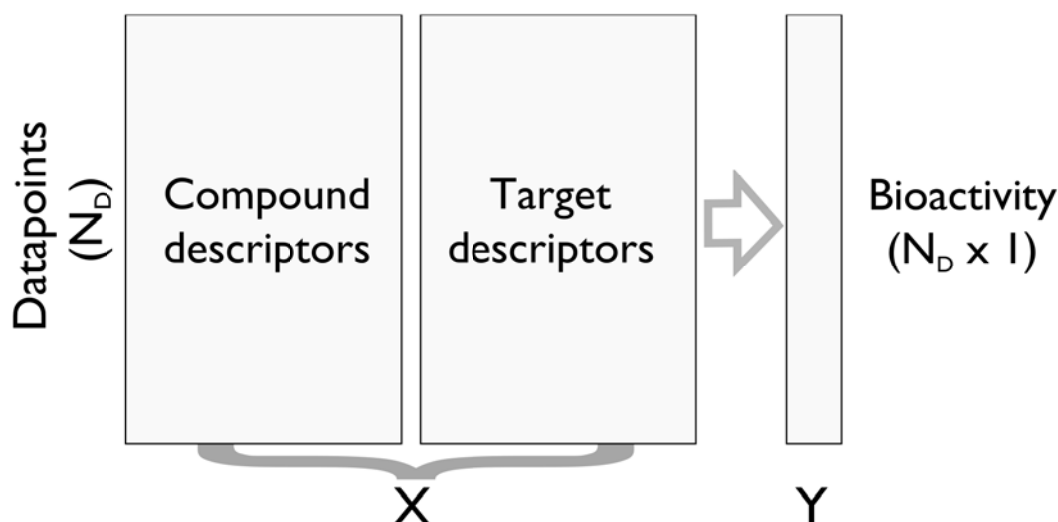


Figure .2.1: **Illustration of the descriptors used as input features in PCM.** Compound-target pairs are encoded by the horizontal stacking of compound and target descriptors. The resulting matrix is used as input data to train a single machine learning model.

.2.2 Descriptors

The ligand-target interaction space can be visualized as a matrix containing the activities of all possible ligand-target combinations (Figure .1.1 on page 10) [J E S Wikberg (2004)]. PCM attempts to predict the activity of a ligand on any target and *vice versa*, the activity of any ligand on a given target. The integration of these independent compound-target interactions is however possible in PCM due to the combination of chemical and target information in a single machine learning model. In practice, this is performed by horizontally stacking compound and target descriptors (Figure .2.1). This combination of chemical and biological information permits the extrapolation in either (or both) the chemical or target space (to the extent the training data allow).

.2.2.1 Target descriptors

As was touched upon above, PCM is rather flexible and can deal with a multitude of different target descriptors. Here, we will summarize some of the more common descriptors and later on in the review focus on novel descriptor types, for a full overview of established descriptors please see [Westen et al. (2011)]. By far the most common descriptors depend on a multiple sequence alignment of the sequences [Lapinsh et al. (2005)]. This type of protein descriptors are usually obtained from

a concatenation of physicochemical descriptors of the amino acids composing the binding site of the proteins considered, to which we refer in the following as amino acid descriptor sets.

Most amino acid descriptor sets have been developed by applying principal component analysis (PCA) over a matrix comprising properties for individual amino acids. The resulting principal components (PCs) are then used as amino acid descriptors, which generally explain $\geq 80\%$ of the variance present in the original matrix [Westen et al. (2013a)]. These amino acid descriptor sets can be categorized depending on the amino acid properties from which they are derived (Table .2.1). The first category comprises descriptor sets obtained by applying PCA on a matrix of physicochemical descriptors. The most common set of this category are 3 and 5 z-scales [Sandberg et al. (1998)], where 3 and 5 refer to the number of PCs kept. This descriptor set was derived by applying PCA on a matrix comprising physicochemical properties of 87 amino acids such as thin-layer chromatography and Nuclear Magnetic Resonance (RMN) data. Although the interpretation of z4 and z5 is not obvious, the PCs of z-scales are related to the following properties: z1 lipophilicity, z2 bulk, z3 charge and polarity, z4 and z5 electronegativity, electrophilicity, hardness and heat of formation.

Similarly, Protein Fingerprint (ProtFP) were also derived from PCA analysis of 58 physicochemical descriptors obtained from the AAindex database [Westen et al. (2013a)], but only considering the 20 natural amino acids. The FASGAI descriptor set [Liang and Li (2007)], from Factor analysis scales of generalized amino acid information, was derived from 335 physicochemical properties of these 20 amino acids. However, the dimensionality was reduced with factor analysis instead of PCA [Liang and Li (ibid.)]. The second category is composed by ST-Scales [Yang et al. (2010)] and T-Scales [Tian, Zhou, and Li (2007)], which were derived from a PCA analysis on 167 and 135 amino acid topological properties, respectively.

The third category is more heterogeneous and is composed of the following descriptor sets: MS-WHIM, VHSE, and BLOSUM. The MS-WHIM (MS-WHIM) set is derived from 36 electrostatic potential properties calculated from the three-dimensional structure of the amino acids [Zaliani and Gancia (1999)]. The VHSE descriptor set (Principal Components Score Vectors of Hydrophobic, Steric, and Electronic properties) are computed by PCA analysis of 50 physicochemical properties, comprising 18 hydrophobic, 17 steric and 15 electronic properties [Mei et al. (2005)]. Finally, the BLOSUM descriptor set was derived by a VARIMAX analysis of physicochemical properties of the 20 natural amino acids and from the BLOSUM62 alignment matrix [Georgiev. (2009)]. Further details about these descriptors and their predictive signal in PCM can be found in two recent publications [Westen et al. (2013a,b)].

Given that in PCM the information from several proteins is combined, it is

Descriptor set	Type	Derived by	Number of PCs	Variance explained	AAs covered	ref
BLOSUM*	Physicochemical and substitution matrix	VARIMAX	10	n/a	20	Georgiev. (2009)
FASGAI*	Physicochemical	Factor Analysis	6	84%	20	Liang and Li (2007)
MSWHIM*	3D electrostatic potential	PCA	3	61%	20	Zaliani and Gancia (1999)
ProtFP (PCA 3 PCs)	Physicochemical	PCA	3	75%	20	Westen et al. (2012a)
ProtFP (PCA 5 PCs)	Physicochemical	PCA	5	83%	20	Westen et al. (2012a)
ProtFP* (PCA 8 PCs)	Physicochemical	PCA	8	92%	20	Westen et al. (2012a)
ProtFP (Feature)	Feature based	Hashing	n/a	n/a	20	Westen et al. (2013a)
ST-scales*	Topological	PCA	5	91%	167	Yang et al. (2010)
T-scales*	Topological	PCA	8	72%	135	[Tian, Zhou, and Li (2007)]
VHSE*	Physicochemical	PCA	8	77%	20	Mei et al. (2005)
z-scales* (3 PCs)	Physicochemical	PCA	3	n/a	87	Sandberg et al. (1998)
z-scales* (5 PCs)	Physicochemical	PCA	5	87%	87	Sandberg et al. (1998)
z-scales (Binned)	Physicochemical	PCA followed by binning	n/a	n/a	20	Sandberg et al. (1998)

Table .2.1: **Amino acid descriptor sets used in PCM (adapted from Westen et al. (2013a)).** Those descriptor sets marked with * can be computed with the R package *camdb*.

normally required to align the sequence of the difference binding sites. This can be done using multiple sequence alignment by established tools such as ClustalW [Sievers et al. (2011)]. In this way, each column in the descriptor matrix corresponds to the same amino acid position in the alignment. A value of zero is normally used to describe gaps appearing in the multiple sequence alignment [Murrell et al. (2014)]. The reader is referred to a pair of benchmarking studies recently published for more information on this type of descriptors [Westen et al. (2013a,b)].

When no reliable alignment is possible, target descriptors can be only calculated using the whole protein sequence [Rao et al. (2011)]. The usage of only primary sequence descriptors to predict protein-protein interactions was shown to be efficient by Shen et al. (2007) who were able to train a SVM model based on more than 16,000 protein-protein pairs described with conjoint triad feature amino acid descriptors. Similarly, analyses of sequence variability among targets exhibiting uncorrelated bioactivity profiles, enabled the characterization of binding pocket residues energetically important for ligand binding and selectivity for GPCRs and kinases [Kuhn et al. (2007); Sheinerman, Giraud, and Laoui (2005); Surgand et al. (2006)].

If present, structural information from crystallographic structures can be used by selecting residues near the ligand binding site (e.g. 5 or 10 Å sphere around the co-crystallized ligand) [Kruger and Overington (2012); Lapinsh et al. (2005); Murrell et al. (2014)]. Subsequently, the corresponding residues for other targets can be obtained from sequence alignment. This semi-structural method is less reliable than

a full structural superposition and alignment gaps might appear. Paradoxically, the former appears in practice to have better performance, which might be due to the fact that domains not involved in ligand binding are not considered [Cortes-Ciriano et al. (2015); De Bruyn et al. (2013); Horst et al. (2011); Westen et al. (2012b)]. To date, binding sites in PCM models have been derived from single crystallographic structures [Ain et al. (2015); Cortes-Ciriano et al. (2015); Paricharak et al. (2015); Westen et al. (2012b)], thus ignoring the intrinsically dynamic nature of proteins. However, databases such as Pocketome [Kufareva, Ilatovskiy, and Abagyan (2012)] might facilitate the introduction of dynamic properties of protein binding sites in PCM models, as they contain ensembles of conformations for druggable binding sites extracted from co-crystal structures in the PDB. Similarly, the identification of protein cavities and the analysis of their dynamics could also contribute to better describe protein binding sites [Desdouits, Nilges, and Blondel (2015)]. To the best of my knowledge, descriptors accounting for the dynamic properties of binding site amino acids have not been reported in the literature. Including this dynamic information might lead to a better description of protein targets in cases where small molecule binding is dependent on the binding site conformation, *e.g.* kinases.

Beyond sequence similarity, targets have also been described in different ways to model compound bioactivities on multiple targets [Kalinina and Wichmann (2011); Meslamani and Rognan (2011); Weill et al. (2011); Willighagen et al. (2011); Yabuuchi et al. (2011)]. Among others, targets have been characterized by:

- Structural pocket similarity analyses
- Topology analyses of both compound-target and protein-protein interaction networks
- The combination of pharmacophoric and interaction fingerprints
- 3D alignment-free methods of binding sequences [Gloriam et al. (2009); Kinnings and Jackson (2009); Mestres et al. (2009); Subramanian et al. (2013); Weill and Rognan (2009)]

The availability of a plethora of target descriptors enables the application of PCM to targets families where, for instance, little structural information is available. In cases where targets are not proteins, but more complex biological systems, such as cell-lines (chapter .6), the target space can be described with *omics* data, namely: CNV data, gene expression levels, exome sequencing data, target fingerprints, protein abundance, and miRNA expression levels [Cortes-Ciriano, I et al. (2015); Menden et al. (2013)].

In *camb*, 8 amino acid descriptor sets can be computed (indicated with * in Table .2.1). Multiple sequence alignment gaps are supported by this *camb* functionality.

Descriptor values for these gaps, indicated with '-', are encoded with zeros. Similarly, the function *SeqDescs* permits the calculation of 13 types of whole protein sequence descriptors from UniProt identifiers or from amino acid sequences [Xiao and Xu (2014)], namely:

- Amino Acid Composition (AAC)
- Dipeptide Composition (DC)
- Tripeptide Composition (TC)
- Normalized Moreau-Broto Autocorrelation (MoreauBroto)
- Moran Autocorrelation (Moran)
- Geary Autocorrelation (Geary)
- CTD (Composition/Transition/Distribution) (CTD)
- Conjoint Traid (CTriad)
- Sequence Order Coupling Number (SOCN)
- Quasi-sequence Order Descriptors (QSO)
- Pseudo Amino Acid Composition (PACC)
- Amphiphilic Pseudo Amino Acid Composition (APAAC)

Further details about these descriptors can be found in Xiao and Xu ([ibid.](#)).

.2.2.2 Ligand descriptors

Similarly, from the ligand side a large number of descriptors have been employed in *PCM* in the last decade [Karelson (2000); Todeschini and Consonni (2008)]. Circular fingerprints are the most commonly applied due to both their consistent good performance and interpretability when using the unhashed (keyed) version [Glenn et al. (2006); Rogers and Hahn (2010)]. In addition, they have been shown to provide high retrieval rates in comparative studies [Bender et al. (2009); Koutsoukas et al. (2013)]. These fingerprints encode compound structures by considering radial atom neighborhoods [Bender, Mussa, and Glen (2004)].

In the calculation of circular Morgan fingerprints (Figure .2.2) [Cortes-Ciriano (2013)], all substructures in a molecule, with a maximal user-defined bond diameter, are assigned an unambiguous integer identifier. These identifiers are then mapped either into an unhashed or hashed array. This process is repeated for all molecules

comprised in a given data set. For the hashed array, the position in the array where the substructures will be mapped is given by the modulo of the division of the substructure identifier by the fingerprint size. In the case of unhashed (keyed) fingerprints, each bit in the fingerprint is associated to only one substructure, producing a length of the unhashed fingerprints equal to the number of distinct substructures present in the data set [Cortes-Ciriano (2013); Murrell et al. (2014)]. Both hashed and unhashed fingerprints can be stored in binary and count format. Keyed circular fingerprints enable the interpretation of models and the identification of chemical substructures implicated in compound potency and selectivity. The performance of models trained on hashed and unhashed circular Morgan fingerprints do not vary significantly [Cortes-Ciriano et al. (2015); Cortes-Ciriano, I et al. (2015)]. Therefore, we advocate for the customary usage of unhashed fingerprints in order to enhance the interpretability of PCM models.

Next to the circular fingerprints, physicochemical descriptors, such as DRAGON or PaDEL [A Mauri (2006); Yap (2011)], have been widely used in recent years (Table 1). Other ligand descriptors, such as atom types, topological indices, MACCs keys or ligand shape descriptors, have been also applied in the context of PCM.

In my experience, the description of compounds with circular Morgan fingerprints permits the generation of statistically validated PCM models but in several occasions the addition of physicochemical properties to fingerprints has been demonstrated to improve performance [De Bruyn et al. (2013)]. This was especially true on data sets with a large chemical diversity, *e.g.* resulting from screening a diverse set or resulting from covering a group of targets with diverse ligands.

In *camb* physicochemical PaDEL descriptors can be calculated with the function *GeneratePadelDescriptors*, whereas the function *MorganFPs* permits the calculation of hashed and unhashed Morgan fingerprints in binary and count format.

.2.2.3 Cross-term descriptors

Thirdly, some PCM studies have defined an additional class of descriptors, called cross-terms, by multiplying ligand and target descriptors. These descriptors serve as descriptors for the non-linear components of the interaction between ligand and target (*e.g.* a hydrogen bond that can be formed in one target but not in another) [Lapinsh et al. (2005); Prusis et al. (2006)]. Therefore, its application is advisable when using linear modelling techniques (such as Partial Least Squares (PLS)). In the case of non-linear techniques, cross-terms are not essential as the models should be able to capture this information [Doddareddy et al. (2009); Westen et al. (2012b)]. Nonetheless, my experience indicates that cross-terms might be nevertheless useful to improve model

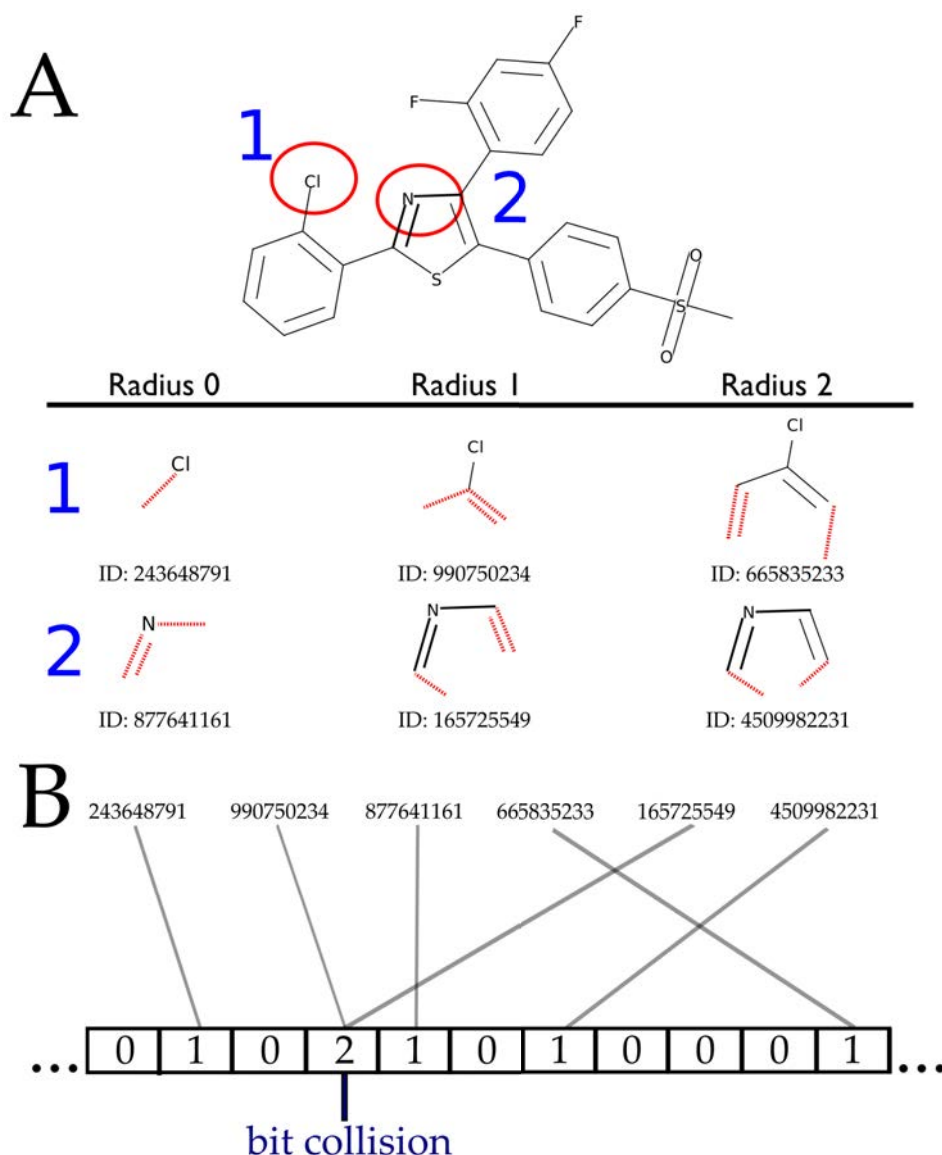


Figure .2.2: **Illustration of the computation of circular Morgan fingerprints.** **A.** In the calculation of circular Morgan fingerprints, all substructures in a molecule, with a maximal user-defined bond radius (2 in the example), are assigned an unambiguous integer identifier. In the figure, 1 and 2 correspond to two arbitrary atoms. The atoms, namely Cl and N, are taken as root atoms to illustrate how the adjacent atom layers are considered in the definition of the substructures present in the molecule. **B.** The position in the array where a substructure will be mapped in the fingerprint is given by the modulo obtained by dividing the fingerprint integer identifier by the fingerprint size. The substructure IDs shown in the Figure are arbitrary.

performance when using SVM or GP even though their interpretability might not be straightforward. For further reading on different types of descriptors applied in PCM we refer the reader to Westen et al. (2011).

.2.3 Statistical Preprocessing

Descriptors with constant value across all data-points do not provide any predictive signal, and thus can be removed with the function *RemoveNearZeroVarianceFeatures* from the R package *camb* [Kuhn (2008); Kuhn and Json (2013); Murrell et al. (2014)]. Correlated descriptors provide the same amount of predictive signal. Thus, they can be considered as redundant. In order to remove these descriptors (except for one which is kept) the function *RemoveHighlyCorrelatedFeatures* from the R package *camb* [Murrell et al. (2014)]. Subsequently, the remaining descriptors are centered to zero mean and scaled to unit variance with the function *PreProcess* from the R package *camb*. This is done given that each descriptor has a different range of values. For instance, the range of the descriptor accounting for the number of *e.g.* Carbon atoms is much smaller than that of the descriptor molecular weight. The reason for normalizing is that these differences in the range of values can have an influence on the importance (the weight) a model can assign to a descriptor, irrespective of its amount of predictive signal.

.2.4 Generation of PCM Models

The values of the model parameters can be optimized with multivariate optimization algorithms such as Monte Carlo Sampling or conjugate gradient. However, in practice, it is computationally less expensive to optimize the value of the parameters by grid search and k-fold Cross-Validation (CV) [Hawkins, Basak, and Mills (2003)]. In this way, the optimized values are sufficiently close to the optimal value to permit the generation of highly predictive models.

Recent studies recommend the usage of nested cross-validation (NCV) to report model performance [Krstajic et al. (2014); Pahikkala et al. (2014); Park and Marcotte (2012); Varma and Simon (2006)]. In NCV, two validation loops are nested: the inner one serves to optimize the values of the hyperparameters through traditional k-fold cross-validation, whereas the outer loop serves to assess the predictive ability of the model trained on the whole training set. This procedure is repeated k times, each time changing the composition of the training and the test sets. Thus, NCV does not provide the best parameter combination, as in each k round the best values of the hyperparameters might change due to the variance of the different training sets. Still,

it provides the best estimate of the CV error as it provides an error interval [Krstajic et al. (2014)].

In CV, the training set, which normally comprises 70% of the data, is split into k folds by e.g. stratified or random sampling of the bioactivity values. The remaining fold, generally 30% of the data, constitutes the test set. The values of the parameters are optimized in the following way. For each combination of parameters, a model is trained on $k - 1$ folds, and the values for the remaining fold are then predicted. This procedure is repeated k times, each time holding out a different fold. The values of the parameters exhibiting the lowest average RMSE value along the k folds is considered as optimal. Subsequently, a model is trained on the whole training set, using the optimized values for the parameters. The predictive power of this model can be assessed on the test set. To significantly compare the quality of the modeling with different machine learning algorithms, the same folds are used to train all models.

The R package *caret* provides a common interface to the most popular machine learning packages that exist in R, and, as such, *camp* invokes *caret*, a common interface to the most popular machine learning packages that exist in R, to set up cross validation frameworks and train machine learning models (see the package tutorial for more details).

.2.5 Commonly used Algorithms

This section briefly presents the theory underlying some of the most widely used algorithms in predictive modelling, which will also be applied throughout this thesis.

Given a data set $D = \{\mathbf{X}, \mathbf{y}\}$ where $\mathbf{X} = \{\mathbf{x}^i\}_{i=1}^n$ is the set of compound descriptors, and $\mathbf{y} = \{y^i\}_{i=1}^n$ is the vector of observed bioactivities, the aim of supervised learning is to find the function (model) underlying D , which can be then used to predict the bioactivity for new data-points, \mathbf{x}_{new} . In the following subsections we briefly summarize the theory behind the algorithms explored in this study.

Kernel Methods

Kernel functions, statistic covariances [Genton (2002)], or simply kernels permit the computation of the dot product between two vectors, $\mathbf{x} \in \mathbf{X}$, in a higher dimensional space, F (potentially of infinite dimension), without explicitly computing the coordinates of these vectors in F :

$$\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = K(\mathbf{x}, \mathbf{x}') \quad (.2.1)$$

where $\phi()$ is a mapping function from X to F , $\phi : X \rightarrow F$. This is known as the kernel trick [Scholkopf, Koji, and Vert (2004)]. Thus the value of the kernel applied on the input vectors is equivalent to their dot product in F . In practice, this permits to apply linear methods based on dot products, *e.g.* SVM, in F while using X in the calculations (thus, not requiring to compute the coordinates of the input data in F). This is computationally less expensive than the explicit calculation of the coordinates of X in F , which in some cases might not even be possible. The linear relationships learnt in F are non-linear in the input space. Moreover, kernel methods are extremely versatile, as the same linear method, *e.g.* SVM, can learn diverse complex non-linear functions in the input space because the functional form is controlled by the kernel, which in turn can be adapted to the data by the user.

The formulae for the kernels used throughout this thesis are:

Bessel Kernel	$K(\mathbf{x}, \mathbf{x}') = \frac{J_{v+1}(\sigma \ \mathbf{x} - \mathbf{x}'\)}{\ \mathbf{x} - \mathbf{x}'\ ^{-n(v+1)}}$
Laplacian Kernel	$K(\mathbf{x}, \mathbf{x}') = e^{-\sigma \ \mathbf{x} - \mathbf{x}'\ }$
Normalized Polynomial Kernel	$K(\mathbf{x}, \mathbf{x}') = \frac{(scale \mathbf{x}^T \mathbf{x}' + offset)^{degree}}{\sqrt{\mathbf{x} \mathbf{x}^T \mathbf{x}' \mathbf{x}'^T}}$
Polynomial Kernel	$K(\mathbf{x}, \mathbf{x}') = (scale \mathbf{x}^T \mathbf{x}' + offset)^{degree}$
PUK Kernel	$K(\mathbf{x}, \mathbf{x}') = \frac{1}{[1 + (\frac{2\sqrt{\ \mathbf{x} - \mathbf{x}'\ ^2} \sqrt{2(\frac{1}{\omega}) - 1}}}{\sigma})^2]^\omega}$
Radial Kernel	$K(\mathbf{x}, \mathbf{x}') = e^{-\frac{\ \mathbf{x} - \mathbf{x}'\ ^2}{2l^2}}$

Table .2.2: Covariance functions (kernels) formula.

where $\|\mathbf{x} - \mathbf{x}'\|$ is the Euclidean distance, J the Bessel function of first kind, and \mathbf{x}^T the transpose of \mathbf{x} .

1. Gaussian Process (GP) In Bayesian inference [MacKay (2003)], the experimental data is used to update the *a priori* knowledge assumed for a certain problem. In the context of supervised learning, we *a priori* assume a distribution over the functions candidate to model the data, *i.e.* the *prior* distribution. The *prior* is then updated with the training examples, which yields the posterior probability distribution:

$$P(\text{GP}(\mathbf{x})|\mathbf{D}) \propto P(\mathbf{y}|\text{GP}(\mathbf{x}), \mathbf{X}) P(\text{GP}(\mathbf{x})) \quad (.2.2)$$

where: (i) $P(\text{GP}(\mathbf{x})|\mathbf{D})$ is the *posterior* probability distribution giving the bioactivity predictions; (ii) the likelihood $P(\mathbf{y}|\text{GP}(\mathbf{x}), \mathbf{X})$ is the probability of the observations, \mathbf{y} , given the training set, \mathbf{X} and the model $\text{GP}(\mathbf{x})$; and (iii) $P(\text{GP}(\mathbf{x}))$ is the *prior*.

GP [Rasmussen and Ws (2006)] are a stochastic process that, similar to a multi-variate Gaussian distribution, defined by its mean value and covariance matrix, is fully specified by its mean function, μ , (usually the zero function) and its covariance function, \mathbf{C}_X :

$$\text{GP}(\mathbf{x}) \sim \mathcal{N}(\mu, \mathbf{C}_X + \sigma_d^2 \delta(\mathbf{x}_i, \mathbf{x}_j)) \quad (i, j \in 1, \dots, n) \quad (.2.3)$$

where $\delta(\mathbf{x}_i, \mathbf{x}_j)$ is the Kronecker delta function and σ_d^2 is the noise of the input data, which is assumed to be normally distributed with mean zero. \mathbf{C}_X is obtained by applying a positive definite kernel function to \mathbf{X} , $\mathbf{C}_X = \text{Cov}(\mathbf{X})$. The function values for any set of input vectors follow a multidimensional normal distribution, and, therefore, the bioactivity value, y_{new} , for a new input vector, \mathbf{x}_{new} , will also follow a Gaussian distribution defined by MacKay (2003); Puntanen and Styan (2005):

$$P(y_{\text{new}}) \sim \mathcal{N}(\mu_{y_{\text{new}}} = \mathbf{k}^T \mathbf{C}_X^{-1} \mathbf{y}, \sigma_{y_{\text{new}}}^2 = m - \mathbf{k}^T \mathbf{C}_X^{-1} \mathbf{k}) \quad (.2.4)$$

where the best estimate for the bioactivity of \mathbf{x}_{new} is the average value of y_{new} , $\mu_{y_{\text{new}}} = \langle P(y_{\text{new}}) \rangle$.

As it will be seen in Eq. 3.5, those input vectors in \mathbf{X} similar to \mathbf{x}_{new} , contribute more to the prediction of y_{new} , as \mathbf{y} is weighted by \mathbf{k}^T . The predicted variance, $\sigma_{y_{\text{new}}}^2$, corresponds to the difference between the *a priori* knowledge about \mathbf{x}_{new} : $m = \text{Cov}(\mathbf{x}_{\text{new}}, \mathbf{x}_{\text{new}}^T)$, and what can be inferred about \mathbf{x}_{new} from similar input vectors: $\mathbf{k}^T \mathbf{C}_X^{-1} \mathbf{k}$.

2. Support Vector Machines (SVM) SVM [Ben-Hur et al. (2008); Cortes and Vapnik (1995)] fit a linear model in a higher dimensional dot product feature space, F , of the form:

$$f(\mathbf{x}|\mathbf{w}) = \langle \mathbf{w}^T \phi(\mathbf{x}) \rangle + \alpha_0 \quad (.2.5)$$

where \mathbf{w} is a vector of weights $\in F$. The kernel trick can be applied if \mathbf{w} can be expressed as a linear combination of the training examples, namely $\mathbf{w} = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i)$.

Given the definition of kernel given above, Eq. 2.5 can be rewritten as:

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i y_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle + \alpha_0 = \sum_{i=1}^n \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + \alpha_0 \quad (.2.6)$$

The optimization of the α_i values is usually performed by applying Lagrangian multipliers (dual formulation):

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (.2.7)$$

subject to $\sum_{i=1}^n \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$. C is a regularization parameter that penalizes for incorrect predictions during training. Thus, the larger the value of C , the larger this penalty.

3. Relevant Vector Machines (RVM) RVM [Tipping (2000)] follow a similar formulation to SVM with the exception that the weights are inferred from the data in a Bayesian framework by defining an explicit prior probability distribution, normally Gaussian, on the parameters α_i :

$$f(\mathbf{x}|\alpha) = \sum_{i=1}^n \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + \alpha_0 \quad (.2.8)$$

This formulation leads to sparse models as a large fraction of the weights are sharply distributed around zero. Thus, only a small fraction of the examples from \mathbf{X} (the Relevance Vectors) are used when making predictions using Eq. .2.8.

Ensemble Methods

Ensemble methods use multiple weak simple models (base learners) to get a meta-model attaining a predictive power higher than that of the models used individually. Thus, building a model ensemble consists of (i) training individual models on (subsets) of the training examples, and (ii) integrating them to generate a combined prediction. Although it is possible to build model ensembles using different machine learning algorithms as base learners, *e.g.* model stacking [Cortes-Ciriano et al. (2015); Hastie, Tibshirani, and Friedman (2001)], decision tree-based ensembles are predominant in the literature. The following subsection briefly presents the ensemble methods used in this study.

1. Bagging: Bagged CART Regression Trees (Tree bag) Bootstrap aggregating or Bagging is a technique that averages the prediction of a set of high-variance base learners (normally regression trees, *e.g.* CART) [Breiman et al. (1984)], each trained on a bootstrap sample, b , drawn with replacement from the training data. Thus,

bagging leads to higher stability and predictive power with respect to the individual base learners, and reduces overfitting [Hastie, Tibshirani, and Friedman (2001)]. In practice, high-variance and low-bias algorithms, such as regression trees, have proved to be very well suited for bagging [Hastie, Tibshirani, and Friedman (ibid.)]. The model can be formulated as:

$$f(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B T_b(\mathbf{x}) \quad (.2.9)$$

where $T_b(\mathbf{x})$ corresponds to the tree base learner trained on the b -th bootstrap sample.

2. Boosting: Gradient Boosting Machines (GBM) Boosting [Breiman (1998); Hastie, Tibshirani, and Friedman (2001); Natekin and Knoll (2013)] differs from bagging in that the base learners, here regression trees, are trained and combined sequentially. At each iteration, a new base-learner is trained on the b -th bootstrap sample, G_b , and added to the ensemble trying to minimize the loss function associated to the whole ensemble. The loss function, Ψ , can be *e.g.* the squared error loss, *i.e.* the average of the square of the training residuals: $\Psi(y, f) = \frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2$.

The final model is given by:

$$f(\mathbf{x}) = \sum_{b=1}^B w_b G_b(\mathbf{x}) \quad (.2.10)$$

where $G_b(\mathbf{x})$ is the base learner trained on the b -th bootstrap sample, w_b its weight, and B the total number of iterations and trees. The weight for a base learner, w_b , is, thus, proportional to its prediction accuracy. The update rule for the model can be written as:

$$f_b(\mathbf{x}) = f_{b-1}(\mathbf{x}) + \nu w_b G_b(\mathbf{x}); 0 < \nu \leq 1 \quad (.2.11)$$

where $f_b(\mathbf{x})$ corresponds to the ensemble at iteration b , and ν to the learning rate (see below). Deepest gradient descent is applied at each iteration to optimize the weight for the new base learner as follows:

$$w_b = \min_w \sum_{i=1}^n \Psi(y_i, f_{b-1}(\mathbf{x}) + G_b(\mathbf{x})) \quad (.2.12)$$

To minimize the risk of overfitting of GBM, several procedures have been proposed. The first one consists of training the individual base learners on bootstrap samples of smaller size than the training set. The relative size of these samples with respect

to that of the training set is controlled by the parameter *bag fraction* or η . A second procedure is to reduce the impact of the last tree added to the ensemble on the minimization of the loss function by adding a regularization parameter, shrinkage or learning rate (ν). The underlying rationale is that sequential improvement by small steps is better than improving the performance substantially in a single step. Likewise, the effect of an inaccurate learner on the whole ensemble is thus reduced. Another way to reduce overfitting is to control the complexity of the trees by setting the maximum number of nodes of the base learners with the parameter tree complexity (t_c). Finally, the number of iterations, *i.e.* number of trees in the ensemble (n_{trees}), also needs to be controlled. The training error decreases with the number of trees, although a high number of iterations might lead to overfitting [Natekin and Knoll (2013)].

3. Random Forest (RF) Like in bagging, RF [Breiman (2001)] build an ensemble (forest) of regression trees and average their predictions:

$$f(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \text{TF}_b(\mathbf{x}) \quad (.2.13)$$

where $\text{TF}_b(\mathbf{x})$ corresponds to the tree base learner in the forest trained on the b -th bootstrap sample. The difference with bagging is that the node splitting is performed using only a subset of descriptors randomly chosen. This additional level of randomization decorrelates the trees in the forest leading, in practice, to a predictive power comparable to boosting [Hastie, Tibshirani, and Friedman (2001)].

In QSAR, RF have been shown to be robust with respect to the parameter values. In practice, a suitable choice of the number of trees (n_{trees}) was shown to be 100, as higher values do not generally lead to significantly higher predictive power [Sheridan (2012, 2013)].

Partial Least Squares Regression (PLS)

Partial least squares or *projection to latent structures* [Wold, Sjöström, and Eriksson (2001)], is a multivariate modeling technique capable to extract quantitative relationships from data sets where the numbers of descriptors, P , is much larger than the number of training examples, N . Multiple linear regression fails to model this type of data sets since for small (N/P) ratios, \mathbf{X} is not a full rank matrix and it will be probably collinear (the "small N large P problem") [Abdi (2010)]. In Principal Components Regression (PCR), the principal components of \mathbf{X} are taken as predictors, thus reducing P (dimensionality reduction) and the problem of multicollinearity. PLS extends this idea by simultaneously projecting both \mathbf{X} and \mathbf{y} to latent variables, with

the constraint of maximizing the covariance of the projections of \mathbf{X} and \mathbf{y} . Subsequently, the response variable is obtained on the latent vectors obtained on \mathbf{X} . We refer the reader to Abdi and Williams [Abdi and Williams (2010)] for further details on PLS.

k-Nearest Neighbours (k-NN)

The k-NN algorithm averages the response value over the k closest neighbours to estimate the response value for a data-point as:

$$f(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^k y_i \quad (.2.14)$$

The Euclidean distance is used to find the k closest neighbours.

.2.6 Validation of PCM Models

Due to the previously mentioned bias in bioactivity data (both from a chemical point of view and target point of view) the ligand-target interaction matrix is virtually never complete [Gregori-Puigjané and Mestres (2008a,b); Rognan (2007)]. I have trained PCM models on sparse data sets with a degree of matrix completeness in the 2-3% range that demonstrated good performance on the test set (interpolation) [Cortes-Ciriano et al. (2014)]. However, the degree of completeness of the ligand-target interaction matrix is only one parameter influencing the predictive ability of a model. The variability on the chemical and the target side are the other two factors that need to be considered both in model validation and to assess its applicability domain [Cortes-Ciriano et al. (ibid.)]. Hence, I strongly suggest validating PCM models following a number of basic guidelines, which are in line with the recommendations from Park and Marcotte (2012).

- Validation on the test set, which would correspond to completing the ligand-target interaction matrix, *i.e.* *interpolation*. A model is trained on *e.g.* 70% percent of the data (training set) and the bioactivity for the remaining 30% (test set) is predicted. In this case, all targets and compounds are present in both the training and the test set. This method corresponds to a Park and Marcotte C1 validation and serves to determine if a reliable model can be fit on the data set.
- Leave-One-Target-Out (LOTO) validation: all bioactivity data annotated on a target is excluded from the training set. A model is subsequently trained on the training set, and the model is used to predict the bioactivities for the compounds annotated on the hold-out target. This process is repeated for each

target. This validation scheme corresponds to a Park and Marcotte C2 validation and reflects the common situation in prospective validation where there is no information for a given target for which we intend to find hits.

- Leave-One-Compound-Out (LOCO) validation: the bioactivity data for a compound on all targets is excluded from the training. Similarly to the LOTO validation, the PCM model trained on the remaining data is used to predict the bioactivity for the hold-out compound on each target. This data availability scenario corresponds to a Park and Marcotte C2 validation and resembles the situation where a PCM model is applied to novel chemistry in a *e.g.* prospective validation screening campaign. If the number of compounds in the training data set is large, compound clusters can be used instead of single compounds, thus leading to the Leave-Once-Compound-Cluster-Out validation scenario (LOCCO) [Cortes-Ciriano et al. (2014)].

We note in particular that in chapter .6 the following abbreviations change: (i) LOTO: Leave-One-Tissue-Out -extrapolation to cell-lines originated from tissues not present in the training set-, (ii) LOCO: Leave-One-Cell-Line-Out -extrapolation to novel cell-lines-, and (iii) LOCCO: Leave-One-Compound-Cluster-Out -extrapolation to novel chemical clusters (see subsection .6.2.5)-.

In addition to the above scenarios, I suggest to compare the performance of the PCM model trained on all data to single-target QSAR models.

- Individual QSAR models. The goal of this validation is two-fold. Firstly a direct comparison to QSAR can determine if it is wise to apply PCM to a data set. Secondly, as was touched upon above, bias in the data can be the cause of some targets being reliably modeled and some targets being poorly modeled [Gregori-Puigjané and Mestres (2008a,b); Rognan (2007)]. When calculating validation parameters (such as the correlation coefficient) on the full test set, poorly modeled targets can be masked. In order to notice discontinuities, I recommend to not only calculate the validation parameters on the full test set, but to also calculate them on the test set data points grouped *per* target and points that are grouped *per* ligand [Cortes-Ciriano et al. (2014); Westen et al. (2013b)]. The values of the statistical metrics calculated *per* target can be directly compared with those obtained with single QSAR models. A direct comparison with the values calculated on the full test set would not be appropriate, as the effect of targets badly modelled might be masked by targets modelled with high accuracy.
- Extrapolation on both the *ligand* and the *target* spaces. Ideally, the final valida-

tion is one where a target and all compounds that have been tested on this (and other targets) are iteratively excluded from the training set. This approach corresponds with a Park and Marcotte C₃ validation. C₃ validation is considered extrapolation rather than interpolation, as both parts of the pair (the ligand and the target) have not been seen in the training set by the model.

- Family Quantitative Structure-Activity Relationship (**QSAR F**): models are trained on all data-points in the data set using exclusively compound descriptors as input features. A **QSAR F** model learns on the bioactivity values, and predicts the average likelihood for a compound of being active on the targets considered. In this way, a **QSAR F** model serves to assess whether the explicit inclusion of target information improves the prediction of compound activity for those compounds exhibiting variable bioactivity profiles across the target panel. If a compound is not selective against particular targets, and thus displays a comparable activity value across the target panel, a **QSAR F** model would suffice to predict the average likelihood of that molecule to be active against any target. However, in the case of a compound displaying selectivity towards particular targets, *i.e.* being active against particular targets and inactive against others, a **QSAR F** model would fail to predict the activities of that compound across the target panel, as compound activity would depend to a large extent on the biological side and not much on the chemical side.
- Family Quantitative Structure-Activity Modelling (**QSAM F**): these models are trained on all data-points in the data set using exclusively target descriptors as input features. This validation scheme assesses whether compound bioactivities are correlated on a given target, *i.e.* a diverse compound set displays the same activity on a given target. Therefore, high predictive ability of a **QSAM F** model indicates that bioactivity prediction depends to a large extent on the target, and to a much lesser extent on the compound structures. In that case, the inclusion of compound descriptors would not provide any predictive signal.
- Inductive Transfer (**Inductive Transfer (IT)**): the idea underlying **IT** is that the knowledge acquired in a given task, *e.g.* the prediction of compound activity on a given target, is used to solve similar problems, *e.g.* to predict the activity of the same compound set on a new target. In **IT**, two sources of information were input to the model, namely: (i) compound descriptors, and (ii) **Target Identity Fingerprints (TIFP)**. TIFP are binary descriptors, of length equal to the number of different targets considered, where each bit position corresponds to one target. To describe a given target, all bits were set to zero except for the bit corresponding to that target. Therefore, targets are located in a high dimensional space where they are equidistant. Formally, TIFP are defined as:

$$\text{TIFP}(i, j) = \delta(i - j)(i, j \in 1, \dots, N_{\text{cells}}) \quad (.2.15)$$

where δ is the Kronecker delta function and N_{cells} the number of distinct targets. In cases where the targets are cell-lines, these fingerprints are known as **Cell-Line Identity Fingerprints (CLIFP)** (chapter .6). This setting can also be regarded as a multi-task learning approach [Brown et al. (2014)].

Taken together, these validation scenarios enable a thorough and earnest validation of **PCM** models. Finally, I also suggest to calculate the statistical metrics on, at least, the predictions calculated with three models trained on different subsets of the complete data set, and to accompany them with the standard deviation observed over the repetitions [Cortes-Ciriano et al. (2014)]. The bootstrap [Efron and Tibshirani (1993)] method can also be used to estimate the standard deviation for the statistical metrics. Similarly, it is advisable to carefully estimate the maximum achievable performance given the uncertainty of the data [Cortes-Ciriano et al. (2014); Cortes-Ciriano, I et al. (2015)] (section .2.7).

.2.6.1 Statistical metrics

The statistical metrics proposed by Golbraikh and Tropsha (2002a) can be used (similar to **QSAR**) to validate models using observed and predicted values on a test (or external) set:

Internal validation (predictions on the cross-validation hold-out folds):

$$q_{\text{int}}^2 \text{ or } R_{\text{int}}^2 = 1 - \frac{\sum_{i=1}^{N_{\text{tr}}} (y_i - \tilde{y}_i)^2}{\sum_{i=1}^{N_{\text{tr}}} (y_i - \bar{y}_{\text{tr}})^2} \quad (.2.16)$$

$$\text{RMSE}_{\text{int}} = \sqrt{\frac{\sum_{i=1}^{N_{\text{tr}}} (y_i - \tilde{y}_i)^2}{N}} \quad (.2.17)$$

where N_{tr} , y_i , \tilde{y}_i and \bar{y}_{tr} represent, respectively, the size of the training set, the observed, the predicted and the averaged values of the dependent variable for those data-points included in the training set. The i th position within the training set is defined by i .

External validation (predictions on test or external sets):

$$Q_{1 \text{ test}}^2 = 1 - \frac{\sum_{j=1}^{N_{\text{test}}} (y_j - \tilde{y}_j)^2}{\sum_{j=1}^{N_{\text{test}}} (y_j - \bar{y}_{\text{tr}})^2} \quad (.2.18)$$

$$Q_{2\ test}^2 = 1 - \frac{\sum_{j=1}^{N_{test}} (y_j - \tilde{y}_j)^2}{\sum_{j=1}^{N_{test}} (y_j - \bar{y}_{test})^2} \quad (.2.19)$$

$$Q_{3\ test}^2 = 1 - \frac{[\sum_{j=1}^{N_{test}} (y_j - \tilde{y}_j)^2]/N_{test}}{[\sum_{j=1}^{N_{tr}} (y_j - \bar{y}_{tr})^2]/N_{tr}} \quad (.2.20)$$

$$RMSE_{test} = \sqrt{\frac{(y_j - \tilde{y}_j)^2}{N}} \quad (.2.21)$$

$$R_{test} = \frac{\sum_{j=1}^{N_{test}} (y_j - \bar{y}_{test})(\tilde{y}_j - \bar{\tilde{y}}_{test})}{\sqrt{\sum_{j=1}^{N_{test}} (y_j - \bar{y}_{test})^2 \sum (\tilde{y}_j - \bar{\tilde{y}}_{test})^2}} \quad (.2.22)$$

$$R_{0\ test}^2 = 1 - \frac{\sum_{j=1}^{N_{test}} (y_j - \tilde{y}_j^{r0})^2}{\sum_{j=1}^{N_{test}} (y_j - \bar{y}_{test})^2} \quad (.2.23)$$

where N_{tr} , N_{test} , y_j , \tilde{y}_j , and \bar{y}_{test} represent the size of the training and test sets, the observed, the predicted, and the averaged values of the dependent variable for those data-points comprising the test set, respectively. \bar{y}_{tr} represents the averaged values of the dependent variable for those data-points comprising the training set. The j th position within the training set is defined by j .

$R_{0\ ext}^2$ is the square of the coefficient of determination through the origin, being $\tilde{y}_j^{r0} = k\tilde{y}_j$ the regression through the origin (observed versus predicted) and k its slope. For a detailed discussion of both the evaluation of the predictive ability through the external set and about the three different formulations for Q_{ext}^2 , namely $Q_{1\ ext}^2$, $Q_{2\ ext}^2$, and $Q_{3\ ext}^2$, see Consonni, Ballabio, and Todeschini (2010). Although the predictive power of a model needs to be put into context (e.g. models with low predictive ability might be useful in hit identification, whereas not in lead optimization), it is generally acknowledged that to be considered as predictive, a model must satisfy the following criteria [Golbraikh and Tropsha (2002b); Tropsha and Gramatica (2003)]:

1. $q_{int}^2 > 0.5$
2. $R_{test}^2 > 0.6$
3. $\frac{(R_{test}^2 R_{0\ test}^2)}{R_{test}^2} < 0.1$
4. $0.85 \leq k \leq 1.15$

All these metrics can be calculated with the function *Validation* of the R package *camb* [Murrell et al. (2014)].

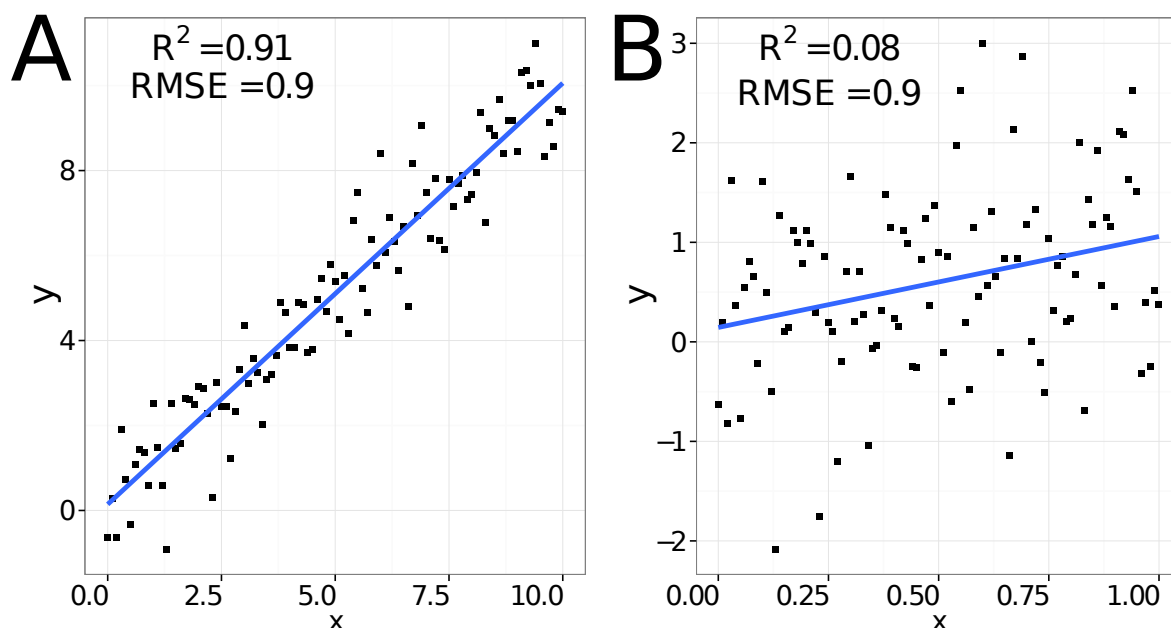


Figure .2.3: **Toy example showing the influence of the range of the response variable (e.g. bioactivities) on R^2 values.** **A.** R^2 and RMSE values of 0.91 and 0.90, respectively, are obtained when the response values range from 0 to 10 (arbitrary units). **B.** By contrast, the R^2 drops to 0.08 when the response value ranges from 0 to 1. Note that in both cases the RMSE values are the same, namely 0.90. To simulate y , random noise with mean 0 and standard deviation equal to 1 was added to x . The noise added was the exactly the same in both cases, namely (A) and (B).

It is important to note that the correlation metrics, e.g. R^2_{test} or Q^2_{test} , are very sensitive to the range of the response variable, as narrow ranges might lead to low R^2 or Q^2 values even if the predictions closely match the observed values. This effect is illustrated with a toy example in Figure .2.3. This example illustrates that low R^2 values obtained with LOTO, LOCO and, especially LOCCO, do not necessarily imply that the predictions are inaccurate. LOCCO and Leave-One-Compound-Out are particularly prone to this situation, as in many cases, the activities of a given compound across a target or cell-line panel do not cover a wide range of bioactivity values. Thus, in these cases the comparison across models should be mainly based on RMSE values, as RMSE values are not affected by the range of the response variable.

.2.7 Assessment of Maximum and Minimum Achievable Model Performance

To assess the maximum and the minimum achievable $\text{RMSE}_{\text{test}}$ and R^2_{test} values according to the experimental uncertainty of the bioactivity values (e.g. pIC_{50} or pGI_{50}), simulated data can be used in the following way:

- **Maximum performance.** A sample, A , of a size equal to the size of the test set is randomly extracted from the vector containing the whole set of bioactivity values. Subsequently, the noise corresponding to the experimental errors (uncertainty) is added to each data-point in A , thus defining the sample B . Finally, the $\text{RMSE}_{\text{test}}$ and R^2_{test} values (or the values for other metrics, e.g. Q^2_{test}) are calculated for A with respect to B . These steps are repeated a large number of times (e.g. 1,000), leading to the definition of the distributions of maximum and minimum achievable $\text{RMSE}_{\text{test}}$ and R^2_{test} values.
- **Minimum performance.** The procedure is the same as in the previous case except for the fact that sample A is randomly permuted before calculating the $\text{RMSE}_{\text{test}}$ and R^2_{test} values.

These calculations can be performed with the functions *MaxPerf* and *MinPerf* from the R package *camb* [Murrell et al. (2014)]. If the uncertainty of individual data-points is not available, estimated average uncertainties from ChEMBL for pIC_{50} and K_i data can be used instead [Kalliokoski et al. (2013); Kramer et al. (2012)] (chapter .5).

.2.8 Conformal Prediction

In the following section, the conformal prediction framework is presented. An entire section is devoted to this method given that: (i) it is used in the case study presented in chapter .6, (ii) it has been implemented in R, leading to the publication of the R package *conformal* [Cortes-Ciriano, I, Bender, A, and Malliavin (2015)].

Assessing the reliability of individual predictions is foremost in machine learning to determine the applicability domain of a predictive model whatever is the modelling task: classification or regression. The applicability domain is usually defined as the amount (or the regions) of descriptor space to which a model can be reliably applied. Conformal prediction is an algorithm-independent technique, *i.e.* it works with any predictive method such as Support Vector Machines or Random Forests, which outputs confidence regions for individual predictions in the case of regression, and P values for categories in a classification setting.

2.8.1 Regression

In the conformal prediction framework [Norinder et al. (2014); Shafer and Vovk (2008)], the data-points in the training set are used to define how unlikely a new data-point is with respect to the data presented to the model in the training phase. The *unlikelihood* (nonconformity) for a given data-point, x , with respect to the training set is quantified with a nonconformity score, α , calculated with a nonconformity measure, which here we define as:

$$\alpha = \frac{|y - \tilde{y}|}{\tilde{\rho}} \quad (.2.24)$$

where y is and \tilde{y} are respectively the observed and the predicted value calculated with a point prediction model, and $\tilde{\rho}$ is the predicted error for x calculated with an error model.

In order to calculate confidence intervals, we need a point prediction model, to predict the response variable, and an error model, to predict errors in prediction ($\tilde{\rho}$). The point prediction and error models can be generated with any machine learning algorithm. Both the point prediction and error models need to be trained with cross-validation in order to calculate the vector of nonconformity scores for the training set, $D_i = \{x_i\}_i^{N_{tr}}$.

The cross-validation predictions generated when training the point prediction model serve to calculate the errors in prediction for the data-points in the training set, $y_i - \tilde{y}_i$. The error model is then generated by training a machine learning model on the training set using these errors as the dependent variable. The (i) cross-validated predictions from the point prediction model, and (ii) the cross-validated errors in prediction from the error model, are used to generate the vector of nonconformity scores for the training set. This vector, after being sorted in increasing order, can be defined as:

$$\alpha_{tr} = \{\alpha_{tr\ i}\}_i^{N_{tr}} \quad (.2.25)$$

where N_{tr} is the number of data-points in the training set.

To generate the confidence intervals for an external set, $D_{ext} = \{x_{ext}\}_j^{N_{ext}}$, we have to define a confidence level, ϵ . The α value associated to the user-defined confidence level, α_ϵ , is calculated as:

$$\alpha_\epsilon = \alpha_{tr\ i} \text{ if } i \equiv |N_{tr} * \epsilon| \quad (.2.26)$$

where \equiv indicates equality. Next, the errors in prediction, $\tilde{\rho}_{ext}$, and the value for the response variable, y_{ext} , for the data-points in the external dataset are predicted with the error and the point prediction models, respectively.

Individual confidence intervals (CI) for each data-point in the external set are derived from:

$$\text{CI}_{\text{ext } j} = |y_{\text{ext } j} - \tilde{y}_{\text{ext } j}| = \alpha_{\epsilon} * \tilde{\rho}_{\text{ext } j} \quad (.2.27)$$

The confidence region (CR) is finally defined as:

$$\text{CR} = \tilde{y}_{\text{ext } j} + / - \text{CI}_{\text{ext } j} \quad (.2.28)$$

The interpretation of the confidence regions is straightforward. For instance, if we choose a confidence level of 80% the true value for new data-points will lie outside the predicted confidence regions in at most 20% of the cases.

.2.8.2 Classification

Initially, a Random Forest classifier is trained on the training set using k-fold cross-validation. In the case of classification, the nonconformity scores are calculated on a per class basis as the ratio between the number of trees in the forest voting for a given class divided by the total number of trees (label-wise Mondrian off-line inductive conformal prediction -MICP-) [Norinder et al. (2014)]. For instance, in a binary classification example, if 87 trees from a Random Forest model comprising 100 trees classify a data-point as belonging to class A, the nonconformity score (or probability) for this class would be 0.87 (87%), whereas its value for class B would be 0.13. This process generates a matrix (nonconformity scores matrix) which rows correspond to the data-points in the training set, and its columns to the number of distinct classes (two in the binary classification example) (Figure .2.4A). Here, we have implemented the pipeline proposed by [Norinder et al. (ibid.)] using Random Forest models. Nevertheless, other ensemble methods could be used to calculate the nonconformity scores.

Next, each column of the matrix is sorted in increasing order. These columns are called Mondrian class lists (MCL) (Figure .2.4A). As in regression, a confidence level, ϵ , needs to be specified. We define significance as $1 - \epsilon$. The model trained on the whole dataset is used to classify the data-points comprised in the external dataset (Figure .2.4). Let's consider one data-point from the external set, namely $x_{\text{ext } j}$. The number of trees in the Random Forest voting for each class is computed, which enables the calculation of the nonconformity scores or probabilities (p) for that point, $x_{\text{ext } j}$. In the binary case, this is defined as:

$$p(x_{\text{ext } j}; A) = \frac{N_{\text{trees voting A}}}{N_{\text{trees}}}; \quad p(x_{\text{ext } j}; B) = \frac{N_{\text{trees voting B}}}{N_{\text{trees}}} \quad (.2.29)$$

To calculate the P values for each class, the number of elements in the corresponding Mondrian class list smaller than the probability values, i.e. $p(x_{\text{ext } j}; A)$ and $p(x_{\text{ext } j}; B)$, is divided by the number of data-points in the training set, N_{tr} :

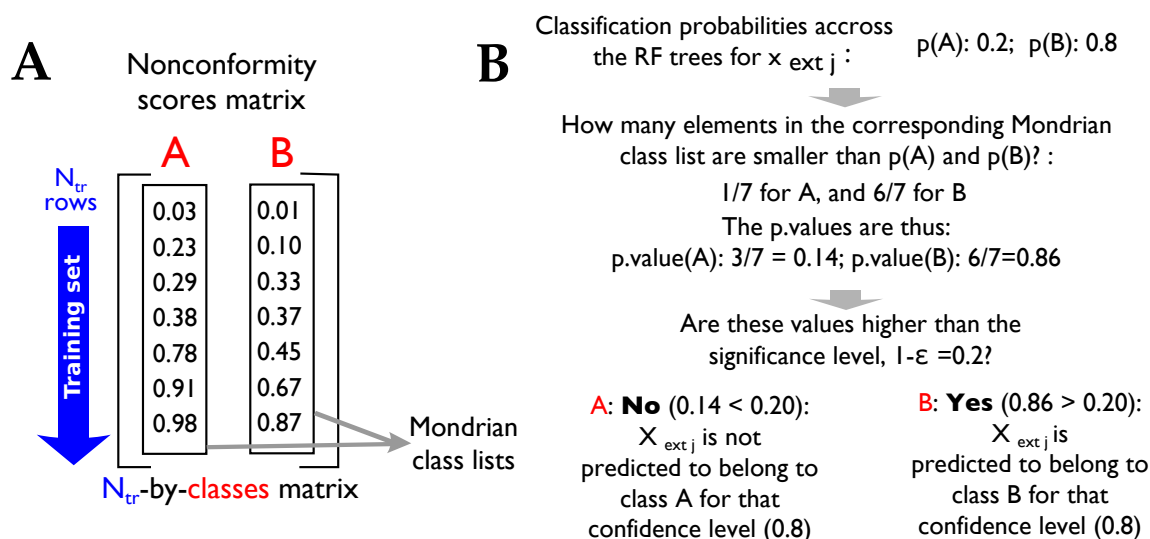


Figure .2.4: Calculation of conformal prediction errors (P values) in a binary classification example considering a confidence level of 0.80.

$$Pvalue(x_{ext j}; A) = \frac{|\{MCL(A) < P(x_{ext j}; A)\}|}{N_{tr}}$$

$$Pvalue(x_{ext j}; B) = \frac{|\{MCL(B) < P(x_{ext j}; B)\}|}{N_{tr}} \quad (.2.30)$$

Finally, these P values are compared to the significance level defined by the user ($1 - \epsilon$). For a data-point to be predicted to belong to a given class, the P value needs to be higher than the significance level. For instance, if $Pvalue(x_{ext j}; A) = 0.46$ and $Pvalue(x_{ext j}; B) = 0.18$, with a significance level of 0.2, $x_{ext j}$ would be predicted to belong to class A, but not to B. If both $Pvalue(x_{ext j}; A)$ and $Pvalue(x_{ext j}; B)$ were higher than the significance level, $x_{ext j}$ would be predicted to belong to both classes. Similarly, if both P values were smaller than the significance level, $x_{ext j}$ would be predicted to belong to neither class A nor to class B.

Bibliography

- A Mauri, VC (2006). "DRAGON software: An easy approach to molecular descriptor calculations". In: *MATCH Commun. Math. Comput. Chem.* 56, pp. 237–248 (cit. on p. 60).
- Abdi, H (2010). "Partial least squares regression and projection on latent structure regression (PLS Regression)". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.1, pp. 97–106 (cit. on p. 68).
- Abdi, H and LJ Williams (2010). "Wiley Interdisciplinary Reviews: Computational Statistics". In: *Volume 2*, pp. 433–459 (cit. on p. 69).
- Ain, QU, O Mendez-Lucio, IC Ciriano, TE Malliavin, GJP van Westen, and A Bender (2015). "Modelling ligand selectivity of serine proteases using integrative proteochemometric approaches improves model performance and allows the multi-target dependent interpretation of features". In: *Integr. Biol.* 6, pp. 24–50 (cit. on p. 58).
- Ben-Hur, A, CS Ong, S Sonnenburg, B Schölkopf, and G Rätsch (Oct. 2008). "Support Vector Machines and Kernels for Computational Biology". In: *PLoS Comput. Biol.* 4.10. Ed. by F Lewitter, e1000173 (cit. on p. 65).
- Bender, A, HY Mussa, and RC Glen (2004). "Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance". In: *J. Chem. Inf. Model.* 44.5, pp. 1708–1718 (cit. on p. 59).
- Bender, A, JL Jenkins, J Scheiber, SCK Sukuru, M Glick, and JW Davies (2009). "How similar are similarity searching methods? A principal component analysis of molecular descriptor space". In: *J. Chem. Inf. Model.* 49.1, pp. 108–119 (cit. on p. 59).
- Breiman, L (June 1998). "Arcing classifier (with discussion and a rejoinder by the author)". In: *Ann. Statist.* 26.3, pp. 801–849 (cit. on p. 67).
- (Oct. 2001). "Random Forests". en. In: *Mach. Learn.* 45.1, pp. 5–32 (cit. on p. 68).
- Breiman, L, JH Friedman, R Olshen, and C Stone (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks (cit. on p. 66).
- Brown, J, Y Okuno, G Marcou, A Varnek, and D Horvath (2014). "Computational chemogenomics: Is it more than inductive transfer?" In: *J. Comput. Aided Mol. Des.* Pp. 1–22 (cit. on p. 72).
- Consonni, V, D Ballabio, and R Todeschini (2010). "Evaluation of model predictive ability by external validation techniques". In: *J. Chemometrics* 24.3-4, pp. 194–201 (cit. on p. 73).
- Cortes, C and V Vapnik (1995). "Support-vector networks". In: *Mach. Learn.* 20.3, pp. 273–297 (cit. on p. 65).

- Cortes-Ciriano, I (2013). *FingerprintCalculator*. URL: <http://github.com/isidro/FingerprintCalculator> (cit. on pp. 59, 60).
- Cortes-Ciriano, I, GJP van Westen, EB Lenselink, DS Murrell, A Bender, and TE Malliavin (2014). "Proteochemometric modeling in a Bayesian framework". In: *J. Cheminf.* 6.1, p. 35 (cit. on pp. 69, 70, 72).
- Cortes-Ciriano, I, DS Murrell, GJP van Westen, A Bender, and TE Malliavin (2015). "Prediction of the Potency of Mammalian Cyclooxygenase Inhibitors with Ensemble Proteochemometric Modeling". In: *J. Cheminf.* 7, p. 1 (cit. on pp. 58, 60, 66).
- Cortes-Ciriano, I, Bender, A, and TE Malliavin (2015). "Prediction of PARP inhibition with Proteochemometric modelling and conformal prediction". In: *In revision at Mol. Inform.* URL: <http://cran.r-project.org/package=conformal> (cit. on p. 75).
- Cortes-Ciriano, I, van Westen, G J P, Bouvier, G, Nilges, M, Overington, J P, Bender, A, and TE Malliavin (2015). "Improved Large-scale prediction of growth inhibition patterns on the NCI60 cancer cell-line panel". In: *In revision Bioinformatics* (cit. on pp. 58, 60, 72).
- De Bruyn, T, GJP van Westen, AP IJzerman, B Stieger, P de Witte, PF Augustijns, and PP Annaert (2013). "Structure-Based Identification of OATP1B1/3 Inhibitors". In: *Mol. Pharmacol.* 83.6, pp. 1257–1267 (cit. on pp. 58, 60).
- Desdouits, N, M Nilges, and A Blondel (2015). "Principal Component Analysis reveals correlation of cavities evolution and functional motions in proteins". In: *J. Mol. Graph. Model.* 55.0, pp. 13 –24 (cit. on p. 58).
- Doddareddy, MR, GJP van Westen, E van der Horst, JE Peironcelly, F Corthals, AP IJzerman, M Emmerich, JL Jenkins, and A Bender (2009). "Chemogenomics: Looking at biology through the lens of chemistry". In: *Stat. Anal. Data Min.* 2.3, pp. 149–160 (cit. on p. 60).
- Efron, B and R Tibshirani (1993). *An introduction to the bootstrap*. New York : Chapman & Hall (cit. on p. 72).
- Genton, MG (2002). "Classes of kernels for machine learning: a statistics perspective". In: *J. Mach. Learn. Res.* 2, pp. 299–312 (cit. on p. 63).
- Georgiev., AG (2009). In: *J. Comp. Biol.* 16 (5), pp. 703–723 (cit. on pp. 56, 57).
- Glenn, RC, A Bender, CH Arnby, L Carlsson, S Boyer, and J Smith (2006). "Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME". In: *I. Drugs* 9.3, pp. 199–204 (cit. on p. 59).
- Gloriam, DE, SM Foord, FE Blaney, and SL Garland (2009). "Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design". In: *J. Med. Chem.* 52.14, pp. 4429–4442 (cit. on p. 58).
- Golbraikh, A and A Tropsha (2002a). "Beware of q2!" In: *J. Mol. Graphics Modell.* 20.4, pp. 269–276 (cit. on p. 72).
- (2002b). "Beware of q2!" In: *J. Mol. Graphics Modell.* 20.4, pp. 269–276 (cit. on p. 73).

- Gregori-Puigjané, E and J Mestres (2008a). "A ligand-based approach to mining the chemogenomic space of drugs". In: *Comb. Chem. High Throughput Screen.* 11.8, pp. 669–676 (cit. on pp. 69, 70).
- (2008b). "Coverage and bias in chemical library design". In: *Curr. Opin. Chem. Biol.* 12.3, pp. 359–365 (cit. on pp. 69, 70).
- Hastie, T, R Tibshirani, and J Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc. (cit. on pp. 66–68).
- Hawkins, DM, SC Basak, and D Mills (Mar. 2003). "Assessing model fit by Cross-Validation". In: *J. Chem. Inf. Model.* 43.2, pp. 579–586 (cit. on p. 62).
- Horst, E van der, JE Peironcelly, GJP van Westen, OO van den Hoven, WRJD Galloway, DR Spring, JK Wegner, HWT van Vlijmen, AP IJzerman, JP Overington, and A Bender (2011). "Chemogenomics approaches for receptor deorphanization and extensions of the chemogenomics concept to phenotypic space". In: *Curr. Top. Med. Chem.* 11.15, pp. 1964–1977 (cit. on p. 58).
- InChI (2013). *IUPAC - International Union of Pure and Applied Chemistry: The IUPAC International Chemical Identifier (InChI)*. URL: <http://www.iupac.org/home/publications/e-resources/inchihtml> (cit. on p. 54).
- Indigo (2013). "Indigo Cheminformatics Library". In: URL: <http://ggasoftwarecom/opensource/indigo/> (cit. on p. 54).
- J E S Wikberg, ML (2004). *Chemogenomics in Drug Discovery*. Ed. by H Kubinyi and G Müller. Methods and Principles in Medicinal Chemistry. Weinheim, FRG: Wiley-VCH Verlag GmbH & Co KGaA, pp. 289–309 (cit. on p. 55).
- Kalinina, O and O Wichmann (2011). "Combinations of protein-chemical complex structures reveal new targets for established drugs". In: *PLoS Comput. Biol.* 7.5, e1002043 (cit. on p. 58).
- Kalliokoski, T, C Kramer, A Vulpetti, and P Gedeck (2013). "Comparability of mixed IC₅₀ data - a statistical analysis". In: *PLoS One* 8.4, e61007 (cit. on p. 75).
- Karelson, M (2000). *Molecular descriptors in QSAR/QSPR, Volume 1*. Wiley-Interscience, p. 430 (cit. on p. 59).
- Kinnings, SL and RM Jackson (2009). "Binding site similarity analysis for the functional classification of the protein kinase family". In: *J. Chem. Inf. Model.* 49.2, pp. 318–329 (cit. on p. 58).
- Koutsoukas, A, R Lowe, Y KalantarMotamedi, HY Mussa, W Klaffke, JBO Mitchell, RC Glen, and A Bender (2013). "In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt Window". In: *J. Chem. Inf. Model.* 53.8, pp. 1957–1966 (cit. on p. 59).
- Kramer, C, T Kalliokoski, P Gedeck, and A Vulpetti (2012). "The experimental uncertainty of heterogeneous public K(i) data". In: *J. Med. Chem.* 55.11, pp. 5165–5173 (cit. on p. 75).

- Krstajic, D, LJ Buturovic, DE Leahy, and S T (2014). "Cross-validation pitfalls when selecting and assessing regression and classification models". In: *J. Cheminf.* 6.1, p. 10 (cit. on pp. 62, 63).
- Kruger, FA and JP Overington (2012). "Global Analysis of Small Molecule Binding to Related Protein Targets". In: *PLoS Comput. Biol.* 8.1, e1002333 (cit. on p. 57).
- Kufareva, I, AV Ilatovskiy, and R Abagyan (2012). "Pocketome: an encyclopedia of small-molecule binding sites in 4D". In: *Nucleic Acids Res.* 40.Database issue, pp. D535–40 (cit. on p. 58).
- Kuhn, D, N Weskamp, E Hüllermeier, and G Klebe (2007). "Functional classification of protein kinase binding sites using Cavbase". In: *ChemMedChem* 2.10, pp. 1432–1447 (cit. on p. 57).
- Kuhn, M (2008). "Building predictive models in R using the caret package". In: *J. Stat. Softw.* 28.5, pp. 1–26 (cit. on p. 62).
- Kuhn, M and K Json (2013). *Applied Predictive Modeling*. New York, NY: Springer New York (cit. on p. 62).
- Lapinsh, M, S Veiksina, S Uhlén, R Petrovska, I Mutule, F Mutulis, S Yahorava, P Prusis, and JES Wikberg (2005). "Proteochemometric mapping of the interaction of organic compounds with melanocortin receptor subtypes". In: *Mol. Pharmacol.* 67.1, pp. 50–59 (cit. on pp. 55, 57, 60).
- Liang, G and Z Li (2007). "Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides". In: *QSAR Comb. Sci.* 26.6, pp. 754–763 (cit. on pp. 56, 57).
- MacKay, DJC (2003). *Information Theory, Inference and Learning Algorithms*. en. Cambridge University Press (cit. on pp. 64, 65).
- Mei, H, ZH Liao, Y Zhou, and SZ Li (2005). "A new set of amino acid descriptors and its application in peptide QSARs". In: *Biopolymers* 80.6, pp. 775–786 (cit. on pp. 56, 57).
- Menden, MP, F Iorio, MJ Garnett, U McDermott, CH Benes, PJ Ballester, and J Saez-Rodriguez (2013). "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties". In: *PLoS One* 8.4, e61318 (cit. on p. 58).
- Mente, S and M Kuhn (2012). "The use of the R language for medicinal chemistry applications". In: *Curr. Top. Med. Chem.* 12.18, pp. 1957–1964 (cit. on p. 53).
- Meslamani, J and D Rognan (2011). "Enhancing the Accuracy of Chemogenomic Models with a Three-Dimensional Binding Site Kernel". In: *J. Chem. Inf. Model.* 51.7, pp. 1593–1603 (cit. on p. 58).
- Mestres, J, E Gregori-Puigjané, S Valverde, and RV Solé (2009). "The topology of drug-target interaction networks: implicit dependence on drug properties and target families". In: *Mol. BioSyst.* 5.9, pp. 1051–1057 (cit. on p. 58).
- Murrell, DS, I Cortes-Ciriano, GJP van Westen, IP Stott, TE Malliavin, A Bender, and RC Glen (2014). "Chemistry Aware Model Builder (camb): an R Package for

- Predictive Bioactivity Modeling". In: <http://github.com/cambDI/camb> (cit. on pp. 53, 54, 57, 60, 62, 73, 75).
- Natekin, A and A Knoll (2013). "Gradient boosting machines, a tutorial". In: *Front. Neurobot.* 7, p. 21 (cit. on pp. 67, 68).
- Norinder, U, L Carlsson, S Boyer, and M Eklund (2014). "Introducing Conformal Prediction in Predictive Modeling A Transparent and Flexible Alternative To Applicability Domain Determination". In: *J. Chem. Inf. Model.* 54.6, pp. 1596–1603 (cit. on pp. 76, 77).
- Pahikkala, T, A Airola, S Pietilä, S Shakyawar, A Szwajda, J Tang, and T Aittokallio (2014). "Toward more realistic drug-target interaction predictions". In: *Brief. Bioinform.* (Cit. on p. 62).
- Paricharak, S, I Cortes-Ciriano, AP IJzerman, TE Malliavin, and A Bender (2015). "Proteochemometric modeling coupled to in silico target prediction: an integrated approach for the simultaneous prediction of polypharmacology and binding affinity of small molecules". In: *J. Cheminf.* 7, p. 15 (cit. on p. 58).
- Park, Y and EM Marcotte (2012). "Flaws in evaluation schemes for pair-input computational predictions". In: *Nat. Methods* 9.12, pp. 1134–1136 (cit. on pp. 62, 69).
- Prusis, P, S Uhlén, R Petrovska, M Lapinsh, and JES Wikberg (2006). "Prediction of indirect interactions in proteins". In: *BMC Bioinformatics* 7.1, p. 167 (cit. on p. 60).
- Puntanen, S and GPH Styan (Jan. 2005). "Schur complements in statistics and probability". In: *The Schur Complement and Its Applications*. Ed. by F Zhang. Numerical Methods and Algorithms 4. Springer US, pp. 163–226 (cit. on p. 65).
- Rao, HB, F Zhu, GB Yang, ZR Li, and YZ Chen (2011). "Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence". In: *Nucleic Acids Res.* 39.suppl, W385–W390 (cit. on p. 57).
- Rasmussen, CE and CKI Ws (2006). *Gaussian Processes for machine learning*. MIT Press (cit. on p. 65).
- Rogers, D and M Hahn (2010). "Extended-connectivity fingerprints". In: *J. Chem. Inf. Model.* 50.5, pp. 742–754 (cit. on p. 59).
- Rognan, D (2007). "Chemogenomic approaches to rational drug design". In: *Br. J. Pharmacol.* 152.1, pp. 38–52 (cit. on pp. 69, 70).
- Sandberg, M, L Eriksson, J Jonsson, M Sjöström, and S Wold (1998). "New chemical descriptors relevant for the design of biologically active peptides A multivariate characterization of 87 amino acids". In: *J. Med. Chem.* 41.14, pp. 2481–2491 (cit. on pp. 56, 57).
- Scholkopf, B, T Koji, and JP Vert (2004). *Kernel Methods in Computational Biology*. MIT Press, p. 400 (cit. on p. 64).
- Shafer, G and V Vovk (2008). "A tutorial on conformal prediction". In: *JMLR* 9, pp. 371–421 (cit. on p. 76).

- Sheinerman, FB, E Giraud, and A Laoui (2005). "High affinity targets of protein kinase inhibitors have similar residues at the positions energetically important for binding". In: *J. Mol. Biol.* 352.5, pp. 1134–1156 (cit. on p. 57).
- Shen, J, J Zhang, X Luo, W Zhu, K Yu, K Chen, Y Li, and H Jiang (2007). "Predicting protein-protein interactions based only on sequences information". In: *Proc. Natl. Acad. Sci. U. S. A.* 104.11, pp. 4337–4341 (cit. on p. 57).
- Sheridan, RP (2012). "Three useful dimensions for domain applicability in QSAR models using random forest". In: *J. Chem. Inf. Model.* 52.3, pp. 814–823 (cit. on p. 68).
- (2013). "Using Random Forest to model the domain applicability of another Random Forest model". In: *J. Chem. Inf. Model.* 53.11, pp. 2837–2850 (cit. on p. 68).
- Sievers, F, A Wilm, D Dineen, TJ Gibson, K Karplus, W Li, R Lopez, H McW, M Remmert, J Söding, JD Thompson, and DG Higgins (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega". In: *Mol. Syst. Biol.* 7.1, p. 539 (cit. on p. 57).
- Subramanian, V, P Prusis, L Pietila, H Xhaard, and G Wohlfahrt (2013). "Visually Interpretable Models of Kinase Selectivity Related Features Derived from Field-Based Proteochemometrics". In: *J. Chem. Inf. Model.* 53.11, pp. 3021–3030 (cit. on p. 58).
- Surgand, JS, J Rodrigo, E Kellenberger, and D Rognan (2006). "A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors". In: *Proteins* 62.2, pp. 509–538 (cit. on p. 57).
- Tian, F, P Zhou, and Z Li (2007). "T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides". In: *Journal of Molecular Structure* 830.1-3, pp. 106–115 (cit. on pp. 56, 57).
- Tipping, ME (2000). *The Relevance Vector Machine* (cit. on p. 66).
- Todeschini, R and V Consonni (2008). *Handbook of Molecular Descriptors*. John Wiley & Sons, p. 688 (cit. on p. 59).
- Tropsha, A and VK Gramatica PaOand Gombar (2003). "The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models". In: *QSAR Comb. Sci.* 22.1, pp. 69–77 (cit. on p. 73).
- Varma, S and R Simon (2006). "Bias in error estimation when using cross-validation for model selection". In: *BMC Bioinformatics* 7.1, p. 91 (cit. on p. 62).
- Wang, Y, J Xiao, TO Suzek, J Zhang, J Wang, Z Zhou, L Han, K Karapetyan, S Dracheva, BA Shoemaker, E Bolton, A Gindulyte, and SH Bryant (2012). "PubChem's BioAssay Database". In: *Nucleic Acids Res.* 40.Database issue, pp. 400–412 (cit. on p. 54).
- Weill, N and D Rognan (2009). "Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands". In: *J. Chem. Inf. Model.* 49.4, pp. 1049–1062 (cit. on p. 58).
- Weill, N, C Valencia, S Gioria, P Villa, M Hibert, and D Rognan (2011). "Identification of Nonpeptide Oxytocin Receptor Ligands by Receptor-Ligand Fingerprint Similarity Search". In: *Mol. Inf.* 30.6-7, pp. 521–526 (cit. on p. 58).

- Westen, GJP van, JJ Wegner, AP IJzerman, HWT van Vlijmen, and A Bender (2011). "Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets". In: *Med. Chem. Comm.* 2.1, pp. 16–30 (cit. on pp. 55, 62).
- Westen, GJP van, OO van den Hoven, R van der Pijl, T Mulder-Krieger, H de Vries, AP Wegner Jorg K an IJzerman, HWT van Vlijmen, and A Bender (Aug. 2012a). "Identifying novel adenosine receptor ligands by simultaneous proteochemometric modeling of rat and human bioactivity data". In: *J. Med. Chem.* 55.16, pp. 7010–7020 (cit. on p. 57).
- Westen, GJP van, OOvd Hoven, Rvd Pijl, T Mulder-Krieger, and A Bender (2012b). "Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data". In: *J. Med. Chem.* 55.16, pp. 7010–7020 (cit. on pp. 58, 60).
- Westen, GJP van, R Swier, JK Wegner, AP IJzerman, HWT van Vlijmen, and A Bender (2013a). "Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets". In: *J. Cheminf.* 5.1, p. 41 (cit. on pp. 56, 57).
- Westen, GJP van, RF Swier, I Cortes-Ciriano, JK Wegner, JP Overington, AP IJzerman, HWT van Vlijmen, and A Bender (2013b). "Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets". In: *J. Cheminf.* 5.1, p. 42 (cit. on pp. 56, 57, 70).
- Willighagen, EL, J Alvarsson, A Andersson, M Eklund, S Lampa, M Lapinsh, O Spjuth, and JES Wikberg (2011). "Linking the Resource Description Framework to cheminformatics and proteochemometrics". In: *J. Biomed. Semant.* 2 Suppl 1, pp. 1–24 (cit. on p. 58).
- Wold, S, M Sjöström, and L Eriksson (2001). "PLS-regression: a basic tool of chemometrics". In: *Chemometr. Intell. Lab.* 58.2, pp. 109–130 (cit. on p. 68).
- Xiao, N and Q Xu (2014). *protr: Protein sequence descriptor calculation and similarity computation with R*. R package version 2-1. URL: <http://cran.r-project.org/web/packages/protr/index.html> (cit. on p. 59).
- Yabuuchi, H, S Niiijima, H Takematsu, T Ida, T Hirokawa, T Hara, T Ogawa, Y Minowa, G Tsujimoto, and Y Okuno (2011). "Analysis of multiple compound-protein interactions reveals novel bioactive molecules". In: *Mol. Syst. Biol.* 7, pp. 472–484 (cit. on p. 58).
- Yang, L, M Shu, K Ma, H Mei, Y Jiang, and Z Li (2010). "ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues." In: *Amino acids* 38, pp. 805–816 (cit. on pp. 56, 57).
- Yap, CW (2011). "PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints". In: *J. Comput. Chem.* 32.7, pp. 1466–1474 (cit. on p. 60).
- Zaliani, A and E Gancia (1999). "MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies". In: *J. Chem. Inform. Comput. Sci.* 39.3, pp. 525–533 (cit. on pp. 56, 57).

Proteochemometric Modelling in a Bayesian Framework

.3 Proteochemometric Modelling in a Bayesian Framework

.3.1 Introduction

THE applicability domain (AD) of a bioactivity model is defined as the range of chemical (and target in PCM) space to which the model can be reliably applied [Bosnić and Kononenko (2009); Netzeva et al. (2005); Tetko et al. (2006)]. Therefore, the AD is a measure of the generalization properties of a given model: the volume of chemical and target space that can be reliably predicted [Sahigara et al. (2012)]. Given that compounds are encoded with descriptors when training predictive models, it is important to distinguish between the chemical space (referring to chemical structures) and the chemical descriptor space. This distinction is important as in the calculation of some popular descriptors (*e.g.* Morgan fingerprints [Rogers and Hahn (2010)]), chemical substructures are hashed: different chemical substructures are mapped at the same descriptor position. Consequently, two different structures in the chemical space can be represented by the same descriptor values. A detailed discussion of the different methods proposed to assess models AD can be found in [Sahigara et al. (2012)], to which the interested reader is referred. In PCM, the AD is an essential feature, as extrapolation has to be used to predict the bioactivity for *new* chemicals on *new* targets [Westen et al. (2011a)].

In parallel to the concern about the evaluation of individual bioactivity predictions, recent publications have aimed at establishing the level of uncertainty in public bioactivity databases [Kalliokoski et al. (2013); Kramer and L (2012); Kramer et al. (2012); Tiikkainen et al. (0)]. In this vein, Brown, Muchmore, and Hajduk (2009) highlighted the importance of including the uncertainty of bioactivity data into the evaluation of models quality. Hence, predictive models should be assessed through: the analysis of the experimental error of the data, the evaluation of the models AD as well as the definition of intervals of confidence for the predictions. However, acceptable levels of prediction errors are also determined by the context in which the model will be applied. Indeed, models exhibiting high prediction errors can be nevertheless useful in a high-throughput (HTS) campaign while not being suitable in lead optimization [Brown, Muchmore, and Hajduk (*ibid.*)].

Bayesian inference provides a reliable theoretical framework to handle all previously mentioned aspects, *i.e.* AD and uncertainty on bioactivity data) within a unique bioactivity model. Gaussian Processes (GP) are a non-parametric machine learning method based upon Bayesian inference: they thus permit an evaluation of the AD of a given model as well as providing the most objective estimation of the predictions uncertainty. Furthermore, the experimental bioactivity error can be used as model input. A GP prediction of a given compound-target combination is a Gaussian distribution whose variance defines intervals of confidence: in principle, this variance measures the distance of the compound-target pair to the training set. GP models can be globally validated by traditional statistical metrics (*e.g.* R^2 or Q^2) [Golbraikh and Tropsha (2002); Tropsha and Golbraikh (2010); Tropsha and Gramatica (2003)] while also providing individual assessment for predictions. GP were firstly introduced in the field of QSAR modelling by Burden (2001). Later on, GP were also used for: (i) the modelling of ADMET properties [Obrezanova and Segall (2010); Obrezanova et al. (2007)], (ii) the prediction of electrolyte solubility [Schwaighofer et al. (2007)], (iii) the bioactivity prediction of small peptide data sets [Ren et al. (2011); Zhou et al. (2008, 2010)], (iv) protein engineering [Romero, Krause, and Arnold (2013)], and (v) the bioactivity prediction of bioactivity-focused (GPCRs) combinatorial chemolibraries [Reutlinger et al. (2014)].

The purpose of this chapter is to propose Gaussian Process (GP) to simultaneously model chemical and multispecies protein information in the frame of PCM. GP models are validated by comparing their performance to that of SVM using a panel of kernels. on two PCM data sets extracted from ChEMBL database [Gaulton et al. (2011)], involving adenosine receptors (10,999 data points, 8 sequences) and aminergic GPCRs (24,593 data points, 91 sequences), and on a third data set extracted from the literature concerning the catalytic activity of four dengue virus NS3 proteases (199 data points, 4 sequences).

.3.2 Materials and Methods

.3.2.1 Data sets

Aminergic GPCRs

The aminergic GPCRs data set was assembled by gathering bioactivity information of 91 different receptors (9 species) from ChEMBL 15 [Gaulton et al. (*ibid.*)], producing a total number of data-points of 24,593. A high quality bioactivity data set was assembled by keeping only assay-independent bioactivity information, namely: the constant of inhibition, K_i , and the constant of dissociation, K_d . In those cases where

a given compound-target pair had multiple bioactivity values annotated, the mean value was used. Moreover, annotations with anything other than '=' were discarded. Agonist, antagonist and partial agonist ligands were included. Bioactivity values in the data set range from 2.030 to 11.570 pK_i units. The component amino acids of the transmembrane binding site were taken from Gloriam et al. (2009) Further information about the data set can be found in Table .3.1 and Table .3.3.

Adenosine receptors

This data set previously published by Westen et al. (2012) is composed of 10,999 bioactivity data points measured on the rat and human adenosine receptors, A₁, A_{2A}, A_{2B} and A₃. The data set was extracted from ChEMBL2 [Gaulton et al. (2011)]. Only compounds tested on rat or human receptors by radio-ligand binding assays and for which pK_i bioactivity values were annotated with a '=' relationship were included in the final data set. Bioactivity values range from 4.50 to 10.52 pK_i units. Compounds were normalized and ionized at pH 7.4. Subsequently, they were assigned 2D coordinates and converted to fingerprints. See Table .3.1 for further details about the data set.

Dengue virus NS3 proteases

This data set was collected from the proteochemometric study published by Prusis et al. (2008), which modeled the catalytic activity of the dengue virus NS3 proteases from four viral serotypes using data-points measured on 56 different tetra-peptide substrates (Table .3.1). These substrates were designed to evaluate the role amino acid residues located at P1'-P4' in the sequence. The catalytic efficiency was measured as the turnover number (k_{cat}) for the cleavage of the substrate. In contrast to the two data sets presented above, the number of data points in this case was only 199.

.3.2.2 Descriptors

Chemical compounds were described by Scitegic circular fingerprints (Extended Connectivity Fingerprints (ECFP)₆ type) [Glen et al. (2006); Rogers and Hahn (2010)], calculated in PipelinePilot 8.5.0.200 [Scitegic Accelrys Software Inc. Pipeline Pilot Student Edition, version 615 (San Diego, USA): Scitegic Accelrys Software Inc (2007)]. For the calculation of keyed ECFP₆ fingerprints, each compound substructure, with a maximal diameter of three bonds, is treated as a compound feature. The substructures are then mapped into an unhashed array of counts, thus enabling the estimation of their contribution to bioactivity. The efficiency of these fingerprints to identify chemical features relevant for bioactivity has been previously demonstrated [Bender et al. (2009); Westen et al. (2012)].

	Adenosine Receptors	Dengue Virus NS3 Proteases	Aminergic GPCRs
Data-points	10,999	199	24,593
Sequences	8	4	91
Ligands	4419	56	11,121
Source Organisms	<i>H. sapiens</i> and <i>Rattus norvegicus</i>	<i>dengue virus</i>	<i>H. sapiens</i> , <i>Rattus norvegicus</i> , <i>Mus musculus</i> , <i>Bos taurus</i> , <i>Sus scrofa</i> , <i>Canis familiaris</i> , <i>Cavia porcellus</i> , <i>Chlorocebus aethiops</i> , and <i>Mesocricetus auratus</i>
Bioactivity	pK _i	K _{cat}	pK _i
Matrix Completeness (%)	31.11	88.84	2.43

Table .3.1: **Overview of the proteochemometric data sets modeled in this work.**

Whereas the completeness of the compound-target interaction matrix of the dengue virus NS3 proteases data set is almost complete (88.84%), the adenosine receptors and GPCRs data set are more challenging to model given: (i) their sparsity (31.11 and 2.43% of matrix completeness respectively), and (ii) the consideration of information from human orthologues, being the respective number of different sequences 8 and 91.

Pairwise compound similarity plots were calculated in R using the *vegan* package [Oksanen et al. (2013)]. Protein amino acids of the GPCRs and adenosine receptors binding sites, as well as the dengue virus NS3 proteases substrates, were described with five amino acid extended principal property scales (5 z-scales). The property calculation was conducted in R [R Core Team (2013)] *via* in-house scripts following the work of Sandberg et al. (1998) In the GPCRs data set a descriptor accounting for the amino acids side chain charge at pH 7.4 was also added (with values of: +1 if the charge is positive, -1 if negative and 0 for neutral amino acid). The four dengue virus NS3 protease variants were described with binary descriptors.

.3.2.3 Modelling with Bayesian inference

Gaussian Processes

Given a data set $D = \{\mathbf{X}, \mathbf{y}\}$ where $\mathbf{X} = \{\mathbf{x}^i\}_{i=1}^n$ is the set of compound and target descriptors, and $\mathbf{y} = \{y^i\}_{i=1}^n$ is the vector of observed bioactivities, the aim is to find a Gaussian Process [Rasmussen and Ws (2006)], $\text{GP}(\mathbf{x})$, capable to infer the relationships within D , in order to predict the bioactivity y_{new} for new compound-target combinations \mathbf{x}_{new} . In the frame of Bayesian inference, GP are defined as:

$$P(\text{GP}(\mathbf{x})|D) \propto P(\mathbf{y}|\text{GP}(\mathbf{x}), \mathbf{X}) P(\text{GP}(\mathbf{x})) \quad (.3.1)$$

where: (i) $P(\text{GP}(\mathbf{x})|D)$ is the *posterior* probability distribution giving the bioactivity predictions, (ii) the likelihood $P(\mathbf{y}|\text{GP}(\mathbf{x}), \mathbf{X})$ is the probability of the observations, \mathbf{y} , given the training set, \mathbf{X} and the model $\text{GP}(\mathbf{x})$, and (iii) $P(\text{GP}(\mathbf{x}))$ is the *prior* probability distribution of the functions $\text{GP}(\mathbf{x})$ candidates to model the data set D .

The *prior* probability distribution is updated with the information contained in D via the likelihood, leading to the definition of the *posterior* probability distribution as the set of functions efficiently modelling D . The average of the *posterior* distribution is considered as the bioactivity prediction (Figure .1.2 at page 18). $\text{GP}(\mathbf{x})$ is a random function which functional values follow a centered Gaussian distribution for any set of data-points. Thus, the $P(\text{GP}(\mathbf{x}))$ values for a finite subset of compound-target vectors $\mathbf{x}_i, \dots, \mathbf{x}_n$ follow a multidimensional normal distribution with mean μ (normally set to zero) and covariance matrix \mathbf{C}_X :

$$\text{GP}(\mathbf{x}) \sim \mathcal{N}(0, \mathbf{C}_X + \sigma_d^2 \delta(\mathbf{x}_j, \mathbf{x}_k)) \quad (j, k \in 1, \dots, n) \quad (.3.2)$$

where $\delta(\mathbf{x}_j, \mathbf{x}_k)$ is the Kronecker delta function and σ_d^2 is the noise of the data-points (experimental error), which is assumed to be normally distributed with mean zero. The value of σ_d^2 accounts for the noise in the observed bioactivities, $\mathbf{y} = \text{GP}(\mathbf{x}) + \mathcal{N}(0, \sigma_d^2)$ which in turn reflects the trade-off between the quality and smoothness of the fitting.

\mathbf{C}_X is obtained by applying a positive definite kernel function (also known as *statistic* covariance) [Genton (2002)] to \mathbf{X} , $\mathbf{C}_X = \text{Cov}(\mathbf{X})$. Owing to the fact that the covariance function is based upon dot products, the *kernel trick* can be applied in a similar way as in SVM [Ben-Hur et al. (2008)]. Kernel parameters are called hyperparameters since their values define the probability of each function of the *prior* probability distribution. The different kernels implemented in this study are listed in Table .2.2.

Bioactivity prediction for new data-points

The bioactivity, y_{new} , of a new compound-target combination, \mathbf{x}_{new} , can be predicted from the joint prior probability distribution $P = \begin{pmatrix} \mathbf{y} \\ y_{\text{new}} \end{pmatrix}$ of \mathbf{y} and y_{new} , due to the multivariate Gaussian distribution assumed for \mathbf{y} :

$$\begin{bmatrix} \mathbf{y} \\ y_{\text{new}} \end{bmatrix} \sim \mathcal{N} \left(0, \mathbf{C}_X = \begin{bmatrix} \mathbf{C}_X = \text{Cov}(\mathbf{X}), & \mathbf{k} = \text{Cov}(\mathbf{X}, \mathbf{x}_{\text{new}}) \\ \mathbf{k}^T, & m = \text{Cov}(\mathbf{x}_{\text{new}}, \mathbf{x}_{\text{new}}) \end{bmatrix} \right) \quad (.3.3)$$

where \mathbf{k}^T is the transpose of the matrix \mathbf{k} , which describes the similarity between \mathbf{X} and \mathbf{x}_{new} . The predicted bioactivity is obtained as the mean value of the probability:

$$P(y_{\text{new}} | \mathbf{x}_{\text{new}}, \mathbf{D}, \mathbf{y}) \quad (.3.4)$$

and the uncertainty of the prediction corresponds to the standard deviation of this probability distribution.

To calculate $P(y_{\text{new}} | \mathbf{x}_{\text{new}}, \mathbf{D}, \mathbf{y})$, the joint probability distribution, $P(\begin{smallmatrix} \mathbf{y} \\ y_{\text{new}} \end{smallmatrix})$, is divided by the probability of the observed bioactivities, $P(\mathbf{y})$. Subsequently, the predicted probability for y_{new} is obtained by calculating the Schur complement [Puntanen and Styan (2005)]:

$$P(y_{\text{new}}) \sim N(\mu_{y_{\text{new}}} = \mathbf{k}^T \mathbf{C}_X^{-1} \mathbf{y}, \sigma_{y_{\text{new}}}^2 = m - \mathbf{k}^T \mathbf{C}_X^{-1} \mathbf{k}) \quad (.3.5)$$

where the best estimate for the bioactivity of \mathbf{x}_{new} is the average value of y_{new} , $\mu_{y_{\text{new}}} = \langle P(y_{\text{new}}) \rangle$, $\sigma_{y_{\text{new}}}$, the standard deviation, being its uncertainty.

As can be seen in Eq. 3.5, those compound-target combinations in \mathbf{X} similar to \mathbf{x}_{new} , contribute more to the prediction of y_{new} , as \mathbf{y} is weighted by \mathbf{k}^T . This means that GP, as a kernel method, mainly infers the value of y_{new} from the most similar compound-target combinations in descriptor space present in the training set, \mathbf{X} .

On the other hand, the predicted variance, $\sigma_{y_{\text{new}}}^2$, is equal to the difference between the *a priori* knowledge about \mathbf{x}_{new} : $m = \text{Cov}(\mathbf{x}_{\text{new}}, \mathbf{x}_{\text{new}})$, and what can be inferred about \mathbf{x}_{new} from similar compound-target combinations present in \mathbf{X} : $\mathbf{k}^T \mathbf{C}_X^{-1} \mathbf{k}$.

Thus, in the case of \mathbf{x}_{new} being similar to the compound-target combinations in \mathbf{X} , the value of $\sigma_{y_{\text{new}}}^2$ is small. By contrast, a high value of $\sigma_{y_{\text{new}}}^2$ indicates that \mathbf{x}_{new} is not similar (is distant) to the compound-target combinations in \mathbf{X} . In that case, the GP cannot learn much about \mathbf{x}_{new} from the training set, so the prediction should be considered as less reliable. Consequently, $\sigma_{y_{\text{new}}}^2$ gives an idea of the applicability domain

(AD) of the model and thus serves to evaluate the uncertainty of the prediction.

.3.2.4 Computational details

Determining the kernel hyperparameters

As previously stated (Equation .3.2), the prior distribution of a GP is mainly defined by its covariance, C_X , which is in turn characterized by its hyperparameter values. For the simplest kernel, Radial Basis function kernel (RBF), also known as Squared Exponential or simply Radial (Table .2.2), the hyperparameters are $(\Omega = \{l, \sigma_d^2\})$ where l are the length scales, (one per descriptor) and σ_d^2 the noise variance. In this case, the covariance between two input vectors can be defined as:

$$\text{Cov}(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{1}{2} \sum_{p=1}^P \frac{(\mathbf{x}_p^i - \mathbf{x}_p^j)^2}{l_p^2}} \quad (.3.6)$$

where p is the descriptor index and P the total number of descriptors. Each length scale, l , is treated as a hyperparameter with value needs to be optimized during model training. High length scale values will be assigned to irrelevant features for the model. Therefore, the inverse of the optimized l value obtained for a given descriptor gives an idea of its importance for the model. This inherent ability of Bayesian inference to infer the relevance of each descriptor is the so-called Automatic Relevance Determination (ARD) [Rasmussen and Ws (2006)]. In the context of PCM, ARD can be exploited to provide a biologically meaningful interpretation of the models.

In the frame of Bayesian inference, we search for the hyperparameter values maximizing the probability of having obtained the observed data. Thus, the hyperparameter values should define a prior distribution $P(\text{GP}(\mathbf{x}_{\text{new}}))$ maximizing the probability of the functions along the data. The problem can be rewritten as: the search of hyperparameter values maximizing the posterior probability distribution over the hyperparameters: $P(\Omega|\mathbf{D})$. In a Bayesian line of reasoning, this posterior probability can be expressed as:

$$P(\Omega|\mathbf{D}) \sim P(\mathbf{y}|\Omega, \mathbf{X}) P(\Omega) \quad (.3.7)$$

where $P(\mathbf{y}|\Omega, \mathbf{X})$, is the marginal likelihood:
 $P(\mathbf{y}|\Omega, \mathbf{X}) = \int P(\mathbf{y}|\text{GP}(\mathbf{x}_{\text{new}})) P(\text{GP}(\mathbf{x}_{\text{new}})) d\text{GP}(\mathbf{x}_{\text{new}})$. The hyperparameter values Ω can thus be determined by maximizing the logarithm of the marginal likelihood [MacKay (2003); Rasmussen and Ws (2006)]:

$$\ln P(\mathbf{y}|\Omega, \mathbf{X}) = -\frac{1}{2}\mathbf{y}^\top \mathbf{C}^{-1}\mathbf{y} - \frac{1}{2}\ln |\mathbf{C}| - \frac{n}{2}\ln 2\pi \quad (.3.8)$$

Several methods can be implemented to accomplish this multivariate optimization problem, such as a simplex method, Monte Carlo (MC) Sampling [Neal (1996)], a genetic algorithm, nested sampling [Skilling (2006)], forward variable selection [Obrezanova et al. (2007)] or the conjugate gradient method [Rasmussen and Ws (2006)].

In the present study, kernel hyperparameters were optimized by grid search and k -fold cross-validation (CV) in the case of the adenosine receptors and aminergic GPCRs data sets, because of their large size and high number of descriptors. The experimental error, σ_d^2 , (Equation .3.2) was considered as fixed with a value of 0.29 pK_i units, this value being taken from the work of Kramer et al. (2012) The same length scale value, l , was used for all descriptors to simplify the hyperparameter optimization.

For the dengue virus data set, due to its small size, and to the lack of information concerning the experimental uncertainty, the noise variance, σ_d^2 , was optimized by conjugate gradient as implemented in the GPML toolbox [Rasmussen and Nickisch (2010)]. As the number of descriptors is only 24, we optimized the length scales using the radial kernel. In the frame of Automatic Relevance Determination (ARD), the importance of each descriptor for the model was estimated using the inverse of the optimized l values, in the way described above.

GP tolerance to noise

To better understand the influence of the experimental error in GP modelling, we trained 15 models for each data set with increasing levels of noise with both the radial and the normalized polynomial (NP) kernel, thus leading to a total number of 90 models. Their predictive ability was monitored on the test set. The levels of added noise (noise variance) ranged from 0 to a maximum value of 10, which corresponds to a noise deviation of 3.2 pK_i units for the adenosine receptors and GPCR data sets, and 3.2 log units for the dengue virus NS3 proteases data set.

Machine learning analyses and implementation

Machine learning models were built in R using the *caret* package [Kuhn (2008)]. Non-default kernels for GP were introduced in the *caret* framework by in-house R scripts and by the definition of custom models (*custom* option in the *caret* package) implementing kernel functions from either the *kernlab* [Karatzoglou et al. (2004)]

package or in-house kernel functions. Likewise, The Gaussian Process for Machine Learning (GPML) Toolbox version 3.2 [Rasmussen and Nickisch (2010)] was used to build GP models in Matlab version 7.15 [MATLAB (2013)] to assess the importance of ligand descriptors (Automatic Relevance Determination).

Descriptors were preprocessed as described in section .2.3, whereas models were trained using grid search with CV (section .2.4). Model validation was performed as described in section .2.6.

.3.2.5 Assessment of maximum model performance

The Tropsha validation criteria [Golbraikh and Tropsha (2002); Tropsha and Golbraikh (2010); Tropsha and Gramatica (2003)], were used for accepting or dismissing the model (section .2.6.1). The distributions of maximum $RMSE_{test}$, Q^2_{test} , $R^2_{0\ test}$, and R^2_{test} were calculated for each data set as explained in section .2.7. The distributions of maximum and minimum values for these metrics then used to validate the metrics values obtained when evaluating the bioactivities predicted for the test sets. If the obtained metrics were beyond the maximum values (for Q^2_{test} , $R^2_{0\ test}$, and R^2_{test}) or the minimum values (for $RMSE_{test}$) of the distribution, the model is likely to be over-optimistic.

The experimental errors required to define the random samples B were determined in the following way. For adenosine and GPCR data sets, the experimental error of pK_i data was considered to be approximately 0.29 pK_i units, which corresponds to the average standard deviation value for public K_i data sets estimated by Kramer et al. (2012). The experimental error of the dengue data set was inferred from the data by considering its uncertainty as a hyperparameter of the GP model since we could not find information about the experimental uncertainty in the study of Prusis et al. (2008).

.3.2.6 Interpretation of ligand substructures

To calculate the influence of a given feature (chemical substructure) to pK_i , we iteratively set the count of the feature equal to zero in all compound descriptors presenting it, in order to virtually remove the substructure. Bioactivity values were then predicted using the modified compound descriptors, and the differences between the predicted values in the presence or absence of a given feature were calculated.

The average value of these differences, weighted by the number of counts of the feature in each compound, corresponds to the contribution of that feature to bioactivity. The contribution was estimated for all compound features considered in the model. The sign of the difference ($\{+/-\}$) indicates if the feature is respectively

beneficial or deleterious for compound bioactivity. This approach is closely related to the method proposed by Westen et al. (2011b), although two modifications have been made: (i) the weighting of the average difference between predicted and observed bioactivities, and (ii) the calculation of descriptor importance on a per target basis.

3.3 Results

3.3.1 Model validation

PCM GP models agree with the validation criteria

Overall, the models obtained for the three data sets with Gaussian Process modelling display statistics in agreement with our validation criteria (Table 3.2). To ensure that these results were not the consequence of spurious correlations, we trained GP models with randomized bioactivity values (y-scrambling). [Clark and Fox (2004)] For all data sets, the intercept was negative, thus ensuring the statistical soundness of our modelling. The best GP model for the adenosine receptors data set was obtained with the normalized polynomial (NP) kernel, exhibiting $RMSE_{test}$ and $R_{0\ test}^2$ values of 0.58 pK_i units and 0.75 respectively. Similarly, in the case of the GPCRs data set, the NP kernel led to the best predictive model, with $RMSE_{test}$ and $R_{0\ test}^2$ values of 0.66 pK_i units and 0.72. As these GP models were trained with a noise deviation of 0.54 pK_i units, the subtraction of the experimental uncertainty, 0.54 pK_i units, from the $RMSE_{test}$ gives a residual error arising from the modelling below 0.12 pK_i units. These $RMSE_{test}$ values correspond to 6.05% and 10.88% of the range of bioactivity values in the training set for the GPCRs and the adenosine receptors data sets.

In the case of the dengue virus data set, GP models show better predictive ability than those reported by [Prusis et al. (2008)], as Q_{test}^2 value of 0.92 is obtained here (Table 3.2) for the best GP model based on the Bessel kernel. The optimization of the noise variance, σ_d^2 , as an hyperparameter during the training process led to a value of 0.27 log units, similar to the values of about 0.3 log units reported by Prusis et al. (2013) in a recent study with similar experimental setup.

GP statistics are within the limits of the theoretical maximum model performance

The distributions of maximum R_{test}^2 , $R_{0\ test}^2$, and Q_{test}^2 and minimum $RMSE_{test}$ theoretical values, obtained as described in subsection Assessment of Maximum Model Performance in Materials and Methods, are given in Figure 3.1 for the three data sets.

The mean value of the distribution of maximum $R_{0\ test}^2$ values are equal to 0.80, 0.68 and 0.96 for the adenosine, GPCRs, and dengue virus NS3 proteases data sets, which highlights that the maximum correlation values that can be gathered when modelling public data are far from the optimal maximum correlation value of one.

Adenosine Receptors data set				
	R^2_{int}	RMSE _{int}	$R^2_{0\ test}$	RMSE _{test}
GP Bessel	0.64	0.70	0.70	0.63
GP Laplacian	0.67	0.68	0.67	0.66
GP Norm. Polynomial (NP)	0.69	0.65	0.75	0.58
GP Polynomial	0.70	0.64	0.70	0.63
GP PUK	0.57	0.79	0.56	0.77
GP Radial	0.65	0.69	0.65	0.68
PLS	0.29	0.97	0.30	1.00
SVM Norm. Polynomial (NP)	0.70	0.64	0.73	0.60
SVM Polynomial	0.71	0.63	0.71	0.62
SVM Radial	0.68	0.65	0.70	0.64
QSAR F	0.31	0.70	0.31	0.96

Aminergic GPCRs data set				
	R^2_{int}	RMSE _{int}	$R^2_{0\ test}$	RMSE _{test}
GP Bessel	0.56	0.83	0.56	0.80
GP Laplacian	0.62	0.78	0.63	0.75
GP Norm. Polynomial (NP)	0.69	0.68	0.72	0.66
GP Polynomial	0.68	0.71	0.70	0.68
GP PUK	0.46	0.93	0.46	0.90
GP Radial	0.69	0.69	0.71	0.66
PLS	0.69	0.69	0.27	1.05
SVM Norm. Polynomial (NP)	0.69	0.68	0.72	0.66
SVM Polynomial	0.69	0.69	0.71	0.66
SVM Radial	0.69	0.69	0.72	0.66
QSAR F	0.38	0.98	0.38	0.97

Dengue virus NS3 proteases data set				
	R^2_{int}	RMSE _{int}	$R^2_{0\ test}$	RMSE _{test}
GP Bessel	0.91	0.43	0.92	0.44
GP Laplacian	0.88	0.54	0.91	0.50
GP Linear	0.91	0.45	0.91	0.48
GP Norm. Polynomial (NP)	0.88	0.50	0.91	0.48
GP Polynomial	0.91	0.42	0.92	0.44
GP PUK	0.77	1.10	0.81	1.13
GP Radial	0.91	0.45	0.91	0.45
PLS	0.90	0.45	0.91	0.49
SVM Norm. Polynomial (NP)	0.86	0.54	0.91	0.46
SVM Polynomial	0.89	0.46	0.90	0.51
SVM Radial	0.90	0.48	0.90	0.48
QSAR F	0.29	1.19	0.48	1.13

Table .3.2: **Internal and external validation metrics for the PCM models.** For the three data sets, the best models are obtained with GP, being the lowest RMSE_{test} and highest $R^2_{0\ test}$ values: (i) adenosine receptors: 0.58 and 0.75 with NP kernel, (ii) GPCRs: 0.66 and 0.72 with NP kernel, and (iii) Dengue virus NS3 proteases 0.44 and 0.92 with Bessel kernel. Overall, GP models for the three data sets agree with the validation criteria.

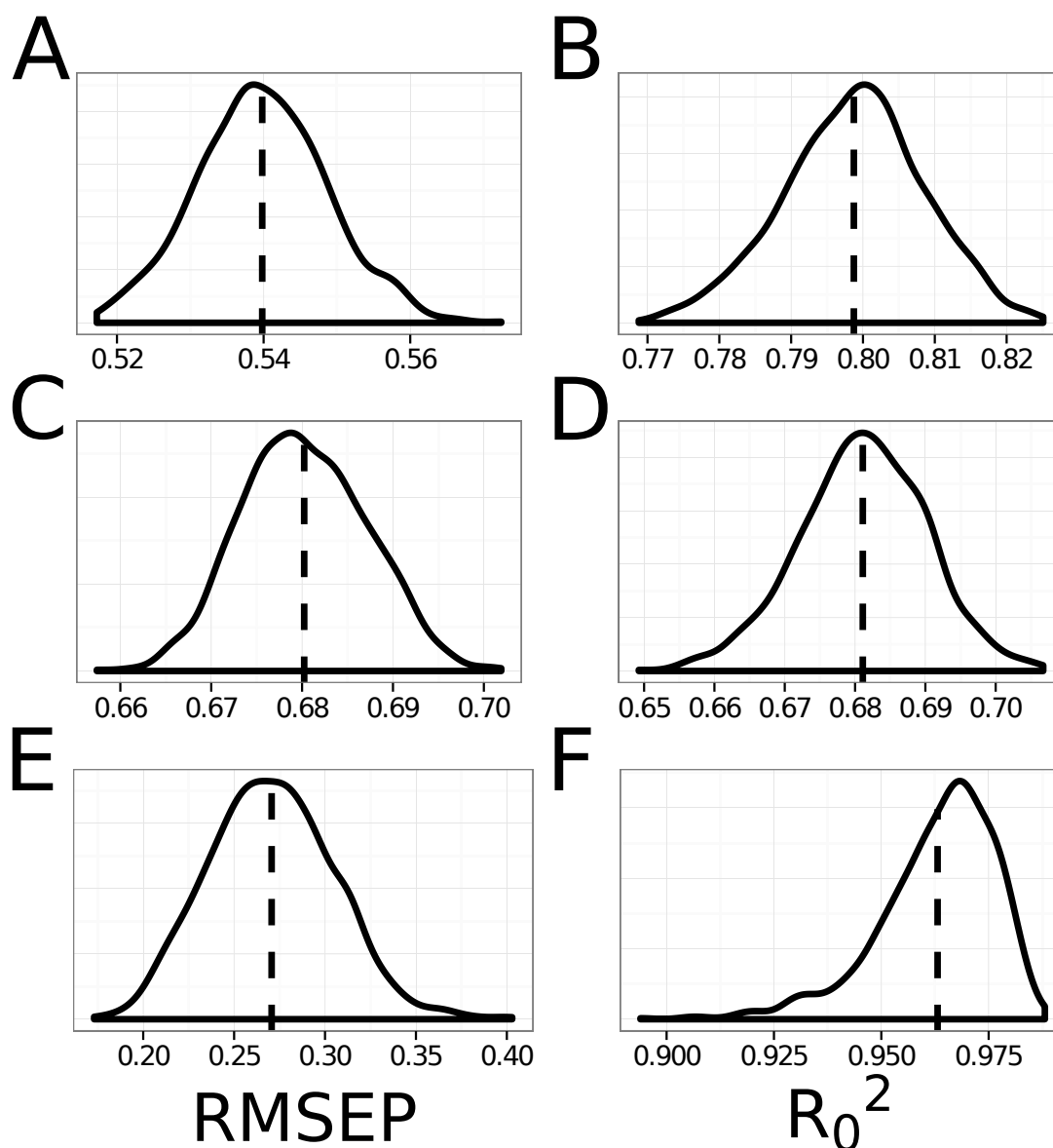


Figure .3.1: Distribution of the maximum theoretical values of $RMSE_{test}$ (A, C and E) and $R_{0\ test}^2$ (B, D and F) for the adenosine receptors (A, B), GPCRs (C, D) and dengue virus NS3 proteases data sets (E, F). These curves permit to estimate the reliability of $R_{0\ test}^2$ and $RMSE_{test}$ obtained for the GP models.

This is not surprising given the noise levels in public bioactivity data [Kalliokoski et al. (2013); Kramer et al. (2012)]. The best $RMSE_{test}$ and $R_{0\ test}^2$ values (Table .3.2) obtained with GP are respectively: 0.58 and 0.75 (adenosine receptors), 0.66 and 0.72

(GPCRs), and 0.44 and 0.92 (dengue virus NS3 proteases), which remain in the limits of these extreme theoretical values (Figure .3.1), thus supporting the suitability of our modelling pipeline to handle data uncertainty. The mean values of the theoretical RMSE distribution were close to the experimental uncertainty on bioactivity, for the adenosine receptors and the dengue virus NS3 proteases data sets, with respective mean $\text{RMSE}_{\text{test}}$ values of 0.54 pK_i units and 0.27 log units (Figure .3.1). However, the mean $\text{RMSE}_{\text{test}}$ value increases up to 0.68 pK_i units for the GPCRs data set owing to its larger size and sparsity.

PCM outperforms QSAR on the studied data sets

A comparison between models trained on only compound descriptors, 'Family QSAR' (QSAR F) [Brown et al. (2014)], and PCM permits to assess whether the use of GP improved the bioactivity modelling, by simultaneously modelling the target and the chemical spaces within a PCM study [Westen et al. (2011a)]. Indeed, radial kernelled Family QSAR models with ligand descriptors (Table .3.2) failed to model the data, being the $\text{RMSE}_{\text{test}}$ and R^2_{test} values respectively: 0.96 and 0.31 (adenosine receptors), 0.97 and 0.38 (GPCRs), and 1.13 and 0.48 (dengue virus NS3 proteases).

Strong mapping power of the normalized polynomial kernel

Radial and polynomial kernels have been traditionally used in QSAR and PCM modelling [Huang et al. (2012); Westen et al. (2012)], but the versatility of other kernels for bioactivity modelling has been recently demonstrated [Huang et al. (2012); Qifu et al. (2009); Wu et al. (2012)]. To investigate this point in the frame of GP models, we compared the performance of various kernels (Bessel, Laplacian, NP, and PUK) with the radial and polynomial kernels.

As described above, in contrast to Huang et al. (2012) we found the normalized polynomial (NP) kernel to have enough mapping power to model the three data sets (Table .3.2). Nonetheless, in the case of the dengue virus NS3 proteases data set, although NP kernel produces a statistically correct modelling with $\text{RMSE}_{\text{test}}$ and R^2_{test} values of 0.48 and 0.91, it is slightly outperformed by the Bessel kernel, which displays respective $\text{RMSE}_{\text{test}}$ and R^2_{test} values of 0.44 and 0.92 (Table .3.2).

The PUK kernel [Qifu et al. (2009)] exhibited strong mapping power in a previous study of HIV-1 proteases and histone deacetylases (HDAC) inhibitors [Huang et al. (2012); Wu et al. (2012)], but in the present study we could not obtain satisfactory models for none of the three data sets. The Laplacian and Bessel kernels allow a proper mapping of the three data sets with R^2_{test} values within the range 0.60-0.90 (see Table .3.2 for further details).

For the adenosine receptors data set, different statistics values are observed between the internal (on the hold-out folds in CV) and external validation (on the test set), as the $RMSE_{test}$ values are larger for the radial kernel (0.68) than for the polynomial and Bessel kernels (0.63 in both cases). Nonetheless, a different picture is observed for $RMSE_{int}$, as the values for the radial, polynomial and Bessel kernels are 0.69, 0.64 and 0.70 pK_i units. Although $RMSE_{test}$ and $RMSE_{int}$ values are similar, the small increase of $RMSE_{test}$ with the Bessel kernel might suggest a slight degree of overfitting [Kubinyi, Hamprecht, and Mietzner (1998)].

GP and SVM perform on par

The performance of the GP and SVM models was compared for each data set using the radial, the polynomial, and the NP kernels, as the first two are the most widespread kernels within the modelling community [Huang et al. (2012); Westen et al. (2012); Westen et al. (2013)]. Using different seed values, we trained ten different models for each modelling technique and data set, resulting in a total of 60 models (Figure .3.2).

To be able to statistically test the difference between the models results, distributions of the $RMSE_{test}$ and $R_{0\ test}^2$ were generated for each kernel / data set combination. Both $RMSE_{test}$ and $R_{0\ test}^2$ statistics were normally distributed in all cases (Shapiro-Wilk normality test, α 0.05), and a two-tailed t-test of independent samples (α 0.05) was applied to compare the behavior of SVM and GP. As it can be seen in Figure .3.2 and from the result of the t-test, both SVM and GP perform on par in the three case studies for radial and NP kernels. Similar results (data not shown) were obtained for the polynomial kernel.

To probe the linearity of the data sets, we trained linear PLS models. For two data sets, PLS appears unable to infer the complex (non-linear) relationships within the data, leading to $RMSE_{test}$ and $R_{0\ test}^2$ of 1.00 and 0.30 for the adenosine receptors, and 1.05 and 0.27 for the GPCRs data sets, respectively (Table .3.2). At contrary, the dengue NS3 proteases data set presents a clearly linear relationship, with $RMSE_{test}$ and $R_{0\ test}^2$ values of the PLS model of 0.49 and 0.91. But, on the same data set, the model obtained with a linear kerneled GP model outperformed PLS, with respective $RMSE_{test}$ and $R_{0\ test}^2$ values of 0.48 and 0.91.

Noise influence on GP depends on the kernel

$RMSE_{test}$ and $R_{0\ test}^2$ were calculated for adenosine receptors, GPCRs, and dengue virus NS3 proteases for different levels of noise σ_a^2 added to the diagonal of the covariance matrix C_X (Equation .3.2). The results obtained for radial kernels (Figure .3.3, upper plots) appear more sensitive to the noise than the ones obtained for NP kernels (Figure .3.3, bottom plots), for which the variations of the $RMSE_{test}$ and $R_{0\ test}^2$ sets

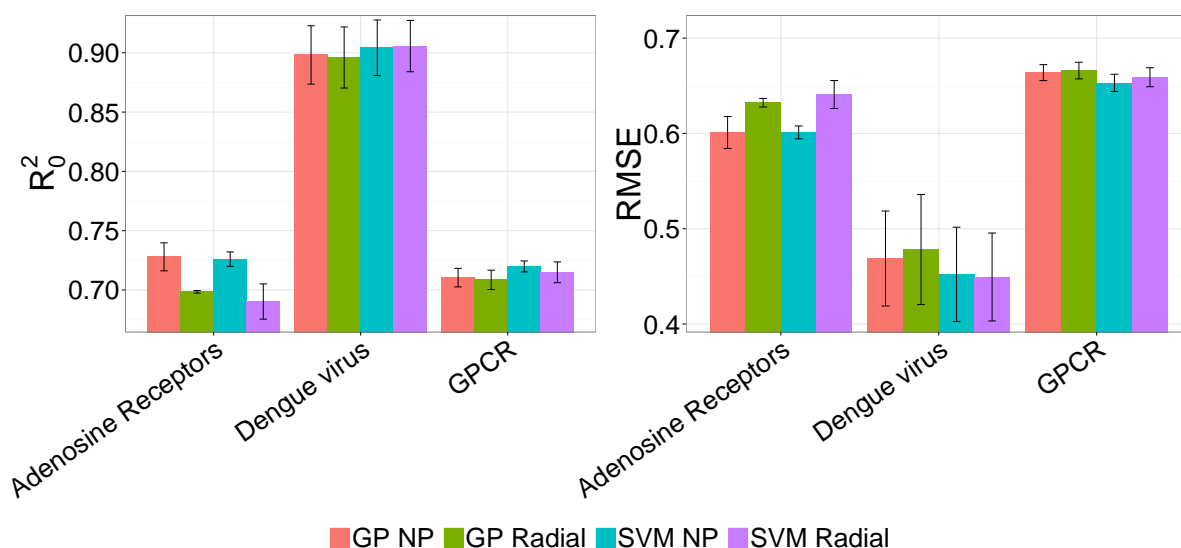


Figure 3.2: **Comparison between the performance of GP and SVM with either the radial or the NP kernel.** Ten models were calculated for each data set and for each combination of modelling technique and kernel, thus resulting in a total of 60 models. The performance of GP and SVM was assessed by kernel for the three data sets. Given that the distributions of $RMSE_{test}$ and $R^2_{0\ test}$ values were normally distributed, a two-tailed t -test of independent samples was applied to statistically evaluate their differences. These analyses let us conclude that SVM and GP perform *on par* for the modelling of the three data sets considered in this study.

are lower than 0.10 pK_i or log units. This trend is the most obvious for the dengue virus NS3 proteases data set, probably originating from the small size of this data set.

The polynomial kernel displayed robustness similar to those of NP kernel. These analyses suggest that NP or polynomial kernels would constitute a reasonable choice when modelling noisy data.

To summarize, GP models perform *on par* with SVM and outperform Family QSAR and PLS on the three data sets. The NP kernel leads to the best GP models being also the most tolerant kernel to noisy bioactivities. GP models trained on the dengue virus NS3 proteases systematically display better metrics than the other data sets, likely due to the high matrix completeness (88.84%) of this data set (Table 3.1).

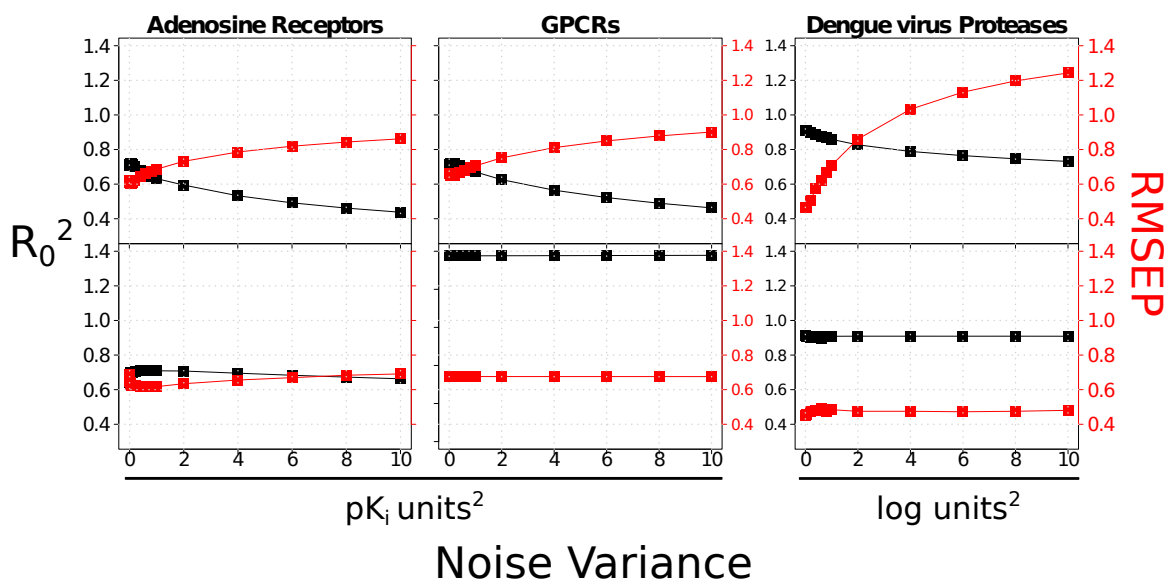


Figure 3.3: **Noise influence in model performance.** $RMSE_{test}$ (red) and R_0^2 ext (black) values obtained when increasing the noise level (noise variance added to the diagonal of the covariance matrix) were calculated for: adenosine receptors (left figure), GPCRs (medium figure) and dengue virus NS3 proteases (right figure). Upper plots correspond to GP models calculated with the radial kernel while the bottom plots refer to GP models with the Normalized Polynomial (NP) kernel. In all cases, the radial kernel appears more sensitive to noise, while the NP kernel performs equally well when noise is added to the data. These data suggest that the NP kernel is more appropriate for the modelling of noisy PCM data sets.

3.3.2 Predicted confidence intervals follow the cumulative density function of the Gaussian distribution

GP predictions mostly follow the cumulative Gaussian distribution

To analyze the reliability of the error bars obtained with GP with the tested kernels, different intervals of confidence (IC) for each predicted bioactivity value on the test set were defined, namely: 68%, 80%, 95%, and 99%. Subsequently, the percentage of compound-target combinations for which the experimental bioactivity value lied within the bounds of each interval was calculated. Following the cumulative density function of the Gaussian distribution (cumulative Gaussian distribution) [Schwaighofer et al. (2007)], the percentage of satisfactory cases should be proportional to the interval size.

To test this hypothesis, the percentages of predicted bioactivities for which the experimental values were within the confidence intervals were compared to the size of these intervals (Figure .3.4). As the small size of the dengue virus NS3 proteases did not allow a good sampling of the Gaussian distribution, this data set was not included in the comparison. This analysis was thus performed for the adenosine receptors and GPCRs data sets with the Bessel, Laplacian, NP, PUK, and radial kernels. It is noteworthy that the predicted variance obtained with the polynomial kernel is much larger than the range of bioactivity values, thus making impossible to evaluate their concordance with the cumulative distribution. However, the NP kernel allows to obtain values within the interval $\{0, 1\}$ for the predicted variance thanks to its normalized formulation.

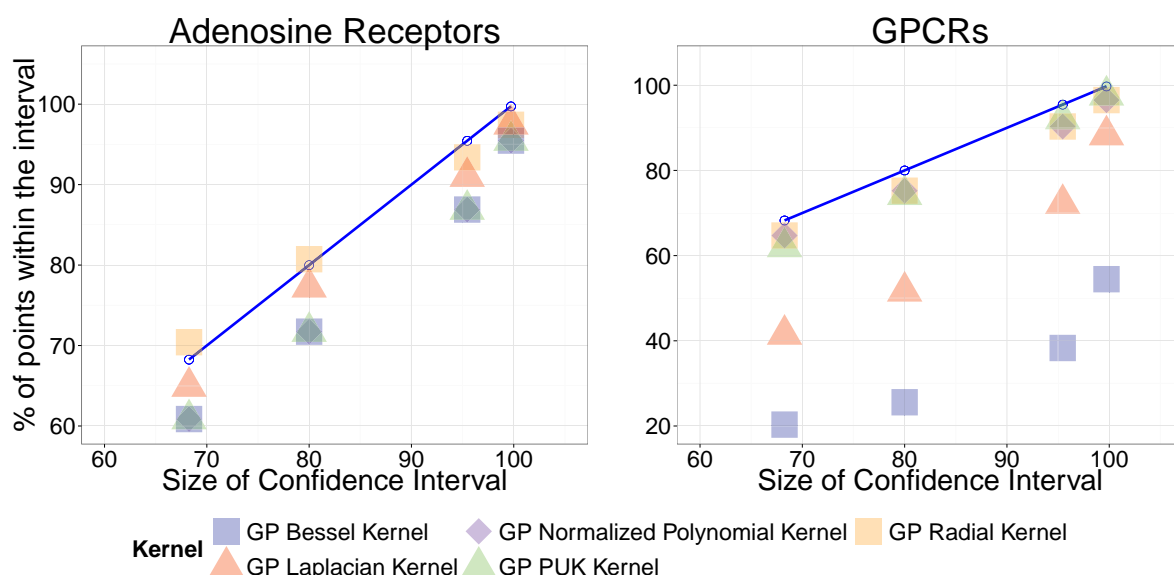


Figure .3.4: **Analysis of the confidence intervals predicted on (left) the adenosine receptors and (right) aminergic GPCRs test sets.** The percentage of annotated values lying within the intervals of confidence of 68%, 80%, 95%, and 99% (ordinate axis) are depicted versus the size of the intervals. The blue line defines the theoretical proportionality between the size of confidence intervals and the number of points within the intervals, in the frame of the Gaussian cumulative function. The radial, PUK, and Normalized Polynomial (NP) kernels are in close conformity with the cumulative Gaussian distribution in both data sets, while the Laplacian and Bessel exhibit a diverse behavior depending on the data set. Therefore, GP provide prediction errors in agreement with the Cumulative Gaussian distribution which can be reliably used to define intervals of confidence for the predictions.

The experimental values for the radial kernel match the theoretically expected behavior, represented on Figure .3.4 by bullet points connected by a blue line, and calculated in the context of a Gaussian cumulative function. The match between experiment and theory holds for the PUK and NP kernels for both data sets. The difference between the cumulative Gaussian distribution and the different intervals of confidence calculated for the Adenosine receptors data set is around 10% for the other kernels (Figure .3.4, left plot). By contrast the Bessel and Laplacian kernels do not provide informative intervals of confidence for the GPCRs data set (Figure .3.4, right plot).

GP determine the AD of the model

The variance predicted with GP models, $\sigma_{y_{new}}^2$, quantifies how much information the model can infer from the data (Eq. .3.5). Therefore, we hypothesized that: the distribution of the differences between the predicted and the observed bioactivity values, are more dispersed for compound-target pairs distant from the training set (high values of $\sigma_{y_{new}}^2$). To verify this hypothesis, we binned the test set into four groups depending on the value of the predicted variance: {0.25, 0.5, 0.75, 1}. The differences between true and predicted bioactivities were compared (Figure .3.5) to the bioactivity errors predicted in the GP model. This analysis was done on the adenosine receptors and GPCR data sets for the predicted variances obtained with the NP and the radial kernels. As the dispersion of the distribution of the differences increases with the errors predicted by GP, irrespective of the kernel or data set considered, this error can be thus considered as a reliable estimate of the applicability domain (AD).

Interestingly, while the average differences between predicted and observed bioactivities are close to zero for the subsets of GP errors of 0.25, 0.5, and 0.75, this average value is biased towards few tenths of a pK_i unit (Figure .3.5) for the subset displaying the largest GP error. This observation indicates that errors on bioactivities are underestimated by the GP model for compound-target pairs distant from the training set.

GP models with the NP and radial kernels provide prediction errors in agreement with the cumulative Gaussian distribution, which is the maximum theoretical precision attainable. Furthermore, the applicability domain of GP models can be determined from the errors predicted by GP.

.3.3.3 Analysis of GP performance *per target*

To further understand the predictive capability of GP models on each analyzed target, we trained ten GP models with the NP kernel. Different seed values were used for the generation of the training and the test sets. Once the GP predictions have been obtained, we divided the test set into subsets grouped by target, and calculated

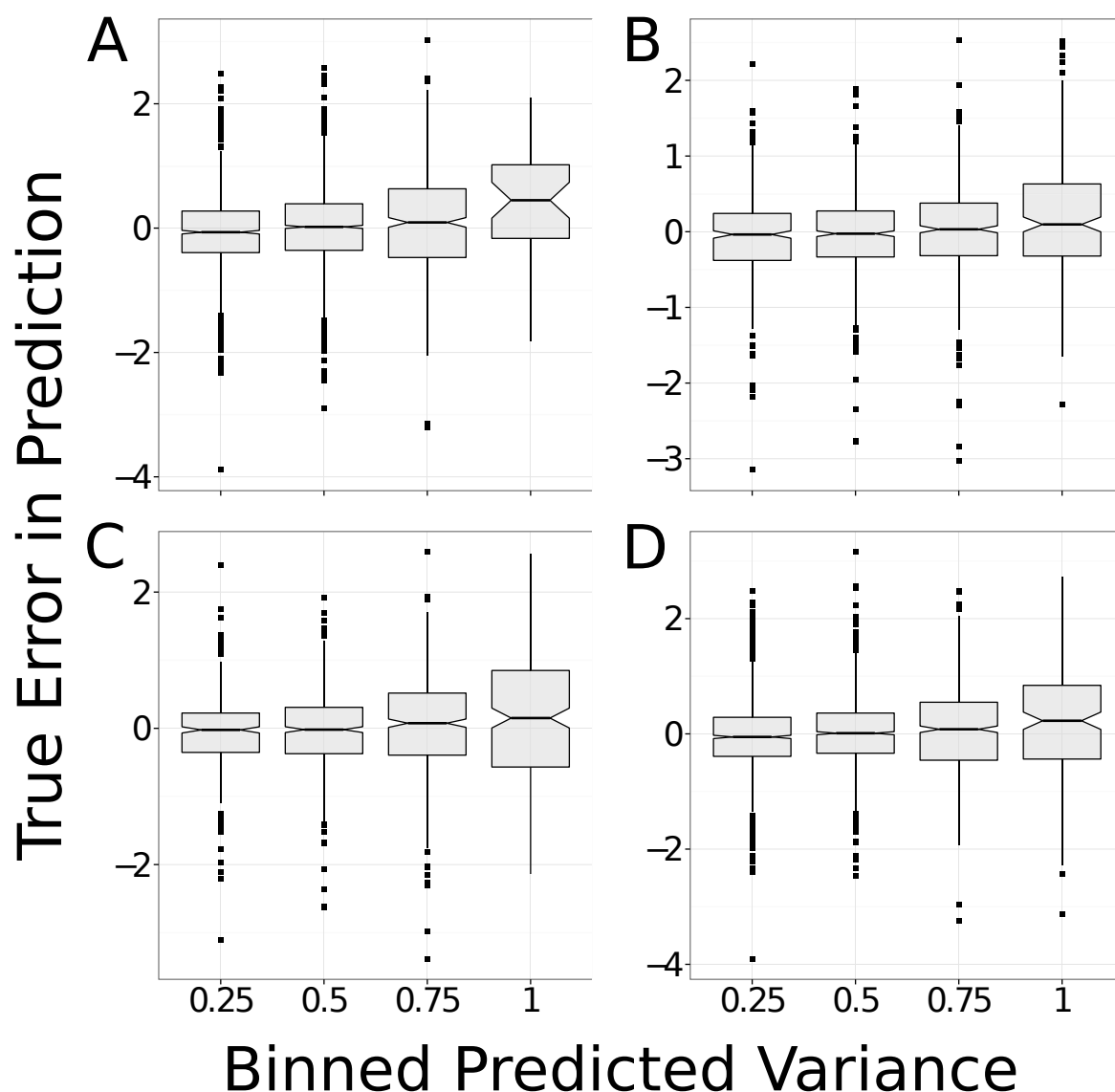


Figure .3.5: **GP** determine models applicability domain. The differences between the true and predicted bioactivities (y axis) and the errors on predictions estimated by the **GP** model (x axis) are compared for the adenosine receptor data set with radial (A) and **NP** (B) kernel, and for the **GPCRs** data set with radial (C) and **NP** (D) kernels. The distribution of the differences between true and predicted bioactivities increases with the **GP** error on the prediction. This validates the **GP** error is a measurement of the **AD** of the model.

average $R_{0\text{ test}}^2$ and $\text{RMSE}_{\text{test}}$ values on these subsets. This analysis per target was conducted only on the data sets of adenosine receptors and GPCRs, because of their large sizes and numbers of involved targets.

Adenosine receptors

The highest mean $\text{RMSE}_{\text{test}}$ value is between 0.70 and 0.75 pK_i units, and the lowest mean $R_{0\text{ test}}^2$ value is 0.62 (Figure .3.6). In this data set, the performance is not directly related to the number of compounds annotated *per* target. Indeed, the best result is obtained on the rat A_{2b} receptor (AA2BR RAT, 803 compounds) whereas one of the worst results is displayed by the human A_1 receptor (AA1R HUMAN, 1635 compounds).

On the other hand, the results cannot be related to the chemical diversity of the compounds, analyzed with pairwise Tanimoto similarity (Figure .3.7). Indeed, the two targets displaying the largest variability in the range of 0.50-0.75 Tanimoto similarity are rat A_3 (AA3R RAT) and human A_{2b} (AA2BR HUMAN), for which quite different performances are observed ($\text{RMSE}_{\text{test}}$ in the 0.70-0.75 range and in the 0.59-0.61 range respectively: Figure .3.6). Similarly, human A_1 (AA1R HUMAN) and A_{2a} (AA2AR HUMAN) receptors, display the smallest variability for compounds, and show quite different levels of performance ($R_{0\text{ test}}^2$ in the 0.56-0.60 range and in the 0.70-0.74 range respectively).

The lack of connection between the performance and the chemical diversity could arise from the binding site residue selection, which might not be equally suited for all adenosine receptors. This is supported by two other facts, namely: (i) the differences in extracellular loop length that are known for the adenosine receptor paralogues; and (ii) secondly the knowledge that these loops are important for ligand binding [Jaakola et al. (2008); Peeters et al. (2011a); Peeters et al. (2011b)].

GPCRs

In the GPCR data set, the best $\text{RMSE}_{\text{test}}$ (Figure .3.8) and $R_{0\text{ test}}^2$ (Figure .3.9) values are obtained on target subsets with a number of annotated compounds larger than 200 (in grey in Figures .3.8 and .3.9). Between the subsets, no major differences in performance are observed for an amount of annotated compounds between several hundreds and over 1500. It is however noticeable that the predictive ability of the models increased as the target space included in the training data set broadened. Indeed, a bioactivity selection previously done including information from 26 human aminergic GPCRs (4,951 data-points), marked with an asterisk in Table .3.3, did not produce any sound statistical metrics, as $R_{0\text{ test}}^2$ values lower than 0.40 were obtained whatever the kernel or machine learning algorithm used. But, the addition to the first selection of the bioactivities measured on mammal orthologues improved

the prediction, although some of the additional bioactivity sets were singletons (Table .3.3).

A large diversity of performance with $RMSE_{test}$ values in the range of 0.00-2.50 pK_i units is observed for the targets annotated with one compound (Figure .3.8). A relationship can be nevertheless established between these performances and the number of annotated compounds on orthologues proteins. For example, the 5-HT₂C mouse receptor (5HT₂C MOUSE) annotated with three compounds exhibits a mean $RMSE_{test}$ value between 0.00 and 0.20 pK_i units (Figure .3.8), because 345 and 558 compounds are respectively annotated on the orthologue rat and human 5-HT₂C receptors. The good performance obtained for this mouse receptor is probably due to the similarity of the 345 and 558 compounds to the ones annotated to the 5-HT₂C mouse receptor.

Protein ID	Frequency	Protein ID	Frequency	Protein ID	Frequency
5HT _{1A} HUMAN*	1152	ACM ₁ HUMAN*	379	ADRB ₂ HUMAN*	122
5HT _{1A} MOUSE	28	ACM ₁ MOUSE	9	ADRB ₂ MOUSE	21
5HT _{1A} RAT	1953	ACM ₁ RAT	443	ADRB ₃ HUMAN*	190
5HT _{1B} HUMAN*	436	ACM ₂ HUMAN*	573	ADRB ₃ MOUSE	1
5HT _{1B} RAT	103	ACM ₂ MOUSE	7	DRD ₁ BOVIN	182
5HT _{1D} HUMAN*	446	ACM ₂ RAT	275	DRD ₁ HUMAN*	315
5HT _{1D} MOUSE	5	ACM ₃ HUMAN*	421	DRD ₁ MOUSE	7
5HT _{1D} PIG	3	ACM ₃ RAT	109	DRD ₁ PIG	97
5HT _{1D} RAT	1	ACM ₄ HUMAN	105	DRD ₁ RAT	366
5HT _{1E} HUMAN	10	ACM ₄ MOUSE	1	DRD ₂ BOVIN	111
5HT _{1F} HUMAN	90	ACM ₅ HUMAN	42	DRD ₂ CHLAE	14
5HT _{1F} RAT	1	ADA _{1A} BOVIN	97	DRD ₂ HUMAN*	2340
5HT _{2A} BOVIN	5	ADA _{1A} HUMAN*	619	DRD ₂ MOUSE	8
5HT _{2A} HUMAN*	699	ADA _{1A} RAT	253	DRD ₂ RAT	1778
5HT _{2A} PIG	12	ADA _{1B} HUMAN*	624	DRD ₃ HUMAN*	1349
5HT _{2A} RAT	669	ADA _{1B} MESAUI	17	DRD ₃ RAT	348
5HT _{2B} HUMAN*	162	ADA _{1B} RAT	36	DRD ₄ HUMAN*	1326
5HT _{2C} HUMAN*	558	ADA _{1D} HUMAN	544	DRD ₄ RAT	35
5HT _{2C} MOUSE	3	ADA _{1D} RAT	205	DRD ₅ HUMAN*	134
5HT _{2C} RAT	345	ADA _{2A} BOVIN	76	DRD ₅ RAT	11
5HT _{4R} CAVPO	11	ADA _{2A} HUMAN*	276	HRH ₁ CAVPO	18
5HT _{4R} HUMAN	139	ADA _{2A} PIG	3	HRH ₁ HUMAN*	162
5HT _{4R} RAT	115	ADA _{2A} RAT	50	HRH ₁ RAT	82
5HT _{5A} HUMAN	57	ADA _{2B} HUMAN	144	HRH ₂ HUMAN	37
5HT _{5A} MOUSE	38	ADA _{2B} RAT	45	HRH ₃ CAVPO	73
5HT _{5A} RAT	19	ADA _{2C} HUMAN*	236	HRH ₃ HUMAN*	857
5HT _{5B} RAT	1	ADA _{2C} RAT	4	HRH ₃ RAT	612
5HT _{6R} HUMAN*	638	ADRB ₁ HUMAN*	111	HRH ₄ HUMAN*	70
5HT _{7R} HUMAN*	234	ADRB ₁ RAT	69	HRH ₄ RAT	2
5HT _{7R} MOUSE	1	ADRB ₂ BOVIN	21		
5HT _{7R} RAT	216	ADRB ₂ CANFA	26		

Table .3.3: **Number of datapoints per GPCR.** Those receptors highlighted by a '*' symbol correspond to those present in a subset of human GPCRs which was first modeled with GP (see subsection GP performance per Target). GPCRs are named according to UniProtKB/ Swiss-Prot database [Magrane and Consortium (2011)].

The importance of various targets for GP prediction was assessed for the adenosine receptors and GPCRs data sets. To obtain statistically validated models, a balance has to be found between two trends: (i) the inclusion of bioactivity information from orthologues improves the predictive ability of the models for both data sets, but (ii) an increase of the chemical diversity might hamper the acquisition of sound models as shown for the adenosine receptors data set.

.3.3.4 Model interpretation of ligand descriptors

Compounds bioactivity depends on multiple weak contributions of chemical substructures

The influence of the substructures on compound bioactivities, for both the adenosine receptors and the GPCRs, was analyzed as described in section Interpretation of ligand substructures. In the present study, the contribution of more than 90% of substructures to the pK_i values is close to zero (black regions in Figure .3.10). We observed similarly that chemical substructures contributing in a very variable way to the pK_i values (average contribution equal to zero and standard deviations in the range of 0.50 - 1.00 pK_i units), are present in sets of compounds displaying large variability in experimental bioactivity on a given target.

Hence, more than 90% of the substructures from the data sets analyzed here, display alternatively the following properties: (i) they are not implicated in compound bioactivity as their presence or absence does not influence compounds bioactivity, (ii) their contribution to the pK_i values, is conditioned to the presence or absence of other substructures [Klekota and Roth (2008)].

The highest contributions to the pK_i values, on both the GPCRs and the adenosine receptors data sets, is close to 1 pK_i units (Figure .3.10), in the range similar to those obtained by Westen et al. (2013) Therefore, even those few substructures with a large contribution, highlighted in Figure .3.10, do not explain a large proportion of the bioactivity.

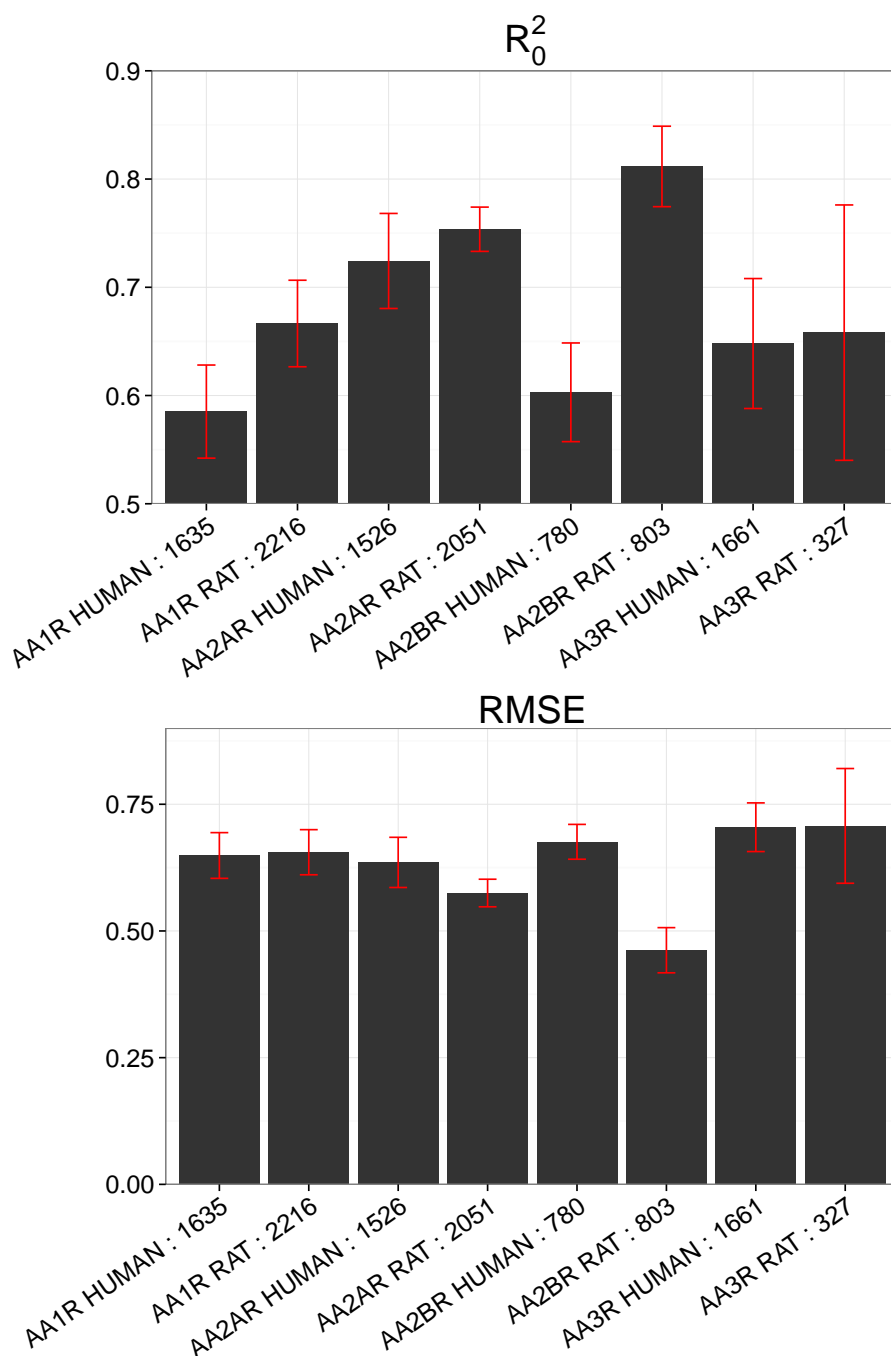


Figure .3.6: **Model performance per target on the test set for the adenosine receptors data set.** The upper panel corresponds to $R^2_{0\text{ test}}$, while the lower panel to $\text{RMSE}_{\text{test}}$. These values were averaged for ten models trained on each subset corresponding to a given target. The best modeled target is the rat adenosine A_{2b} receptor (AA2BR RAT), while the worst is the rat A_3 receptor (AA3R RAT). In all cases, the mean $\text{RMSE}_{\text{test}}$ values are below 0.75 pK_i units, indicating that GP modelling can predict compound bioactivity on subsets corresponding to a given target.

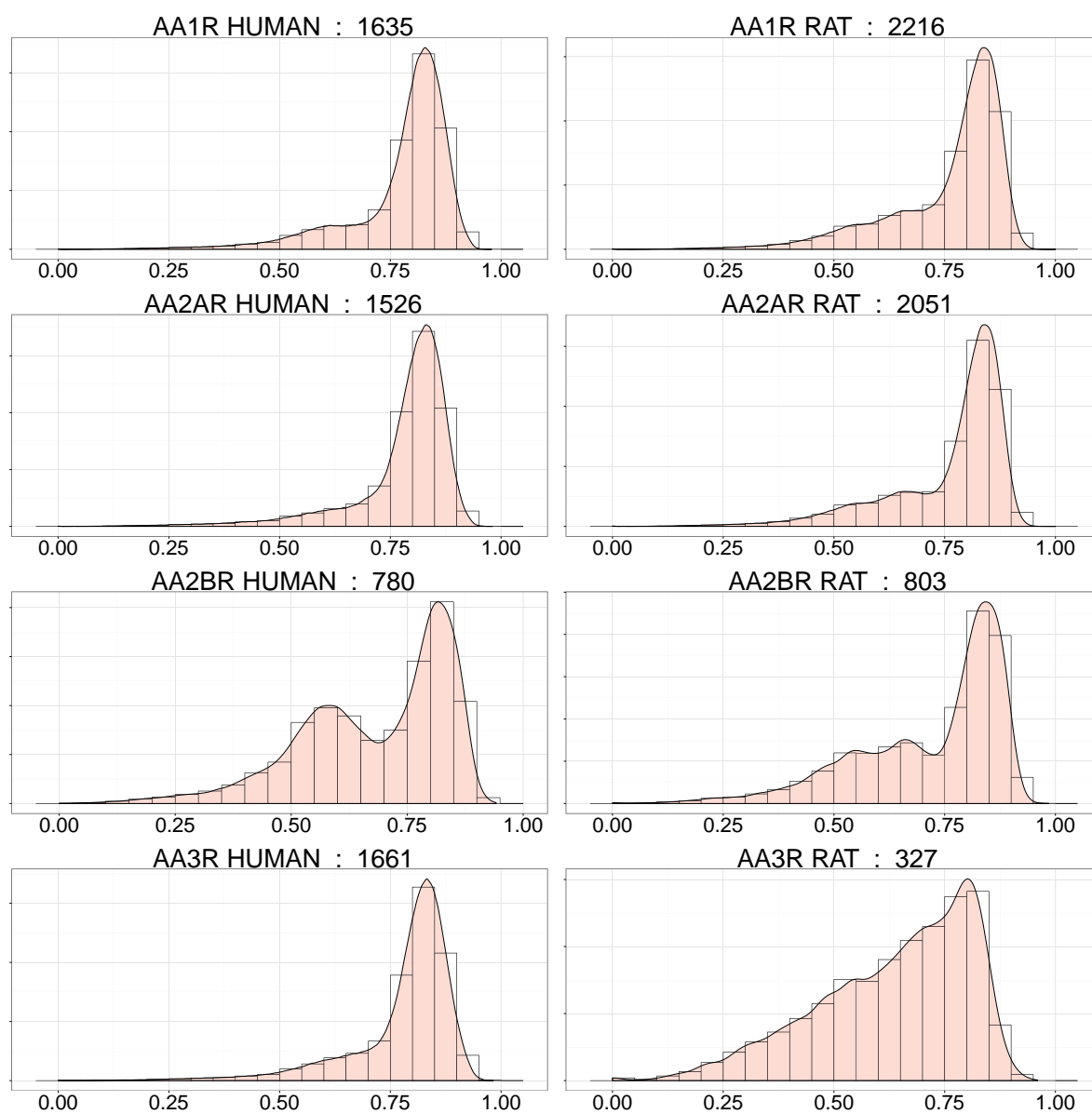


Figure .3.7: **Distribution of pairwise compound Tanimoto similarity calculated on the target subsets extracted from the adenosine receptors data set. The overall mean pairwise similarity is around 0.8.**

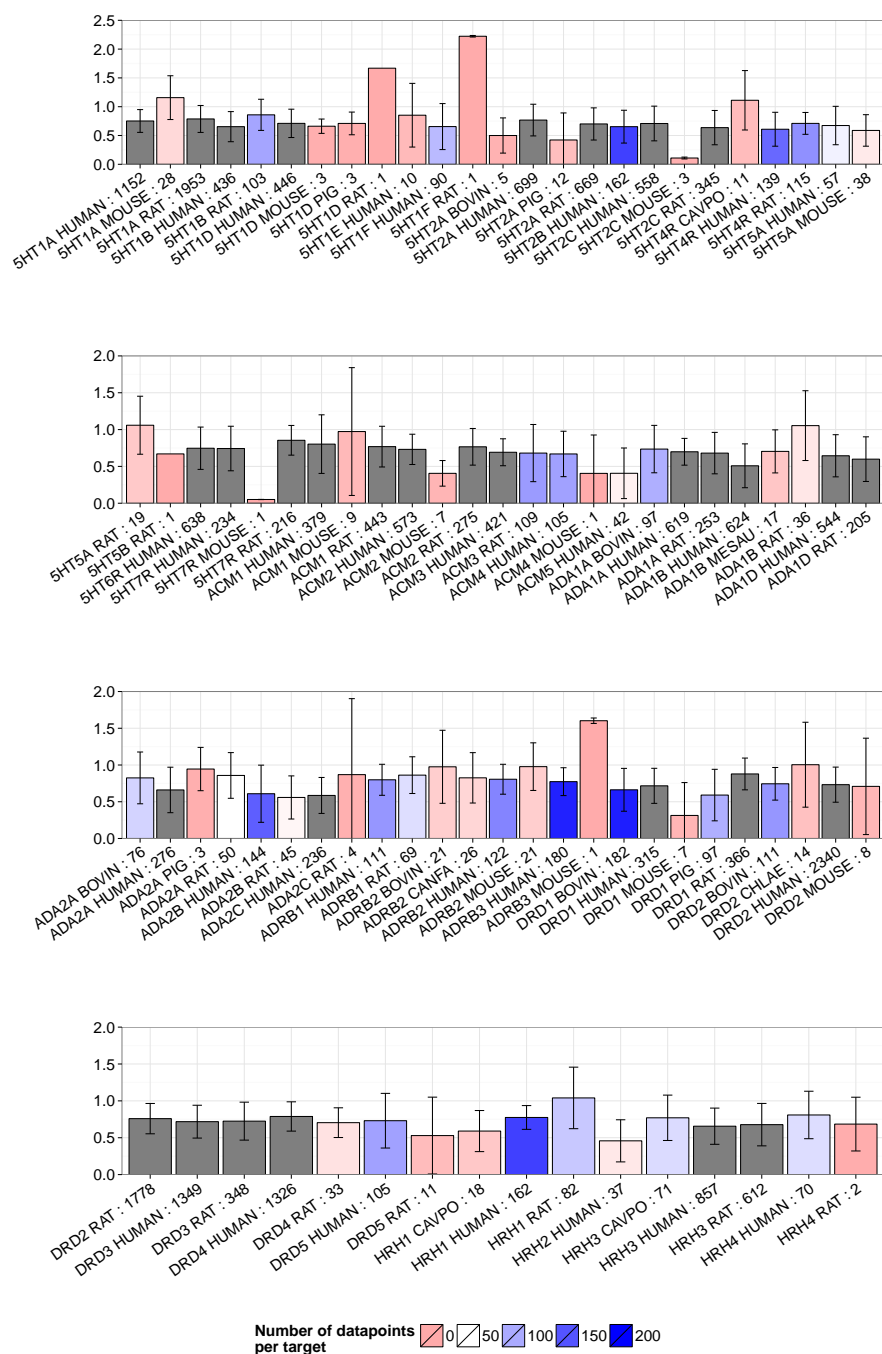


Figure .3.8: Evaluation of model performance per target on the **GPCRs** dataset. $RMSE_{test}$ values, averaged on ten models trained on different resamples of the dataset, are represented by bars, colored according to the number of datapoints per target. The standard deviations on $RMSE_{test}$ are shown as error bars. Dark grey bars correspond to targets with more than two hundred annotated compounds.

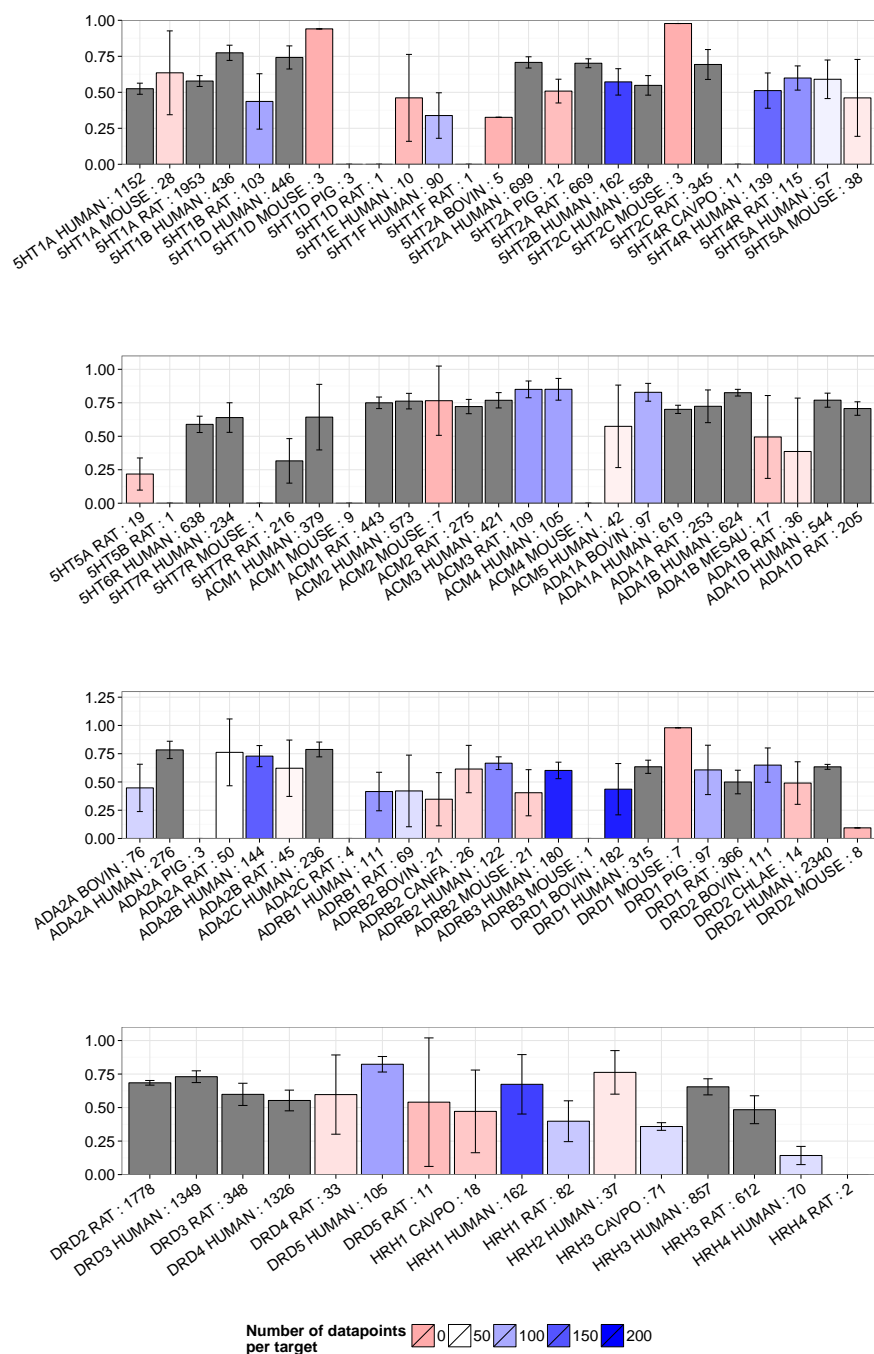


Figure 3.9: Evaluation of model performance per target for GPCRs dataset on the test set. $R^2_{0\text{ test}}$ values, averaged on ten models trained on different resamples of the dataset, are represented by bars, colored according to the number of datapoints per target. Dark grey bars correspond to targets with more than two hundred annotated compounds. Both negative and infinite $R^2_{0\text{ test}}$ values were set to zero.

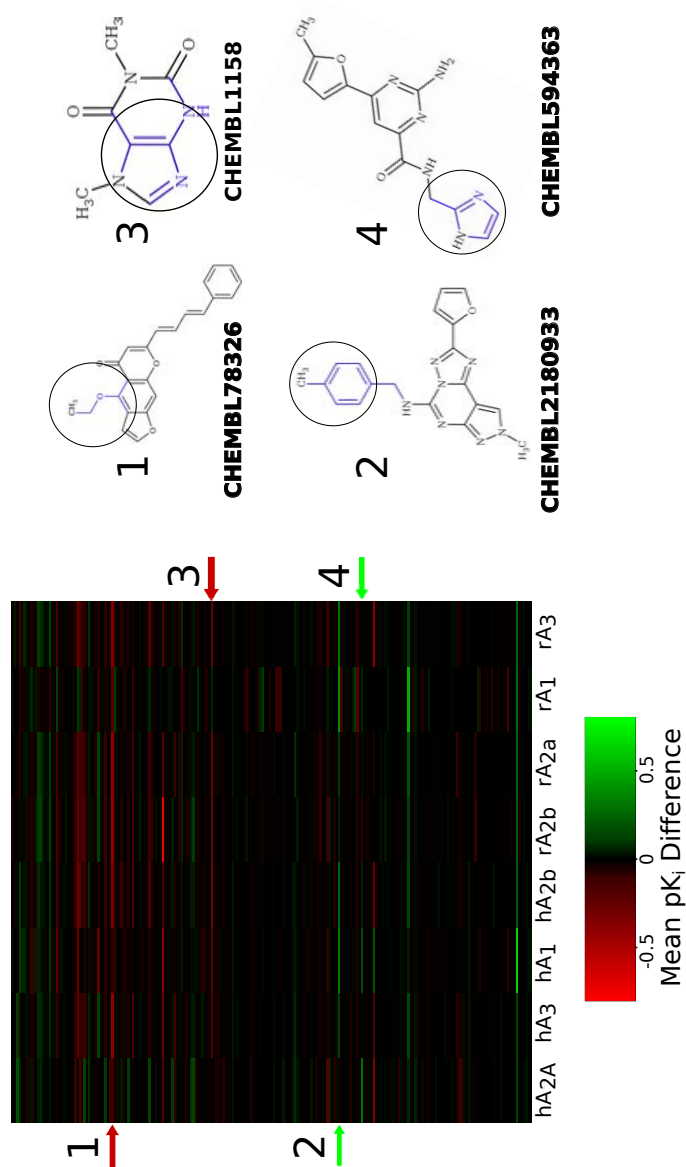


Figure 3.10: Heatmap representing the contribution of each chemical substructure to compound bioactivity on each adenosine receptor. Columns are indexed by targets and rows by compound substructures. Depicted are some examples of compounds containing features beneficial (green) or deleterious (red) for bioactivity. Although a few substructures are predicted to have a beneficial or deleterious influence on the pK_i , there are others for which the effect depends on the target considered or on the rest of substructures present in a given compound. Therefore, over 90% of the substructures (black) are not implicated in compound bioactivity or their contribution depends on the other substructures present in a given compound.

ARD provides a biologically meaningful interpretation of PCM models

The substrates in the dengue virus NS3 proteases data set are tetra-peptides. The relative importance of the four residues of these tetra-peptides was deconvoluted in the frame of [ARD](#), described in Materials and Methods, by taking the inverse of the optimized l value of the radial kernel (Figure .3.11).

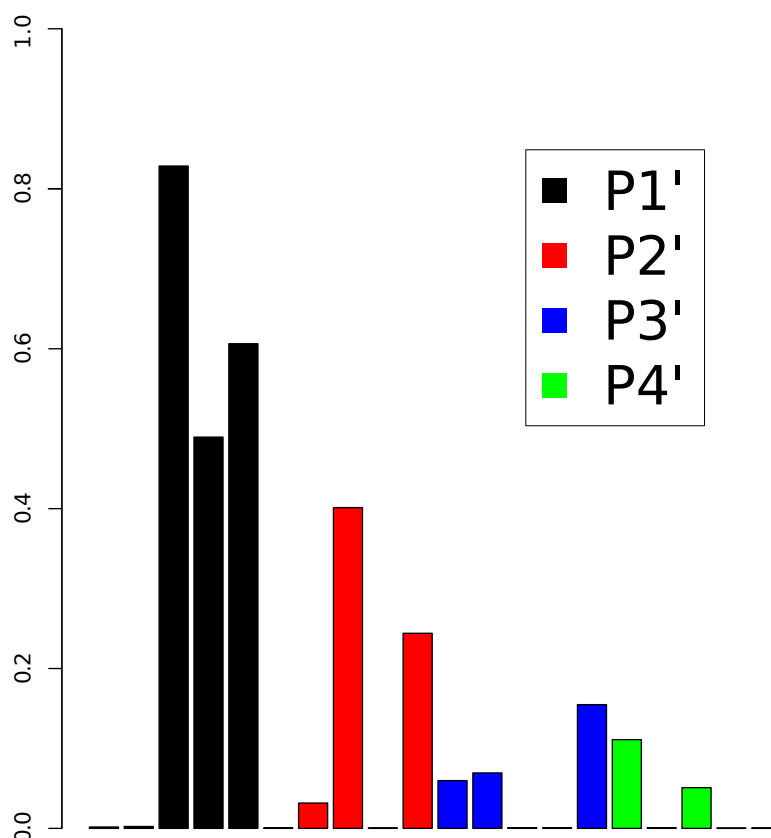


Figure .3.11: **Descriptor importance for the dengue virus NS3 proteases data set.** Descriptor importance is calculated in the frame of Bayesian [ARD](#) as the inverse of the value of the length scale of each descriptor. The descriptors of the first and second residues of the tetra-peptides (positions P1' and P2') are the most relevant for the model. This is in agreement with the higher influence of these two substrate positions for the cleavage rates of the proteases.

The largest inverse values are obtained for P1' followed by P2', P3' and P4' dis-

playing similar values. Thus, the first amino acid (P1') is the most relevant for the model followed by the second one (P2'), in contrast to the third and fourth ones (P3' and P4'). In the study of Prusis et al. (2008), the PLS coefficients with the highest values correspond to the first and second amino acids, as it is also the case here. A further detailed comparison of the PLS and the presented GP model is beyond the scope of this study. However, it should be noticed that the descriptors used in the present study and in [Prusis et al. (ibid.)] differ: 5 z-scales in our case versus 3 z-scales, C7.4, t1-Rig, and t2-Flex [Gottfries (2006)] in the PLS model. Although the PLS and GP models might assign different weights to the different descriptors, they both identify the first amino acid position as having the largest influence on K_{cat} , in agreement with experimental results [Prusis et al. (2008)].

GP models were interpreted on the basis of ligand descriptors. For data sets where ligands are compound descriptors (GPCRs and adenosine receptors data sets), the interpretation was not conclusive. By contrast, the interpretation of GP models according to the amino acids of the tetra-peptide ligands in the dengue data sets gave biologically meaningful results, in agreement with the scientific literature [Prusis et al. (ibid.)]. In that way, ARD can be applied to biologically interpret systems: e.g. identify residues responsible for compound binding. Additionally, ARD with the radial kernel can model non-linear relationships, which is not possible with PLS without the introduction of (not easily interpretable) cross-terms [Prusis et al. (2008); Westen et al. (2011a)].

.3.4 Discussion

In this chapter, we have demonstrated that Gaussian Processes (GP) allow to predict compound bioactivities on biomolecular targets. The statistical soundness of GP models is observed for a broad panel of kernels, among which the NP and radial kernels display the best results. GP and SVM display statistically similar performance for the modelling of multispecies proteochemometric data sets of different sizes. Moreover, Family QSAR and PCM models were trained on the same number of data-points and PCM produce much better results than Family QSAR, due to the introduction of target descriptors.

GP were applied on the following data sets: two large data sets involving GPCRs and adenosine receptors and one small data set (199 data-points) comprising four dengue NS3 proteases. The dengue data set exhibits a high degree of linearity, as demonstrated by the high performance of both PLS and GP with a linear kernel on this data set. Unsurprisingly, a better performance of GP is observed with different kernels for the dengue data set than for the two other ones, due to the high matrix completeness in the dengue data set and to its linearity. The satisfactory results

obtained for the dengue data set encourages the application of GP to model relatively small data sets issued from a single laboratory. The use of such in-house data sets would reduce the bias introduced by the annotation errors and by the use of non-normalized experimental conditions.

The inclusion of chemical and target information from several organisms (orthologues) increases model performance and the applicability of models to predict bioactivity for new compound target-combinations. These observations are in favor for the routine inclusion of multispecies bioactivity information in PCM settings. These results disagree with Gao et al. (2013), who stated that the addition of orthologues to human aminergic GPCRs would reduce the AD. Our understanding of the results obtained here is that the incorporation of bioactivity data from a wide range of species led to a significant increase of models performance given that binding patterns tend to be conserved among orthologues [Kruger and Overington (2012)].

We have seen on the GPCR data set, that the inclusion of singletons compounds bioactivities on human orthologues helps to increase models performance. This may be of tremendous relevance in the often encountered cases where limited bioactivity information is known on a given human target, but a much larger number of bioactivities have been measured on orthologues of this target [Fredholm et al. (2001); Kruger and Overington (2012); Westen et al. (2012)]. Our results suggest that the chemical diversity considered and the number of data-points have to be balanced to obtain sound models while exhibiting proper predictive abilities.

An additional outcome of GP with respect to SVM is the estimation of the uncertainty of predictions. Indeed, the Bayesian formulation of GP permits to obtain intervals of confidence for individual predictions defined from the GP predicted variance. These intervals were shown to be in agreement with the cumulative Gaussian distribution when using the radial and NP kernels, but not always for the Bessel or Laplacian kernels, highlighting that the kernel choice has to be made in the light of both models performance and reliability of the predicted variances.

We have also shown here that GP using as covariance function the polynomial or the NP kernel can handle noisy data sets, as the GP performance is only slightly affected when noise is introduced in the data. Nonetheless, each kernel should be chosen in the light of underlying structure of the data set, as the kernel controls the prior distribution over functions, and thus the models generalization properties [Duvenaud et al. (2013); Rasmussen and Ws (2006)]. It is noteworthy to mention that we have implemented a broad, though not exhaustive, panel of kernels, which is susceptible to be further completed with other base kernels or kernel combinations (composite kernels) [Duvenaud et al. (2013); Kronberger and Kommenda (2013); Rasmussen and Ws (2006)].

GP can consider individual experimental errors as input for the probabilistic model which may constitute a preeminent advantage when gathering information from diverse sources, each of which including distinct levels of experimental uncertainty [Schwaighofer et al. (2007)]. In the present study, an approximation of the experimental uncertainty of heterogeneous pK_i values, recently reported by Kramer et al. (2012) to exhibit a standard deviation of 0.54 pK_i units, has been introduced in the model. Nonetheless, GP allow the inclusion of the uncertainty of each individual datapoint into the model, which might lead to a more accurate modelling pipeline in cases where the experimental uncertainty of each datapoint is available.

Traditionally, the application of GP to model large data sets has been limited since the inversion of the covariance matrix scales with the cube of its dimension, *i.e.* GP is an algorithm of complexity $O(N^3)$ [Obrezanova et al. (2007); Rasmussen and Ws (2006)]. In the present study, we have not reported training times since models have been trained with GP implementations coded in different programming languages (subsection Machine Learning Analyses and Implementation). In the experience of the authors, the application of ARD is limited by the size of the data sets, being not applicable in practice to data sets with more than several thousands of data-points, or with more than several hundreds of descriptors. Nevertheless, new GP implementations have proved to seemingly decrease calculation times [Csato and Opper (2002); Paciorek et al. (2013); Tresp (2000)], which might increase the applicability of GP to large PCM data sets in the future.

Overall, we have shown here that GP simultaneously provides bioactivity predictions and assessment of their reliability. The application of GP to PCM data sets, gives the insight that GP could also be very useful in the drug discovery for personalized medicine, when the target space includes several mutants of a given target [Lapinsch et al. (2008); Westen et al. (2013)]. In the same way, GP could even be used in the context of decision making in clinics [Spjuth et al. (2011)].

.3.5 Conclusion

Gaussian Processes (GP) have been proposed and tested for the prediction of bioactivity measurements, and found to perform at the same level of statistical significance as Support Vector Machines (SVM). In addition, GP is the only method, up to now, to give predictions as probability distributions, thus permitting the estimation of errors on the bioactivity predictions as well as an estimation of the applicability domain. Moreover, GP are tolerant to noisy bioactivities. GP models trained on PCM data sets can also be used to analyze the effect of ligand features (compound substructures or peptide residues).

Bibliography

- Ben-Hur, A, CS Ong, S Sonnenburg, B Schölkopf, and G Rätsch (Oct. 2008). "Support Vector Machines and Kernels for Computational Biology". In: *PLoS Comput. Biol.* 4.10. Ed. by F Lewitter, e1000173 (cit. on p. 93).
- Bender, A, JL Jenkins, J Scheiber, SCK Sukuru, M Glick, and JW Davies (Jan. 2009). "How similar are similarity searching methods? A principal component analysis of molecular descriptor space". In: *J. Chem. Inf. Model.* 49.1, pp. 108–119 (cit. on p. 91).
- Bosnić, Z and I Kononenko (2009). "An overview of advances in reliability estimation of individual predictions in machine learning". In: *Intelligent Data Analysis* 13.2, pp. 385–401 (cit. on p. 89).
- Brown, J, Y Okuno, G Marcou, A Varnek, and D Horvath (2014). "Computational chemogenomics: Is it more than inductive transfer?" In: *J. Comput. Aided Mol. Des.* Pp. 1–22 (cit. on p. 101).
- Brown, SP, SW Muchmore, and PJ Hajduk (Apr. 2009). "Healthy skepticism: assessing realistic model performance". In: *Drug Discov. Today* 14.7-8, pp. 420–427 (cit. on p. 89).
- Burden, FR (2001). "Quantitative structure-activity relationship studies using Gaussian processes". In: *J. Chem. Inform. Comput. Sci.* 41.3, pp. 830–835 (cit. on p. 90).
- Clark, R and P Fox (2004). "Statistical variation in progressive scrambling". In: *J. Comput. Aided Mol. Des.* 18.7-9, pp. 563–576 (cit. on p. 98).
- Csato, L and M Opper (2002). "Sparse on-line Gaussian Processes". In: *Neural Comput.* 14.3, pp. 641–668 (cit. on p. 119).
- Duvenaud, D, JR Lloyd, R Grosse, JB Tenenbaum, and Z Ghahramani (2013). "Structure discovery in nonparametric regression through compositional kernel search". In: *ArXiv e-prints* 13024922 (cit. on p. 118).
- Fredholm, BB, AP IJzerman, KA Jacobson, KN Klotz, and J Linden (2001). "International Union of Pharmacology XXV Nomenclature and classification of adenosine receptors". In: *Pharmacol. Rev.* 53.4, pp. 527–552 (cit. on p. 118).
- Gao, J, Q Huang, D Wu, Q Zhang, Y Zhang, T Chen, Q Liu, R Zhu, Z Cao, and Y He (Apr. 2013). "Study on human GPCR-inhibitor interactions by proteochemometric modeling". In: *Gene* 518.1, pp. 124–131 (cit. on p. 118).
- Gaulton, A, LJ Bellis, AP Bento, J Chambers, M Davies, A Hersey, Y Light, S McGlinchey, D Michalovich, B Al-Lazikani, and JP Overington (2011). "ChEMBL: a large-scale bioactivity database for drug discovery". In: *Nucleic Acids Res.* 40.D1, pp. 1100–1107 (cit. on pp. 90, 91).

- Genton, MG (2002). "Classes of kernels for machine learning: a statistics perspective". In: *J. Mach. Learn. Res.* 2, pp. 299–312 (cit. on p. 93).
- Glen, RC, A Bender, CH Arnby, L Carlsson, S Boyer, J Smith, and RC Glenn (2006). "Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME". In: *IDrugs* 9.3, pp. 199–204 (cit. on p. 91).
- Gloriam, DE, SM Foord, FE Blaney, and SL Garland (July 2009). "Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design". In: *J. Med. Chem.* 52.14, pp. 4429–4442 (cit. on p. 91).
- Golbraikh, A and A Tropsha (2002). "Beware of q²!" In: *J. Mol. Graphics Modell.* 20.4, pp. 269–276 (cit. on pp. 90, 97).
- Gottfries, J (2006). "The drug designers guide to selectivity". In: *Chemometrics and Intelligent Laboratory Systems* 83.2, pp. 148–156 (cit. on p. 117).
- Huang, Q, H Jin, Q Liu, Q Wu, H Kang, Z Cao, and R Zhu (July 2012). "Proteochemometric modeling of the bioactivity spectra of HIV-1 protease inhibitors by introducing protein-ligand interaction fingerprint". In: *PLoS ONE* 7.7, e41698 (cit. on pp. 101, 102).
- Jaakola, VP, MT Griffith, MA Hanson, V Cherezov, EYT Chien, JR Lane, AP IJzerman, and RC Ss (2008). "The 2.6 Angstrom crystal structure of a human A_{2A} adenosine receptor bound to an antagonist". In: *Science* 322.5905, pp. 1211–1217 (cit. on p. 108).
- Kalliokoski, T, C Kramer, A Vulpetti, and P Gedeck (2013). "Comparability of mixed IC₅₀ Data - A statistical analysis". In: *PloS ONE* 8.4, e61007 (cit. on pp. 89, 100).
- Karatzoglou, A, A Smola, K Hornik, and A Zeileis (2004). "kernlab – An S₄ package for kernel methods in R". In: *J. Stat. Soft.* 11.9, pp. 1–20 (cit. on p. 96).
- Klekota, J and FP Roth (Nov. 2008). "Chemical substructures that enrich for biological activity". In: *Bioinformatics* 24.21, pp. 2518–2525 (cit. on p. 110).
- Kramer, C and R L (2012). "QSARs, data and error in the modern age of drug discovery". In: *Curr. Top. Med. Chem.* 12.17, pp. 1896–1902 (cit. on p. 89).
- Kramer, C, T Kalliokoski, P Gedeck, and A Vulpetti (June 2012). "The experimental uncertainty of heterogeneous public Ki data". In: *J. Med. Chem.* 55.11, pp. 5165–5173 (cit. on pp. 89, 96, 97, 100, 119).
- Kronberger, G and M Kommenda (2013). "Evolution of covariance functions for Gaussian Process regression using genetic programming". In: *CoRR* abs/13053 (cit. on p. 118).
- Kruger, FA and JP Overington (2012). "Global analysis of small molecule binding to related protein targets". In: *PLoS Comput. Biol.* 8.1, e1002333 (cit. on p. 118).
- Kubinyi, H, FA Hamprecht, and T Mietzner (1998). "Three-Dimensional Quantitative Similarity-Activity Relationships (3D QSiAR) from SEAL similarity matrices". In: *J. Med. Chem.* 41.14, pp. 2553–2564 (cit. on p. 102).
- Kuhn, M (2008). "Building predictive models in R using the caret package". In: *J. Stat. Softw.* 28.5, pp. 1–26 (cit. on p. 96).

- Lapinsh, M, M Eklund, O Spjuth, P Prusis, and JES Wikberg (Apr. 2008). "Proteochemometric modeling of HIV protease susceptibility". In: *BMC Bioinformatics* 9.1, p. 181 (cit. on p. 119).
- MacKay, DJC (2003). *Information Theory, Inference and Learning Algorithms*. en. Cambridge University Press (cit. on p. 95).
- Magrane, M and U Consortium (2011). "UniProt Knowledgebase: a hub of integrated protein data". In: *Database* 2011 (cit. on p. 109).
- MATLAB (2013). *version 7.1 (R2013b)*. Natick, Massachusetts: The MathWorks Inc (cit. on p. 97).
- Neal, RM (1996). *Bayesian Learning for Neural Network*. Springer-Verlag (cit. on p. 96).
- Netzeva, TI, A Worth, T Aldenberg, R Benigni, MTD Cronin, P Gramatica, J Jaworska, S Kahn, G Klopman, CA Marchant, G Myatt, N Nikolova-Jeliazkova, GY Patlewicz, R Perkins, D Rs, T Schultz, DW Stanton, JJM van de Sandt, W Tong, G Veith, and C Yang (Apr. 2005). "Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships The report and recommendations of ECVAM Workshop 52". In: *ATLA: Altern. Lab. Anim.* 33.2, pp. 155–173 (cit. on p. 89).
- Obrezanova, O and MD Segall (2010). "Gaussian processes for classification: QSAR modeling of ADMET and target activity". In: *J. Chem. Inf. Model.* 50.6, pp. 1053–1061 (cit. on p. 90).
- Obrezanova, O, G Csányi, JMR Gola, and MD Segall (2007). "Gaussian Processes: A method for automatic QSAR modeling of ADME properties". In: *J. Chem. Inf. Model.* 47.5, pp. 1847–1857 (cit. on pp. 90, 96, 119).
- Oksanen, J, FG Blanchet, R Kindt, P Legendre, PR Minchin, RB O'Hara, GL Simpson, P Solymos, MHH Ss, and H Wagner (2013). *vegan: Community ecology package*. URL: <http://cran.r-project.org/web/packages/vegan/index.html> (cit. on p. 92).
- Paciorek, CJ, B Lipshitz, W Zhuo, Prabhat, CG Kaufman, and RC T (2013). *Parallelizing Gaussian Process calculations in R*. arXiv e-print 13054886 (cit. on p. 119).
- Peeters, MC, GJP van Westen, Q Li, and AP IJzerman (2011a). "Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation". In: *Trends. Pharmacol. Sci.* 32.1, pp. 35–42 (cit. on p. 108).
- Peeters, MC, GJP van Westen, D Guo, LE Wisse, CE Müller, MW Beukers, and AP IJzerman (2011b). "GPCR structure and activation: an essential role for the first extracellular loop in activating the adenosine A2B receptor". In: *The FASEB Journal* 25.2, pp. 632–643 (cit. on p. 108).
- Prusis, P, M Lapinsh, S Yahorava, R Petrovska, P Niyomrattanakit, G Katzenmeier, and JES Wikberg (Oct. 2008). "Proteochemometrics analysis of substrate interactions with dengue virus NS3 proteases". In: *Bioorg. Med. Chem.* 16.20, pp. 9369–9377 (cit. on pp. 91, 97, 98, 117).
- Prusis, P, M Junaid, R Petrovska, S Yahorava, A Yahorau, G Katzenmeier, M Lapinsh, and JES Wikberg (2013). "Design and evaluation of substrate-based octapeptide and

- non substrate-based tetrapeptide inhibitors of dengue virus NS2B-NS3 proteases". In: *Biochem. Biophys. Res. Commun.* 434.4, pp. 767–772 (cit. on p. 98).
- Puntanen, S and GPH Styan (Jan. 2005). "Schur complements in statistics and probability". In: *The Schur Complement and Its Applications*. Ed. by F Zhang. Numerical Methods and Algorithms 4. Springer US, pp. 163–226 (cit. on p. 94).
- Qifu, Z, H Haifeng, Z Y, and S Guodong (2009). "Support vector machine based on universal kernel function and its application in Quantitative Structure - Toxicity Relationship model". In: *Proceedings of the 2009 International Forum on Information Technology and Applications - Volume 03*. IFITA '09. Washington, DC, USA: IEEE Computer Society, pp. 708–711 (cit. on p. 101).
- R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria (cit. on p. 92).
- Rasmussen, CE and CKI Ws (2006). *Gaussian Processes for machine learning*. MIT Press (cit. on pp. 93, 95, 96, 118, 119).
- Rasmussen, CE and H Nickisch (2010). "Gaussian Processes for Machine Learning (GPML) Toolbox". In: *JMLR* 11, pp. 3011–3015 (cit. on pp. 96, 97).
- Ren, Y, B Wu, Y Pan, F Lv, X Kong, X Luo, Y Li, and Q Yang (2011). "Characterization of the binding profile of peptide to transporter associated with antigen processing (TAP) using Gaussian process regression". In: *Comput. Biol. Med.* 41.9, pp. 865–870 (cit. on p. 90).
- Reutlinger, M, T Rodrigues, P Schneider, and G Schneider (2014). "Combining on-chip synthesis of a focused combinatorial library with computational target prediction reveals imidazopyridine GPCR ligands". In: *Angew. Chem. Int. Ed.* 53.2, pp. 582–585 (cit. on p. 90).
- Rogers, D and M Hahn (May 2010). "Extended-connectivity fingerprints". In: *J. Chem. Inf. Model.* 50.5, pp. 742–754 (cit. on pp. 89, 91).
- Romero, PA, A Krause, and FH Arnold (2013). "Navigating the protein fitness landscape with Gaussian processes". In: *Proc. Natl. Acad. Sci. U.S.A.* 110.3, E193–E201 (cit. on p. 90).
- Sahigara, F, K Mansouri, D Ballabio, A Mauri, V Consonni, and R Todeschini (2012). "Comparison of different approaches to define the applicability domain of QSAR models". In: *Molecules (Basel, Switzerland)* 17.5, pp. 4791–4810 (cit. on p. 89).
- Sandberg, M, L Eriksson, J Jonsson, M Sjöström, and S Wold (1998). "New chemical descriptors relevant for the design of biologically active peptides A multivariate characterization of 87 amino acids". In: *J. Med. Chem.* 41.14, pp. 2481–2491 (cit. on p. 92).
- Schwaighofer, A, T Schroeter, S Mika, J Laub, A ter Laak, D Sulzle, U Ganzer, N Heinrich, and KR MÄijller (Mar. 2007). "Accurate solubility prediction with error bars for electrolytes: A machine learning approach". In: *J. Chem. Inf. Model.* 47.2, pp. 407–424 (cit. on pp. 90, 104, 119).
- Scitegic Accelrys Software Inc. *Pipeline Pilot Student Edition, version 615* (San Diego, USA): Scitegic Accelrys Software Inc (2007) (cit. on p. 91).

- Skilling, J (2006). "Nested sampling for general Bayesian computation". In: *Bayesian Analysis* 1.4, pp. 833–859 (cit. on p. 96).
- Spjuth, O, M Eklund, M Lapinsh, M Junaid, and JES Wikberg (June 2011). "Services for prediction of drug susceptibility for HIV proteases and reverse transcriptases at the HIV drug research centre". In: *Bioinformatics* 27.12, pp. 1719–1720 (cit. on p. 119).
- Tetko, IV, P Bruneau, H Mewes, DC Rohrer, and GI Poda (Aug. 2006). "Can we estimate the accuracy of ADME-Tox predictions?" In: *Drug Discov. Today* 11.15-16, pp. 700–707 (cit. on p. 89).
- Tiikkainen, P, L Bellis, Y Light, and L Franke (o). "Estimating error rates in bioactivity databases". In: *J. Chem. Inf. Model.* o.o, null (cit. on p. 89).
- Tresp, V (2000). "A Bayesian Committee Machine". In: *Neural Comput.* 12.11, pp. 2719–2741 (cit. on p. 119).
- Tropsha, A and A Golbraikh (2010). "Predictive Quantitative Structure-Activity Relationships modeling". In: *Handbook of Chemoinformatics Algorithms* 33, p. 211 (cit. on pp. 90, 97).
- Tropsha, A and VK Gramatica PaOand Gombar (2003). "The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models". In: *QSAR Comb. Sci.* 22.1, pp. 69–77 (cit. on pp. 90, 97).
- Westen, GJP van, JK Wegner, AP IJzerman, HWT van Vlijmen, and A Bender (2011a). "Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets". In: *Med. Chem. Commun.* 2, pp. 16–30 (cit. on pp. 89, 101, 117).
- Westen, GJP van, JK Wegner, P Geluykens, L Kwanten, I Vereycken, A Peeters, AP IJzerman, HWT van Vlijmen, and A Bender (2011b). "Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development". In: *PLoS ONE* 6.11, e27518 (cit. on p. 98).
- Westen, GJP van, OO van den Hoven, R van der Pijl, T Mulder-Krieger, H de Vries, AP Wegner Jorg K an IJzerman, HWT van Vlijmen, and A Bender (Aug. 2012). "Identifying novel adenosine receptor ligands by simUeous proteochemometric modeling of rat and human bioactivity data". In: *J. Med. Chem.* 55.16, pp. 7010–7020 (cit. on pp. 91, 101, 102, 118).
- Westen, GJP van, A Hendriks, JK Wegner, AP IJzerman, HWT van Vlijmen, and A Bender (2013). "Significantly improved HIV inhibitor efficacy prediction employing proteochemometric models generated from antivirogram data". In: *PLoS Comput. Biol.* 9.2. Ed. by SL Kosakovsky Pond, e1002899 (cit. on pp. 102, 110, 119).
- Wu, D, Q Huang, Y Zhang, Q Zhang, Q Liu, J Gao, Z Cao, and R Zhu (2012). "Screening of selective histone deacetylase inhibitors by proteochemometric modeling". In: *BMC Bioinformatics* 13.1, p. 212 (cit. on p. 101).
- Zhou, P, F Tian, X Chen, and Z Shang (2008). "Modeling and prediction of binding affinities between the human amphiphysin SH3 domain and its peptide ligands

- using genetic algorithm-Gaussian processes". In: *J. Pept. Sci.* 90.6, pp. 792–802 (cit. on p. 90).
- Zhou, P, X Chen, Y Wu, and Z Shang (2010). "Gaussian process: an alternative approach for QSAM modeling of peptides". In: *Amino Acids* 38.1, pp. 199–212 (cit. on p. 90).

Benchmarking the Influence of Simulated Experimental Errors in QSAR

.4 Benchmarking the Influence of Simulated Experimental Errors in QSAR

To date, no systematic study has assessed the effect of random experimental errors on the predictive power of QSAR models. To address this shortage, we have benchmarked the noise sensitivity of 12 learning algorithms on 12 data sets (15,840 models in total), namely: Support Vector Machines (SVM) with radial and polynomial (Poly) kernels, Gaussian Process (GP) with radial and polynomial kernels, Relevant Vector Machines (radial kernel), Random Forest (RF), Gradient Boosting Machines (GBM), Bagged Regression Trees, Partial Least Squares and k-Nearest Neighbours. Model performance on the test set was used as a proxy to monitor the relative noise sensitivity of these algorithms as a function of the level of simulated noise added to the bioactivities from the training set. The noise was simulated by sampling from Gaussian distributions with increasingly larger variances, which ranged from zero to the range of pIC_{50} values comprised in a given data set. General trends were identified by designing a full-factorial experiment, which was analyzed with a normal linear model.

Overall, GBM displayed low noise tolerance, although its performance was comparable to RF, SVM Radial, SVM Poly, GP Poly and GP Radial at low noise levels. Of practical relevance, we show that the *bag fraction* parameter has a marked influence on the noise sensitivity of GBM, suggesting that low values (*e.g.* 0.1-0.2) for this parameter should be set when modelling noisy data. The remaining 11 algorithms display a comparable noise tolerance, as a smooth and linear degradation of model performance is observed with the level of noise. However, SVM Poly and GP Poly display significant noise sensitivity at high noise levels in some cases. Overall, these results provide a practical guide to make informed decisions about which algorithm and parameter values to use according to the noise level present in the data.

.4.1 Introduction

Computational chemogenomic [Cortes-Ciriano et al. (2015a)] techniques capitalize on bioactivity data to (quantitatively) predict and better understand unknown interac-

tions between small molecules and their biomolecular targets. The development of these techniques has been mainly fostered by (i) the increase of computing resources and the availability of scalable machine learning software, and (ii) the advent of high-throughput technologies, which have contributed to a vast increase of proprietary and public bioactivity data [Gaulton et al. (2011); Wang et al. (2012)]. Although public bioactivity databases have grown in size steadily over the last decade, detailed information about the assays used and the experimental errors of the measurements are generally not reported. The question then arises how the experimental errors (or the lack thereof) should be included in the generation and validation of *in silico* predictive models, and to which extent the quality of the data affects models predictive ability on new molecules. These issues need to be addressed prior to model training, *e.g.* which algorithms are robust to noisy input data and to which extent?, and further downstream in the modelling pipeline, *e.g.* how should bioactivity models be validated in the light of the noise of the data?. In this manuscript, we consider experimental error, or simply noise, as the random error of a measured variable, *e.g.* IC₅₀ values [Fuller (2008)].

The quality of the data can be determined by the divergence of the average value of the experimental replicates (*i.e.* sample mean) with respect to the true bioactivity value, which would correspond to the average value of an infinite number of replicates (*i.e.* the population mean). The sample standard deviation decreases with the number of replicates. Thus, the difference between the population and the sample mean will decrease as the number of replicates increases, leading to a more precise estimation of the true bioactivity value. Therefore, in practice, the number of replicates and their standard deviation can serve to determine the quality of a data set. In this line, Wenlock and Carlsson (2014) have benchmarked the influence of the quality of the data in the generation of drug metabolism and pharmacokinetic models using data sets from AstraZeneca. The authors defined high-quality data as those bioactivity values measured in replicates with a standard deviation below a given threshold. This study showed that the quality of the training data is correlated to model performance on external molecules.

Whereas discarding those data-points measured only once would constitute a reasonable cleaning step in the data collection phase [Wenlock and Carlsson (*ibid.*)], in practice, this might lead to a marked decrease in the number of data-points available for modelling, thus compromising the generation of statistically robust models. Therefore, it is paramount to control the trade-off between the quality and the size of the data. Two recent publications [Kalliokoski et al. (2013); Kramer et al. (2012)] have analyzed the variability of pK_i and pIC₅₀ values from ChEMBL. The authors reported standard deviations for heterogeneous pIC₅₀ and pK_i values of 0.68 and 0.54, respectively. In practice, these average experimental errors for public pIC₅₀ and pK_i data, or the experimental errors of each data-point when available [Cortes-Ciriano, I, Bender,

A, and Malliavin (2015); Cortes-Ciriano, I et al. (2015)], can serve to assess whether the predictive power of the models is realistic or not [Brown, Muchmore, and Hajduk (2009); Cortes-Ciriano et al. (2015b); Cortes Ciriano et al. (2014); Cortes-Ciriano, I, Bender, A, and Malliavin (2015); Cortes-Ciriano, I et al. (2015)]. In this line, Brown, Muchmore, and Hajduk (2009) provided practical rules-of-thumb to evaluate the maximum R^2 (coefficient of determination) values attainable for the observed against the predicted pIC_{50} values for a set of compounds as a function of the range of pIC_{50} values considered and of the number of data-points. This scheme was extended by Cortes Ciriano et al. (2014) by also considering the distribution of these pIC_{50} values, and by proposing to calculate the distribution of minimum RMSE and maximum R^2 values given the quality of the data. These distributions of the maximum values for correlation metrics (e.g. R^2 , R_0^2 or Q^2), and of the minimum values in the case of RMSE, can serve to assess whether the predictive power of the models is justified by the quality of the underlying training data, as well as to quantify the probability of obtaining a given R^2 or RMSE value. Thus, the distributions of maximum and minimum values can be regarded as sampling distributions for the validation metrics.

Although there exist algorithms to handle noisy input data [Ge, Xia, and Tu (2010); Qin, Xia, and Li (2010); Rasmussen and Ws (2006); Tsang et al. (2009); Zhang (2004)], the vast majority of the models reported in the medicinal chemistry literature are still based on algorithms that (i) treat the dependent variable as a definite point value, and (ii) that do not consider the experimental errors of the input data. Most machine learning algorithms have been developed assuming noise-free input data [Atla et al. (2011)]. Thus, their application to real-world problems, where noisy data sets are prevalent, might lead to overfitting [Hawkins, Basak, and Mills (2003)], and thus to a decrease of model performance on external data. Assessing the magnitude of this decrease and the robustness to noise of different learning paradigms has been subject of intense investigation in the machine learning community [Angluin and Laird (1988); Atla et al. (2011); Kearns (1998); Manolopoulos and Spirakis (2003); Natarajan et al. (2013); Teytaud (2001); Zhu and Wu (2004); Zhu, Wu, and Chen (2003)]. Most of these studies have dealt with classification problems [Manolopoulos and Spirakis (2003); Nettleton, Orriols-Puig, and Fornells (2010); Zhu and Wu (2004); Zhu, Wu, and Chen (2003)]. Nettleton, Orriols-Puig, and Fornells (2010) compared the tolerance to noise, both on the descriptors and on the class labels, of the following classifiers on 13 highly unbalanced data sets: (i) Naive Bayes [John and Langley (1995)], (ii) C4.5 decision trees [Quinlan (1993)], (iii) IBk instance-based learner [Aha, Kibler, and Albert (1991)], and (iv) Sequential Minimal Optimization (SMO) Support Vector Machines (SVM) [Platt (1998)]. The authors found Naive-Bayes as the most robust algorithm, and SMO SVM the most sensitive, in agreement with Atla et al. (2011). Interestingly, the authors showed that noise in the labels affects to a greater extent the performance of the learners when compared to noise in the descriptors. These results are reminiscent of the work by Norinder and Bostrom

(2012). Therein, the authors benchmarked the tolerance to noisy chemical descriptors of decision tree ensembles across 16 QSAR data sets, finding that, in practice, the introduction of uncertainty in chemical descriptors does not reduce model performance.

To date, no systematic study has assessed the effect of random experimental errors of bioactivities on the predictive power of commonly used learning methods in QSAR. The present contribution aims at addressing this shortage. We recently compared the influence of the experimental errors on the predictive power of regression Gaussian Process (GP) models, finding that the radial kernel appears more robust to noisy input data than polynomial kernels [Cortes Ciriano et al. (2014)], what agrees with the machine learning literature [Steinwart (2002)]. Here, we extend this study by evaluating the influence of the experimental errors on the predictive power of 8 commonly used machine learning algorithms in a robust statistical manner (Table .4.1). The 8 machine learning algorithms, covering 5 learning paradigms, gave rise to 12 models as some parameters vary for a given method, *e.g.* kernel type. For the sake of clarity, these 12 models will be referred to as algorithms or models throughout the rest of the manuscript. The learning paradigms and algorithms are, respectively (Table .4.1): (i) kernel methods: GP (radial and polynomial kernel), SVM (radial and polynomial kernel), and Relevant Vector Machines (RVM) (radial kernel), (ii) ensemble bagging methods: Random Forest (RF) and Bagged CART Regression Trees (Tree bag), (iii) ensemble boosting methods: Gradient Boosting Machines (GBM), (iv) linear methods: Partial Least Squares (PLS), and (v) k-Nearest Neighbour (NN) learning (5-NN, 10-NN and 20-NN). We used 12 QSAR data sets reporting compound potency as pIC_{50} values (Table .4.2). Chemical structures were encoded with Morgan fingerprints and 1-D and 2-D physicochemical descriptors. For each triplet (data set, algorithm, descriptor type) we trained each of the 12 models 11 times, each time with an increasingly higher level of simulated noise added to the pIC_{50} values from the training set. Model performance on the test set, quantified by the RMSE values for the observed against the predicted pIC_{50} values, was used as a proxy to assess the noise sensitivity of the 12 algorithms explored here. In order to identify general trends in a statistically sound manner, and thus to assess the robustness of these algorithms with respect to the level of noise in the input data, we designed a balanced fixed-effect full-factorial experiment with replications. This experimental design was analyzed with a normal linear model using the RMSE values on the test set as the dependent variable.

Learning Paradigm	Algorithm	Parameters and values used in CV	ref
Kernel	Gaussian Process Radial Kernel (GP Radial)	$\sigma \in \{2^{-10}, 2^{-4}..2^2, 2^4\}$; σ_d^2 (noise variance): 0.001	Rasmussen and Ws (2006)
Kernel	Gaussian Process Polynomial Kernel (GP Poly)	scale $\in \{2^{-10}, 2^{-4}..2^2, 2^4\}$; degree $\in (k)_{k=2}^6$; σ_d^2 : 0.001	Rasmussen and Ws (2006)
Kernel	Relevant Vector Machines Radial Kernel (RVM Radial)	$\sigma \in \{2^{-6}, 2^{-4}..2^2, 2^4\}$	Tipping (2000)
Kernel	Support Vector Machines Radial Kernel (SVM Radial)	$\sigma \in \{2^{-10}, 2^{-4}..2^2, 2^4\}$; $C \in \{2^{-10}, 2^{-4}..2^2, 2^4, 10, 100\}$	Cortes and Vapnik (1995)
Kernel	Support Vector Machines Polynomial Kernel (SVM Poly)	scale $\in \{2^{-10}, 2^{-4}..2^2, 2^4\}$; offset: 0; degree $\in (k)_{k=2}^6$; $C \in \{2^{-10}, 2^{-4}..2^2, 2^4, 10, 100\}$	Cortes and Vapnik (1995)
Ensemble: bagging	Bagged CART Regression Trees (Tree bag)	-	Breiman et al. (1984)
Ensemble: boosting	Gradient Boosting Machines (GBM)	Learning rate (ν) $\in \{0.04, 0.08, 0.12, 0.16\}$; n_{trees} : 500; tree complexity (t_c): 25; bag fraction (η): 0.5	Friedman (2001); Natekin and Knoll (2013)
Ensemble: bagging	Random Forest (RF)	n_{trees} : 500	Breiman (2001)
Linear	Partial Least Squares (PLS)	-	Wold, Sjöström, and Eriksson (2001)
k-Nearest (NN)	5-NN	$N_{\text{neighbours}}$: 5	Fix and Hodges (1989)
k-Nearest Neighbours	10-NN	$N_{\text{neighbours}}$: 10	Fix and Hodges (1989)
k-Nearest Neighbours	20-NN	$N_{\text{neighbours}}$: 20	Fix and Hodges (1989)

Table .4.1: Algorithms benchmarked in this study. The third column indicates the parameters that were tuned using grid search and cross-validation (CV). The default values were used for those parameters not indicated therein.

.4.2 Materials and Methods

.4.2.1 Data sets

.4.2.2 Data sets

We gathered a total of 12 QSAR data sets from the literature (references given in Table .4.2) and from ChEMBL database version 19 [Gaulton et al. (2011)]. All data sets report compound potency as IC_{50} values. These values were modelled in a logarithmic scale ($pIC_{50} = -\log_{10} IC_{50}$). The size of the data sets range from 334 datapoints (Human Factor 7) to 2,312 (Cyclooxygenase 2). Detailed information about these data sets can be found in Table .4.2.

.4.2.3 Molecular Representation

The function *StandardiseMolecules* from the R package *camb* [Murrell et al. (2014)] was used to normalize all chemical structures using the default options. This normalization step is crucial for the generation of compound descriptors, as the value of most

of them (except for *e.g.* heavy atom counts) depend on a consistent representation of molecular properties such as aromaticity of ring systems, tautomer representation or protonation states.

.4.2.4 Molecular Representation

The function *StandardiseMolecules* from the R package *camb* [Murrell et al. (2014)] was used to normalize all chemical structures using the default options. This normalization step is crucial prior to the generation of compound descriptors, as the value of most of them (except for *e.g.* heavy atom counts) depend on a consistent representation of molecular properties such as aromaticity of ring systems, tautomer representation or protonation states.

.4.2.5 Compound Descriptors

Compounds were encoded with circular Morgan fingerprints [Glen et al. (2006); Rogers and Hahn (2010)] calculated using RDKit (release version 2013.03.02) [Cortes-Ciriano (2013); Landrum (2006)]. Morgan fingerprints encode compound structures by considering radial atom neighborhoods [Rogers and Hahn (2010)]. The choice of Morgan fingerprints as compound descriptors was motivated by the high retrieval rates obtained with these fingerprints in benchmarking studies of compound descriptors [Bender et al. (2009); Koutsoukas et al. (2013)]. The size of the fingerprints was set to 256 bits, whereas the maximum radius of the substructures considered was set to 2 bonds. To calculate the fingerprints for a given compound, each substructure in that compound, with a maximal diameter of four bonds, was assigned to an unambiguous identifier. Subsequently, these substructures were mapped into a hashed array of counts. The position in the array where each substructure was mapped was given by the modulo of the division of the substructure identifier by the fingerprint size. A total of 188 1-D and 2-D physicochemical descriptors was computed with RDKit (release version 2013.03.02) [Landrum (2006)].

Data set label	Biological endpoint	Datapoints	Activity range (pIC ₅₀)	Bioactivity standard deviation (pIC ₅₀)	Data source	Reference
COX-1	Human cyclooxygenase-1	1,347	4.00 - 9.00	0.90	ChEMBL 16	Cortes-Ciriano et al. (2015b)
COX-2	Human cyclooxygenase-2	2,312	4.00 - 10.70	1.20	ChEMBL 16	Cortes-Ciriano et al. (2015b)
DHFR rat	Rat dihydrofolate reductase	760	4.00 - 9.50	1.25	ChEMBL 16	Paricharak et al. (2015)
DHFR homo	Human dihydrofolate reductase	744	4.00 - 9.72	1.32	ChEMBL 16	Paricharak et al. (2015)
F7	Human factor-7	344	4.04 - 8.18	0.97	US20050043313	Chen et al. (2013)
IL4	Human interleukin-4	599	4.67 - 8.30	0.57	WO 2006/133426 A2	Chen et al. (2013)
MMP2	Human matrix metalloproteinase-2	443	5.10 - 10.30	0.99	WO2005042521	Chen et al. (2013)
P61169	Rat D ₂ dopamine receptor	1,968	4.00 - 10.22	1.02	ChEMBL 19	Gaulton et al. (2011)
P41594	Human metabotropic glutamate receptor 5	1,381	4.00 - 9.40	1.14	ChEMBL 19	Gaulton et al. (2011)
P20309	Human muscarinic acetylcholine receptor M3	779	4.00 - 10.50	1.70	ChEMBL 19	Gaulton et al. (2011)
P25929	Human neuropeptide Y receptor type 1	501	4.11 - 10.70	1.40	ChEMBL 19	Gaulton et al. (2011)
P49146	Human neuropeptide Y receptor type 2	561	4.00 - 10.15	1.27	ChEMBL 19	Gaulton et al. (2011)

Table .4.2: Data sets modelled in this study. These data sets can be found in the references indicated in the last two columns.

.4.2.6 Model generation

The function *RemoveNearZeroVarianceFeatures* from the R package *camb* was used to remove those descriptors displaying constant values across all compounds (near-zero variance descriptors) using a cut-off value equal to 30/1 [Kuhn (2008); Kuhn and Json (2013); Murrell et al. (2014)]. Subsequently, the remaining descriptors were centered to zero mean and scaled to unit variance (z-scores) with the function *PreProcess* from the R package *camb*.

Grid-search with 5-fold cross validation (CV) was used to optimize the model parameters [Hawkins, Basak, and Mills (2003)]. The whole data set was split into 6 folds of the same size by stratified sampling of the pIC₅₀ values. One fold, 1/6, was withheld as the test set and served to assess the predictive power of the models. The remaining folds, 5/6, constituted the training set and were used to optimize the values of the parameters in the following way. For each combination of parameter values in the grid, a model was trained on 4 folds from the training set, and the values for the remaining fold were then predicted. This procedure was repeated 5 times, each time holding out a different fold. The values of the parameters exhibiting the lowest average RMSE value across these 5 repetitions was considered as optimal. Subsequently, a model was trained on the whole training set, using the optimized values for the parameters. The predictive power of this model was assessed on the test set by calculating the RMSE value for the observed against the predicted bioactivities.

We run five replicates (models) for all factor level combinations. The training and test sets for each replicate were composed of different subsets of the complete data set. In order to make the results comparable on a given data set for a given replicate, all models were trained using the same fold composition. Thus, for a given data set and replicate the same test set was used to assess the predictive power of all algorithms at all noise levels. We note in particular that simulated noise was never added to the test sets.

.4.2.7 Machine Learning Implementation

Machine learning models were built in R using the wrapper packages *caret* [Kuhn (2008)] and *camb* [Murrell et al. (2014)]. The following R packages were used to train the machine learning algorithms considered here: (i) *kernlab* [Karatzoglou et al. (2004)] for Support Vector Machines (SVM) [Ben-Hur et al. (2008)], Relevance Vector Machines (RVM) [Tipping (2000)], and Gaussian Processes (GP) [Rasmussen and Ws (2006)], (ii) *gbm* [Ridgeway (2013)] for Gradient Boosting Machines (GBM) [Friedman (2001)], (iii) *class* [Venables and Ripley (2002)] for k-Nearest Neighbours (KNN) [Fix and Hodges (1989)], (iv) *pls* [Mevik, Wehrens, and Liland (2013)] for Partial Least Squares (PLS) [Wold, Sjöström, and Eriksson (2001)], (v) *randomForest*

[Liaw and Wiener (2002)] for Random Forest (RF) [Breiman (2001)], and (vi) *ipred* [Ps and Hothorn (2013)] for bagged Classification And Regression Trees (CART) [Breiman et al. (1984)].

.4.2.8 Simulation of Noisy Bioactivities

To assess the effect of random experimental errors on the predictive power of QSAR, 11 models *per* triplet (data set, algorithm, descriptor type) were trained, each of them with an increasingly larger level of noise, ϵ , added to the pIC_{50} values from the training set. Noise levels were simulated by sampling from a Gaussian distribution with zero mean and corresponding larger variance, σ_{noise}^2 . The value of the variance across the 11 noise levels was defined as a function of the range of bioactivities considered in each data set:

$$\{\sigma_{\text{noise } i}^2\}_{i=0}^{10} = (\max(\text{pIC}_{50}) - \min(\text{pIC}_{50})) * \text{Noise}_{\text{level } i}; \quad (.4.1)$$

where i is the index of the noise level, and $(\max(\text{pIC}_{50}) - \min(\text{pIC}_{50}))$ corresponds to the range of pIC_{50} values comprised in a given data set. $\text{Noise}_{\text{level}}$ was defined as:

$$\text{Noise}_{\text{level}} = \{i/10\}_{i=0}^{10} \quad (.4.2)$$

The first noise level corresponds to a variance of 0, *i.e.* no noise was added and, therefore, the bioactivity values corresponded to the reported pIC_{50} values. The bioactivity values for the training set, $\mathbf{Y}_{\text{tr Noise } i}$, were calculated as:

$$\mathbf{Y}_{\text{tr Noise } i} = \mathbf{Y}_{\text{tr}} + \epsilon \sim N(0, \sigma_{\text{noise } i}^2) \quad (.4.3)$$

where $\mathbf{Y}_{\text{tr Noise } i}$ corresponds to the noisy pIC_{50} values for noise level i , \mathbf{Y}_{tr} to the reported pIC_{50} values, and ϵ to the vector containing the simulated noise. Therefore, $\mathbf{Y}_{\text{tr Noise } i}$ was used as the dependent variable during model training.

The simulated noise is thus sampled from a Gaussian distribution with variance: $\sigma_{\text{noise } i}^2$. This choice makes the results comparable across data sets irrespective of the range of pIC_{50} values comprised in each of them. For convenience, noise levels will be reported in the following as percentage points (*e.g.* noise level 0.5 would correspond to 50%). We note in particular that this noise simulation method is sensitive to outliers (*i.e.* highly active or inactive compounds), as the range of pIC_{50} values would be considerably enlarged in their presence. Thus, we advise to remove outliers for the generation of the range of noise levels.

We are aware that the reported pIC_{50} values, which would correspond to noise level 0%, already contain random experimental errors. We are not overly concerned

about this issue given that all models trained on a given data set need to deal with that base level of noise.

4.2.9 Experimental Design

In order to investigate the relative noise sensitivity of the 12 algorithms, a balanced fixed-effect full-factorial experiment with replications was designed [Winer, Brown, and Michels (1991)]. The following four factors were considered, namely: data set (Data set), noise level (Noise), descriptor type (Descriptor type), and learning algorithm (Algorithm). Given that the factor Data set has an influence on model performance, as some data sets are better modelled than others, but it is irrelevant to the effect of interest (*i.e.* the effect of noise across learning algorithms irrespective of the data set), it was considered as a blocking factor [Winer, Brown, and Michels (*ibid.*)]. Similarly, the factor Descriptor type was also added as a blocking factor.

This factorial design was studied with the following linear model:

$$\text{RMSE}_{i,j,k,l,m \text{ test}} = \text{Data set}_i + \text{Descriptor type}_j + \text{Noise}_k + \text{Algorithm}_l + (\text{Noise} * \text{Algorithm})_{kl} + \mu_0 + \epsilon_{i,j,k,l,m} \quad (.4.4)$$

$$(i \in \{1, \dots, N_{\text{data sets}} = 12\}; j \in \{1, \dots, N_{\text{Descriptor type}} = 2\}; k \in \{1, \dots, N_{\text{noise levels}} = 11\}; l \in \{1, \dots, N_{\text{algorithms}} = 12\}; m \in \{1, \dots, N_{\text{resamples}} = 100\})$$

where the response variable, $\text{RMSE}_{i,j,k,l,m \text{ test}}$, corresponds to the RMSE values on the test set. Data set_i , Descriptor type_j , Noise_k and Algorithm_l are the main effects, and $\text{Noise} * \text{Algorithm}$ corresponds to the interaction term between the learning algorithm and the noise level. The factor levels Random Forest (Algorithm), F7 (Data set), Morgan fingerprints (Descriptor type), and Noise : 0 (Noise level) were used as reference factor levels to calculate the intercept term of the linear model, μ_0 , which is simply the mean $\text{RMSE}_{\text{test}}$ value for this combination of factor levels. The coefficients (slopes) for the other factor level combinations, *e.g.* GP:Noise 20%, correspond to the difference between their mean $\text{RMSE}_{\text{test}}$ value and the intercept. The error term, $\epsilon_{i,j,k,l,m}$, corresponds to the random error of each $\text{RMSE}_{\text{test}}$ value, which are defined as: $\epsilon_{i,j,k,l,m} = \text{RMSE}_{i,j,k,l,m \text{ test}} - \overline{\text{RMSE}_{i,j,k,l}}$. These errors are assumed to (i) be mutually independent, (ii) have zero expectation, and (iii) have constant variance.

One model was trained for all factor level combinations, giving rise to 15,840 models (12 learning algorithms x 11 noise levels x 12 data sets x 2 descriptor types x 5 replications). The predictive power of the models, which serves as a proxy to evaluate the

noise sensitivity of the algorithms, was assessed on the test set and quantified by the RMSE value, $RMSE_{i,j,k,l,m=1 \text{ test}}$, for the observed against the predicted bioactivities. Bootstrapping [Efron and Tibshirani (1993)] was used to generate 100 resamples ($N_{\text{replications}}$) for these $RMSE_{i,j,k,l,m \text{ test}}$ values (*i.e.* $RMSE_{i,j,k,l,m=2:100 \text{ test}}$), thus ensuring a balanced experimental design. Therefore, the total number of observations considered in the linear model was 1,584,000 (15,840 trained models \times 100 resamples each). The significance level was set to 5%. The normality and homoscedasticity assumptions of the linear model were respectively assessed with (i) quantile-quantile (Q-Q) plots, and (ii) by visual inspection of the $RMSE_{\text{test}}$ distributions, and by plotting the $RMSE_{\text{test}}$ values against the residuals [Winer, Brown, and Michels (1991)].

The interaction term was introduced to assess the interaction effects between the factors Algorithm and Noise. Figure .4.1 illustrates the concept of two-way interactions (*i.e.* between two factors) with a toy example where both the factor Algorithm and the factor Noise have only two levels, namely: Algorithm: RF and GP; and Noise: 0 and 100. There is no interaction between two factors when the difference of the mean values of the dependent variable (in this case RMSE) across the levels of a factor (*e.g.* Algorithm) does not change across the levels of a second factor (*e.g.* Noise). In the example, this means that the difference in RMSE between RF and GP is the same irrespective of the level of noise (Figure .4.1A). Therefore, it could be concluded that the difference in performance between RF and GP is not affected by the level of noise. This can be easily showed in an interaction plot (Figure .4.1A right panel) by plotting the levels of the factor Noise against the RMSE values for GP and RF. In the absence of interaction, the lines connecting the points corresponding to the RMSE values for each algorithm along the levels of the factor Noise are parallel.

By contrast, in the presence of interaction (Figure .4.1B), the difference in RMSE between RF and GP would vary across the levels of the factor Noise. Therefore, the performance of the algorithms would depend on the level of noise, and the lines in the interaction plot would not be parallel (Figure .4.1B right panel). Consequently, it would not be possible to conclude about the effect of a single independent variable (*e.g.* factor Algorithm) on the RMSE (termed *main effects*), as the RMSE values would depend on the level of other factors (*e.g.* Noise). For instance, in Figure .4.1B, it would not be possible to state that RF perform better than GP because the RMSE values corresponding to each algorithm depend on the level of noise. Therefore, in the presence of interaction, the influence of a factor on the dependent variable has to be analyzed for each level of the second factor, which is known as analysis of *simple effects*. In the example (Figure .4.1B), this would correspond to stating that GP perform better than RF when no noise is present in the input data, whereas RF perform better than GP when the level of noise equals 100%.

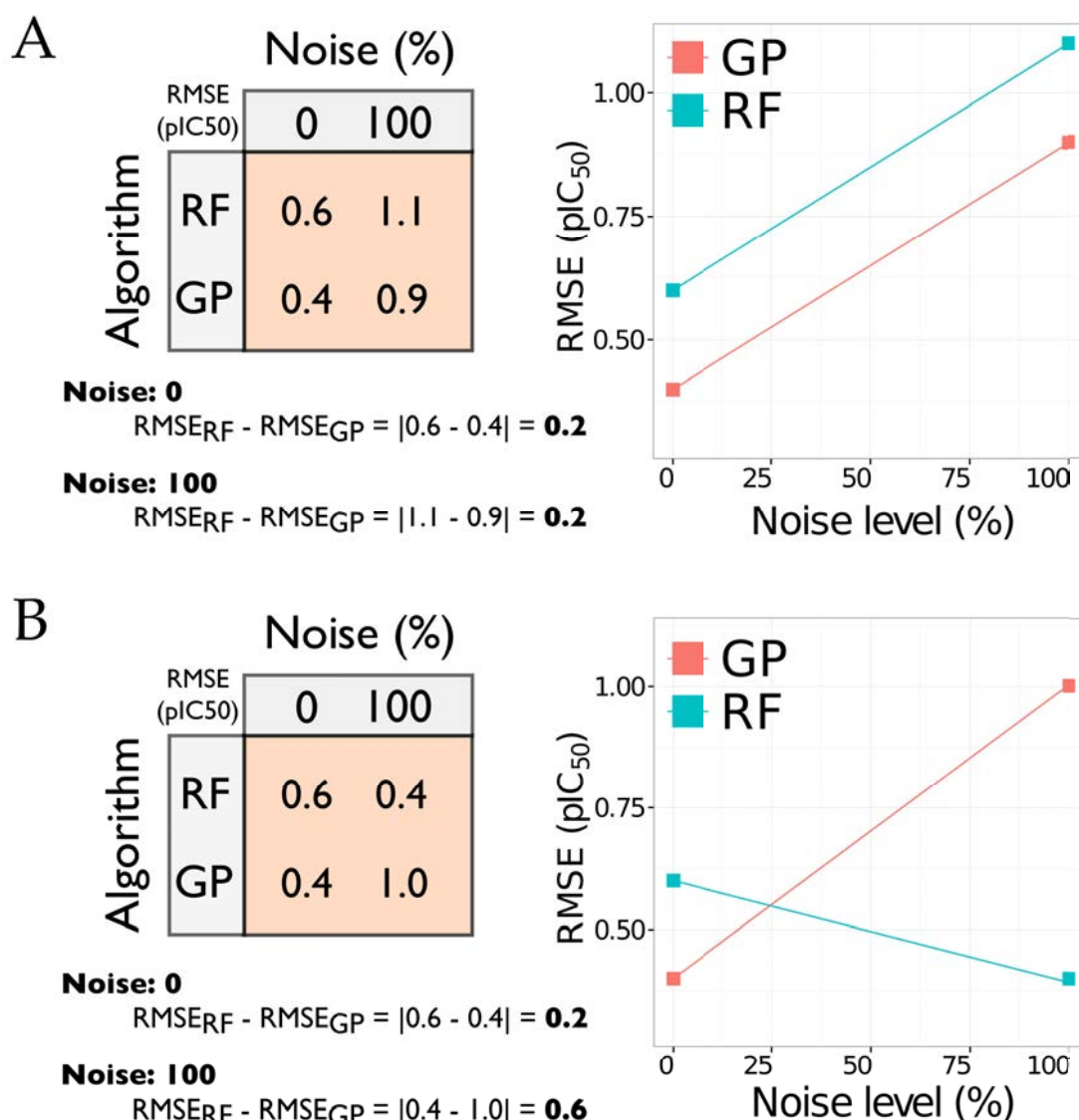


Figure 4.1: Illustration of two-way interactions between two-level factors, namely: Algorithm: RF and GP; and Noise: 0 and 100. There is interaction between two factors (**A**) when the difference of the mean RMSE values (response variable) across the levels of a factor (*e.g.* Algorithm) does not change across the levels of a second factor (*e.g.* Noise). In the example, the difference in performance between GP and RF is the same across all levels of factor Noise. This is illustrated by the presence of parallel lines. By contrast, the presence of non parallel lines (**B**) indicates that the performance of GP and RF changes depending on the noise level. Thus, GP outperforms RF at noise level 0%, whereas RF outperforms GP at noise level 100%.

.4.3 Results

The fitted linear model displayed an adjusted R^2 value (adjusted by the complexity -number of parameters- of the model) of 0.71, and a standard error for the residuals of 0.27. This analysis showed significant interaction between the factors Noise and Algorithm (P-value < 0.001), thus indicating that the effect of Noise on the $RMSE_{test}$ values is not constant across the algorithms studied, and *vice versa*. This is illustrated by the presence of non-parallel lines in the interaction plots (Figure .4.2A,B). Figure .4.3A shows the distributions of $RMSE_{test}$ values for the 12 learning algorithms across all replications, data sets, descriptor types, and noise levels, whereas Figure .4.3B reports the distributions of $RMSE_{test}$ values for the 11 noise levels considered across all data sets, descriptor types, algorithms and replications. The values for the coefficients, namely slopes and intercept, and their P-values are reported in Table .4.3.

RVM Radial constantly displays the worst predictive power (Figure .4.2A), followed by GBM, Tree bag, 5-NN, 10-NN and 20-NN. This effect can also be inferred from the high value for the slope corresponding to the factor level *RVM Radial*, namely 0.34 pIC_{50} units (Table .4.3, first column). Although low, the predictive power of RVM exhibits a smooth degradation with the level of noise comparable to the other methods, with the exception of GBM. This indicates that RVM are less sensitive to noise than GBM. Interestingly, GBM displays mean $RMSE_{test}$ values comparable to those obtained with RF, SVM Radial, GP Radial, GP Poly and SVM Poly for noise levels 0 and 10% (Figure .4.2A), which is also indicated by the slope value for GBM, namely -0.09 (Table .4.3, first column). However, the mean $RMSE_{test}$ values significantly increase from noise level 20% upwards, thus indicating that the performance of GBM highly depends on the noise level. The sensitivity to noise of the remaining 11 algorithms displays a smooth and linear degradation with the level of noise (Figure .4.2B), showing that these algorithms exhibit a comparable tolerance to noise, and, thus, are less prone to overfitting than GBM.

The sensitivity to noise of boosting algorithms has been previously reported [Long and Servedio (2010)]. Introducing randomness in ensemble modelling by subsampling has proved efficient to increase the generalization ability of a model and to reduce its susceptibility to overfitting [Breiman (2001); Dietterich (2000); Natekin and Knoll (2013)]. In GBM, randomness during training is introduced by controlling the fraction (*bag fraction*) of the training data randomly sampled without replacement that will be used to fit the next base tree learner at each consecutive learning iteration. The value for *bag fraction* is set to 0.5 by default [Natekin and Knoll (2013)]. To further understand the effect of this parameter on the noise tolerance of GBM in QSAR, we trained a model using Morgan fingerprints as compound descriptors for all Algorithm-Noise combinations across a wide range of *bag fraction* values (*bag*

$fraction \in (k/10)_{k=1}^{10}$) for 7 data sets of diverse size, thus giving rise to 840 models (7 data sets * 12 algorithms * 10 *bag fraction* values). Figure .4.4 reports the mean $RMSE_{test}$ values for these models. Up to a noise level of 30%, the performance of all models is comparable across the range of *bag fraction* values explored. By contrast, from 30% onwards the difference increases abruptly. In all cases, the mean $RMSE_{test}$ difference between models trained with *bag fraction* values of 1 and 0.1-0.2 increases with the noise level, reaching a difference value of ~ 1.5 pIC_{50} units at noise level 100%. Taken together, these data evidence that a proper tuning of the *bag fraction* parameter is required to palliate the noise sensitivity of GBM.

SVM Poly displayed low noise sensitivity at low noise levels, although the tolerance to noise conspicuously decreased at noise levels higher than 50%. This is illustrated by the inter-point distance in Figure .4.2A. This phenomenon was less marked for GP Poly, and was not observed for GP Radial nor SVM Radial. Overall, the noise sensitivity of GP Poly is comparable to that of SVM Radial, GP Radial and RF. Nevertheless, GP Poly and SVM Poly exhibit high noise sensitivity in some cases, as indicated by the corresponding tails in the violin plots in Figure .4.3A.

Similar to GBM, the machine learning community has reported the propensity of k-Nearest Neighbours to overfitting when modelling noisy data in classification settings [Kononenko and Kukar (2007); Sánchez, Luengo, and Herrera (2013, 2014); Wu, Ianakiev, and Govindaraju (2002)]. The performance of 5-NN, 10-NN and 20-NN was found comparable to that of Tree bag and PLS, and lower than that of RF, SVM Radial, SVM Poly, GP Poly and GP Radial. The sensitivity to noise of Tree bag, 5-NN, 10-NN and 20-NN decreased more sharply at high noise levels, which can be observed by the inter-point distance in Figure .4.2A. As a rule of thumb, it is considered that the sensitivity to noise decreases with the increase of the number of neighbours (k) [Everitt et al. (2011)]. Here, we did not observe this trend, as the noise sensitivity of 5-NN, 10-NN and 20-NN remains constant across all noise levels, as illustrated by the parallel lines observed in Figure .4.2A. Therefore, k-NN appears robust to noise in the context of QSAR across the data sets studied. The performance of k-NN models was slightly lower than that of Tree bag across all noise levels (Figure .4.2A), whereas RF constantly displayed comparable predictive power (Figure .4.2A,B) to Tree bag. It is known that bagging reduces the variance of the final model [Hastie, Tibshirani, and Friedman (2001)]. In the case of RF, the additional layer of randomization is expected to decrease the sensitivity to noise, and, thus, lead to higher predictive power on the test set. However, this effect was not observed across the algorithms, noise levels, data sets and descriptors explored in this study.

Worth of mention is the fact that at the highest noise levels explored, PLS displays the highest predictive power, as well as the smoothest degradation in performance as the level of simulated noise increases. This can be observed by the low inter-point

distance in Figure .4.2A, and by the low slope of the line corresponding to PLS (orange) in Figure .4.2B. Therefore, PLS appears more robust to noise than more algorithmically complex techniques such as RF.

Factor level	Slope	P-value	Factor level	Slope	P-value	Factor level	Slope	P-value
RF with Noise 0	0.60	<0.01	SVM Radial : Noise 10	0.02	<0.01	Tree bag : Noise 70	0.07	<0.01
			SVM Poly : Noise 10	0.03	<0.01	GBM : Noise 70	-0.07	<0.01
SVM Radial	-0.07	<0.01	GP Radial : Noise 10	-0.01	0.03	PLS : Noise 70	-0.07	<0.01
SVM Poly	-0.03	<0.01	GP Poly : Noise 10	0.02	<0.01	SVM Radial : Noise 80	-0.07	<0.01
GP Radial	-0.06	<0.01	RVM Radial : Noise 10	0.00	0.58	GP Poly : Noise 80	-0.03	<0.01
GP Poly	-0.04	<0.01	5 NN : Noise 10	-0.02	<0.01	Tree bag : Noise 80	-0.03	<0.01
RVM Radial	0.34	<0.01	10 NN : Noise 10	-0.02	<0.01	GBM : Noise 80	0.54	<0.01
5 NN	0.05	<0.01	20 NN : Noise 10	-0.02	<0.01	PLS : Noise 80	-0.09	<0.01
10 NN	0.05	<0.01	Tree bag : Noise 10	-0.02	<0.01	SVM Radial : Noise 80	0.03	<0.01
20 NN	0.05	<0.01	GBM : Noise 10	0.04	<0.01	5 NN : Noise 80	0.08	<0.01
Tree bag	-0.02	<0.01	PLS : Noise 10	-0.02	<0.01	10 NN : Noise 80	-0.09	<0.01
GBM	-0.09	<0.01	SVM Radial : Noise 20	0.03	<0.01	20 NN : Noise 80	-0.09	<0.01
PLS	0.09	<0.01	SVM Poly : Noise 20	0.06	<0.01	Tree bag : Noise 80	-0.01	0.01
			GP Radial : Noise 20	-0.01	<0.01	GBM : Noise 80	1.39	<0.01
Noise 10	0.03	<0.01	GP Poly : Noise 20	0.02	<0.01	PLS : Noise 80	-0.27	<0.01
Noise 20	0.08	<0.01	RVM Radial : Noise 20	0.02	<0.01	SVM Radial : Noise 90	-0.08	<0.01
Noise 30	0.14	<0.01	5 NN : Noise 20	-0.05	<0.01	SVM Poly : Noise 90	0.11	<0.01
Noise 40	0.20	<0.01	10 NN : Noise 20	-0.05	<0.01	GP Radial : Noise 90	-0.17	<0.01
Noise 50	0.28	<0.01	20 NN : Noise 20	-0.05	<0.01	GP Poly : Noise 90	-0.10	<0.01
Noise 60	0.35	<0.01	Tree bag : Noise 20	-0.03	<0.01	RVM Radial : Noise 90	0.05	<0.01
Noise 70	0.45	<0.01	GBM : Noise 20	0.17	<0.01	5 NN : Noise 90	-0.01	<0.01
Noise 80	0.54	<0.01	PLS : Noise 20	-0.04	<0.01	10 NN : Noise 90	-0.01	<0.01
Noise 90	0.63	<0.01	SVM Radial : Noise 30	0.05	<0.01	20 NN : Noise 90	-0.01	<0.01
Noise 100	0.73	<0.01	SVM Poly : Noise 30	0.07	<0.01	Tree bag : Noise 90	0.01	0.06
			GP Radial : Noise 30	0.00	<0.01	GBM : Noise 90	1.58	<0.01
Physicochemical Descs.	-0.03	<0.01	GP Poly : Noise 30	0.02	<0.01	PLS : Noise 90	-0.28	<0.01
			RVM Radial : Noise 30	0.06	<0.01	SVM Radial : Noise 100	-0.17	<0.01
			5 NN : Noise 30	-0.05	<0.01	SVM Poly : Noise 100	0.18	<0.01
			10 NN : Noise 30	-0.05	<0.01	GP Radial : Noise 100	-0.23	<0.01
			20 NN : Noise 30	-0.05	<0.01	GP Poly : Noise 100	-0.10	<0.01
			Tree bag : Noise 30	-0.03	<0.01	RVM Radial : Noise 100	0.03	<0.01
			GBM : Noise 30	0.34	<0.01	5 NN : Noise 100	-0.08	<0.01
			PLS : Noise 30	-0.08	<0.01	10 NN : Noise 100	-0.08	<0.01
			SVM Radial : Noise 40	0.04	<0.01	20 NN : Noise 100	-0.08	<0.01
			SVM Poly : Noise 40	0.07	<0.01	Tree bag : Noise 100	0.00	0.58
			GP Radial : Noise 40	-0.02	<0.01	GBM : Noise 100	1.76	<0.01
			GP Poly : Noise 40	0.00	<0.01	PLS : Noise 100	-0.38	<0.01

Table 4.3: Values for the slopes (coefficients) and P-values for the fitted linear model. The factor levels Random Forest (Algorithm), F7 (Data set), Morgan fingerprints (Descriptor type), Replicate 1 (Replicate) and Noise : 0 (Noise level) were used to calculate the intercept term of the linear model. Noise levels are reported as percentage points. The significance level was set to 5%, whereas all P-values are two-sided.

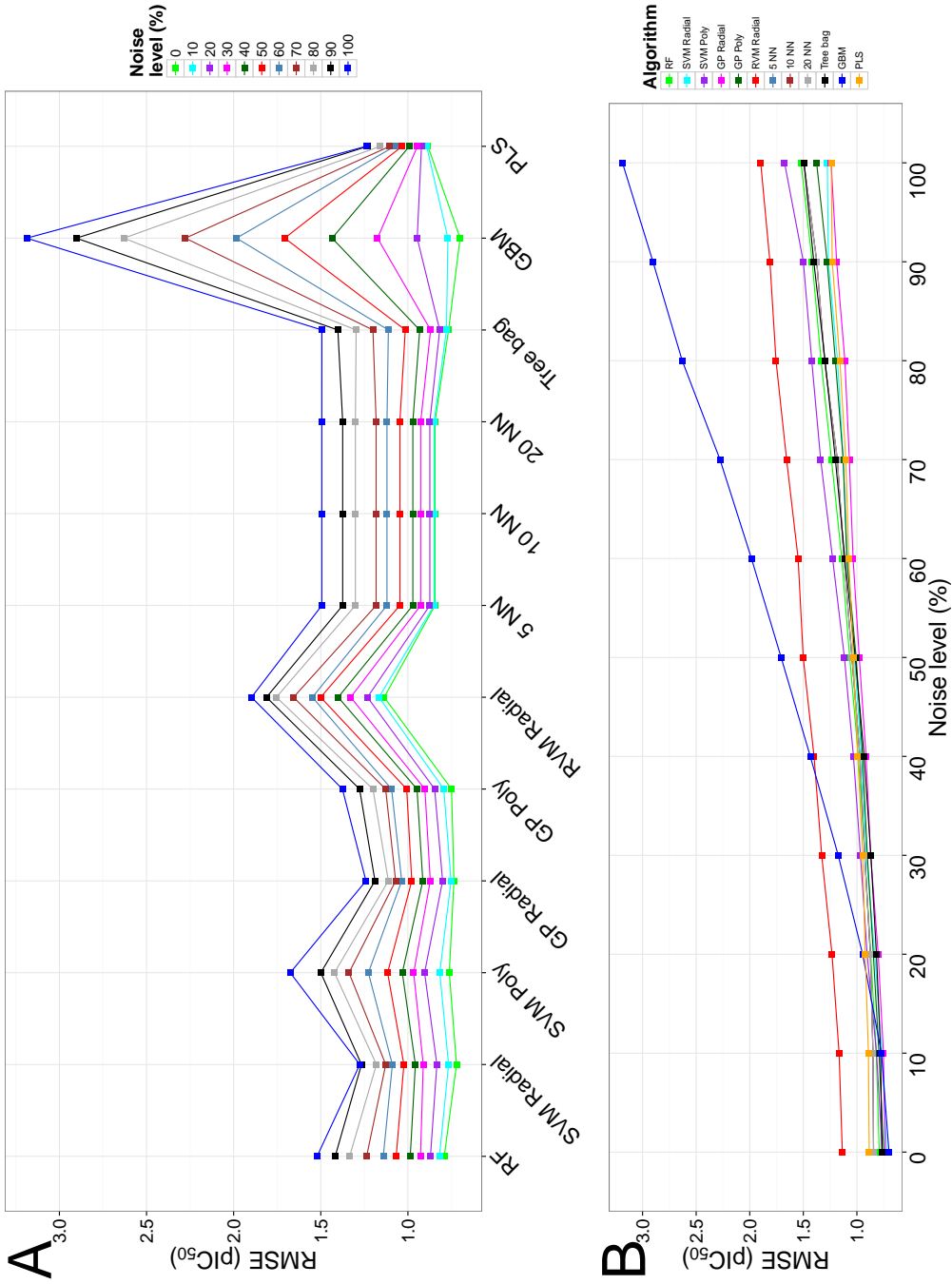


Figure 4.2: **A,B.** Interaction plots. Median RMSE_{test} values across all data sets for two-way combinations of factors. The data set-specific intercept was subtracted from the RMSE_{test} values in order to make the results comparable across the seven data sets.

4.4 Discussion

We have benchmarked the noise sensitivity of 12 learning algorithms commonly used in QSAR, comprising 5 learning paradigms, on 12 data sets using two descriptor types, namely: Morgan fingerprints (topological descriptors) and physicochemical-property-based descriptors. Model performance on the test set was used as a proxy to monitor the relative noise sensitivity of these algorithms as a function of the level of noise added to the bioactivities from the training set. The noise was simulated by sampling from Gaussian distributions with increasingly larger variances, which ranged from zero to the range of pIC_{50} values comprised in a given data set. Although the exploration of machine learning algorithms (and data sets) reported is not exhaustive, we have conducted robust statistical analyses, which have evidenced general trends about the behaviour of these algorithms across different noise levels and, in the case of kernel methods, across kernel types. Overall, GBM displayed low tolerance to noisy bioactivities although its performance was comparable to RF, SVM Radial, SVM Poly, GP Poly and GP Radial for low noise levels.

We note in particular that at noise level 0%, the lowest predictive power, excluding RVM Radial, is displayed by 5-NN, 10-NN and 20-NN and PLS, which are the least algorithmically complex methods among the learning strategies explored. This also indicates that the relationship between the pIC_{50} values and the molecular properties presents a certain degree of non-linearity, thus making the data sets used here suitable to benchmark the noise sensitivity of non-linear algorithms. Similarly, it is important to note that the aim here is to assess the noise sensitivity of a set of algorithms covering diverse learning paradigms, but not to compare their relative performance on these data sets (although all models displayed sufficient predictive power to be considered as statistically robust) [Golbraikh and Tropsha (2002)].

In a previous publication [Cortes Ciriano et al. (2014)], it was reported the differential tolerance to noise of Gaussian Process models depending on the kernel chosen. Here, we found that both GP models with radial (GP Radial) and polynomial (GP Poly) kernels displayed high predictive power on the test set for low noise levels (0 and 10%). By contrast, the $RMSE_{test}$ values slightly increased in the case of GP Poly with the level of noise, which agrees with the machine learning literature [Atla et al. (2011); Cortes Ciriano et al. (2014); Hastie, Tibshirani, and Friedman (2001); Steinwart (2002)]. This effect was more evident in the case of SVM models. In practice, it is advisable to use a low degree for the polynomial kernels when used with SVM, as polynomial kernels of higher degree are prone to overfitting and are less robust to noise [Hastie, Tibshirani, and Friedman (2001)]. From a practical standpoint, we observed that the noise tolerance of SVM Poly and GP Poly could be improved (data not shown) if the grid used to optimize the model parameters in cross-validation covers a wide range of values (Table 4.1). Therefore, we advocate to perform grid

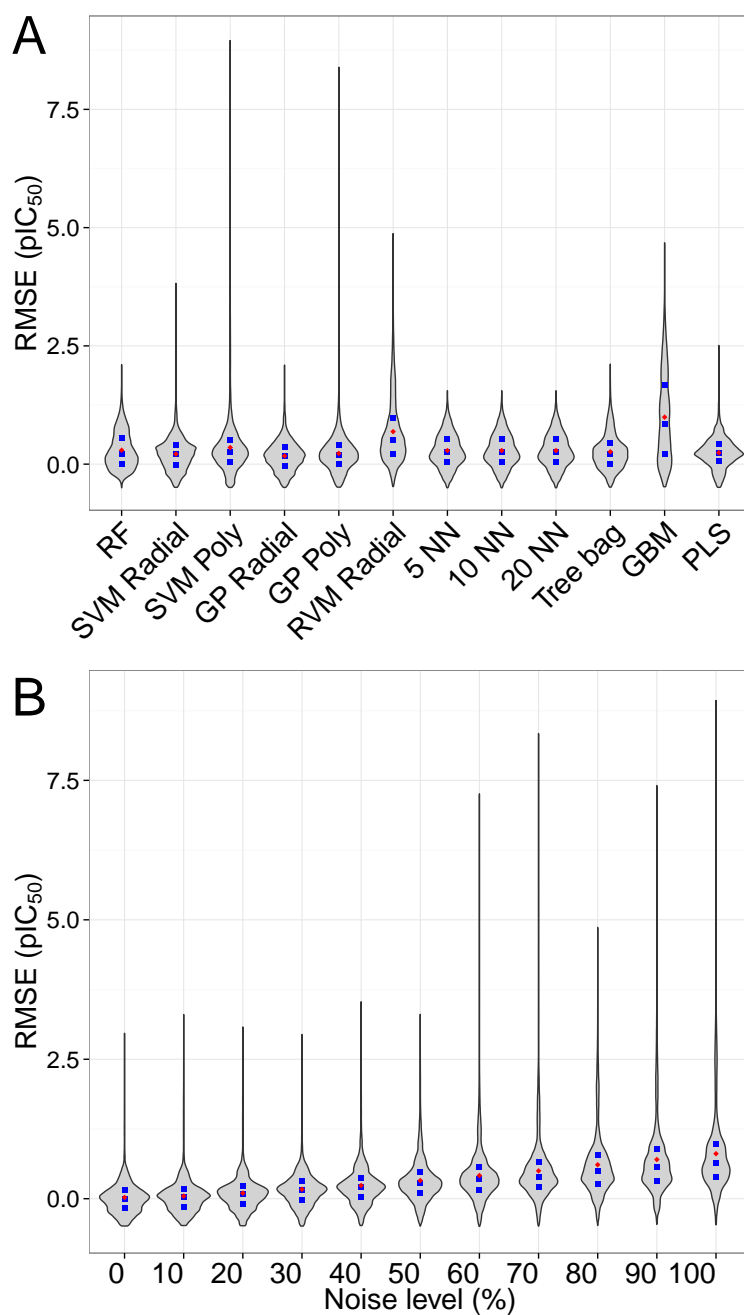


Figure .4.3: **A.** $RMSE_{test}$ values across all data sets and noise levels for the 12 algorithms. **B.** $RMSE_{test}$ values across all data sets and models for the 11 noise levels studied. The data set-specific intercept was subtracted from the $RMSE_{test}$ values in order to make the results comparable across the seven data sets.

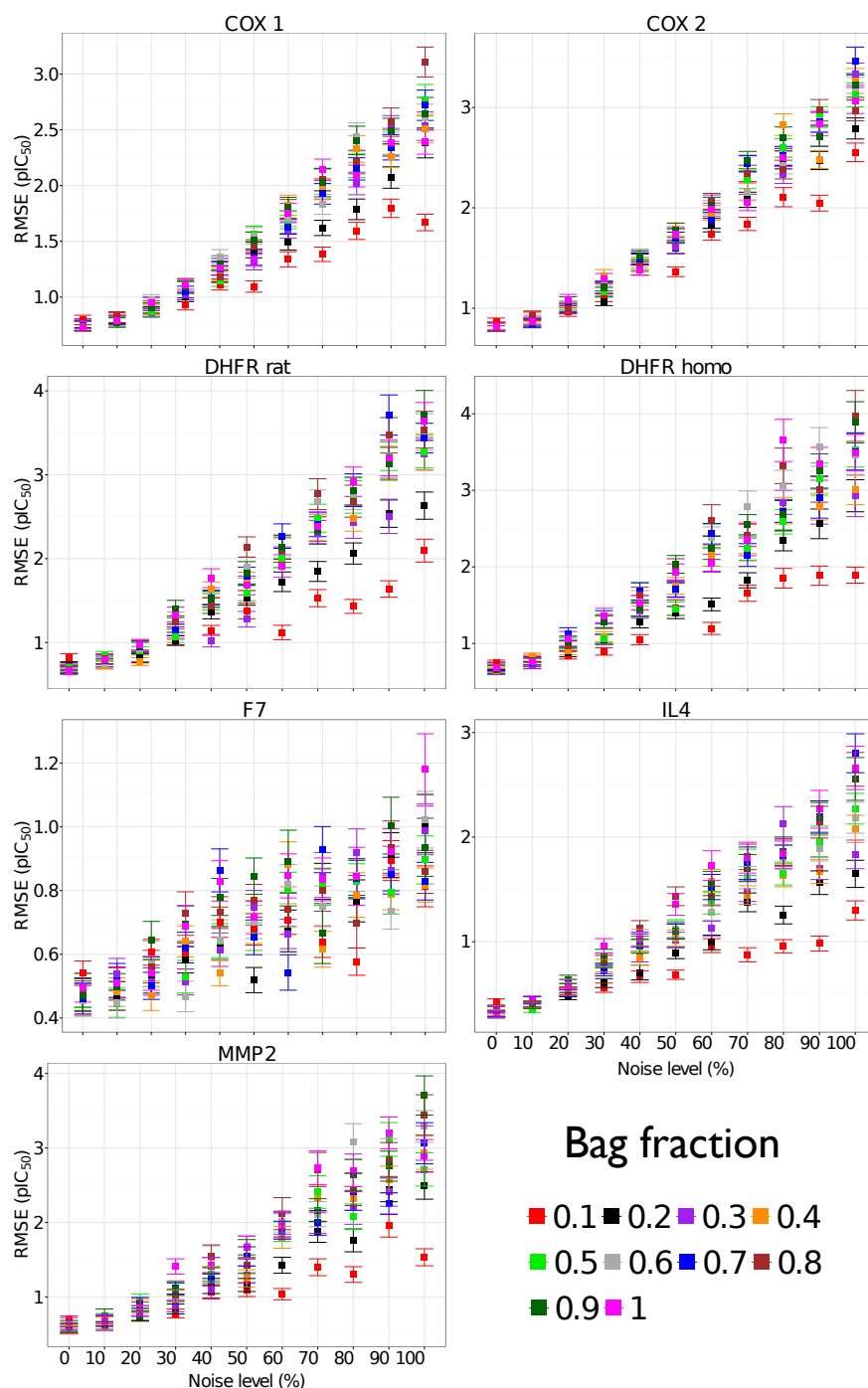


Figure 4.4: RMSE_{test} values for GBM models trained with increasingly higher bagfraction values across all Noise – Algorithm – dataset combinations. For low noise levels (up to 30%) the performance of all models is comparable irrespective of the bagfraction value. However, from noise level 30% upwards, the mean RMSE_{test} difference between models trained with bag fraction values of 1 and 0.1-0.2 increases with the noise level. Overall, these data suggest that the noise sensitivity of GBM highly depends on the bagfraction value.

search across a broad range of parameter values for GP Poly (scale) and SVM Poly (scale and C).

Interestingly, the noise sensitivity of SVM Radial and GP Radial was comparable, thus indicating that the variability in noise sensitivity for SVM and GP is more dependent on the kernel choice than on the machine learning technique. Although GP and SVM display comparable predictive power [Cortes Ciriano et al. (2014)], the Bayesian formulation of GP enables the inclusion of the experimental error of each datapoint as input to the model. This property of GP might be useful when modelling data sets in which the experimental errors for individual datapoints are reported.

Another interesting observation is the low noise tolerance of GBM under the default parameter settings, *i.e.* the bag fraction $\eta = 0.5$. These results are in line with Dietterich (2000), who reported that boosting displays higher performance than bagging in classification on noise-free data, whereas in the presence of noise bagging outperforms boosting. Therefore, these data indicate that careful attention should be given to the choice of parameter values when using GBM in QSAR.

It is important to note that noise in QSAR does not always correspond to random experimental errors. For instance, the purity of compounds can degrade over time and their solubility in the assay medium is not always verified. Similarly, IC_{50} values depend on the assay conditions. Therefore, it is advisable to consider these sources of error prior to building QSAR models whenever possible. For a detailed review of the sources of errors in bioactivity data see ref. Kramer and L (2012). Concerning the quantification of the level of noise (quality) of bioactivity data, Wenlock and Carlsson (2014) considered as high quality data those datapoints whose replicate standard deviation was below an arbitrary cut-off value. Whereas information about the quality of the data might be available when using in-house data obtained in normalized experimental conditions, this is not generally possible for academic laboratories, as public databases lack detailed information about the assay protocols and the variation across experimental replicates. Therefore, assessing the quality of public data might not always be possible. In these cases, it is reasonable to consider respective standard deviations of 0.68 and 0.54 for heterogeneous pIC_{50} and pK_i values [Kalliokoski et al. (2013); Kramer et al. (2012)].

Overall, this study provides a practical guide to make informed decisions about which algorithm and parameter values to use according to the noise level present in the data. As we have shown here, an inappropriate algorithmic (or kernel) choice can have a significant impact on the predictions on external molecules when modelling low quality data, even for low noise levels. Therefore, the quality of the data, be it estimated from replicates or from the literature [Kalliokoski et al. (2013); Kramer et al. (2012)], should become a customary criterion to guide how to approach a given

bioactivity modelling task from an algorithmic standpoint and how to validate the resulting models.

Bibliography

- Aha, DW, D Kibler, and MK Albert (Jan. 1991). "Instance-Based Learning Algorithms". In: *Mach. Learn.* 6.1, pp. 37–66 (cit. on p. 131).
- Angluin, D and P Laird (1988). "Learning From Noisy Examples". In: *Mach. Learn.* 2.4, pp. 343–370 (cit. on p. 131).
- Atla, A, R Tada, V Sheng, and N Singireddy (2011). "Sensitivity of Different Machine Learning Algorithms to Noise". In: *J. Comput. Sci. Coll.* 26.5, pp. 96–103 (cit. on pp. 131, 146).
- Ben-Hur, A, CS Ong, S Sonnenburg, B Schölkopf, and G Rätsch (Oct. 2008). "Support Vector Machines and Kernels for Computational Biology". In: *PLoS Comput. Biol.* 4.10. Ed. by F Lewitter, e1000173 (cit. on p. 136).
- Bender, A, JL Jenkins, J Scheiber, SCK Sukuru, M Glick, and JW Davies (2009). "How similar are similarity searching methods? A principal component analysis of molecular descriptor space". In: *J. Chem. Inf. Model.* 49.1, pp. 108–119 (cit. on p. 134).
- Breiman, L (Oct. 2001). "Random Forests". en. In: *Mach. Learn.* 45.1, pp. 5–32 (cit. on pp. 133, 137, 141).
- Breiman, L, JH Friedman, R Olshen, and C Stone (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks (cit. on pp. 133, 137).
- Brown, SP, SW Muchmore, and PJ Hajduk (Apr. 2009). "Healthy skepticism: assessing realistic model performance". In: *Drug Discov. Today* 14.7-8, pp. 420–427 (cit. on p. 131).
- Chen, H, L Carlsson, M Eriksson, P Varkonyi, U Norinder, and I Nilsson (2013). "Beyond the scope of Free-Wilson analysis: Building interpretable QSAR models with machine learning algorithms". In: *J. Chem. Inf. Model.* 53.6, pp. 1324–1336 (cit. on p. 135).
- Cortes, C and V Vapnik (1995). "Support-vector networks". In: *Mach. Learn.* 20.3, pp. 273–297 (cit. on p. 133).
- Cortes-Ciriano, I (2013). *FingerprintCalculator*. URL: <http://github.com/isidro/FingerprintCalculator> (cit. on p. 134).
- Cortes-Ciriano, I, QU Ain, V Subramanian, EB Lenselink, O Mendez-Lucio, AP IJzerman, G Wohlfahrt, P Prusis, TE Malliavin, GJP van Westen, and A Bender (2015a). "Polypharmacology modelling using Proteochemometrics: recent developments and future prospects". In: *Med. Chem. Comm.* (Cit. on p. 129).

- Cortes-Ciriano, I, DS Murrell, GJP van Westen, A Bender, and TE Malliavin (2015b). "Prediction of the Potency of Mammalian Cyclooxygenase Inhibitors with Ensemble Proteochemometric Modeling". In: *J. Cheminf.* 7, p. 1 (cit. on pp. 131, 135).
- Cortes Ciriano, I, G van Westen, EB Lenselink, DS Murrell, A Bender, and T Malliavin (2014). "Proteochemometrics Modeling in a Bayesian Framework". In: *J. Cheminf.* 6, p. 35 (cit. on pp. 131, 132, 146, 149).
- Cortes-Ciriano, I, Bender, A, and TE Malliavin (2015). "Prediction of PARP inhibition with Proteochemometric modelling and conformal prediction". In: *In revision at Mol. Inform.* URL: <http://cran.r-project.org/package=conformal> (cit. on pp. 130, 131).
- Cortes-Ciriano, I, van Westen, G J P, Bouvier, G, Nilges, M, Overington, J P, Bender, A, and TE Malliavin (2015). "Large-scale prediction of growth inhibition patterns on the NCI60 cancer cell-line panel". In: *In revision at Bioinformatics* (cit. on p. 131).
- Dietterich, TG (2000). "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization". In: *Mach. Learn.* 40.2, pp. 139–157 (cit. on pp. 141, 149).
- Efron, B and R Tibshirani (1993). *An introduction to the bootstrap*. New York : Chapman & Hall (cit. on p. 139).
- Everitt, BS, S Landau, M Leese, and D Stahl (2011). "Miscellaneous Clustering Methods". In: *Cluster Analysis*. John Wiley & Sons, Ltd, pp. 215–255 (cit. on p. 142).
- Fix, E and JL Hodges (1989). "(1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on Fix and Hodges (1951)". In: *Int. Stat. Rev.* 57, pp. 233–247 (cit. on pp. 133, 136).
- Friedman, JH (2001). "Greedy function approximation: A gradient boosting machine". In: *Ann. Stat.* 29.5, pp. 1189–1232 (cit. on pp. 133, 136).
- Fuller, WA (2008). *Wiley Series in Probability and Mathematical Statistics*. John Wiley & Sons, Inc., pp. 441–445 (cit. on p. 130).
- Gaulton, A, LJ Bellis, AP Bento, J Chambers, M Davies, A Hersey, Y Light, S McGlinchey, D Michalovich, B Al-Lazikani, and JP Overington (2011). "ChEMBL: a large-scale bioactivity database for drug discovery". In: *Nucleic Acids Res.* 40.D1, pp. 1100–1107 (cit. on pp. 130, 133, 135).
- Ge, J, Y Xia, and Y Tu (2010). "A discretization algorithm for uncertain data". In: *Database and Expert Systems Applications*, pp. 485–499 (cit. on p. 131).
- Glen, RC, A Bender, CH Arnby, L Carlsson, S Boyer, J Smith, and RC Glenn (2006). "Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME". In: *IDrugs* 9.3, pp. 199–204 (cit. on p. 134).
- Golbraikh, A and A Tropsha (2002). "Beware of q²!" In: *J. Mol. Graphics Modell.* 20.4, pp. 269–276 (cit. on p. 146).
- Hastie, T, R Tibshirani, and J Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc. (cit. on pp. 142, 146).

- Hawkins, DM, SC Basak, and D Mills (Mar. 2003). "Assessing model fit by Cross-Validation". In: *J. Chem. Inf. Model.* 43.2, pp. 579–586 (cit. on pp. 131, 136).
- John, GH and P Langley (1995). "Estimating Continuous Distributions in Bayesian Classifiers". In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. UAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 338–345 (cit. on p. 131).
- Kalliokoski, T, C Kramer, A Vulpetti, and P Gedeck (2013). "Comparability of mixed IC50 Data - A statistical analysis". In: *PloS ONE* 8.4, e61007 (cit. on pp. 130, 149).
- Karatzoglou, A, A Smola, K Hornik, and A Zeileis (2004). "kernlab – An S4 package for kernel methods in R". In: *J. Stat. Soft.* 11.9, pp. 1–20 (cit. on p. 136).
- Kearns, M (1998). "Efficient noise-tolerant learning from statistical queries". In: *J. ACM*. ACM Press, pp. 392–401 (cit. on p. 131).
- Kononenko, I and M Kukar (2007). *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Horwood Publishing Limited (cit. on p. 142).
- Koutsoukas, A, S Paricharak, WRJD Galloway, DR Spring, AP IJzerman, RC Glen, D Marcus, and A Bender (2013). "How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space". In: *J. Chem. Inf. Model.* 54.1, pp. 230–242 (cit. on p. 134).
- Kramer, C and R L (2012). "QSARs, data and error in the modern age of drug discovery". In: *Curr. Top. Med. Chem.* 12.17, pp. 1896–1902 (cit. on p. 149).
- Kramer, C, T Kalliokoski, P Gedeck, and A Vulpetti (June 2012). "The experimental uncertainty of heterogeneous public Ki data". In: *J. Med. Chem.* 55.11, pp. 5165–5173 (cit. on pp. 130, 149).
- Kuhn, M (2008). "Building predictive models in R using the caret package". In: *J. Stat. Softw.* 28.5, pp. 1–26 (cit. on p. 136).
- Kuhn, M and K Json (2013). *Applied Predictive Modeling*. New York, NY: Springer New York (cit. on p. 136).
- Landrum, G (2006). *RDKit Open-source cheminformatics*. URL: <http://rdkit.org/> (cit. on p. 134).
- Liaw, A and M Wiener (2002). "Classification and Regression by randomForest". In: *R News* 2.3, pp. 18–22. URL: <http://cran.r-project.org/doc/Rnews/> (cit. on p. 137).
- Long, PM and RA Servedio (2010). "Random classification noise defeats all convex potential boosters". In: *Mach. Learn.* 78.3, pp. 287–304 (cit. on p. 141).
- Manolopoulos, Y and P Spirakis (2003). "Machine Learning Algorithms: a study on noise sensitivity". In: *1st Balkan Conference on Informatics BCI*. Springer-Verlag, pp. 356–365 (cit. on p. 131).
- Mevik, BH, R Wehrens, and KH Liland (2013). *pls: Partial Least Squares and Principal Component regression*. R package version 2.4-3. URL: <http://cran.r-project.org/package=pls> (cit. on p. 136).
- Murrell, DS, I Cortes-Ciriano, GJP van Westen, IP Stott, TE Malliavin, A Bender, and RC Glen (2014). "Chemistry Aware Model Builder (camb): an R Package for

- Predictive Bioactivity Modeling". In: <http://github.com/cambDI/camb> (cit. on pp. 133, 134, 136).
- Natarajan, N, IS Dhillon, PK Ravikumar, and A Tewari (2013). "Learning with Noisy Labels". In: *Adv. Neural. Inf. Process. Syst.* Curran Associates, Inc., pp. 1196–1204 (cit. on p. 131).
- Natekin, A and A Knoll (2013). "Gradient boosting machines, a tutorial". In: *Front. Neurorobot.* 7, p. 21 (cit. on pp. 133, 141).
- Nettleton, D, A Orriols-Puig, and A Fornells (2010). "A study of the effect of different types of noise on the precision of supervised learning techniques". In: *Artif. Int. R.* 33-4, pp. 275–306 (cit. on p. 131).
- Norinder, U and H Bostrom (2012). "Introducing uncertainty in predictive modeling—friend or foe?" In: *J. Chem. Inf. Model.* 52.11, pp. 2815–2822 (cit. on p. 131).
- Paricharak, S, I Cortes-Ciriano, AP IJzerman, TE Malliavin, and A Bender (2015). "Proteochemometric modeling coupled to in silico target prediction: an integrated approach for the simultaneous prediction of polypharmacology and binding affinity of small molecules". In: *J. Cheminf.* 7, p. 15 (cit. on p. 135).
- Platt, JC (1998). "Fast Training of Support Vector Machines Using Sequential Minimal Optimization". In: *Advances in Kernel Methods - Support Vector Learning*. MIT Press (cit. on p. 131).
- Ps, A and T Hothorn (2013). *ipred: Improved Predictors*. R package version 0.9-3. URL: <http://cran.r-project.org/package=ipred> (cit. on p. 137).
- Qin, B, Y Xia, and F Li (2010). "A Bayesian classifier for uncertain data". In: *Proceedings of the 2010 ACM Symposium on Applied Computing*, pp. 1010–1014 (cit. on p. 131).
- Quinlan, JR (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. (cit. on p. 131).
- Rasmussen, CE and CKI Ws (2006). *Gaussian Processes for machine learning*. MIT Press (cit. on pp. 131, 133, 136).
- Ridgeway, G (2013). *gbm: Generalized Boosted Regression Models*. R package version 2.1. URL: <http://cran.r-project.org/package=gbm> (cit. on p. 136).
- Rogers, D and M Hahn (May 2010). "Extended-connectivity fingerprints". In: *J. Chem. Inf. Model.* 50.5, pp. 742–754 (cit. on p. 134).
- Sánchez, JA, J Luengo, and F Herrera (2013). "Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification". In: *Pattern Recogn.* 46.1, pp. 355–364 (cit. on p. 142).
- (2014). "Improving the Behavior of the Nearest Neighbor Classifier against Noisy Data with Feature Weighting Schemes". In: *Hybrid Artificial Intelligence Systems*. Vol. 8480. Lecture Notes in Computer Science. Springer International Publishing, pp. 597–606 (cit. on p. 142).
- Steinwart, I (2002). "On the Influence of the Kernel on the Consistency of Support Vector Machines". In: *J. Mach. Learn. Res.* 2, pp. 67–93 (cit. on pp. 132, 146).
- Teytaud, O (2001). "Robust Learning: Regression Noise". In: *Proceedings of IJCNN 2001 Conference*, pp. 1787–1792 (cit. on p. 131).

- Tipping, ME (2000). *The Relevance Vector Machine* (cit. on pp. 133, 136).
- Tsang, S, B Kao, KY Yip, WS Ho, and SD Lee (2009). "Decision trees for uncertain data". In: *IEEE 25th International Conference on Data Engineering, 2009 ICDE '09*, pp. 441–444 (cit. on p. 131).
- Venables, WN and BD Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: <http://www.statsox.ac.uk/pub/MASS4> (cit. on p. 136).
- Wang, Y, J Xiao, TO Suzek, J Zhang, J Wang, Z Zhou, L Han, K Karapetyan, S Dracheva, BA Shoemaker, E Bolton, A Gindulyte, and SH Bryant (2012). "PubChem's BioAssay Database". In: *Nucleic Acids Res.* 40.Database issue, pp. 400–412 (cit. on p. 130).
- Wenlock, MC and L Carlsson (2014). "A study of how experimental errors influence DMPK QSAR/QSPR models". In: *J. Chem. Inf. Model.* 0.0, p. 0 (cit. on pp. 130, 149).
- Winer, B, D Brown, and K Michels (1991). *Statistical Principles in Experimental Design*. McGraw-Hill series in psychology. McGraw-Hill. URL: <http://books.google.fr/books?id=OqppAAAAMAAJ> (cit. on pp. 138, 139).
- Wold, S, M Sjöström, and L Eriksson (2001). "PLS-regression: a basic tool of chemometrics". In: *Chemometr. Intell. Lab.* 58.2, pp. 109–130 (cit. on pp. 133, 136).
- Wu, Y, K Ianakiev, and V Govindaraju (2002). "Improved k-nearest neighbor classification". In: *Pattern Recogn.* 35.10, pp. 2311–2318 (cit. on p. 142).
- Zhang, JBT (2004). "Support vector classification with input data uncertainty". In: *Adv. Neural Inf. Process. Syst.* 17, pp. 161–169 (cit. on p. 131).
- Zhu, X and X Wu (2004). "Class Noise vs. Attribute Noise: A Quantitative Study". In: *Artif. Int. R.* 22.3, pp. 177–210 (cit. on p. 131).
- Zhu, X, X Wu, and Q Chen (2003). "Eliminating Class Noise in Large Datasets". In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. Ed. by T Fawcett and N Mishra, pp. 920–927 (cit. on p. 131).

Prediction of the Potency of Mammalian Cyclooxygenase Inhibitors with Ensemble Proteochemometric Modelling

.5 Isoform Selectivity Prediction: COX

.5.1 Introduction

Cyclooxygenases (EC 1.14.99.1), also known as endoperoxidases, prostaglandin G/H synthases or simply **COX**, are involved in the biosynthesis of prostaglandin H₂ from arachidonic acid [Luo, He, and Bohlin (2005)]. Prostaglandin H₂ is further converted into prostanoids which play a key role in inflammation. Thus, since the development of aspirin® in 1899 [Vane (1971)], the inhibition of the cyclooxygenase activity with non-steroidal anti-inflammatory drugs (**NSAIDs**) has been exploited to treat inflammation. Nonetheless, kidney failure and gastrointestinal side-effects, such as peptic ulcer, have been correlated to long-term intake of **NSAIDs** [Fine (2013)]. Until 1991, only one form of the enzyme (**COX-1**) was thought to be responsible for both the constitutive and the local biosynthesis of prostaglandins. In that year [Xie et al. (1991)], an inducible cyclooxygenase (**COX-2**) was discovered and the different roles of both isoenzymes were revealed. There does exist some overlap: **COX-1** is constitutively expressed serving as the source of housekeeping prostaglandins, whereas the expression of **COX-2** increases in pathophysiological situations such as acute pain, inflammation or cancer [Sostres, Gargallo, and Lanas (2014)]. From this it is thought that efficacy and side-effects can, to some extent, be delineated when blocking the prostaglandin synthesis pathway associated with inflammation and pain.

In the last two decades, research in both the pharmaceutical industry and academic laboratories has been driven by the hypothesis that selective **COX-2** inhibitors would exhibit strong anti-inflammatory and analgesic properties without leading to the unwanted gastrointestinal side effects [Warner et al. (1999)]. Nevertheless, a few organs, *e.g.* the brain cortex and renal glomeruli, express **COX-2** constitutively [Luo, He, and Bohlin (2005)]. The association between the inhibition of **COX-2** in these organs with cardiovascular hazard (**CVH**) was ratified in 2004 and 2005 [Bresalier et al. (2005); Nussmeier et al. (2005)]. These findings led the US Food and Drug Agency (**FDA**) to retrieve rofecoxib (Vioxx) and valdecoxib (Bextra) from the market, and to include boxed warnings for all selective **COX-2** inhibitors. Higher risk of heart attack and hypertension have also been reported for non-selective **NSAIDs**, thus highlighting that cardiovascular risk might not be related to the degree of **COX** selectivity [Howes (2007)]. In 2012, Yu et al. (2012) demonstrated that the cardiovascular risk originates from **COX-2** inhibition by selective and not selective **NSAIDs**

and is taking place in blood vessels. These authors have shown that COX-2 inhibition leads to a decrease in prostaglandin (mainly PGI₂) and to increased nitric oxide (NO) production which is sufficient to increase the risk of heart failure, hypertension and thrombosis [Yu et al. (2012)].

Nevertheless, there are still niche populations which can benefit from selective COX-2 inhibitors, *e.g.* patients who cannot afford to take non-selective COX inhibitors, due to an increased risk of peptic ulcers or cancer. In addition, selective COX-2 inhibitors continue to be the common treatment for chronic inflammatory and pain disorders [Crofford (2013); Fine (2013)], and NSAIDs are known to reduce the risk of (among others) [Hson et al. (2013); Robak, Smolewski, and Robak (2008); Zhang et al. (2014)]: colon cancer [Chen et al. (2008); Moore and Simmons (2000); Soh et al. (2008); Thun, Henley, and Patrono (2002)], Alzheimer's disease, and platelet aggregation [Jouzeau et al. (1997); Sostres, Gargallo, and Lanas (2014)]. Overall, NSAIDs are still one of the most commonly prescribed drugs in the world [Jones et al. (2008)], and this trend is likely to increase owing to the aging of the population. Therefore, the administration of NSAIDs in clinics is currently subject to a benefit-risk assessment between the patients clinical profile and potential drugs side-effects [Curiel and Katz (2013)], always aiming at optimizing both the dosage and the duration of the drug regimen [Fine (2013)].

The isoform selectivity of COX inhibitors stems from a structural difference in the binding site. The binding site of both cyclooxygenases is highly conserved except for the substitution of an isoleucine at position 523 in COX-1 with a valine in COX-2 [Blobaum and Marnett (2007)]. This substitution results in a larger binding site in COX-2, as the smaller size of valine allows access to a side-pocket. This structural difference has been exploited for the rational design of potent and selective COX-2 inhibitors by both medicinal and computational chemistry [Blobaum and Marnett (2007); Dannhardt and Laufer (2000); Leval et al. (2000)].

To date, a plethora of *in silico* studies have been published with the aim of better understanding and predicting the potency of COX inhibitors on either COX-1 or COX-2 using molecular docking and QSAR models [Dube et al. (2014); Gupta and Kumar (2012); Kim et al. (2004); Narsinghani and Chaturvedi (2006); Reddy et al. (2007)]. Nonetheless, none of these studies was able to integrate bioactivity information from multiple mammalian COX in the frame of a single machine learning model. Given that the bioactivity profiles of selective COX inhibitors on COX-1 and COX-2 are highly uncorrelated, thus presenting high selectivity ratios [Dannhardt and Laufer (2000); Leval et al. (2000)], only a predictive model trained on both the chemical and the target space would be able to simultaneously predict compound potency on a panel of cyclooxygenases, as well as to predict the activity of a given compound on a yet untested isoform. In that way, new potent, selective and safe COX

inhibitors could be discovered.

Here, we apply the principles of **PCM** to model the potency of 3,228 compounds on 11 mammalian cyclooxygenases. To this aim, we have trained **PCM** models with different machine learning algorithms on public IC_{50} values from ChEMBL 16 [Gaulton et al. (2011)], including data on human **COX-1**, **COX-2**, and on 9 orthologues. In an attempt to increase model performance, these models have been combined in ensembles (ensemble modelling), thus constituting the first **PCM** study where ensemble **PCM** modelling is applied. Additionally, the description of compounds with keyed fingerprints has enabled the deconvolution of the chemical space to rationalize both the potency and the selectivity of **COX** inhibitors towards a particular isoenzyme.

.5.2 Materials and Methods

.5.2.1 Data set

IC_{50} values for 11 mammalian cyclooxygenases, listed in Table .5.1, were retrieved from ChEMBL 16 [Gaulton et al. (ibid.)]. To ensure the reliability of the bioactivity values, only IC_{50} values corresponding to small molecules and satisfying the following criteria were kept: (i) activity relationship equal to '=', (ii) assay score confidence ≥ 8 , and (iii) activity unit equal to 'nM'. The average pIC_{50} value was calculated when multiple IC_{50} values were annotated on the same compound-target combination. The application of these filters led to a final data set composed of 3,228 distinct compounds and 11 sequences, being the total number of data-points 4,937 (13.9% matrix completeness). The negative logarithm with base 10 of the IC_{50} values (pIC_{50}) was used as the response variable to train all models. We decided to mix bioactivity data from different assays given that Kalliokoski et al. (2013) reported that the standard deviation of public IC_{50} data is 25% larger than the standard deviation corresponding to public K_i data, and thus mixing IC_{50} data from different assays adds a moderate level of noise. The crystallographic structure of the ovine **COX-1** complexed with celecoxib (PDB ID: 3KK6 [Berman et al. (2000); Rimon et al. (2010)]) was used to extract the residues in the binding site. Those residues within a sphere of radius equal to 10\AA centered in the ligand were selected. The corresponding residues for the other 10 sequences were identified by multiple sequence alignment [Sievers et al. (2011)].

.5.2.2 Descriptors

Chemical structures were standardized (section .2.1) with the function *StandardiseMolecules* from the R package *camb* [Murrell et al. (2014)] with the following options:

UniProt ID	Type	Organism	Number of Bioactivities	% Compounds Annotated
P23219	1	<i>Homo sapiens</i>	1,346	41.7
O62664	1	<i>Bos taurus</i>	48	1.5
P22437	1	<i>Mus musculus</i>	50	1.5
O97554	1	<i>Oryctolagus cuniculus</i>	11	0.3
P05979	1	<i>Ovis aries</i>	442	13.7
Q63921	1	<i>Rattus Norvegicus</i>	23	0.7
P35354	2	<i>Homo sapiens</i>	2,311	71.6
O62698	2	<i>Bos taurus</i>	21	0.7
Q05769	2	<i>Mus musculus</i>	305	9.4
P79208	2	<i>Ovis aries</i>	341	10.6
P35355	2	<i>Rattus Norvegicus</i>	39	1.2

Table .5.1: **Composition of the COX data set.** The total number of bioactivities, after duplicate removal and selected from ChEMBL as described in Materials and Methods, and of distinct compounds are 4,937 and 3,228 respectively. The last column indicates the percentage of the total number of distinct compounds (3,228) annotated on each target.

(i) inorganic molecules were removed, and (ii) molecules were selected irrespectively of the number of fluorines, chlorines, bromines or iodines present in their structure, or of their molecular mass. Morgan fingerprints [Glen et al. (2006); Rogers and Hahn (2010)] were calculated using RDkit (release version 2013.03.02) [Cortes-Ciriano (2013); Landrum (2006)]. Physicochemical descriptors (PaDEL) [Yap (2011)] were calculated with the function *GeneratePadelDescriptors* from the R package *camb*. The R package *vegan* was used to generate the distributions of pairwise compound similarities (Jaccard distance) [Oksanen et al. (2013)].

The amino acids composing the binding site of the mammalian cyclooxygenases considered in this study (Table .5.1), were described with five amino acid extended principal property scales (5 z-scales) [Sandberg et al. (1998)]. z-scales were calculated with the R package *camb* [Murrell et al. (2014)].

.5.2.3 Machine learning implementation

Machine learning models were built in R using the packages *caret* [Kuhn (2008)] and *camb* [Murrell et al. (2014)]. Model ensembles were created with the help of the R package *caretEnsemble* [Mayer and E (2015)]. Both the data set and the modelling

pipeline coded in R is available in the documentation of the R package *camb* [Murrell et al. (2014)].

.5.2.4 Model generation

Descriptors with a variance close to zero were removed with the function *RemoveNearZeroVarianceFeatures* from the R package *camb* using a cut-off value equal to $30/1$ [Kuhn (2008); Kuhn and Json (2013); Murrell et al. (2014)]. Subsequently, the remaining descriptors were centered to zero mean and scaled to unit variance with the function *PreProcess* from the R package *camb*.

The values of the model parameters were optimized by grid search and 5-fold cross validation (CV) [Hawkins, Basak, and Mills (2003)] (section .2.4). To significantly compare the quality of the modelling with different machine learning algorithms, the same folds were used to train all models.

Both single PCM models and model ensembles were used to predict the bioactivities for the test set, and their error in prediction compared. The bioactivity values corresponding to the data-points in the test set were not considered when building neither the single PCM models nor the model ensembles.

In order to assess whether merging the chemical and the target space in a single PCM model enhances model performance, we trained two Random Forest (RF) models using either: (i) only compound descriptors (Family Quantitative Structure-Activity Relationship -QSAR-) [Brown et al. (2014)], or (ii) only target descriptors (Family Quantitative Sequence-Activity Modelling -QSAM-) [Brown et al. (ibid.)]. Obtaining a high performance with a Family QSAR model would indicate that the bioactivities of a given compound on different targets are correlated. Thus, target descriptors would not contribute to increase model performance. On the other hand, high performance observed for a Family QSAM model would indicate that the bioactivity values only depend on the targets and not on the compounds, *i.e.* the bioactivities of a set of diverse compounds are correlated on a given target. In this case, compound descriptors would not be required to predict compounds affinity, as target descriptors alone would be sufficient.

.5.2.5 Model validation

Both internal and external validation were performed according to the criteria proposed by Golbraikh and Tropsha (2002); Tropsha and Golbraikh (2010); Tropsha and Gramatica (2003), and to the RMSE values. These criteria and the formulae of the statistical metrics are given in sections .2.6.1.

.5.2.6 Assessment of maximum model performance

To further assess the reliability of the models in the light of the uncertainty of the bioactivity values [Kalliokoski et al. (2013); Kramer and L (2012); Kramer et al. (2012)], we established the maximum $R_{0\text{ test}}^2$ and q_{test}^2 , and minimum $\text{RMSE}_{\text{test}}$ values achievable given: (i) the uncertainty of public IC_{50} data, and (ii) the number of data-points in both the training and the test set. The distributions of minimum $\text{RMSE}_{\text{test}}$, and maximum q_{test}^2 , and $R_{0\text{ test}}^2$ values were calculated as explained in section .2.7.

The maximum and minimum values of respectively $R_{0\text{ test}}^2 / q_{\text{test}}^2$ and $\text{RMSE}_{\text{test}}$ were then used to validate model performance on the test set (section .2.7). If the obtained metrics were beyond the maximum values (for q_{test}^2 and $R_{0\text{ test}}^2$) or the minimum values (for $\text{RMSE}_{\text{test}}$) of the corresponding distributions, the model is likely to be over-optimistic [Hawkins, Basak, and Mills (2003)]. This estimation of the maximum achievable model performance takes into account the range and distribution of the bioactivities present in the data. This is of particular importance as it has been recently reported by Sheridan [Sheridan (2012)] that (i) certain bioactivity ranges are better predicted than others, and (ii) R_0^2 values might be very low if the bioactivity range considered is too narrow, even if the predictions closely match the observed values.

.5.2.7 Ensemble modelling

Gradient-boosting machines (GBM) [Friedman (2001)], Random Forest (RF) [Breiman (2001)], and Support Vector Machines (SVM) [Ben-Hur et al. (2008)] were implemented to train a model library. The resulting models were combined in model ensembles using two techniques, namely: greedy optimization and model stacking. Depending on the models considered when training an ensemble, two types of model ensembles were defined: (i) homo-ensembles: the same algorithm was used to train all models of the ensemble, though the parameter values were different in each model, (ii) hetero-ensembles: the number of algorithms used to train the models combined in the ensemble was greater or equal than 2.

Greedy optimization

Greedy optimization, based on the work of Caruana et al. (2004), optimizes the RMSE on the cross-validation predictions on the hold-out folds. These predictions were calculated with the model library trained on a training set with identical fold composition. Each model was assigned a weight in the following manner. Initially, all models had a weight equal to zero. Afterwards, the weight of a given model was repeatedly incremented by 1 if the subsequent normalized weight vector allowed a

closer match between the weighted combination of cross-validated predictions and the observed values. This repetition was carried out n times, $n = 1,000$ in the present work, and the resulting weight vector was normalized to obtain the final models weighting. The final model ensemble was used to predict the activities on the test set, and the error in prediction compared to that of single PCM models on same set.

Model stacking (MS)

The concept of model stacking is illustrated in Figure .5.1. In this case, the predictions on the training set calculated with the model library during cross-validation served as descriptors. Thus, a training matrix was defined where rows were indexed by the data-points in the training set used to train the model library, and columns by the models in the aforesaid library.

A machine learning model was trained on this matrix, irrespective of the algorithms used to generate the model library. This model is then used to predict the bioactivities for the test set, and the error in prediction compared to that of single PCM models on the test set. The bioactivity values corresponding to the data-points in the test set are not considered when building the ensemble. If the selected algorithm has the inherent capability to determine the importance of each descriptor, as for Elastic Net, a vector of weights for the models can be defined. Given that each descriptor corresponds to a particular model, this vector will determine its contribution to the generated ensemble. In the present study we used the following algorithms: linear model, Elastic Net, SVM with linear and radial kernels, and RF.

.5.2.8 Estimation of the error of individual predictions

In order to estimate errors for individual predictions, we used the standard deviation of the predictions of the individual models composing a given model ensemble, *i.e.* ensemble standard deviation (E_{std}). Previous studies [Dragos, Gilles, and A (2009); Sheridan (2012, 2013); Wood et al. (2013)] have highlighted the usefulness of considering the ensemble standard deviation as a domain applicability (DA) measure, specially in the case of RF models, where the calculation of the standard deviation along the trees is straightforward [Sheridan (2012, 2013)]. Here, we extend this idea to ensembles composed of models trained with different algorithms (hetero-ensembles). For each data-point in either the test set or in the hold-out fold in the case of cross-validation, we calculated the standard deviation of the predictions generated with each model conforming the model ensemble. Subsequently, the ensemble standard deviation was scaled with the parameter β . This permits to obtain individual confidence intervals for each prediction, which are thus defined as:

$$IC = \tilde{y} \pm E_{\text{std}} \beta \{ \beta \in \mathbb{R} \mid \beta > 0 \} \quad (.5.1)$$

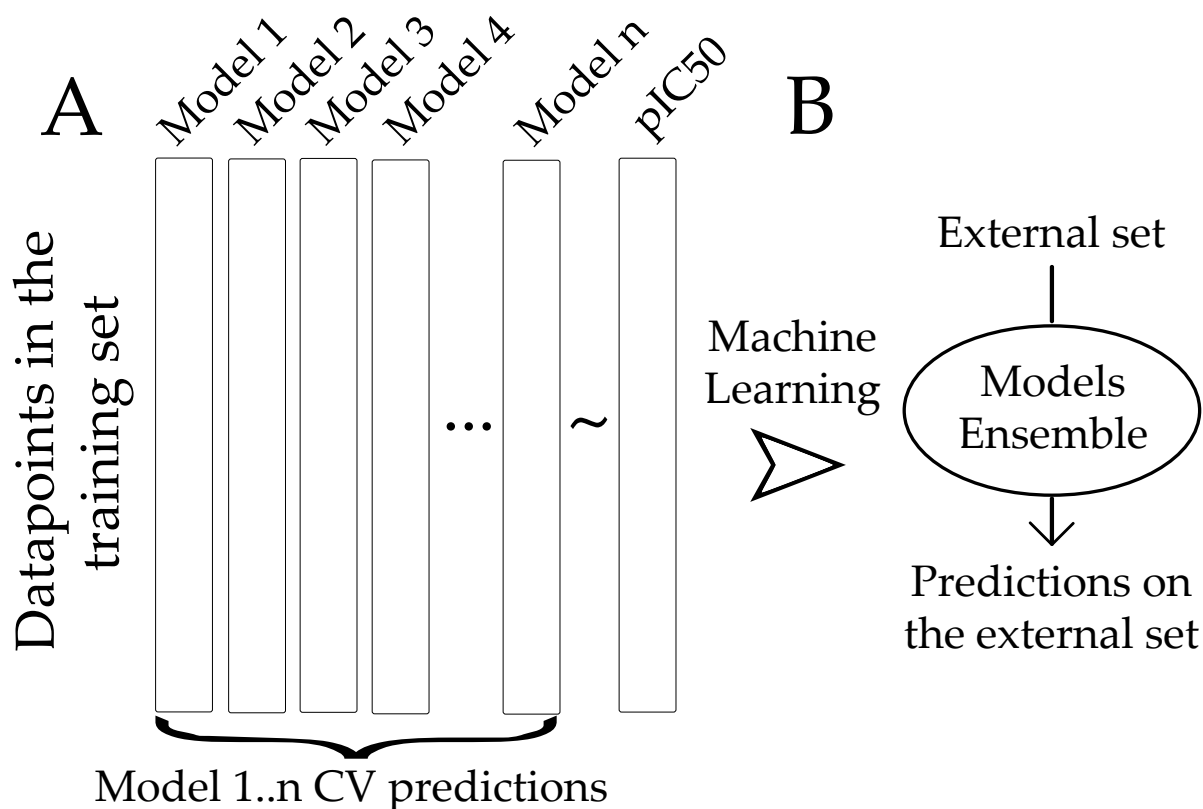


Figure .5.1: **Ensemble modelling with model stacking.** A. A set of models are trained with diverse machine learning algorithms (*Model*₁ .. *Model*_{*n*} in the Figure). The predictions of these models on each data-point in the training set calculated during cross validation, are used as descriptors to create a new training matrix, which rows are indexed by the data-points in the training set and columns by the models in the library. A machine learning model is trained on this matrix. The resulting model is the model ensemble. B. The model ensemble is then applied on the test set.

To assess the practical usefulness of the derived confidence intervals, the percentage of data-points for which the predicted values lied within IC ($0 < \beta < 4$) was calculated. Both the predictions calculated during model training (using the optimal parameter values), *i.e.* cross-validated predictions, as well as the predictions on the test set were used.

.5.2.9 Interpretation of compound substructures

The contribution of chemical substructures to bioactivity on human cyclooxygenases was deconvoluted using a predictive and a Student's method (Figure .5.2):

Prediction of bioactivity values with and without each compound substructure (predictive method, Figure .5.2A)

This first technique quantifies the contribution of each chemical substructure to bioactivity by calculating the distribution of differences between (i) the predicted bioactivity for all compounds containing a given substructure, and (ii) the predicted bioactivity using PCM for these compounds, from which that substructure was virtually removed [pcm_lead_opt; Cortes-Ciriano et al. (2014); Marcou et al. (2012); Polishchuk et al. (2013); Rosenbaum et al. (2011); Spowage, Bruce, and Hirst (2009)]. To virtually remove a substructure, we iteratively set its count equal to zero in all compound descriptors presenting it. The difference between the predicted bioactivity values in the presence or absence of a given substructure was then calculated. The average value of these differences, weighted by the number of counts of the feature in each compound, corresponds to the average contribution of that feature to bioactivity [Cortes-Ciriano et al. (2014)]. The contribution was estimated for all compound features considered in the model. The sign of the difference ({+/-}) indicates whether the feature is respectively beneficial or deleterious for compound bioactivity.

Statistical significance between bioactivity distributions with and without each compound substructure (Student's method, Figure .5.2B)

In order to identify chemical substructures that might not be recognized by the predictive method due to moderate PCM model performance, we also deconvoluted the chemical space in a model-independent way. We created two bioactivity sets, each containing the pIC₅₀ values for either human COX-1 or human COX-2. For each of these bioactivity sets and for each substructure, we defined two distributions, namely: (i) the distribution A of pIC₅₀ values corresponding to the compounds presenting a given substructure in a given bioactivity set, and (ii) the distribution B of pIC₅₀ values for those compounds not presenting that substructure in the same bioactivity set. The normality of these distributions was assessed with the Shapiro-Wilk test

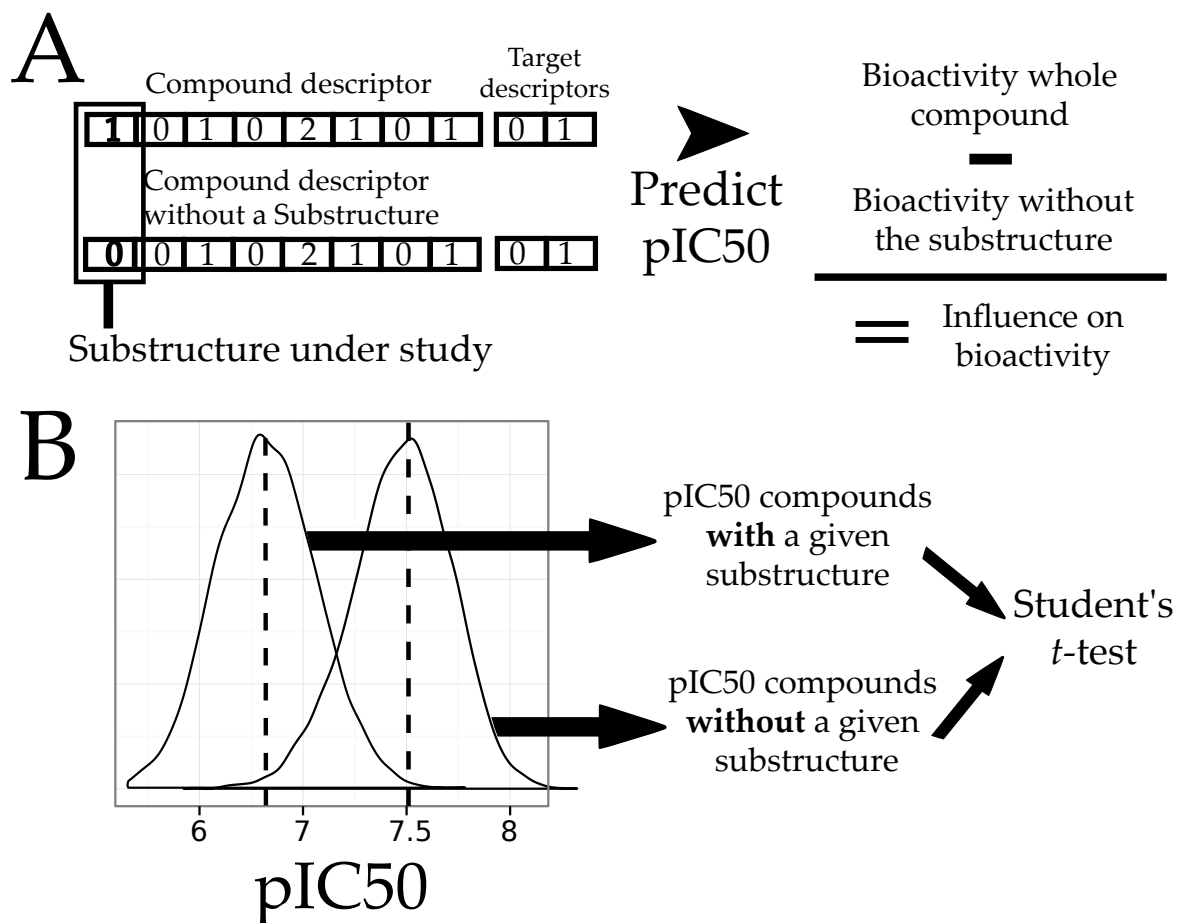


Figure .5.2: **Interpretation of compound substructures.** A. Predictive method. The average influence on bioactivity of a given substructure is calculated as the difference between the distributions corresponding to: (i) the predicted bioactivity for all compounds containing that substructure, and (ii) the predicted bioactivity using **PCM** for these compounds, from which that substructure was virtually removed by setting its count to zero. B. Student's method. In this case, the average substructure influence on bioactivity is evaluated as the difference between the pIC₅₀ distributions for those compounds presenting and not presenting a given substructure.

($\alpha = 0.05$). If both distributions, A and B , followed the Gaussian distribution, a two-tailed t -test for independent samples ($\alpha = 0.05$) was applied to statistically evaluate the difference between them. If the difference was significant, we assumed that the considered substructure has an influence on bioactivity on the isoenzyme associated to the bioactivity set considered.

The sign of the difference between the mean value of A and B indicates whether the presence of the substructure hampers or fosters compound bioactivity on that isoenzyme. Therefore, each substructure was assigned a label, 'deleterious' or 'beneficial', depending on its influence on bioactivity on either COX-1 and COX-2.

Finally, we intended to assess which substructures always increase or decrease compound bioactivity on human COX-1 and COX-2. In that way, substructures identified in the previous step are finally identified as: (i) increasing or decreasing bioactivity on human COX-1, (ii) increasing or decreasing bioactivity on human COX-2, and (iii) increasing or decreasing bioactivity on both human COX-1 and COX-2.

.5.3 Results

.5.3.1 Analysis of the chemical and the target space

Target space

The PCA analysis of the amino acid descriptors of the binding site of the 11 mammalian cyclooxygenases (Table .5.1) is shown in Figure .5.3. Orthologue sequences COX1 and COX2 define two distant clusters. As paralogues display more sequence variability than orthologues, and as small molecules tend to display similar binding within orthologues [Kruger and Overington (2012)], we hypothesize that merging bioactivities from orthologues and paralogues will lead to more predictive models. In addition, these results indicate that the amino acid descriptors account for structural differences between COX-1 and COX-2.

Chemical space

The initial bioactivity selection from ChEMBL 16 [Gaulton et al. (2011)], consisted of 6,804 data-points. As previously highlighted [Kramer et al. (2012)], a large number of target-compound combinations in ChEMBL are annotated with more than one bioactivity value, hence the total number of different compound-target combinations after duplicate removal was 4,937. As in Figure .5.4A, the standard deviations for the bioactivity values annotated on the same compound-target combination are in less than 2% of the cases higher than two pIC_{50} units, whereas more than 90% of

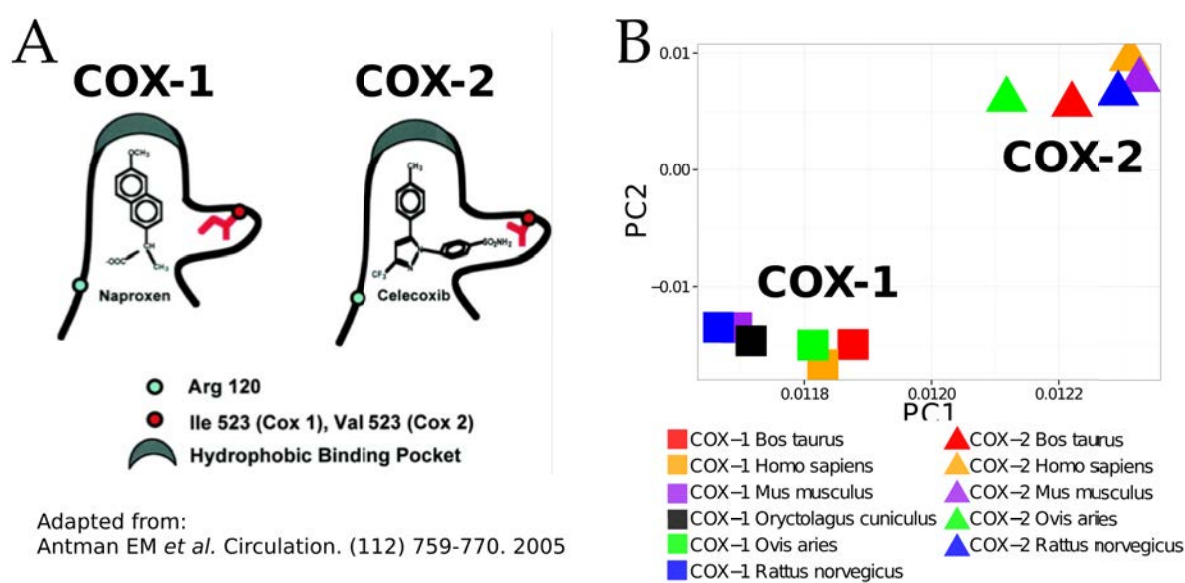


Figure .5.3: **PCA analysis of the target space.** **A.** Schematic overview of the COX binding pockets. **B.** PCA analysis was applied on the binding site descriptors used to train the models. The first two principal components explained more than 80% of the variance, thus indicating that there are mainly two sources of variability in the descriptor space, namely the isoenzyme type. This fact can be seen as **COX-1** (triangles) and **COX-2** (squares) define two distant clusters. Overall, the binding sites of orthologue cyclooxygenases are more similar than those of paralog sequences. These results also indicate that the amino acid descriptors account for structural differences between **COX-1** and **COX-2**, which can be learnt by the models. Thus, it is expected that merging orthologues and paralogues will lead to more predictive models.

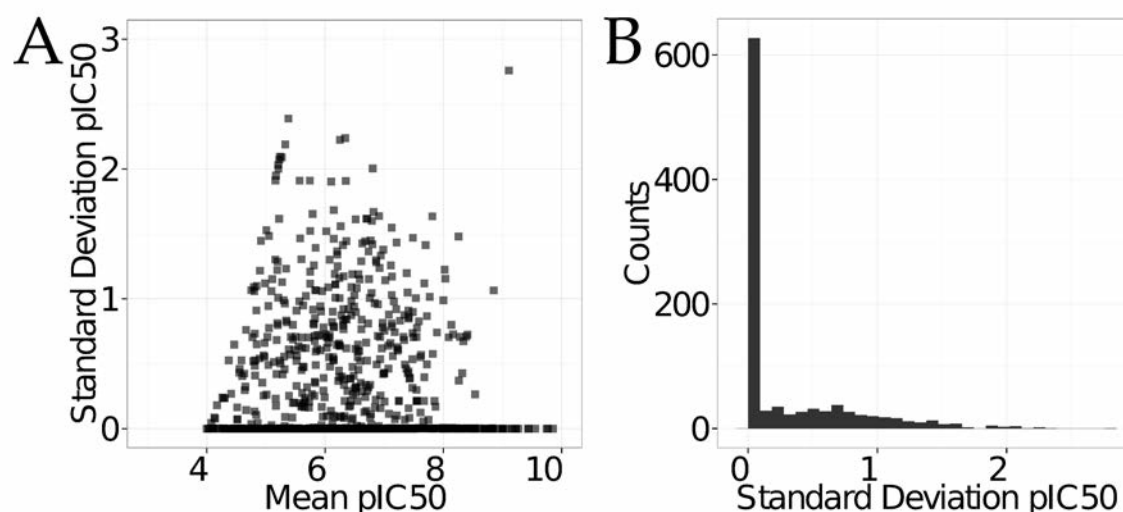


Figure .5.4: **Statistics of the repeated bioactivities for each compound-target combination.** A. The abscissa represents the mean value for the bioactivities repeated for each compound-target combination with more than one annotated bioactivity. The ordinate represents their standard deviations. Repeated bioactivities are equally distributed for low, moderate and high affinity COX inhibitors. B. Histogram of the standard deviation of the repeated bioactivities. The distribution is strongly skewed towards 0, thus indicating that the differences between repeated bioactivities are generally negligible.

the repeated bioactivities exhibit a standard deviation close to zero (Figure .5.4B). Consequently, we decided to take the average of these repeated values instead of the median value: this latter value would be more suitable only if outliers were more abundant.

Selectivity data set

As stated in the introduction, the main advantage of a PCM model applied to mammalian cyclooxygenases would be to anticipate the potency of a given compound towards a particular isoenzyme. To ensure that our data set covered chemical entities with diverse bioactivity profiles on COX-1 and COX-2, we selected all compounds annotated on both human cyclooxygenases. This resulted in a selection of 1,086 compounds, out of a total of 3,228 different inhibitors present in the data set. The scatterplot of the bioactivities of these compounds on human COX-1 against human COX-2 (Figure .5.6A) reveals that the difference in bioactivity for some compounds depending on the isoenzyme is higher than 4 pIC₅₀ units (upper left corner of

Figure .5.6A). $RMSE$ and R_0^2 values for the bioactivities on COX-1 with respect to COX-2 are, respectively, 1.69 pIC₅₀ units and -0.42. As the area above the diagonal of Figure .5.6A is more populated, there are more compounds with higher activity on COX-2 than on COX-1. Therefore, these data let us conclude that the data set comprises compounds exhibiting high selectivity towards COX-2. In addition, the overlap between the data-points in the PCA of the compound descriptors (Figure .5.5) indicates that the compounds annotated on the COX targets cover the same regions of the chemical space.

.5.3.2 PCM validation

Overall, the models obtained with GBM, RF, and SVM (Table .5.2A and Figure .5.7) satisfied our model validation criteria namely: $q_{int}^2 > 0.5$ and, q_{test}^2 and $R_{test\ 0}^2 > 0.6$. The performance of the three algorithms is comparable since $R_{test\ 0}^2$ values range from 0.60 to 0.61, and $RMSE_{test}$ from 0.76 to 0.79 pIC₅₀ units between the different models. Interestingly, the predictive power did not vary when using hashed or unhashed fingerprints, being the $R_{test\ 0}^2$ and $RMSE_{test}$ differences smaller than 0.01 in both cases (data not shown). Thus, we decided to rather use unhashed fingerprints as this choice enables an interpretation of the models according to chemical substructures.

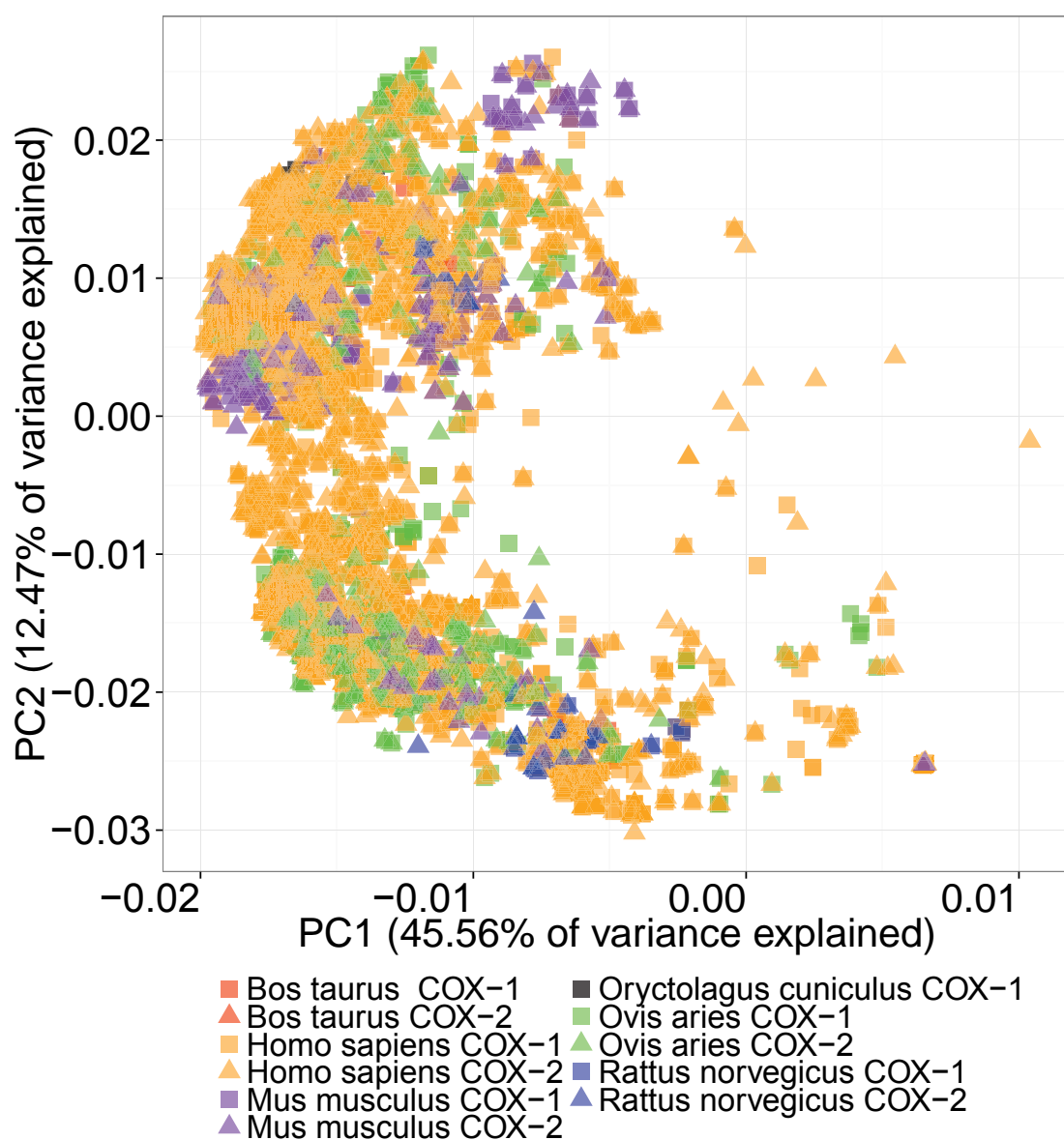


Figure .5.5: **PCA of the compound descriptors used to train the PCM models.** The **PCA** was performed on the pairwise Pearson rank correlation matrix calculated with the compound descriptors used to train the models. The two first principal components (PC) explain 58.03% of the variance. COX-1 and COX-2 are represented with squares and triangles respectively. Overall, the overlap between the datapoints indicate that the compounds annotated on different targets cover the same regions of the chemical space.

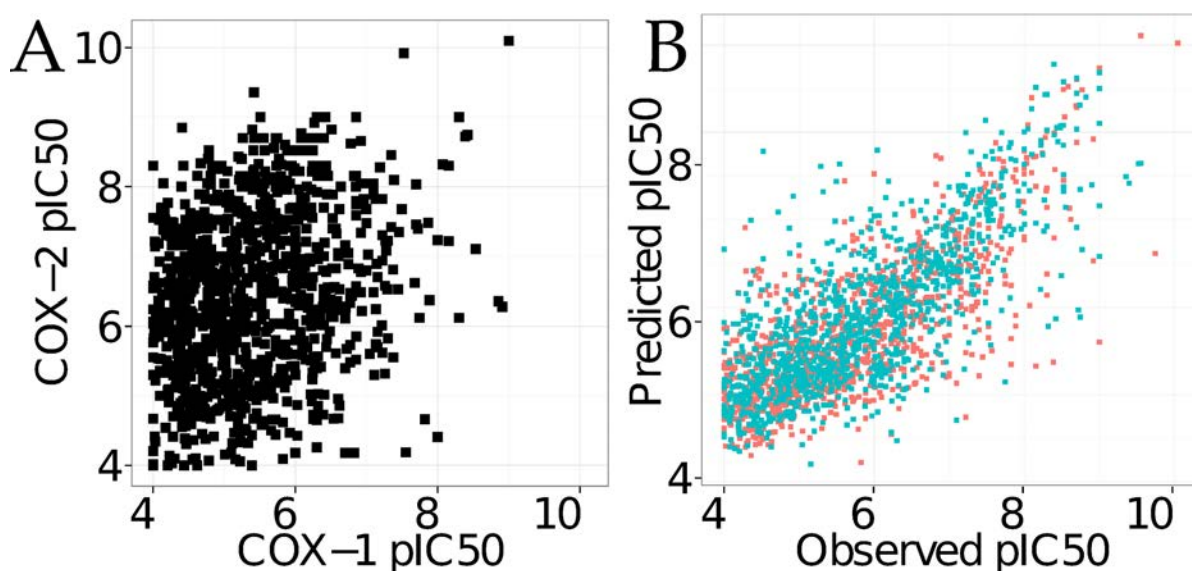


Figure .5.6: **COX** inhibitors selectivity on human **COX-1** and **COX-2**. A. Scatterplot corresponding to the comparison of bioactivities against human **COX-1** and **COX-2** for 1,288 compounds. A large proportion of the compounds present a **COX-2/COX-1** selectivity ratio between 2 and 4 pIC₅₀ units. Therefore, the present data set includes **COX** inhibitors with highly divergent bioactivity profiles for **COX-1** and **COX-2** ($R_0^2 = -0.420$). B. Scatterplot of the observed against the predicted pIC₅₀ values for the compounds described in A. Blue squares correspond to the activity on **COX-1**, whereas orange squares correspond to the activity on **COX-2**. The **PCM** models explain more than 59% of the variance ($R_0^2 = 0.593$), thus highlighting the ability of the **PCM** models to predict the potency of compounds displaying uncorrelated bioactivity profiles on human cyclooxygenases.

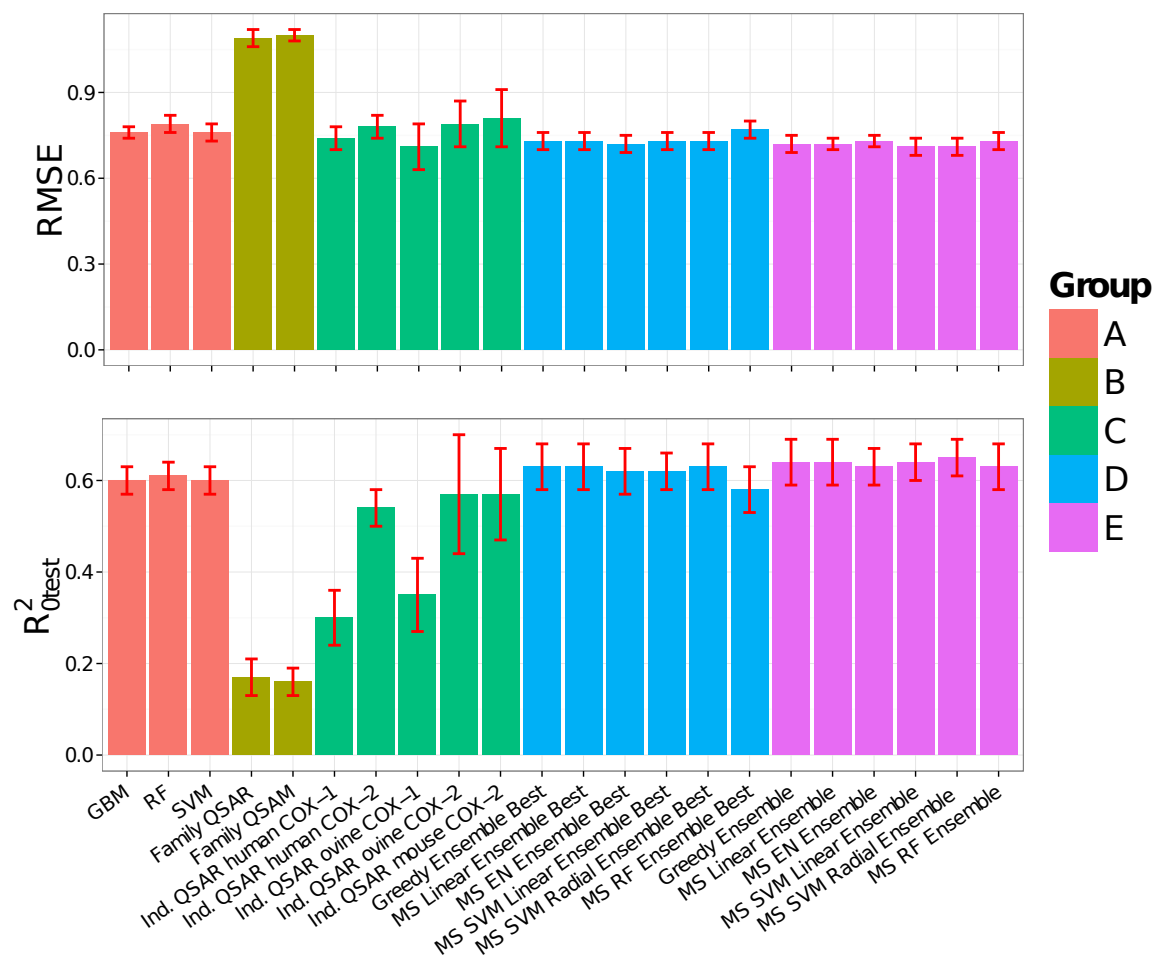


Figure .5.7: **Model performance on the test set.** $RMSE_{test}$ (A) and R^2_{test} (B) values for the following models: (group A) single **PCM**, (group B) Family **QSAR** and Family **QSAM**, (group C) individual **QSAR**, (group D) model ensembles comprising those single **PCM** models exhibiting the highest predictive power, and (group E) model ensembles comprising the whole model library. Bars are colored according to the groups defined in Table .5.2. Confidence intervals correspond to the mean value \pm one standard deviation calculated with bootstrapping [Efron and Tibshirani (1993)].

		q^2_{int}	RMSE _{int}	$R^2_{0 \text{ test}}$	RMSE _{test}	q^2_{test}	CCC
A	GBM	0.59 +/- 0.02	0.77 +/- 0.01	0.60 +/- 0.03	0.76 +/- 0.02	0.60 +/- 0.03	0.76 +/- 0.02
	RF	0.60 +/- 0.03	0.78 +/- 0.02	0.61 +/- 0.03	0.79 +/- 0.03	0.61 +/- 0.03	0.74 +/- 0.02
	SVM	0.61 +/- 0.03	0.75 +/- 0.03	0.60 +/- 0.03	0.76 +/- 0.03	0.60 +/- 0.03	0.76 +/- 0.02
B	Family QSAR	0.17 +/- 0.02	1.13 +/- 0.02	0.17 +/- 0.04	1.09 +/- 0.03	0.17 +/- 0.04	0.43 +/- 0.03
	Family QSAM	0.16 +/- 0.02	1.10 +/- 0.02	0.16 +/- 0.03	1.10 +/- 0.02	0.16 +/- 0.03	0.28 +/- 0.02
C	Ind. QSAR human COX-1	0.31 +/- 0.04	0.75 +/- 0.05	0.30 +/- 0.06	0.74 +/- 0.04	0.30 +/- 0.06	0.45 +/- 0.05
	Ind. QSAR human COX-2	0.60 +/- 0.24	0.78 +/- 0.03	0.54 +/- 0.04	0.78 +/- 0.04	0.53 +/- 0.04	0.68 +/- 0.03
	Ind. QSAR ovine COX-1	0.28 +/- 0.11	0.83 +/- 0.08	0.35 +/- 0.08	0.71 +/- 0.08	0.09 +/- 0.09	0.50 +/- 0.07
	Ind. QSAR ovine COX-2	0.53 +/- 0.07	0.78 +/- 0.06	0.57 +/- 0.13	0.79 +/- 0.08	0.57 +/- 0.13	0.74 +/- 0.09
	Ind. QSAR mouse COX-2	0.49 +/- 0.08	0.84 +/- 0.10	0.57 +/- 0.10	0.81 +/- 0.10	0.57 +/- 0.11	0.71 +/- 0.07
D	Greedy Ensemble Best	-	0.73 +/- 0.01	0.63 +/- 0.05	0.73 +/- 0.03	0.63 +/- 0.05	0.77 +/- 0.02
	MS Linear Ensemble Best	0.63 +/- 0.02	0.73 +/- 0.01	0.63 +/- 0.05	0.73 +/- 0.03	0.63 +/- 0.05	0.78 +/- 0.02
	MS EN Ensemble Best	0.63 +/- 0.02	0.72 +/- 0.02	0.62 +/- 0.05	0.72 +/- 0.03	0.62 +/- 0.05	0.78 +/- 0.02
	MS SVM Linear Ensemble Best	0.63 +/- 0.01	0.73 +/- 0.02	0.62 +/- 0.04	0.73 +/- 0.03	0.63 +/- 0.05	0.78 +/- 0.02
	MS SVM Radial Ensemble Best	0.63 +/- 0.02	0.73 +/- 0.02	0.63 +/- 0.05	0.73 +/- 0.03	0.63 +/- 0.05	0.78 +/- 0.02
	MS RF Ensemble Best	0.61 +/- 0.01	0.76 +/- 0.01	0.58 +/- 0.05	0.77 +/- 0.03	0.58 +/- 0.05	0.75 +/- 0.02
E	Greedy Ensemble	-	0.73 +/- 0.01	0.64 +/- 0.05	0.72 +/- 0.03	0.64 +/- 0.05	0.78 +/- 0.02
	MS Linear Ensemble	0.63 +/- 0.02	0.73 +/- 0.02	0.64 +/- 0.05	0.72 +/- 0.02	0.64 +/- 0.05	0.78 +/- 0.02
	MS EN Ensemble	0.64 +/- 0.01	0.73 +/- 0.01	0.63 +/- 0.04	0.73 +/- 0.02	0.63 +/- 0.04	0.78 +/- 0.02
	MS SVM Linear Ensemble	0.64 +/- 0.03	0.73 +/- 0.04	0.64 +/- 0.04	0.71 +/- 0.03	0.64 +/- 0.04	0.80 +/- 0.02
	MS SVM Radial Ensemble	0.64 +/- 0.02	0.73 +/- 0.02	0.65 +/- 0.04	0.71 +/- 0.03	0.65 +/- 0.04	0.80 +/- 0.02
	MS RF Ensemble	0.64 +/- 0.02	0.73 +/- 0.02	0.63 +/- 0.05	0.73 +/- 0.03	0.63 +/- 0.05	0.78 +/- 0.02

Table .5.2: **Internal and external validation metrics (mean values +/- one standard deviation) for the PCM (A), Family QSAM (B), Family QSAR (B), Individual QSAR models (C), Ensemble PCM models combining the most predictive models (D), and Ensemble PCM models combining the whole model library (E).** "Best" refers to the ensembles trained on only the three most predictive RF, GBM and SVM models. MS of models trained with different algorithms in a models ensemble allows to increase predictive ability, as the highest $R^2_{0 \text{ test}}$ and RMSE_{test} values, 0.652 and 0.706 pIC₅₀ units respectively, were obtained with the "MS SVM Radial Ensemble". The standard deviation for the metrics was calculated with the bootstrap method [Efron and Tibshirani (1993)].

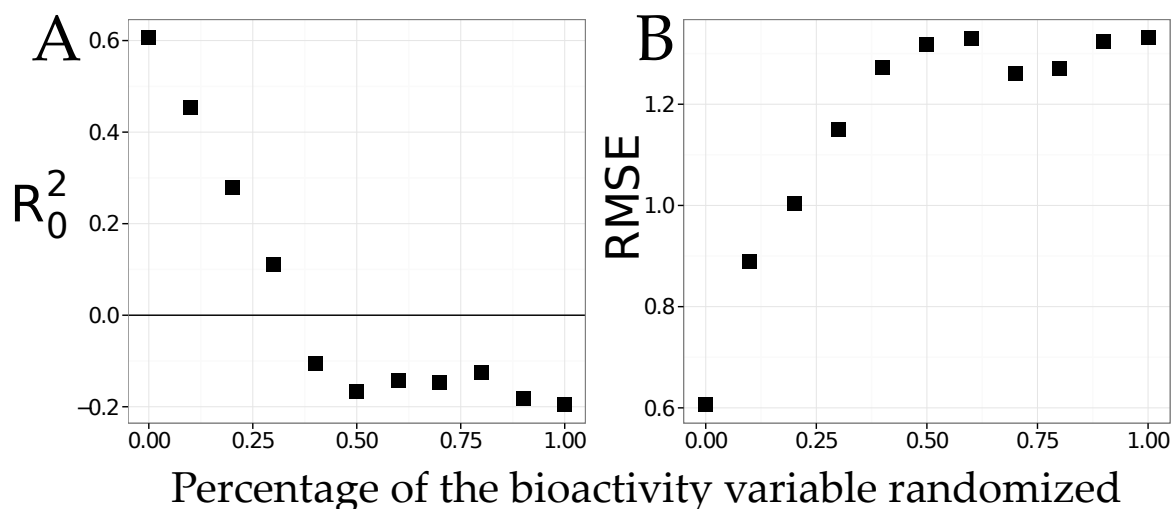


Figure .5.8: **Y-scrambling**. Scatterplots corresponding to the percentage of bioactivities randomized, against (A) $R_{0\text{ test}}^2$ and (B) $\text{RMSE}_{\text{test}}$ values. The intercept in A becomes negative when 25-50% of the bioactivity variable is randomized. This finding indicates that PCM performance is not the consequence of spurious correlations in the descriptor space.

To ensure that our modelling results did not arise from chance correlations, we trained models with an increasingly bigger fraction of randomized bioactivity values (y-scrambling) [Clark and Fox (2004)]. The representation of model performance as a function of the percentage of randomized bioactivities is given in Figure .5.8. When approximately 35% of the bioactivity values are randomized, $R_{0\text{ test}}^2$ become negative, which indicates that the relationships found by our models between both the chemical and the target space, and the bioactivity values are not spurious [Clark and Fox (*ibid.*)].

.5.3.3 PCM models are in agreement with the maximum achievable performance

The distributions of the respectively maximum and minimum achievable $R_{0\text{ test}}^2$ and $\text{RMSE}_{\text{test}}$ values are depicted in Figure .5.9. The maximum correlation values $R_{0\text{ test}}^2$ are far from 1, which agrees with observations previously reported for public data [Brown, Muchmore, and Hajduk (2009); Cortes-Ciriano et al. (2014)]. The mean of the minimum theoretical $\text{RMSE}_{\text{test}}$ values lies between 0.68 and 0.69, which is comparable to the level of uncertainty in public IC_{50} data reported by Kalliokoski et al. (2013). The mean of the distribution of theoretical $R_{0\text{ test}}^2$ values is between 0.67 and 0.69. The minimum $\text{RMSE}_{\text{test}}$ and maximum $R_{0\text{ test}}^2$ values obtained with the individual models,

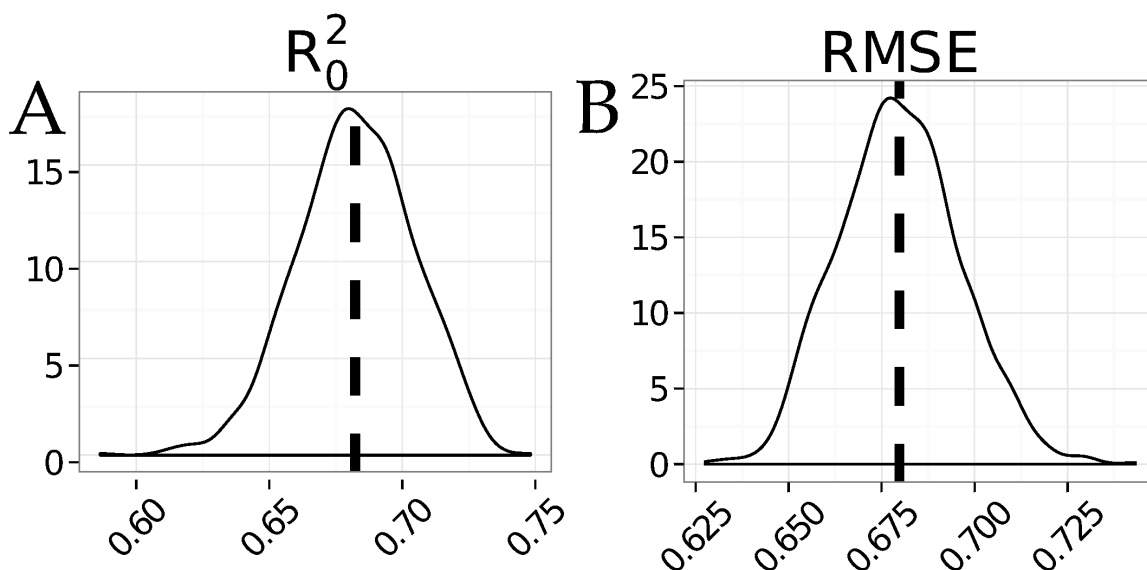


Figure .5.9: **Distribution of theoretical $R^2_{0\text{ test}}$ (A) and $\text{RMSE}_{\text{test}}$ (B) values.** The mean of the $R^2_{0\text{ test}}$ distribution, 0.68, highlights that the uncertainty in public bioactivity data does not permit models with $R^2_{0\text{ test}}$ values close to 1. Similar results were obtained for q^2_{test} . The minimum $\text{RMSE}_{\text{test}}$ value that a model can achieve without exhibiting overfitting is close to the experimental uncertainty.

0.76 and 0.61 respectively (Table .5.2A and Figure .5.7), thus appear consistent with the underlying uncertainty in the present data set.

.5.3.4 PCM outperforms both Family QSAR and Family QSAM on this data set

Interestingly, neither the Family **QSAR** nor the Family **QSAM** model alone could infer the relationships in the data set, as the respective $R^2_{0\text{ test}}$ and $\text{RMSE}_{\text{test}}$ values were: (i) for Family **QSAR**: 0.17 and 1.09 pIC₅₀ units, and (ii) for Family **QSAM**: 0.16 and 1.10 pIC₅₀ units (Table .5.2B and Figure .5.7). Taken together, these results suggest that: (i) compound bioactivities on different targets are not correlated, as indicated by the low performance of the Family **QSAR** model, and (ii) compound bioactivities depend on compounds structure, as highlighted by the low performance of the **QSAM** model.

.5.3.5 PCM outperforms individual QSAR models

We then evaluated on individual targets the usefulness of **PCM** in comparison with **QSAR** models (Table .5.2C and Figure .5.7). Independent **QSAR** models for those targets with more than 100 bioactivities, namely: human **COX-1** and **COX-2**, ovine **COX-1** and **COX-2**, and mouse **COX-2**. The human **COX-2** model exhibits a $\text{RMSE}_{\text{test}}$ value of 0.78 pIC_{50} units, which is 0.03 pIC_{50} units larger than the $\text{RMSE}_{\text{test}}$ value for the data-points annotated on human **COX-2** averaged over ten **PCM** models, namely 0.76 +/- 0.04 pIC_{50} units. By contrast, the R^2_{test} value drops to 0.54, indicating the higher performance of **PCM**. Better correlations are obtained for the individual **QSAR** models corresponding to both the mouse and the ovine **COX-2**, for which the R^2_{test} values are 0.57 in both cases, whereas the $\text{RMSE}_{\text{test}}$ values are 0.81 and 0.79 pIC_{50} units. In contrast, the human and the ovine **COX-1** **QSAR** models cannot relate the descriptor space to the bioactivity values in a statistically sound manner, as they exhibit respective R^2_{test} values of 0.30 and 0.36.

Altogether, these data evidence the versatility of **PCM** to integrate incomplete information from different sequences. Furthermore, **PCM** strongly outperforms one-target and one-space models (Family **QSAR**, individual **QSAR**, and Family **QSAM**) [Cortes-Ciriano et al. (2015)].

.5.3.6 Model ensembles exhibit higher performance than single PCM models

As the most predictive **PCM** model exhibited moderately high R^2_{test} and q^2_{test} values, as well as moderately low $\text{RMSE}_{\text{test}}$ values (Table .5.2A and Figure .5.7), we explored the possibility of enhancing model performance by combining different models into a more predictive model ensemble (.5.2D, E and Figure .5.7). Two ensemble techniques were implemented, namely: greedy optimization and model stacking (**Model Stacking (MS)**), previously described in section "Ensemble Modelling". To gather a library of diverse models, we trained a total of 282 **GBM**, **RF** and **SVM** models. Each of these models was trained with different parameter values. Hence, the performance of single models ranged from very poor to that of the individual models described above (Table .5.2A and Figure .5.7).

Initially, we created ensembles using only the most predictive **GBM**, **RF** and **SVM** models (Table .5.2D and Figure .5.7). Overall, all model ensembles (Table .5.2D) exhibited higher predictive power than single models (Table .5.2A). The best R^2_{test} value, 0.63, was obtained with the greedy and the **MS** linear ensemble. The weights for the three models in the greedy ensemble were: (i) **GBM**: 0.35, (ii) **RF**: 0.12, and (iii) **SVM**: 0.53. The **MS** Elastic Net ensemble displayed the highest predictive power, with a $\text{RMSE}_{\text{test}}$ value of 0.72 (Table .5.2D and Figure .5.7). The small differences in

performance observed between ensembles, with the exception of the RF ensemble are negligible, since, in the experience of the authors [Cortes-Ciriano et al. (2014)], the standard deviation observed for the R^2_{test} and $\text{RMSE}_{\text{test}}$ values when using different samples during model training are between 0.1 and 0.3. The only model that led to worse results was the RF ensemble, with R^2_{test} and $\text{RMSE}_{\text{test}}$ values of 0.58 and 0.77 respectively.

In a second step, ensembles were optimized using all models in the model library, namely 282 (Table .5.2E and Figure .5.7). Interestingly, the values of the statistical metrics of all ensembles increased. The MS SVM ensemble with radial kernel displayed the highest predictive ability, with R^2_{test} and $\text{RMSE}_{\text{test}}$ of 0.65 and 0.71 pIC₅₀ units, which only differs marginally from the minimum theoretical $\text{RMSE}_{\text{test}}$ value, namely 0.68 (Figure .5.9).

Worthy of mention is the lack of performance improvement (data not shown) of homo-ensembles (*i.e.* ensembles created with models trained with a given algorithm but with different parameter values) with respect to the most predictive single models (Table .5.2A and Figure .5.7), as the difference in R^2_{test} and $\text{RMSE}_{\text{test}}$ values was below 0.01 for both metrics. By contrast, the ensembles exhibiting the highest predictive power on the test set were obtained when combining models with high and low predictive ability. This increase in performance is likely to arise from the fact that these models display uncorrelated resampling profiles, *i.e.* the predictions calculated on the hold-out folds during cross-validation are not correlated (Figure .5.10). Overall, these data underline the highest predictive power of hetero-ensembles generated with a model library displaying a comprehensive range of predictive abilities.

.5.3.7 The ensemble standard deviation enables the definition of informative confidence intervals

Figure .5.11 displays the percentage of data-points which predicted values lie within confidence intervals calculated with increasingly larger β values (Equation .5.1). The ensemble model exhibiting the highest predictive power ($\text{RMSE}_{\text{test}}$: 0.71; R^2_{test} : 0.65), namely MS SVM Radial Ensemble, was used to make the predictions and to calculate confidence intervals. Confidence intervals calculated for the cross-validated predictions (shown as squares in Figure .5.11) require larger β values to reach a given level of confidence when compared to those calculated on the test set (shown as triangles in Figure .5.11). This can be seen as the percentage of data-points for which true value is within the confidence interval ($\beta = 1$) for the cross-validated predictions is 40%, whereas this value increases till 70% in the case of the test set. This difference might be due to the fact that predictions on the test set are made with models trained on a larger fraction of the data set. Nevertheless, the error in prediction on the test

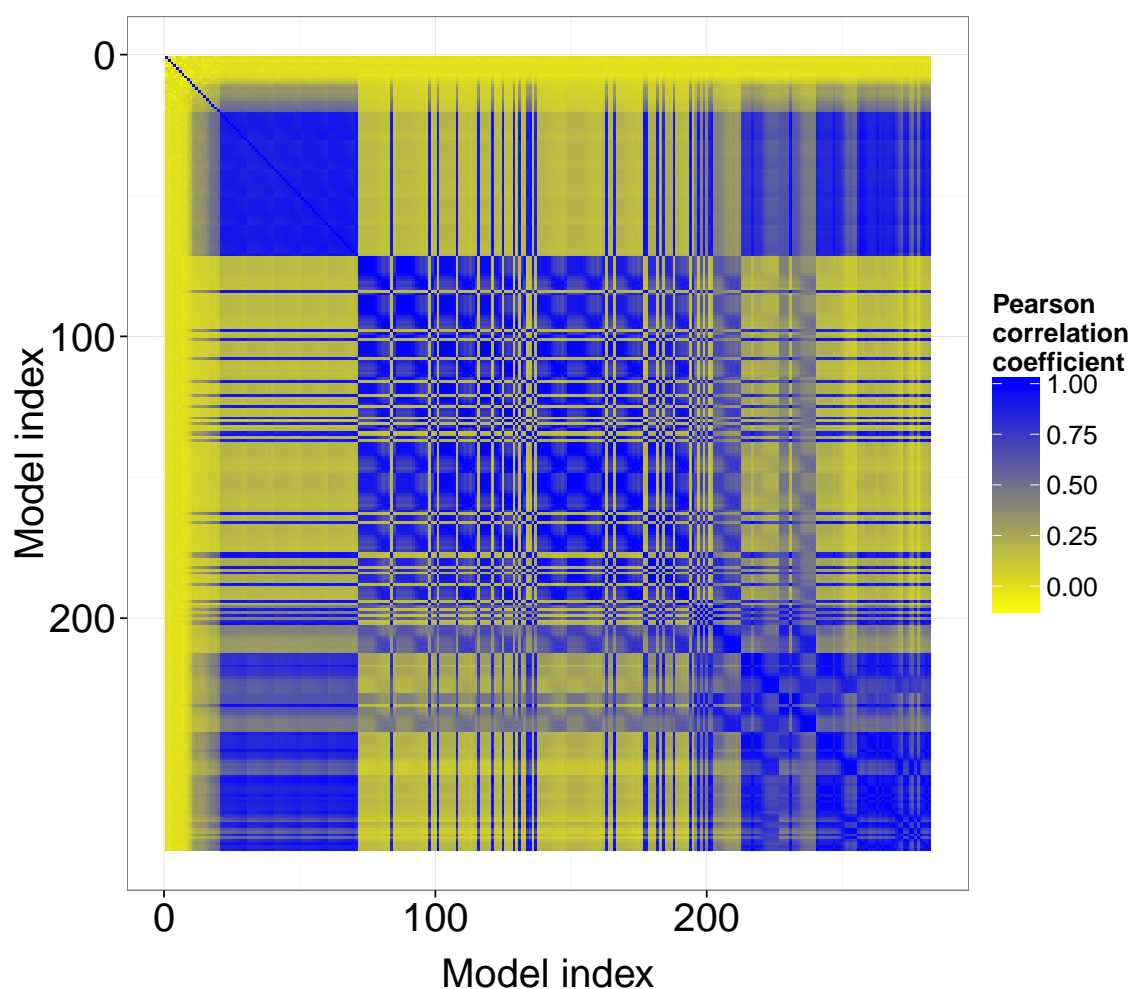


Figure .5.10: **Pairwise Pearson correlation for the cross-validation predictions across the model library.** The predictive power across the model library is not uniformly distributed, as the predicted values for a large fraction of model pairs are uncorrelated (yellow areas). Therefore, the combination of these models in a model ensemble is expected to lead to higher predictive power than individual models ("wisdom of crowds").

set might increase if the compounds present therein were structurally dissimilar. In those cases, a larger β value would be required, with respect to that for the training set, to reach a given confidence level.

Overall, the percentage of true values lying within the confidence interval derived for a given β value is expected to increase with model performance. Figure .5.11 can be used to determine the β value corresponding to the confidence interval required by the user.

.5.3.8 Ensemble modelling enables the prediction of uncorrelated human COX inhibitor bioactivity profiles

As previously stated, selectivity is a crucial aspect in the discovery and optimization of COX inhibitors. To assess whether PCM models were able to predict the pIC₅₀ values for compounds displaying uncorrelated bioactivity profiles on human COX-1 and COX-2, we predicted the bioactivity values for the 1,086 compounds annotated on both human COX-1 and COX-2. Figure .5.6B, which displays the observed against the predicted pIC₅₀ values for these compounds, shows that PCM models are able to predict the potency for compounds displaying uncorrelated bioactivity profiles on human cyclooxygenases. Indeed, the R^2_{test} and RMSE_{test} values calculated for the observed pIC₅₀ values with respect to those predicted by the PCM model are, respectively, 0.59 and 0.76 pIC₅₀ unit.

Subsequently, we analyzed the capability of PCM models to correctly predict the bioactivity for both selective and non-selective compounds. A compound was considered as selective or non selective if the absolute value of the difference between its bioactivity on COX-1 and COX-2 is larger or smaller than 2 pIC₅₀ units. On this basis, 226 compounds were considered as selective, and 860 as non selective. The error in prediction for the non selective compounds was lower than 1 pIC₅₀ unit in 85.4% of the cases, and lower than 0.5 pIC₅₀ unit for 55.6% thereof. On the other hand, the error in prediction was lower than 1 pIC₅₀ unit for 73.23% of the selective compounds, and lower than 0.5 pIC₅₀ unit for 42.9% thereof. When considering a more stringent selectivity cut-off value, namely 3 pIC₅₀ units, we obtained a set of 61 compounds. The error in prediction for this set was lower than 1 pIC₅₀ unit in 66.4% of the cases, and lower than 0.5 pIC₅₀ unit for 40.2% thereof.

Consequently, these data indicate that PCM models are capable to predict the potency for both selective and non selective compounds on human COX-1 and COX-2. In addition, we anticipate that model performance is likely to increase with the inclusion of more bioactivity data in the models.

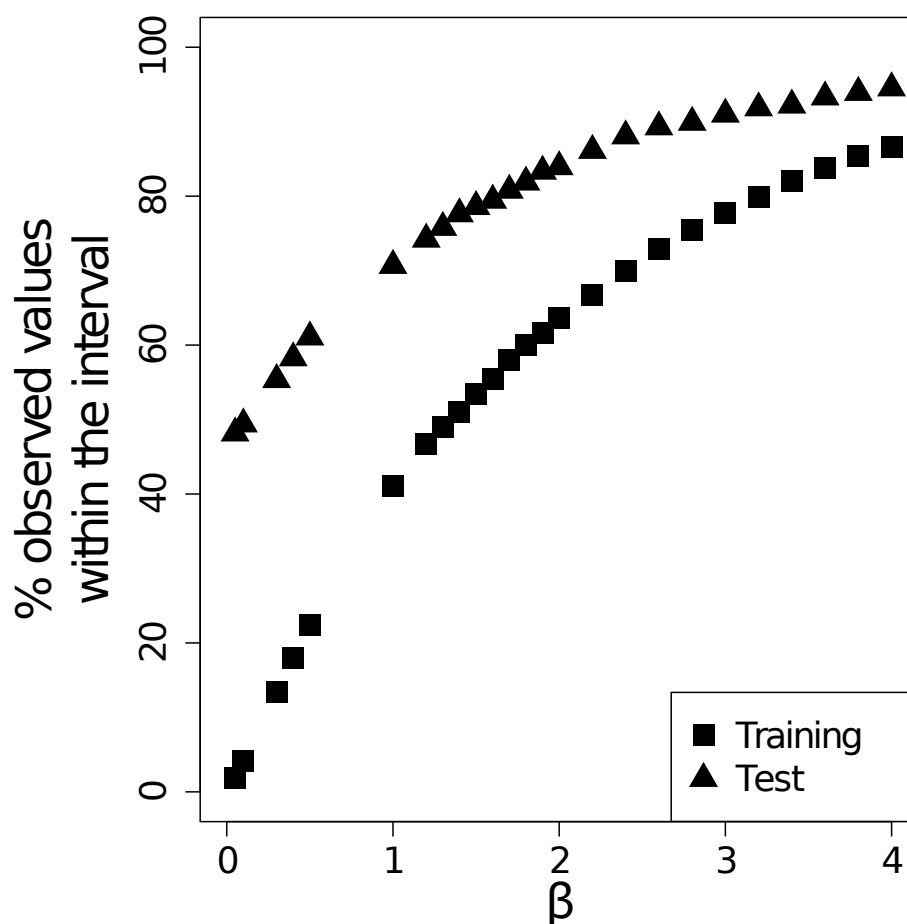


Figure .5.11: **Confidence intervals calculated from the ensemble standard deviation of the models present in the model ensembles.** The percentage of data-points which predicted bioactivities lie within confidence intervals calculated with increasingly larger β values (Equation .5.1), is shown for: (i) the cross validated predictions calculated during model training (*Training* in the Figure), and (ii) for the predictions on the test set (*Test* in the Figure) calculated with the most predictive model ensemble, namely "Stacking SVM Radial Ensemble". The percentage of true values lying within the confidence interval derived for a given β value increases with the number of data-points available during model training. Overall, the confidence intervals derived from the ensemble standard deviation provide an estimation of the reliability of individual predictions, as in practice, this plot can be used to determine the β value corresponding to a given confidence level.

.5.3.9 Model performance per target is related to compound diversity

To further assess model performance on a *per* target basis, we generated 10 RF models each one trained on a different subset of the whole data set. The variation of performance between the protein targets can be also related to the compound diversity (Figure .5.12).

Human cyclooxygenases, with the highest number of annotated compounds (Table .5.1), exhibited average $\text{RMSE}_{\text{test}}$ values between 0.74 and 0.76 pIC_{50} unit. For these proteins, the distributions of pairwise compound similarity (Figure .5.12) are skewed towards high similarity values, with mean values between 0.75 and 0.85.

Likewise, mouse COX-2 and ovine COX-1 display average $\text{RMSE}_{\text{test}}$ values of 0.70 and 0.73 pIC_{50} unit probably related to the smaller number of compounds annotated on these proteins (Table .5.1). High predictive ability on mouse COX-2 was expected given the high R^2_{test} value, 0.57, obtained with the individual QSAR model, whereas low performance was expected for ovine COX-1, as the individual QSAR model displayed a R^2_{test} value of 0.36. Unsurprisingly, skewed distributions in compound diversity are observed for mouse COX-2 and ovine COX-1 (Figure .5.12).

Conversely, ovine COX-2, with 341 annotated compounds, displayed a worse average $\text{RMSE}_{\text{test}}$ value, within the 0.80-0.85 range of pIC_{50} unit (Figure .5.13). This decrease in performance for ovine COX-2 might be ascribed to the higher dispersion of the pairwise compound similarity distribution with respect to those observed for mouse COX-2 and ovine COX-1 (Figure .5.12).

The dependency of model performance on compound diversity is even more contrasted for targets with less than 100 annotated bioactivities. Indeed, the average $\text{RMSE}_{\text{test}}$ value for mouse COX-1, with 50 compounds, lies within the 0.57-0.62 range of pIC_{50} unit and the distribution of compounds diversity is skewed towards high similarity values (Figure .5.12). However, the average $\text{RMSE}_{\text{test}}$ value increases till 0.80-0.90 pIC_{50} unit for bovine COX-1 (Figure .5.12), annotated with 48 bioactivities and for which the pairwise compound similarity distribution presents several peaks, thus highlighting the structural diversity of the compounds. Finally, targets with less than 30 annotated compounds exhibit multimodal pairwise similarity distributions and, consequently, model performance is low, with standard deviations in the 0.50-1.00 range of pIC_{50} unit (Figure .5.13).

Overall, chemical diversity in the training set contributes to enhance the applicability of a PCM model. Nonetheless, a balance needs to be established between this diversity and the number of data-points to ensure model convergence.

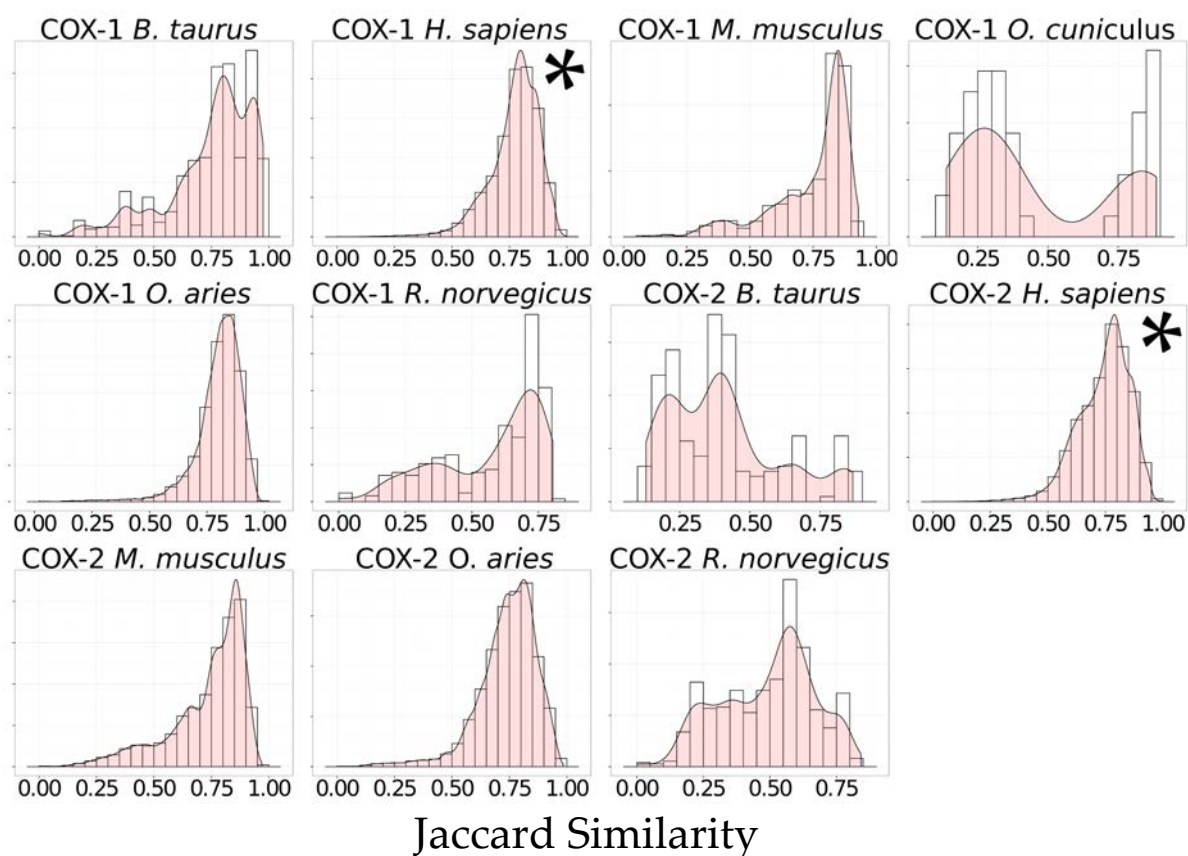


Figure .5.12: **Jaccard pairwise similarity distributions for the compounds annotated on each target.** Compounds annotated on the human cyclooxygenases (annotated with a star in the plots) display compound similarity distributions with mean values skewed towards 1. By contrast, compounds annotated on targets with less than 30 annotated bioactivities display multimodal similarity distributions. A correlation between model performance and both the number of data-points and chemical diversity was established (see main text). Distributions were calculated with the same descriptors than the ones used to train the [PCM](#) models.

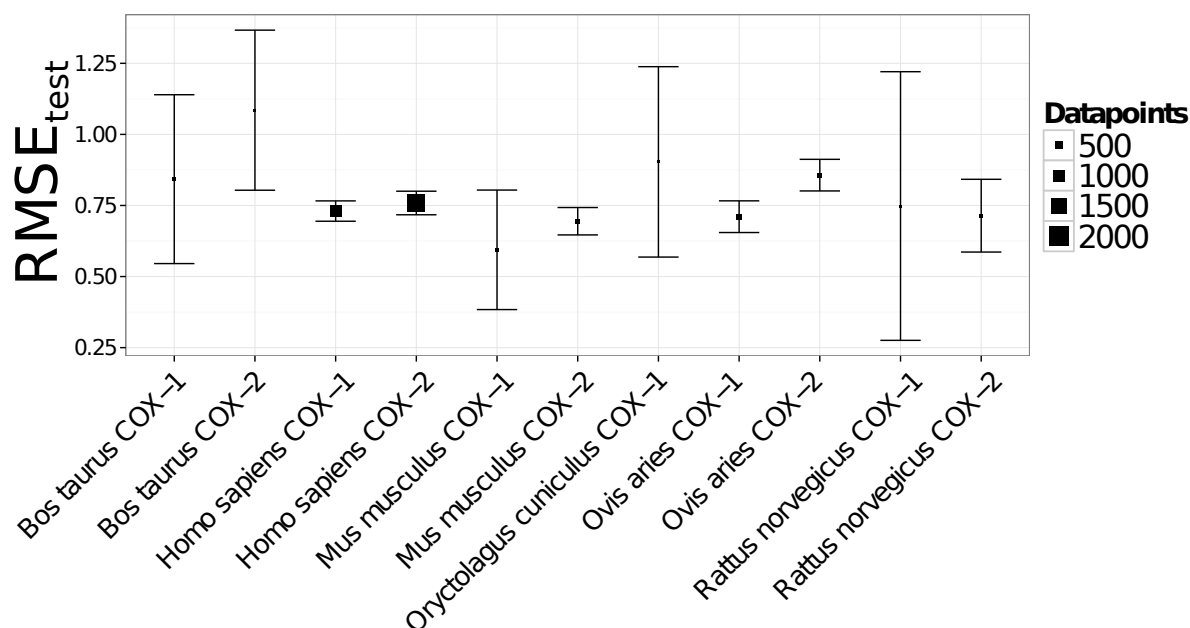


Figure .5.13: **Target-averaged model performance.** The number of data-points is displayed through the size of the squares. A correlation can be established between the number of data-points and model performance, quantified by the standard deviation of the RMSE_{test} values. Targets annotated with less than 30 compounds or with chemical structures displaying high structural diversity (*Oryctolagus cuniculus* COX-1, *Rattus norvegicus* COX-1, *Bos taurus* COX-1, and *Bos taurus*) are produced with high mean RMSE_{test} values. These observations indicate that PCM models are not always able to extrapolate in the chemical or the target space if a given target or compound family is not sufficiently represented in the data set.

.5.3.10 Interpretation of compound substructures

Predictive method

The usage of unhashed fingerprints permitted the deconvolution of the chemical space to determine the influence of compound substructure on bioactivity. Two substructure analysis methodologies were implemented, as described in the section "Interpretation of Compound Substructures". The first approach, predictive method, relies on the PCM model to correctly predict the bioactivity for a compound when a given substructure is virtually removed from a compound descriptor. The second approach, Student's method, is a pipeline designed to statistically assess how the presence of a given substructure influences, on average, bioactivity on the compounds.

Figure .5.14 shows the contribution of each substructure to bioactivity on human COX-1 and COX-2 calculated with the predictive method. Red and blue areas correspond respectively to substructures that, on average, enhance or decrease compound bioactivity. Representative substructures either deleterious or beneficial for bioactivity are also shown. Generally, substructures shown to have an influence on bioactivity display an opposite behaviour depending on the isoenzyme type. For example, a pyrrole ring with aryl substituents in the 2,3-positions (substructure **c** in Figure .5.14) is predicted to have a high influence on bioactivity, increasing it on COX-2 and decreasing it on COX-1. This observation is in agreement with the literature as the 2,3-diarylpyrrole series with an halogen substituent in the 5-position acting as electron withdrawing group have been found as selective COX-2 inhibitors [Wilkerson et al. (1994, 1995)]. The pyrrole moiety with a radical in the 1-position is also found as a selectivity feature towards COX-2 (substructure **b** in Figure .5.14). This agrees with the discovery by Khanna et al. (1997) of a series of 1,2-diarylpyrroles as potent and selective COX-2 inhibitors.

On the other hand, substructures conferring a deleterious effect could also be identified. substructure **e** in Figure .5.14 is represented within compound 3-(1H-indol-5-yloxy)-5,5-dimethyl-4-(4-methylsulfonylphenyl)furan-2-one (ChEMBL322276). This compound is part of a series of 3-heteroaryloxy-4-phenyl-2(5H)-furanones reported as selective COX-2 inhibitors by Lau et al. (1999). Its COX-1/COX-2 selectivity ratio is larger than 4.17, which agrees with the prediction of decreasing bioactivity on COX-1. In general, substructures decreasing bioactivity tend to be small and less informative (e.g. single atoms or substructures with two heavy atoms), than those fostering compound potency.

Student's method

The implementation of the Student's method to deconvolute the chemical space (Figure .5.15), which evaluates the statistical significance between bioactivity distributions

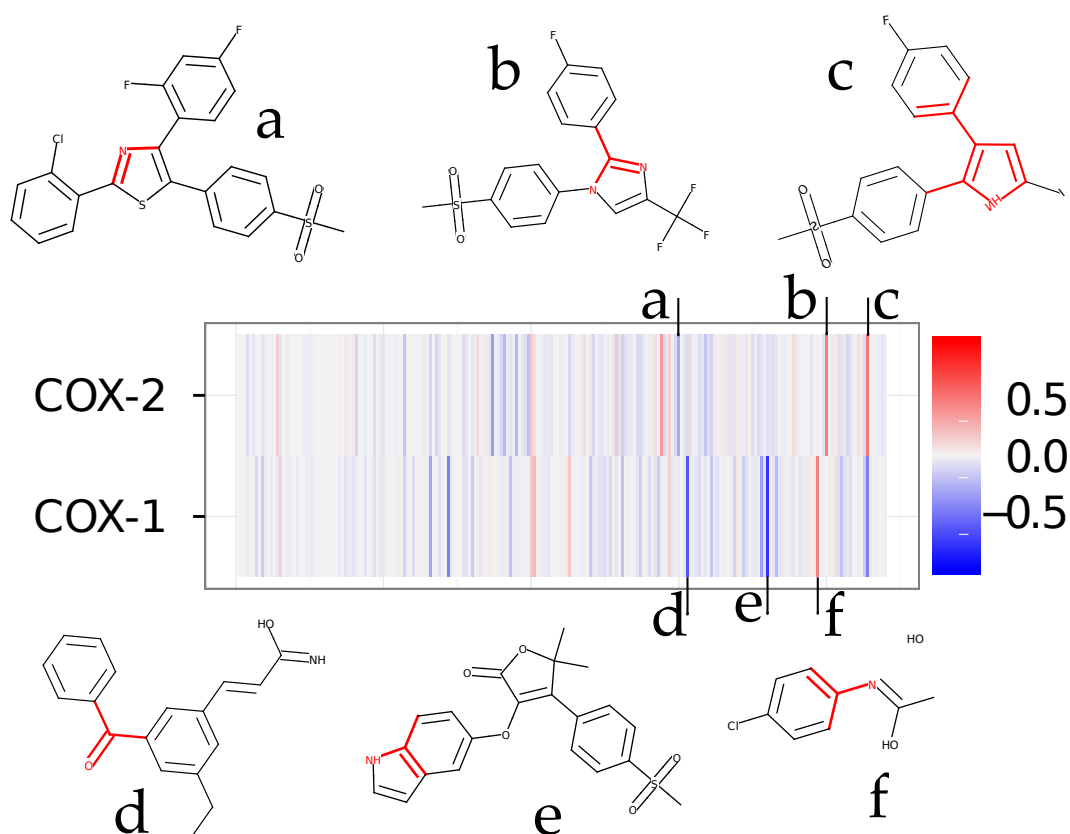


Figure 5.14: **Influence of compound substructures on potency and selectivity on human COX-1 and COX-2.** Rows in the heatmap are indexed by the isoenzyme type whereas columns correspond to compound substructures. Substructures are depicted in red within arbitrary molecules presenting it. The color represents the average influence (pIC₅₀ units) of each substructure on bioactivity. Red corresponds to an average increase in bioactivity, whereas blue indicates the a deleterious effect. Well-known chemical moieties, *e.g.* pyrrole rings (c), were singled out as selectivity determinants. For instance, substructure **d** is present in sulfonamides such as diflumidone, and substructure **B** in selective 1,2-diarylpyrroles COX-2 inhibitors.

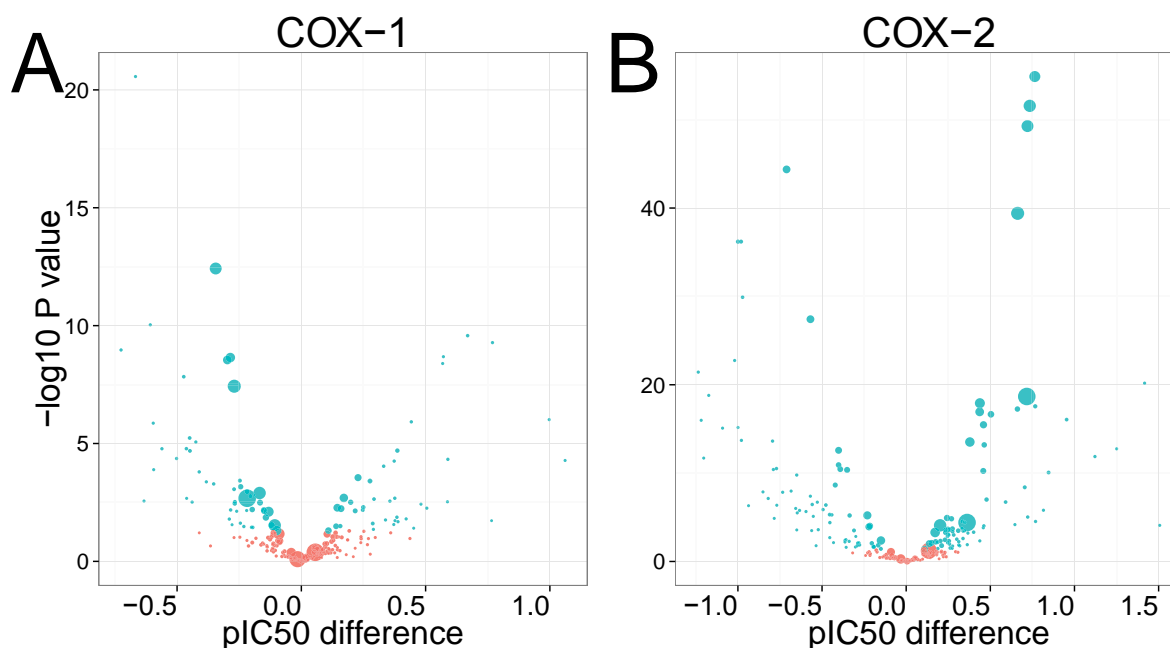


Figure 5.15: Volcano plots corresponding to the results of the Student's method applied on human COX-1 (A) and COX-2 (B). The size of the points is proportional to the number of molecules in the data set containing a given substructure. Significant P values are shown in red (two-tailed t -test, $\alpha = 0.05$).

in the presence or absence of each compound substructure, led to the following observations: (i) 74 substructures increase bioactivity on COX-2, (ii) 64 substructures decrease bioactivity on COX-2, (iii) 9 substructures increase bioactivity on COX-1, (iv) 2 substructures decrease bioactivity on COX-1, (v) 1 substructure increases bioactivity on both COX-1 and COX-2, and (vi) 6 substructures decrease bioactivity on both COX-1 and COX-2.

Well-known chemical moieties conferring selectivity to COX-2 were present in this substructure selection. Figure 5.16 shows the 20 substructures predicted to have the highest influence to increase bioactivity on human COX-2. For instance, substructures containing thiazole, pyrrole, pyrazole and oxazole rings were enriched for COX-2 [Dannhardt and Laufer (2000); Leval et al. (2000)]. Likewise, tri-fluorometil and sulfonamide radicals, which appear in *e.g.* celecoxib, were also enriched [Dannhardt and Laufer (2000)]. Substructures predicted to influence in the same way the compound bioactivity on both COX-1 and COX-2 are small, which makes difficult to extract medicinal chemistry knowledge therefrom (Figure 5.17).

It is nevertheless remarkable that the output of both methods is contradictory

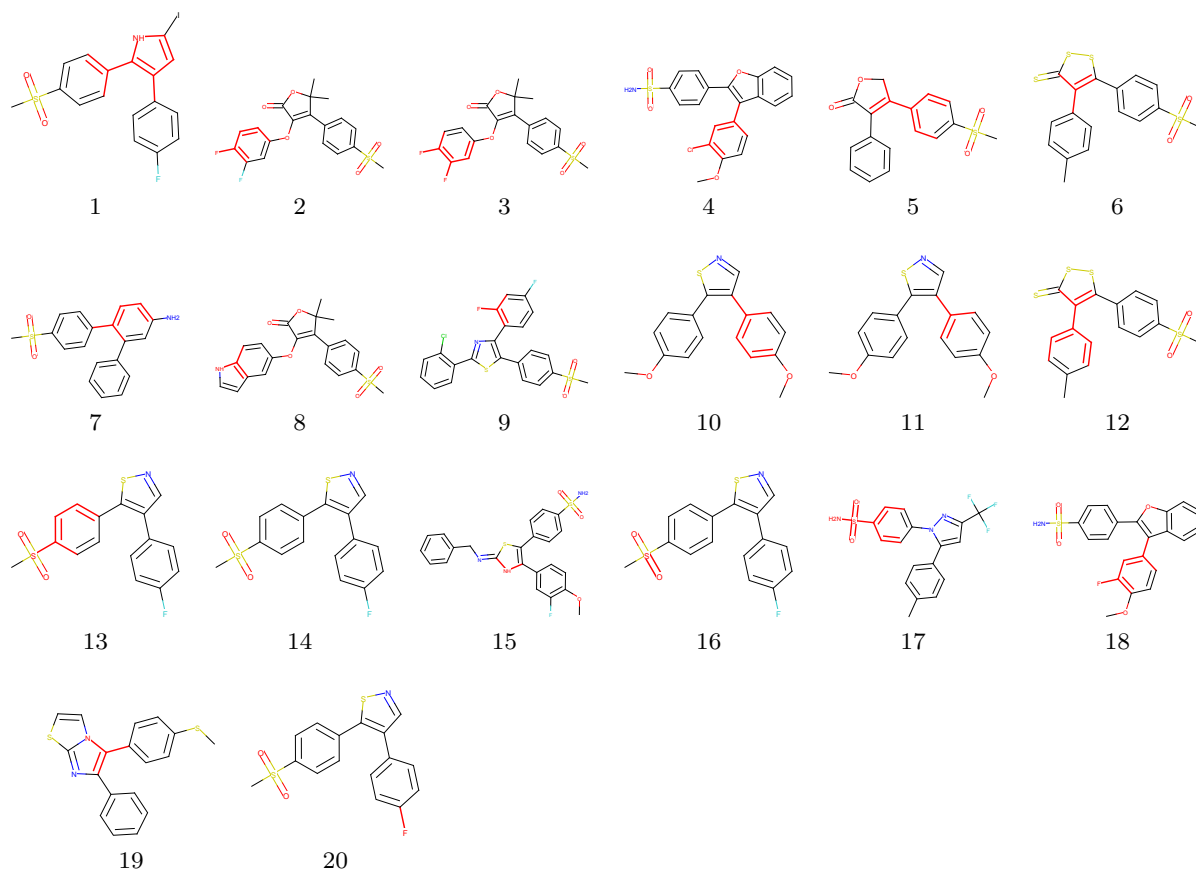


Figure .5.16: **Compound substructures predicted to increase the bioactivity on human COX-2.** The 20 substructures predicted to have the highest influence on bioactivity on human COX-2 (P35354) are plotted. Known chemical moieties such as pyrrole rings (1), aryl substituents (e.g. 4 and 5) or benzylsulfonamide (17) are represented. These substructures appear in diverse NSAIDs such as rofecoxib or etoricoxib, as well as in chemical families of COX-2 inhibitors based on e.g. 1,5-diarylpyrazoles or 3,4-diaryl-substituted furans [Blobaum and Marnett (2007); Dannhardt and Laufer (2000); Leval et al. (2000)]

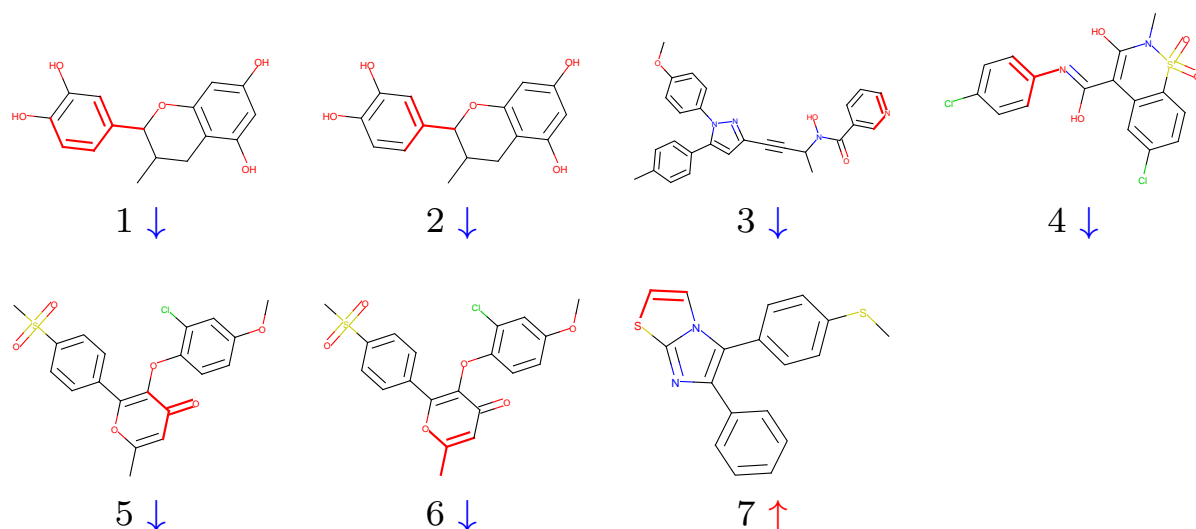


Figure .5.17: **Compound substructures predicted to have the same influence on human COX-1 and COX-2.** Sub-structures predicted to decrease bioactivity are accompanied by a blue arrow, whereas that predicted to increase bioactivity are followed by a red arrow. Smaller substructures are found in this case, predominating substituents on the benzene ring. Therefore, substructure-activity relationships are difficult to be determined.

for some substructures. By way of example, substructure **d** in Figure .5.14 is considered as deleterious for bioactivity on COX-1 by the predictive method, whereas it is regarded as beneficial by the Student's method. Dannhardt, Fiebich, and Schwepenhäuser (2002) highlighted the key role of the carbonyl moiety for the potency of a series of diarylmethanone compounds on both COX isoenzymes. Nonetheless, Scholz et al. (2012) have recently reported a series of *ortho*-carbaborane derivatives of indomethacin as selective COX-2 inhibitors. Furthermore, substructure **d** also appears in a series of [2-[(4-substituted or 4,5-disubstituted)-pyridin-2-yl]carbonyl-(5- or 6-substituted or 5,6-disubstituted)-1H-indol-3-yl]acetic acid analogues identified as COX-2 inhibitors [Hayashi et al. (2012)]. Plausible reasons for this divergence are analyzed in the Discussion section.

Overall, both substructure analysis pipelines have proven to be able to highlight chemical moieties conferring or decreasing potency and selectivity in agreement with the literature.

5.4 Discussion

In this chapter two ensemble modelling techniques, namely greedy optimization and model stacking, have been presented and benchmarked on a PCM data set comprising the bioactivities of COX inhibitors on 11 mammalian cyclooxygenases (Table .5.1). PCM has been shown to relate the target and the chemical spaces to bioactivity in a statistically sound manner (Table .5.2) [Golbraikh and Tropsha (2002); Tropsha and Golbraikh (2010); Tropsha and Gramatica (2003)]. Family QSAR as well as Family QSAM displayed poor performance (Table .5.2B and Figure .5.7).

Three machine learning algorithms (GBM, RF and SVM) have been implemented individually and combined in model ensembles. The application of ensemble modelling has been shown to outperform single machine learning models, the improvement being larger if the three most predictive GBM, RF and SVM models are combined in the same ensemble (Table .5.2D and Figure .5.7). Nonetheless, the model stacking (MS) SVM radial kernel model trained on the predictions of a library of 282 single PCM models (Table .5.2E and Figure .5.7) displayed the lowest RMSE_{test} and the highest $R^2_{0\text{ test}}$ values. This non-linear model combination led to a RMSE_{test} value comparable to the experimental uncertainty of public IC₅₀ data [Kalliokoski et al. (2013)]. It is noteworthy to mention that this ensemble was obtained by combining several hundreds of poor and highly predictive models instead of only the most predictive models of each class, GBM, RF and SVM (Table .5.2D and Figure .5.7). Therefore, these results suggest that if sufficient computing resources are available, higher predictive ability can be obtained with a large and diverse model library. Given that the ensemble concept is not restricted to any particular machine learning algorithm, the pipeline proposed in this study can be further explored.

The variability in the predictions of the individual models composing model ensembles, quantified by the ensemble standard deviation, served to define informative confidence intervals. Previous studies highlighted the usefulness of this variability as a applicability domain metric [Dragos, Gilles, and A (2009); Sheridan (2012, 2013); Wood et al. (2013)]. Here, we have extended this concept to ensembles of models trained on different algorithms (Figure .5.11). The higher performance of model ensembles has already been observed [Costello et al. (2014); Marbach et al. (2012)]. This phenomenon, usually termed 'wisdom of crowds', arises from the fact that different models provide complementary information. Moreover, the combination of a number of models palliates the effect of extreme predictions by averaging them (regression to the mean), and the chances of obtaining erroneous predictions with a single model decrease. Interestingly, it has been recently reported in the context of cell line sensitivity prediction [Costello et al. (2014)] that higher performance was obtained by combining moderate predictive models, instead of the most predictive models of each class. This observation has been corroborated in the present study

(Table .5.2E and Figure .5.7). Overall, the application of ensemble modelling with a model library trained with either the same algorithm but different parameter values (homo-ensemble), or with different algorithms (hetero-ensemble) constitutes a promising alternative to single models in the context of predictive bioactivity modelling.

High predictive ability for compounds displaying uncorrelated bioactivity profiles on COX-1 and COX-2 was attained with both single models and model ensembles (Figure .5.6B). Therefore, the present study illustrates how the combination of the target and the chemical spaces in a single PCM model improves the prediction of compound potency in the context of multi-target systems. The implications of COX-2 in widespread diseases, *e.g.* cancer, has prompted the design of potent and selective COX-2 inhibitors since the early 1990s [Dannhardt and Laufer (2000); Leval et al. (2000)]. Thus, the suitability of PCM to predict COX inhibitor potency and to integrate multispecies bioactivity data opens new avenues for the design of cyclooxygenase inhibitors.

The two approaches presented in this study for the deconvolution of the chemical space, namely: (i) bioactivity prediction with and without a given compound substructure (predictive method), and (ii) assessment of the statistical difference between the bioactivity distributions corresponding to compounds presenting or not a given compound substructure (Student's method), singled out chemical moieties responsible for COX-2 selectivity in agreement with the scientific literature.

The divergent results described for substructure **d** in Figure .5.14, plausibly arise from the following properties of the two methods. As in the predictive method the bioactivity is predicted by calculating the average difference between the predicted value for a compound with and without a given substructure, the (potentially non-linear) relationships between the substructures present in a molecule can be established, and the dependence of bioactivity on additional substructures or scaffolds present in the molecule accounted. On the other hand, the Student's method considers the substructures as independent. The two methods can thus give contrasted results for example in the following case.

We can envision a compound, A, presenting a substructure, S_1 , having no effect on bioactivity, and a second substructure, S_2 , strongly fostering bioactivity on the studied biomolecular target. Additionally, we consider compound B, which only harbors substructure S_2 . Contradictory results would be given by the two methods with respect to the influence of substructure S_1 on bioactivity. The predictive method would predict a similar bioactivity value for compound A with and without substructure S_1 , as the bioactivity depends on substructure S_2 . By contrast, the Student's method would consider substructure S_1 as relevant for bioactivity given that the difference between the bioactivities of compounds A and B, *i.e.* either presenting or

not substructure S_1 , would be significant. It follows from the preceeding that the predictive method is best suited to give insight into the contribution of single substructures to the bioactivity of individual compounds, whereas the Student's method is more suited for the identification of the general relevance of the substructures to bioactivity. Another important consideration is the presence of substructures whose effects on bioactivity are correlated. In the situation where a compound presents two substructures whose influences on bioactivity are correlated, the predictive method would likely predict a similar activity when either of them is deleted. Covering diverse structures in the data set might alleviate this issue, as the probability of finding repeated substructure pairs is likely to decrease with chemical diversity and data set size. Overall, if the general influence of a substructure on bioactivity is assessed with the predictive method, both the mean value and the standard deviation of the differences between the predicted bioactivity values with and without a given substructure should be reported, as the standard deviation indicates whether the influence of that substructure to bioactivity depends on other substructures or not [Cortes-Ciriano et al. (2014)].

In the Student's method, the pIC_{50} difference associated to a significant p-value might be negligible from a medicinal chemistry standpoint. In addition, the capability of the t -test to identify significant differences depends on the sample size. Thus, a small pIC_{50} difference can be detected as significant if the sample size is large, whereas it might not be detected for smaller samples. Therefore, the conclusions extracted from the application of the Student's method depend on the analyzed data set, whereas the predictive method might be less dependent on the data set composition if the models are applied within their applicability domain. In the present study, we have not applied any method to control the family-wise error rate which comes from the multiple comparisons problem [Shaffer (1995)]. However, we anticipate that in other studies comprising a larger number of substructures, it would be advisable to control this problem. For a recent and detailed discussion of the application of the student t -test to assess the statistical significance of bioactivity differences in the context of Matched Molecular Pair Analysis (MMPA), the reader is referred to Kramer et al. (2014). In summary, the application of both methods can help to unravel whether the contribution of a given substructure to compound bioactivity depends exclusively on itself, or on the presence of other substructures or chemical scaffolds [Klekota and Roth (2008)].

5.5 Conclusion

Ensemble modelling has been introduced in the context of PCM to predict the potency of mammalian cyclooxygenase inhibitors. The combination of single models in model ensembles has led to increased predictive ability, as well as to the definition of

confidence intervals for individual predictions. PCM has been shown to enable the prediction of the potency for compounds exhibiting uncorrelated bioactivity profiles with high confidence. Finally, the implementation of two different substructure analysis pipelines, which reliability for different purposes has been pointed out, has permitted the recognition of chemical moieties implicated in potency and selectivity in agreement with the scientific literature.

Bibliography

- Ben-Hur, A, CS Ong, S Sonnenburg, B Schölkopf, and G Rätsch (Oct. 2008). "Support Vector Machines and Kernels for Computational Biology". In: *PLoS Comput. Biol.* 4.10. Ed. by F Lewitter, e1000173 (cit. on p. 164).
- Berman, HM, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne (2000). "The Protein Data Bank". In: *Nucleic Acids Research* 28.1, pp. 235–242 (cit. on p. 161).
- Blobaum, AL and LJ Marnett (2007). "Structural and functional basis of cyclooxygenase inhibition". In: *J. Med. Chem.* 50.7, pp. 1425–1441 (cit. on pp. 160, 190).
- Breiman, L (Oct. 2001). "Random Forests". en. In: *Mach. Learn.* 45.1, pp. 5–32 (cit. on p. 164).
- Bresalier, RS, RS Sandler, H Quan, JA Bolognese, B Oxenius, K Horgan, C Lines, R Riddell, D Morton, A Lanos, MA Konstam, and JA Baron (2005). "Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial". In: *N. Engl. J. Med.* 352.11, pp. 1092–1102 (cit. on p. 159).
- Brown, J, Y Okuno, G Marcou, A Varnek, and D Horvath (2014). "Computational chemogenomics: Is it more than inductive transfer?" In: *J. Comput. Aided Mol. Des.* Pp. 1–22 (cit. on p. 163).
- Brown, SP, SW Muchmore, and PJ Hajduk (Apr. 2009). "Healthy skepticism: assessing realistic model performance". In: *Drug Discov. Today* 14.7-8, pp. 420–427 (cit. on p. 177).
- Caruana, R, A Niculescu-Mizil, G Crew, and A Ksikes (2004). "Ensemble selection from libraries of models". In: *Proceedings of the Twenty-first International Conference on Machine Learning*. ICML '04. New York, NY, USA: ACM, p. 18 (cit. on p. 164).
- Chen, L, Y He, H Huang, H Liao, and W Wei (2008). "Selective COX-2 inhibitor celecoxib combined with EGFR-TKI ZD1839 on non-small cell lung cancer cell lines: in vitro toxicity and mechanism study". In: *Medical Oncology (Northwood, London, England)* 25.2, pp. 161–171 (cit. on p. 160).
- Clark, R and P Fox (2004). "Statistical variation in progressive scrambling". In: *J. Comput. Aided Mol. Des.* 18.7-9, pp. 563–576 (cit. on p. 177).
- Cortes-Ciriano, I (2013). *FingerprintCalculator*. URL: <http://github.com/isidro/FingerprintCalculator> (cit. on p. 162).
- Cortes-Ciriano, I, GJP van Westen, EB Lenselink, DS Murrell, A Bender, and TE Malliavin (2014). "Proteochemometric modeling in a Bayesian framework". In: *J. Cheminf.* 6.1, p. 35 (cit. on pp. 167, 177, 180, 194).

- Cortes-Ciriano, I, QU Ain, V Subramanian, EB Lenselink, O Mendez-Lucio, AP IJzerman, G Wohlfahrt, P Prusis, TE Malliavin, GJP van Westen, and A Bender (2015). "Polypharmacology modelling using Proteochemometrics: recent developments and future prospects". In: *Med. Chem. Comm.* (Cit. on p. 179).
- Costello, JC, LM Heiser, E Georgii, M Gönen, MP Menden, NJ Wang, M Bansal, M Ammad-Ud-Din, P Hintsanen, SA Khan, JP Mpindi, O Kallioniemi, A Honkela, T Aittokallio, K Wennerberg, JJ Cons, D Gallahan, D Singer, J Saez-Rodriguez, S Kaski, JW Gray, and G Stolovitzky (2014). "A community effort to assess and improve drug sensitivity prediction algorithms". In: *Nat. Biotechnol.* advance on (cit. on p. 192).
- Crofford, LJ (2013). "Use of NSAIDs in treating patients with arthritis". In: *Arthritis Research & Therapy* 15.Suppl 3, S2 (cit. on p. 160).
- Curiel, RV and JD Katz (2013). "Mitigating the cardiovascular and renal effects of NSAIDs". In: *Pain medicine (Malden, Mass.)* 14 Suppl 1, S23–28 (cit. on p. 160).
- Dannhardt, G, BL Fiebich, and J Schweppenhäuser (2002). "COX-1/COX-2 inhibitors based on the methanone moiety". In: *European J. Med. Chem.* 37.2, pp. 147–161 (cit. on p. 191).
- Dannhardt, G and S Laufer (2000). "Structural approaches to explain the selectivity of COX-2 inhibitors: is there a common pharmacophore?" In: *Curr. Med. Chem.* 7.11, pp. 1101–1112 (cit. on pp. 160, 189, 190, 193).
- Dragos, H, M Gilles, and V A (July 2009). "Predicting the predictability: a unified approach to the applicability domain problem of QSAR models". In: *J. Chem. Inf. Model.* 49.7, pp. 1762–1776 (cit. on pp. 165, 192).
- Dube, PN, SS Bule, SN Mokale, MR Kumbhare, PR Dighe, and YV Ushir (2014). "Synthesis and biological evaluation of substituted 5-methyl-2-phenyl-1H-pyrazol-3(2H)-one derivatives as selective COX-2 inhibitors: Molecular docking study". In: *Chem. Biol. Drug. Des.* Pp. 409–419 (cit. on p. 160).
- Efron, B and R Tibshirani (1993). *An introduction to the bootstrap*. New York : Chapman & Hall (cit. on pp. 175, 176).
- Fine, M (2013). "Quantifying the impact of NSAID-associated adverse events". In: *Am. J. Manag. Care* 19.14 Suppl, s267–272 (cit. on pp. 159, 160).
- Friedman, JH (2001). "Greedy function approximation: A gradient boosting machine". In: *Ann. Stat.* 29.5, pp. 1189–1232 (cit. on p. 164).
- Gaulton, A, LJ Bellis, AP Bento, J Chambers, M Davies, A Hersey, Y Light, S McGlinchey, D Michalovich, B Al-Lazikani, and JP Overington (2011). "ChEMBL: a large-scale bioactivity database for drug discovery". In: *Nucleic Acids Res.* 40.D1, pp. 1100–1107 (cit. on pp. 161, 169).
- Glen, RC, A Bender, CH Arnby, L Carlsson, S Boyer, J Smith, and RC Glenn (2006). "Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME". In: *IDrugs* 9.3, pp. 199–204 (cit. on p. 162).
- Golbraikh, A and A Tropsha (2002). "Beware of q²!" In: *J. Mol. Graphics Modell.* 20.4, pp. 269–276 (cit. on pp. 163, 192).

- Gupta, GK and A Kumar (2012). "3D-QSAR studies of some tetrasubstituted pyrazoles as COX-II inhibitors". In: *Acta Pol.* 69.4, pp. 763–772 (cit. on p. 160).
- Hawkins, DM, SC Basak, and D Mills (Mar. 2003). "Assessing model fit by Cross-Validation". In: *J. Chem. Inf. Model.* 43.2, pp. 579–586 (cit. on pp. 163, 164).
- Hayashi, S, N Ueno, A Murase, Y Nakagawa, and J Takada (2012). "Novel acid-type cyclooxygenase-2 inhibitors: Design, synthesis, and structure-activity relationship for anti-inflammatory drug". In: *European J. Med. Chem.* 50, pp. 179–195 (cit. on p. 191).
- Howes, LG (2007). "Selective COX-2 inhibitors, NSAIDs and cardiovascular events - is celecoxib the safest choice?" In: *Ther. Clin. Risk Manag.* 3.5, pp. 831–845 (cit. on p. 159).
- Hson, DJ, ND Hartley, J Gamble-G, N Brown, BC Shonesy, PJ Kingsley, RJ Colbran, J Reese, LJ Marnett, and S Patel (2013). "Substrate-selective COX-2 inhibition decreases anxiety via endocannabinoid activation". In: *Nature Neurosci.* 16.9, pp. 1291–1298 (cit. on p. 160).
- Jones, R, G Rubin, F Berenbaum, and J Scheiman (2008). "Gastrointestinal and cardiovascular risks of Nonsteroidal anti-inflammatory drugs". In: *Am. J. Med.* 121.6, pp. 464–474 (cit. on p. 160).
- Jouzeau, JY, B Terlain, A Abid, E Nédélec, and P Netter (1997). "Cyclooxygenase isoenzymes How recent findings affect thinking about nonsteroidal anti-inflammatory drugs". In: *Drugs* 53.4, pp. 563–582 (cit. on p. 160).
- Kalliokoski, T, C Kramer, A Vulpetti, and P Gedeck (2013). "Comparability of mixed IC₅₀ Data - A statistical analysis". In: *PloS ONE* 8.4, e61007 (cit. on pp. 161, 164, 177, 192).
- Khanna, IK, RM Weier, Y Yu, PW COs, JM Miyashiro, CM Koboldt, AW Veenhuizen, JL Currie, K Seibert, and PC Isakson (1997). "1,2-diarylpyrroles as potent and selective inhibitors of cyclooxygenase-2". In: *J. Med. Chem.* 40.11, pp. 1619–1633 (cit. on p. 187).
- Kim, H, CH Chae, KY Yi, K Park, and S Yoo (2004). "Computational studies of COX-2 inhibitors: 3D-QSAR and docking". In: *Bioorg. Med. Chem.* 12.7, pp. 1629–1641 (cit. on p. 160).
- Klekota, J and FP Roth (Nov. 2008). "Chemical substructures that enrich for biological activity". In: *Bioinformatics* 24.21, pp. 2518–2525 (cit. on p. 194).
- Kramer, C and R L (2012). "QSARs, data and error in the modern age of drug discovery". In: *Curr. Top. Med. Chem.* 12.17, pp. 1896–1902 (cit. on p. 164).
- Kramer, C, T Kalliokoski, P Gedeck, and A Vulpetti (June 2012). "The experimental uncertainty of heterogeneous public Ki data". In: *J. Med. Chem.* 55.11, pp. 5165–5173 (cit. on pp. 164, 169).
- Kramer, C, JE Fuchs, S Whitebread, P Gedeck, and KR Liedl (2014). "Matched Molecular Pair Analysis: Significance and the impact of experimental uncertainty". In: *J. Med. Chem.* 57.9, pp. 3786–3802 (cit. on p. 194).

- Kruger, FA and JP Overington (Jan. 2012). "Global analysis of small molecule binding to related protein targets". In: *PLoS Comput. Biol.* 8.1, e1002333 (cit. on p. 169).
- Kuhn, M (2008). "Building predictive models in R using the caret package". In: *J. Stat. Softw.* 28.5, pp. 1–26 (cit. on pp. 162, 163).
- Kuhn, M and K Json (2013). *Applied Predictive Modeling*. New York, NY: Springer New York (cit. on p. 163).
- Landrum, G (2006). *RDKit Open-source cheminformatics*. URL: <http://rdkit.org/> (cit. on p. 162).
- Lau, CK, C Brideau, CC Chan, S Con, WA Cromlish, D Ethier, JY Gauthier, R Gordon, J Guay, S Kargman, C Li, P Prasit, D Reindeau, M Thérien, DM Visco, and L Xu (1999). "Synthesis and biological evaluation of 3-heteroaryloxy-4-phenyl-2(5H)-furanones as selective COX-2 inhibitors". In: *Bioorg. Med. Chem. Letters* 9.22, pp. 3187–3192 (cit. on p. 187).
- Leval, X de, J Delarge, F Somers, P de Tullio, Y Henrotin, B Pirotte, and JM Dogné (2000). "Recent advances in inducible cyclooxygenase (COX-2) inhibition". In: *Curr. Med. Chem.* 7.10, pp. 1041–1062 (cit. on pp. 160, 189, 190, 193).
- Luo, C, M He, and L Bohlin (2005). "Is COX-2 a perpetrator or a protector? Selective COX-2 inhibitors remain controversial". In: *Acta Pharmacol. Sin.* 26.8, pp. 926–933 (cit. on p. 159).
- Marbach, D, JC Costello, R Kuffner, NM Vega, RJ Prill, DM Camacho, KR Allison, M Kellis, JJ Cons, and G Stolovitzky (2012). "Wisdom of crowds for robust gene network inference". In: *Nat. Methods* 9.8, pp. 796–804 (cit. on p. 192).
- Marcou, G, D Horvath, V Solov'ev, A Arrault, P Vayer, and A Varnek (2012). "Interpretability of SAR/QSAR models of any complexity by atomic contributions". In: *Mol. Inform.* 31.9, pp. 639–642 (cit. on p. 167).
- Mayer, Z and KJ E (2015). "caretEnsemble: Framework for combining caret models into ensembles". In: URL: <http://cran.r-project.org/web/packages/caretEnsemble/index.html> (cit. on p. 162).
- Moore, BC and DL Simmons (2000). "COX-2 inhibition, apoptosis, and chemoprevention by nonsTidal anti-inflammatory drugs". In: *Curr. Med. Chem.* 7.11, pp. 1131–1144 (cit. on p. 160).
- Murrell, DS, I Cortes-Ciriano, GJP van Westen, IP Stott, TE Malliavin, A Bender, and RC Glen (2014). "Chemistry Aware Model Builder (camb): an R Package for Predictive Bioactivity Modeling". In: <http://github.com/cambDI/camb> (cit. on pp. 161–163).
- Narsinghani, T and SC Chaturvedi (2006). "QSAR analysis of meclofenamic acid analogues as selective COX-2 inhibitors". In: *Bioorg. Med. Chem. Letters* 16.2, pp. 461–468 (cit. on p. 160).
- Nussmeier, NA, AA Whelton, MT Brown, RM Langford, A Hoeft, JL Parlow, SW Boyce, and KM Verburg (2005). "Complications of the COX-2 inhibitors parecoxib and valdecoxib after cardiac surgery". In: *N. Engl. J. Med.* 352.11, pp. 1081–1091 (cit. on p. 159).

- Oksanen, J, FG Blanchet, R Kindt, P Legendre, PR Minchin, RB O'Hara, GL Simpson, P Solymos, MHH Ss, and H Wagner (2013). *vegan: Community ecology package*. URL: <http://cran.r-project.org/web/packages/vegan/index.html> (cit. on p. 162).
- Polishchuk, PG, VE Kuzmin, AG Artemenko, and EN Muratov (2013). "Universal approach for structural interpretation of QSAR/QSPR models". In: *Mol. Inform.* 32, pp. 843–853 (cit. on p. 167).
- Reddy, RN, R Mutyala, P Aparoy, P Reddanna, and MR Reddy (2007). "Computer aided drug design approaches to develop cyclooxygenase based novel anti-inflammatory and anti-cancer drugs". In: *Curr. Pharm. Des.* 13.34, pp. 3505–3517 (cit. on p. 160).
- Rimon, G, RS Sidhu, DA Lauver, JY Lee, NP Sharma, C Yuan, RA Frieler, RC Trievel, BR Lucchesi, and WL Smith (2010). "Coxibs interfere with the action of aspirin by binding tightly to one monomer of cyclooxygenase-1". In: *Proc. Natl. Acad. Sci. USA* 107.1, pp. 28–33 (cit. on p. 161).
- Robak, P, P Smolewski, and T Robak (2008). "The role of non-steroidal anti-inflammatory drugs in the risk of development and treatment of hematologic malignancies". In: *Leuk. Lymphoma* 49.8, pp. 1452–1462 (cit. on p. 160).
- Rogers, D and M Hahn (May 2010). "Extended-connectivity fingerprints". In: *J. Chem. Inf. Model.* 50.5, pp. 742–754 (cit. on p. 162).
- Rosenbaum, L, G Hinselmann, A Jahn, and A Zell (2011). "Interpreting linear support vector machine models with heat map molecule coloring". In: *J. Cheminf.* 3, p. 11 (cit. on p. 167).
- Sandberg, M, L Eriksson, J Jonsson, M Sjöström, and S Wold (July 1998). "New chemical descriptors relevant for the design of biologically active peptides A multivariate characterization of 87 amino acids". In: *J. Med. Chem.* 41.14, pp. 2481–2491 (cit. on p. 162).
- Scholz, M, AL Blobaum, LJ Marnett, and E Hey-Hawkins (2012). "ortho-carbaborane derivatives of indomethacin as cyclooxygenase (COX)-2 selective inhibitors". In: *Bioorg. Med. Chem.* 20.15, pp. 4830–4837 (cit. on p. 191).
- Shaffer, JP (1995). "Multiple hypothesis testing". In: *Ann. Rev. Psychol.* 46.1, pp. 561–584 (cit. on p. 194).
- Sheridan, RP (2012). "Three useful dimensions for domain applicability in QSAR models using random forest". In: *J. Chem. Inf. Model.* 52.3, pp. 814–823 (cit. on pp. 164, 165, 192).
- (2013). "Using Random Forest to model the domain applicability of another Random Forest model". In: *J. Chem. Inf. Model.* 53.11, pp. 2837–2850 (cit. on pp. 165, 192).
- Sievers, F, A Wilm, D Dineen, TJ Gibson, K Karplus, W Li, R Lopez, H McW, M Remmert, J Söding, JD Thompson, and DG Higgins (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega". In: *Mol. Syst. Biol.* 7.1, p. 539 (cit. on p. 161).

- Soh, J, JU Kazi, H Li, WJ Thompson, and IB Weinstein (2008). "Celecoxib-induced growth inhibition in SW480 colon cancer cells is associated with activation of protein kinase G". In: *Mol. Carcinog.* 47.7, pp. 519–525 (cit. on p. 160).
- Sostres, C, CJ Gargallo, and A Lanas (2014). "Aspirin, cyclooxygenase inhibition and colorectal cancer". In: *World J Gastrointest Pharmacol Ther.* 5.1, pp. 40–49 (cit. on pp. 159, 160).
- Spowage, BM, CL Bruce, and JD Hirst (2009). "Interpretable correlation descriptors for quantitative structure-activity relationships". In: *J. Cheminf* 1.1, p. 22 (cit. on p. 167).
- Thun, MJ, SJ Henley, and C Patrono (2002). "Nonsteroidal anti-inflammatory drugs as anticancer agents: mechanistic, pharmacologic, and clinical issues". In: *J. Natl. Cancer. Inst.* 94.4, pp. 252–266 (cit. on p. 160).
- Tropsha, A and A Golbraikh (2010). "Predictive Quantitative Structure-Activity Relationships modeling". In: *Handbook of Chemoinformatics Algorithms* 33, p. 211 (cit. on pp. 163, 192).
- Tropsha, A and VK Gramatica PaOand Gombar (2003). "The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models". In: *QSAR Comb. Sci.* 22.1, pp. 69–77 (cit. on pp. 163, 192).
- Vane, JR (1971). "Inhibition of prostaglandin synthesis as a mechanism of action for aspirin-like drugs". In: *Nature: New biology* 231.25, pp. 232–235 (cit. on p. 159).
- Warner, TD, F Giuliano, I Vojnovic, A Bukasa, JA Mitchell, and JR Vane (1999). "Nonsteroid drug selectivities for cyclooxygenase-1 rather than cyclo-oxygenase-2 are associated with human gastrointestinal toxicity: a full in vitro analysis". In: *Proc. Natl. Acad. Sci. USA* 96.13, pp. 7563–7568 (cit. on p. 159).
- Wilkerson, WW, W Galbraith, K Gans-Brangs, M Grubb, WE Hewes, B Jaffee, JP Kenney, J Kerr, and N Wong (1994). "Antiinflammatory 4,5-Diarylpyrroles: Synthesis and QSAR". In: *J. Med. Chem.* 37.7, pp. 988–998 (cit. on p. 187).
- Wilkerson, WW, RA Copeland, M Covington, and JM Trzaskos (1995). "Antiinflammatory 4,5-diarylpyrroles 2: Activity as a function of cyclooxygenase-2 inhibition". In: *J. Med. Chem.* 38.20, pp. 3895–3901 (cit. on p. 187).
- Wood, DJ, L Carlsson, M Eklund, U Norinder, and J Stå lring (Mar. 2013). "QSAR with experimental and predictive distributions: an information theoretic approach for assessing model quality". In: *J. Comput. Aided Mol. Des.* 27.3, pp. 203–219 (cit. on pp. 165, 192).
- Xie, WL, JG Chipman, DL Rson, RL Erikson, and DL Simmons (1991). "Expression of a mitogen-responsive gene encoding prostaglandin synthase is regulated by mRNA splicing". In: *Proc. Natl. Acad. Sci. USA* 88.7, pp. 2692–2696 (cit. on p. 159).
- Yap, CW (May 2011). "PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints". In: *J. Comput. Chem.* 32.7, pp. 1466–1474 (cit. on p. 162).

- Yu, Y, E Ricciotti, R Scalia, SY Tang, G Grant, Z Yu, G Landesberg, I Crichton, W Wu, E Pure, CD Funk, and GA FitzGerald (2012). "Vascular COX-2 modulates blood pressure and thrombosis in mice". In: *Sci. Transl. Med.* 4.132, 132ra54–132ra54 (cit. on pp. [159](#), [160](#)).
- Zhang, S, X Zhang, XW Ding, RK Yang, SL Huang, F Kastelein, M Bruno, XJ Yu, D Zhou, and XP Zou (2014). "Cyclooxygenase inhibitors use is associated with reduced risk of esophageal adenocarcinoma in patients with Barretts esophagus: a meta-analysis". In: *Br. J. Cancer* (cit. on p. [160](#)).

Large-scale Cancer Cell-Line Sensitivity Prediction

.6 Large-scale prediction of growth inhibition patterns on the NCI60 cancer cell-line panel

.6.1 Introduction

Cultured cell-lines have proved versatile disease models for cancer drug discovery [Weinstein (2012)]. In the last decades, large-scale multi-omics initiatives have catalogued the somatic alterations of cancer cell-line panels coupled with their pharmacological response to thousands of compounds [Barretina et al. (2012); Garnett et al. (2012); Shoemaker (2006)]. The US National Cancer Institute (NCI) pioneered these efforts by assembling the NCI60 tumour cell-line panel, which, to date, has been assayed for their sensitivity to over 130,000 compounds and extensively profiled at the molecular level [Shoemaker (2006)]. Although these cell-line collections have proved valuable to identify genomic markers of drug sensitivity [Barretina et al. (2012); Garnett et al. (2012)] and to develop new drugs [Adams and Kauffman (2004)], the question now arises how these pharmacogenomic data can be meaningfully mined, both to discover cancer-specific drugs, but also to design personalized cancer treatments.

Previous computational modelling on the NCI60 panel include the identification of drug mechanism of action (MoA) [Weinstein et al. (1992)], visualization tools for drug sensitivity data [Paull et al. (1989); Weinstein et al. (1997)], and drug sensitivity prediction based on cell-line profiling data ([Kutalik, Beckmann, and Bergmann (2008); Riddick et al. (2011); Staunton et al. (2001); Szakacs et al. (2004)]. Beyond algorithmic differences, the conceptual limitation shared by these models was the unfeasibility of extrapolating to novel compounds and cell-lines simultaneously, as cell-line profiling data or growth inhibition patterns were separately used as predictive features. Hence these models were able to do one of the two depending on the information they were trained on. To overcome this limitation, two recent studies have pioneered the combination of drug and cell-line information (gene expression, gene-copy number variation and mutation profiles) on the data from the Genomics of Drug Sensitivity in Cancer project for drug sensitivity prediction [Garnett et al. (2012)]. Menden et al. (2013) modelled the sensitivity of 608 cell-lines to 131 drugs with neural networks and

Random Forest models, obtaining a R^2 value on a blind test of 0.64. Ammad-ud-din et al. (2014) applied kernelized Bayesian matrix factorization to model the sensitivity of 650 cell-lines to 116 drugs ($R^2 = 0.78$). The authors showed that the combination of chemical and cell-line information improved model performance, which permitted to interpolate drug activities in order to complete the missing entries of a cell-line-drug interaction matrix (482 cell-lines x 116 drugs).

Here, we propose the simultaneous modelling of chemical and cell-line information in single machine learning models to predict with error bars the growth inhibition 50% bioassay end-point (GI_{50}) of 17,142 compounds screened against 59 cancer cell-lines from the NCI60 panel. The integration of these different, yet complementary, streams of information is often termed Proteochemometrics (PCM) or pharmacogenomic modelling (PGM) [Cortes-Ciriano et al. (2015a); Wheeler et al. (2013)]. In typical PCM models, although other approaches exist [Jacob and Vert (2008); Yamanishi et al. (2010)], each compound-cell-line interaction is numerically encoded by the concatenation of compound and cell-line descriptors, which are related in single machine-learning models to a specific biological readout of interest [Cortes-Ciriano et al. (2015a); Westen et al. (2011)] Thus, PCM helps in understanding complex relationships, such as compound selectivity towards a given cancer cell-line, and enables the estimation of the bioactivity for (novel) compounds on (novel) cell-lines (Figure .6.1).

We downloaded and curated cell-line profiling data consisting of 59 cell-lines from which we assembled 16 profiling datasets, denoted here as data set views. We benchmark their predictive signal, and demonstrate that the simultaneous modelling of compound and cell-line information enables the prediction of compound potency and cell-line selectivity. Unlike previous methods, our models interpolate and extrapolate compound bioactivities to novel cell-lines and tissues on the NCI60 panel, and to chemically dissimilar compounds. Finally, we demonstrate that the predicted bioactivities can be used to predict growth inhibition patterns across the NCI60 panel and that significant drug-pathway associations are consistent with the experimental data published in the literature.

.6.2 Materials and Methods

.6.2.1 Data sets

Raw pGI_{50} values ($-\log_{10} GI_{50}$), compound concentration necessary to reduce cell growth by 50%, were downloaded from CellMiner (Database version 1.4) [Reinhold et al. (2012)]. The mean value was calculated when several measurements were available for the same compound-cell-line combination. The standard deviation of these

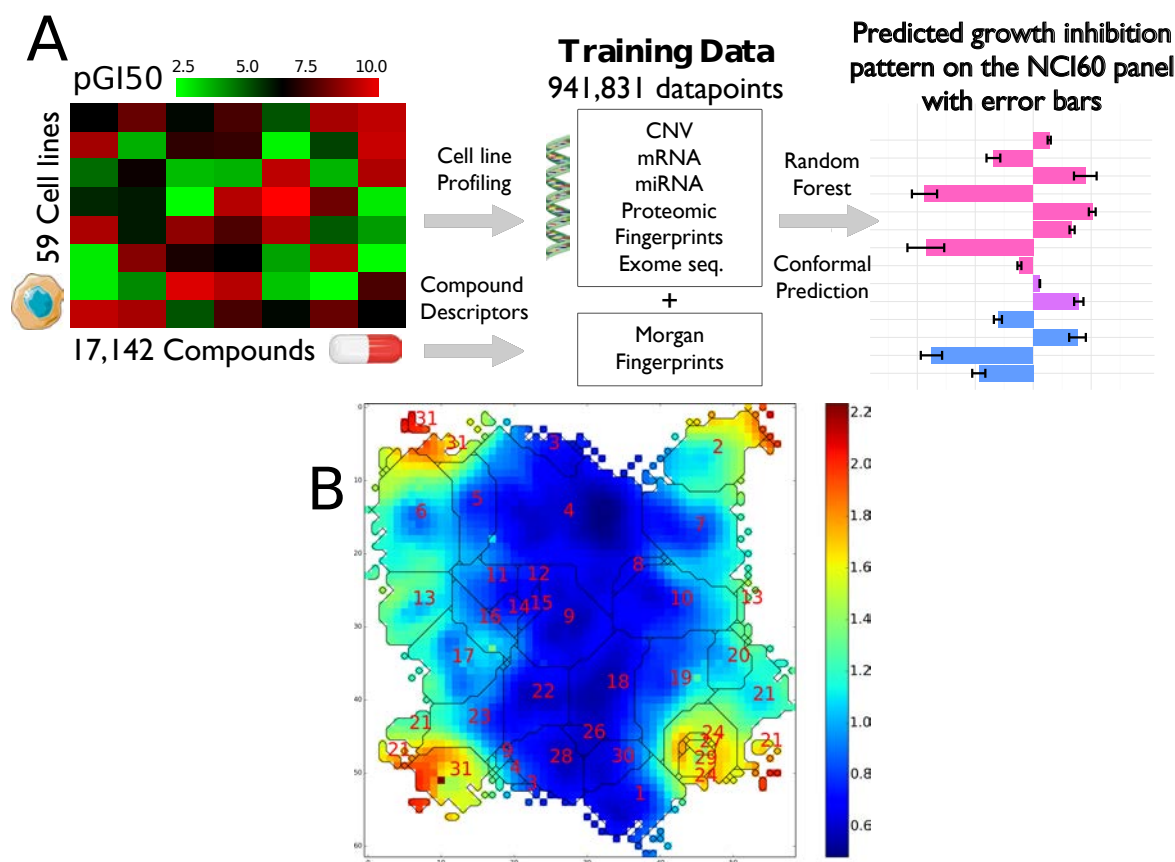


Figure .6.1: **Modelling workflow and compound clustering.** A. pGI₅₀ values for 17,142 compounds on 59 cancer cell-lines (941,831 data-points) were modeled with PCM Random Forests and conformal prediction. B. U-matrix for the SOM used to cluster the compounds. Black lines delimit the 31 clusters defined, whereas red labels indicate the cluster number. The similarity between each neuron and its 8 neighboring neurons defines the color code: blue corresponds to high similarity (homogeneous areas), and red corresponds to low similarity (heterogeneous areas). Therefore, clusters presenting blue and red neurons exhibit higher levels of intra-cluster chemical diversity.

replicates was considered as the experimental uncertainty. Chemical structures were sketched with the same stylistic convention with the function *StandardiseMolecules* of the R package *camb* using the default values [Murrell et al. (2014)]. The final data set consisted of 941,831 data-points, 17,142 compounds and 59 cell-lines (the NCI60 panel except ME.MDA_N), which corresponds to a matrix completeness of 93.08%. Table .6.1 summarizes the profiling data set views used to describe the cell-lines.

.6.2.2 Compound descriptors

Compounds were described with circular Morgan fingerprints in count format. The size of the fingerprints was set to 256 bits, whereas the maximum radius of the substructures considered was set to 2 bonds.

.6.2.3 Compound clustering

Compounds were clustered with periodic two-dimensional Self-Organizing Maps (SOMs) [Bouvier et al. (2014)]. A 2D SOM is defined by a 3D matrix. Two dimensions, here 50x50, determine the map size and were chosen to be periodic, whereas the third dimension contained the compound fingerprints. Each of the vectors along the third dimension is called a neuron, v . The same fingerprints used to train the PCM models served as input vectors to the SOMs. SOM values were initialized from a uniform distribution spanning the values present in the input vectors. At each training step, the most similar neuron to the input vector considered, *i.e.* the Best Matching Unit (BMU), was updated. To delineate the clusters, the conventional Unified distance matrix (U-matrix) was calculated. The U-matrix value associated to a given neuron, $U_{\text{height}}(v)$, is defined as the average Euclidean distance between that neuron and its eight closest neighbours:

$$U_{\text{height}}(v) = \frac{1}{8} \sum_{\mu \in Nv} E_d(v, \mu) \quad (.6.1)$$

where $N(v)$ is the set of neighbors and $E_d(v, \mu)$ the Euclidean distance between neurons. A distance threshold value was then applied to the U-matrix in order to define the contours of the compound clusters [Bouvier et al. (ibid.)].

.6.2.4 Model generation

Random Forest (RF) models [Breiman (2001)] were trained with the module *ensemble.RandomForestRegressor* of the python library scikit-learn [Pedregosa et al. (2011)], using the following values for the parameters:

- Number of trees in the forest: 100 [Sheridan (2013)].
- Criterion to assess the quality of a split: mean squared error.
- Minimum number of data points to split a node: 1.
- Minimum number of data points in a leaf to keep a given node split: 1.
- Maximum number of randomly selected descriptors considered when splitting a node: dimensionality of the input space.

Original profiling data set	Abbreviated data set view name	Details
Cell-line fingerprints [Lorenzi et al. (2009)]	Cell Fingerprints	Number of short tandem repeats at 16 genomic loci.
DNA copy-number variation [Varma et al. (2014)]	CNV	CNV for the 967 genes exhibiting at least two mutations in the NCI60 panel. DNA gain ($>3N$, $\log_2 = 0.58$) was encoded as 1, DNA losses ($<1N$, $\log_2 = -1$) as -1, and the rest ($2N$) with 0.
DNA copy-number variation [Varma et al. (2014)]	CNV Onc. & T. Suppre.	CNV for oncogenes and tumour suppressors.
Global proteome expression [Gholami et al. (2014)]	Cor. Proteome	Spearman's r_s matrix (59 x 59) between the expression levels of 8,113 proteins.
mRNA [Reinhold et al. (2010)]	G.t.l ABC	Transcript levels (\log_2) of 47 ABC transporters.
mRNA [Reinhold et al. (2010)]	G.t.l Onc. & T. Suppre.	Transcript levels (\log_2) of (i) oncogenes, and (ii) tumour suppressors.
mRNA [Reinhold et al. (2010)]	G.t.l Kin.	Transcript levels (\log_2) of 402 human kinases.
mRNA [Reinhold et al. (2010)]	G.t.l 1000 genes	Transcript levels (\log_2) of the 1,000 genes displaying the highest variance across the NCI60 panel.
mRNA [Reinhold et al. (2010)]	G.t.l 1000 pathways	Average transcript levels (\log_2) of the 1,000 pathways displaying the highest variability among the NCI60 panel.
mRNA [Reinhold et al. (2010)]	G.t.l 1000 genes & Kin. & Onco. & T. Suppre.	Transcript levels (\log_2) of (i) the 1,000 genes displaying the highest variability among the NCI60 panel, (ii) the human kinome, (iii) oncogenes, and (iv) tumour suppressors.
mRNA [Reinhold et al. (2010)]	G.t.l Kin. & Onco. & T. Suppre.	Transcript levels (\log_2) of (i) the human kinome, (ii) oncogenes, and (iii) tumour suppressors.
mRNA [Reinhold et al. (2010)]	Cor. Transcriptome	Spearman's r_s matrix (59 x 59) between the transcript levels (\log_2) of 19,965 genes for all cell-line pairs.
miRNA [Reinhold et al. (2010)]	miRNA	Expression (\log_2) of 627 miRNAs.
Reverse-phase lysate arrays [Nishizuka et al. (2003)]	RPLA	Normalized protein abundance levels (\log_2) for 89 proteins.
Whole exome sequencing [Abaan et al. (2013)]	Exome	Mutation status (1: mutated, 0: non mutated) of 112 Type II variants predicted to be deleterious (Polyphen score higher than 0.85)
Whole exome sequencing & DNA copy-number variation	Exome & CNV	Concatenation of data set views exome seq. and CNV.

Table .6.1: **Description of the data set views benchmarked for compound sensitivity prediction on the NCI60 panel.** The abbreviated names used in Figure .6.5 are indicated in the second column. Prior biological knowledge, such as pathway information, was included in some data set views, whereas the gene transcript levels and mutational status for genes implicated in cancer, kinases and ABC transporters were gathered independently and combined in data set views to assess the redundancy of their predictive signal.

Random forests [Breiman (2001)] were chosen to train all models because of (i) the smaller training times required if compared to kernel methods, and because of (ii) the robustness of their performance with respect to the value of their parameters [Sheridan (2013)]. Therefore RF do not require to perform grid search and cross-validation to determine the best values for the hyperparameters. All calculations were conducted in a machine with 16 Intel® Xeon® processors E5-2670 and a total memory of 256 GB. Training times on the complete data set ranged between 6-8 hours.

Prior to model training, the data set was randomly divided into (i) a training set comprising 90% of the data-points, and (ii) a test set comprising the remaining 10% of the data. This process was repeated ten times, each time holding-out a different subset of the data, which enabled the generation of predicted values for all data-points. The predictive power of the models was assessed on the test set according to the $RMSE_{test}$ and R_0^2 test values [Golbraikh and Tropsha (2002a,b); Tropsha and Golbraikh (2007)] and section.2.6.

.6.2.5 Model validation

To quantify the contribution of the target and the chemical spaces to model learning, the following strategies were explored [Brown et al. (2014)]:

- Family Quantitative Structure-Activity Relationship (QSAR F): models were trained on all data-points in the data set using exclusively compound fingerprints as input features. A QSAR F model learns on the 59 pGI_{50} values, and predicts the average likelihood for a compound of being cytotoxic. In this way, a QSAR F model serves to assess whether the explicit inclusion of cell-line information improves the prediction of compound activity for those compounds exhibiting variable growth inhibition profiles across the cell-line panel. If a compound is not selective against particular cell-lines, and thus displays a comparable activity value across the cell-line panel, a QSAR F model would suffice to predict the average likelihood of that molecule to be active against any cell-line, *i.e.* cytotoxic. However, in the case of a compound displaying selectivity towards particular cell lines, *i.e.* being active against particular cancer cell-lines and inactive against others, a QSAR F model would fail to predict the activities of that compound across the cell-line panel, as compound activity would depend to a large extent on the biological side and not much on the chemical side (accounted by the compound descriptors).
- Family Quantitative Cell-Line-Activity Modelling (QCAM F): these models were trained on all data-points in the data set using exclusively cell-line descriptors as input features. This validation scheme assesses whether compound bioactivities are correlated on a given cell-line, *i.e.* a diverse compound set displays the same

activity on a given cell line. Therefore, high predictive ability of a **Quantitative Cell-Line-Activity Modelling (QCAM)** F model indicates that bioactivity prediction depends to a large extent on the cell-line, and to a much lesser extent on the compound structures. In that case, the inclusion of compound descriptors would not provide any predictive signal.

- Individual **QSAR** models per cell-line: one **QSAR** model per cell-line was trained exclusively on compound descriptors. In this case, the comparison was made with the cluster-averaged interpolation power of **PCM** to evaluate whether the integration of compound and cell-line information leads to higher predictive ability with respect to per cell-line **QSAR** models.
- Inductive Transfer (**IT**) **PCM**: the idea underlying **IT** is that the knowledge acquired in a given task, *e.g.* the prediction of compound activity on a given cell-line, is used to solve similar problems, *e.g.* to predict the activity of the same compound set on a new cell-line. In **IT**, two sources of information were input to the model, namely: (i) compound descriptors, and (ii) **CLIFP**. **CLIFP** are binary descriptors, of length equal to the number of different cell-lines considered, where each bit position corresponds to one cell-line. To describe a given cell-line, all bits were set to zero except for the bit corresponding to that cell-line. Therefore, cell-lines are located in a high dimensional space where they are equidistant. Formally, **CLIFP** are defined as:

$$\text{CLIFP}(i, j) = \delta(i - j)(i, j \in 1, \dots, N_{\text{cells}}) \quad (.6.2)$$

where δ is the Kronecker delta function and N_{cells} the number of distinct cell-lines. This setting can also be regarded as a multi-task learning approach [Brown et al. ([ibid.](#))].

- Explicit Learning (**EL**) **PCM**: the models were trained on (i) compound descriptors, and on (ii) one cell-line profiling data set view. Compound and cell-line descriptors were horizontally stacked prior to model training.

Additionally, the extrapolation power of the methods was assessed using the following scenarios [Cortes-Ciriano et al. ([2015a](#))].

- Leave-One-Cell-Line-Out (**LOCO**): all data-points annotated on a given cell-line were held out from the training set, whereas a **PCM** model was trained on the remaining data. Subsequently, the values for the test set were predicted. $\text{RMSE}_{\text{test}}$ and R_0^2 test values were then calculated for the predicted bioactivities with respect to the observed ones. The previous steps were repeated each time holding-out the data-points corresponding to a different cell-line. This procedure intends to describe the situation where a **PCM** model is challenged

to extrapolate to novel cell-lines, although cell-lines originated from the same tissue might be present in the training set.

- **Leave-One-Tissue-Out (LOTO):** PCM models were further challenged to extrapolate to cell-lines which tissue of origin was not present in the training set. This scheme is similar to LOCO, except for the fact that all cancer cell-lines originated from the same tissue were held out from the training set at each time.
- **Leave-One-Compound-Cluster-Out (LOCCO):** all data-points annotated on a given chemical cluster were held out from the training set, whereas a PCM model was trained on the remaining data. This data availability scenario reflects the situation where a model is challenged to predict the bioactivity for dissimilar compounds, and thus permits the assessment of the extrapolation capabilities of PCM on the chemical space.

Finally, the performance of all models was evaluated on a per cell-line and on a per compound cluster basis. To this aim, RMSE and R_0^2 values were calculated on subsets of the test set grouped by cell-line (cell-line-averaged performance) or by compound cluster (compound cluster-averaged performance). The maximum and minimum achievable performance was evaluated as described in .2.7. For this calculation, we defined experimental uncertainty as the standard deviation of replicate pGI₅₀ measurements. In cases where no experimental uncertainty was available for a data-point, the mean of the available replicate-averaged experimental uncertainties, namely 0.272 pGI₅₀ unit, was used.

.6.2.6 Conformal prediction

Conformal prediction was applied to calculate confidence intervals for individual predictions (section .2.8). To calculate and validate the predicted intervals of confidence, the following pipeline was implemented [Norinder et al. (2014)] and section .2.8. Firstly, the whole data set was divided into an external set (20% of the data), and a training set (80%). The latter was subsequently split into a calibration set (30%) and a proper training set (70%). Two models were trained on the proper training set, the first of which predicted pGI₅₀ values (point prediction model), whereas the second predicted errors in prediction (error model). Both models were trained with compound fingerprints and the 'G.t.l 1,000 genes' data set view as input features. The point prediction model was generated by training a RF model on the proper training set with 10-fold cross-validation, and with pGI₅₀ values as the dependent variable.

.6.2.7 Pathway-drug associations

The average of the log₂ gene transcript levels for the genes composing each pathway in the MSigSB C2 Canonical Pathways gene set [Liberzon et al. (2011)] was taken as

the expression level of each of these pathways. To assess the association between drug response (pGI_{50}) and the expression of a given pathway, we fitted a linear model controlled by tissue source (*i.e.* using the tissue of origin as a blocking factor) [Haibe-Kains et al. (2013)], defined as:

$$pGI_{50} = \beta_p P_i + \beta_T T_i + \epsilon \quad (.6.3)$$

where P_i and T_i correspond to the expression of pathway i in a given cell-line or tissue, respectively, and ϵ to the error term. The significance of pathway-drug associations was estimated by the statistical significance of β_p (two-sided t-test, α 0.05).

.6.2.8 Comparison to previous methods

To compare our modeling approach to previous studies, we applied PGM to two additional datasets, namely the CCLE and GDSC. The metrics used to evaluate model performance were RMSE and R^2 as these were the metrics reported by those.

Preparation of the GDSC data set

MAS5-normalized gene transcript levels, measured with HT-HGU133A Affymetrix whole genome array, were downloaded from the GDSC website (<http://www.cancerrxgene.org/>) on February 16th 2015. Compound IC₅₀ values were converted to \log_{10} (IC₅₀ μ M) in order to enable the comparison of our results with previous studies [Ammad-ud-din et al. (2014); Menden et al. (2013)]. In addition, we converted the IC₅₀ values to pIC₅₀ values, *i.e.* $-\log_{10}$ (IC₅₀ M). To describe the cell-lines we selected the transcript levels of the 1,000 genes displaying the highest variance across the cell-line panel. 10-fold cross-validation was used to assess the interpolation power of the models. To assess extrapolation on the cell-line space, we used Leave-One-Tissue-Out (LOTO) validation, whereas Leave-One-Compound-Out validation was used to assess the predictive power on new molecules. We used Leave-One-Compound-Out instead of Leave-One-Compound-Cluster-Out validation given that the total number of distinct compounds was not high, namely 139, and thus permitted to assess the extrapolation power on a per compound basis. All models were trained using: (i) 256-bit hashed Morgan fingerprints in count format using a maximum substructure radius of 2 bonds, and (ii) transcript levels for the 1,000 genes displaying the highest variance across the cell-line panel.

We note in particular that in Ammad-ud-din et al. (2014) the extrapolation power of the models to new chemical structures was assessed by randomly dividing the compounds in 8 sets. A model was trained on 7 sets and this model was then used to predict the bioactivities for the held-out set. This process was repeated 8 times,

each time holding out a different set. In this setting, which is similar to LOCCO except for the fact that compounds are not grouped based on a similarity clustering, it is likely that the distribution of IC₅₀ values for a given set spans a wide range of values, thus permitting to obtain high R² values for the observed against the predicted bioactivities (Figure .2.3). By contrast, the range of IC₅₀ values is likely to be much narrower for individual compounds across the cell-line panel. Therefore, the R² values obtained with Leave-One-Compound-Out validation with PGM models are likely to be smaller than those obtained with LOCCO for the same accuracy in prediction, quantified with the RMSE value for the observed against the predicted bioactivities. From this, it is important to note that although the R² values reported by Ammad-ud-din et al. (2014) when assessing the extrapolation power of the models on new molecules (compound sets in their case), namely 0.52 +/- 0.37, might be higher in some cases than those obtained with Leave-One-Compound-Out validation, this does not necessarily mean higher predictive power (Figure .2.3). Therefore, the comparison between the two studies should be done in terms of RMSE values. We note in particular that we did not apply the same validation as Ammad-ud-din et al. (*ibid.*), namely partitioning the data set in 8 compound sets, as the composition of the 8 different sets was not reported by the authors.

Preparation of the CCLE data set

Gene transcript levels (Affymetrix U133+2 arrays), RMA-processed and normalized using quantile normalization, and compound IC₅₀ values (μM) were downloaded from the CCLE website (<https://www.broadinstitute.org/ccle/home>) on February 16th 2015. IC₅₀ values were converted to pIC₅₀ values, *i.e.* $-\log_{10}(\text{IC}_{50} \text{ M})$, and to $\ln(\text{IC}_{50} \text{ μM})$. We used the natural logarithm (\ln) instead of the logarithm with base 10, as in the GDSC data set, given that the range of values was higher for $\ln(\text{IC}_{50} \text{ μM})$ than for the $\log_{10}(\text{IC}_{50} \text{ μM})$ values. To describe the cell-lines we selected the transcript levels of the 1,000 genes displaying the highest variance across the cell-line panel. The same learning strategies applied to the GDSC data set were applied here, namely: 10-fold cross-validation, LOTO and Leave-One-Compound-Out. All models were trained using: (i) 256-bit hashed Morgan fingerprints in count format using a maximum substructure radius of 2 bonds, and (ii) the transcript levels for the 1,000 genes displaying the highest variance across the cell-line panel. Previous studies have not integrated chemical and cell-line data to predict compound IC₅₀ values using the CCLE data set. We provide the results obtained with these PGM models to allow the benchmarking of future predictive methods.

A word of caution

It is paramount to note that the application of predictive models trained on either

the CCLE or the GDSC should be restricted to those data sets, as models trained on one of these data sets is likely to fail on the other one. This has been previously indicated by Haibe-Kains et al. (2013):

"Ultimately, the poor correlation in these published studies [CCLE and GDSC] presents an obstacle to using the associated resources to build or validate predictive models of drug response. Because there is no clear concordance, predictive models of response developed using data from one study are almost guaranteed to fail when validated on data from another study, and there is no way with available data to determine which study is more accurate. This suggests that users of both data sets [CCLE and GDSC] should be cautious in their interpretation of results derived from their analyses."

.6.3 Results

.6.3.1 Summary of the cell-line profiling data set views

We collected seven profiling data sets for 59 cell-lines from the NCI60 panel, excluding ME.MDA-N due to the lack of gene transcript microarrays (Table .6.1). We then combined these molecular/phenotype data sets in a variety of different ways, what we term, in analogy to database views, data set views [Costello et al. (2014)]. We define data set view as: (i) a profiling data set, (ii) a subset thereof, e.g. gene transcript levels of gene sets, or (iii) as a modification of the data set to which prior knowledge is added, e.g. the calculation of pathway expression levels based on knowledge of cell signalling networks. A total of 16 data set views were defined, which are summarized in Table .6.1.

In addition to the complete data set, comprising all available data, namely: 17,142 distinct compounds and 941,831 data-points, we assembled two additional data sets: (i) variable bioactivity profile data set: comprising 3,641 distinct compounds (199,940 data-points) which bioactivity distribution on the 59 cell-lines exhibit standard deviations higher than 0.5 pGI₅₀ unit. It served to assess model performance on compounds displaying a dynamic range of bioactivities across the cell-line panel; and (ii) high confidence data set: comprising exclusively data-points averaged over at least two experiments (304,212 data-points and 5,302 distinct compounds), which served to evaluate whether model performance improves when using replicate-averaged drug sensitivity data.

Characterization of the chemical space

To assess the chemical diversity in the data, we clustered the 17,142 compounds with Self-organizing maps (SOM) (Figure .6.1B and subsection .6.2.3), which resulted in

the definition of 31 distinct chemical clusters. Several clusters, *e.g.* 4 and 18, are chemically homogeneous, as highlighted by the high inter-neuron similarity (blue areas in Figure .6.1B). By contrast, other clusters comprise more diverse compounds (shown in red in Figure .6.1B). For instance, cluster 2 is composed of polycyclic aromatic compounds with diverse halogen substituents and topologies. The definition of chemical clusters will help to challenge the inter- and extrapolation power of the models to dissimilar structures.

Model validation

PCM models trained on the complete data set, using compound fingerprints and the data set view "G.t.l 1,000 genes" as input features, exhibited respective mean $\text{RMSE}_{\text{test}}$ and R_0^2 test values of 0.40 ± 0.00 pGI₅₀ unit and 0.83 ± 0.00 ($n = 10$). These values were consistent with the theoretical maximum and minimum achievable performance since maximum and minimum mean $\text{RMSE}_{\text{test}}$ and R_0^2 test values of $1.42/0.35$ pGI₅₀ and $0.96/-0.96$ were respectively obtained with the simulated data (Figure .6.2). Moreover, model performance did not stem from chance correlations, as R_0^2 test values became negative when 75% of the bioactivities were randomized [Clark and Fox (2004)] (Figure .6.3).

Modelling the high confidence data set led to similar performance, with $\text{RMSE}_{\text{test}}$ and R_0^2 test values of 0.45 pGI₅₀ unit and 0.84 , respectively. This indicates that the predictive power does not decrease when including data-points measured in only one experiment. PCM was further challenged on the variable bioactivity profile data set, on which it exhibited $\text{RMSE}_{\text{test}}$ and R_{20} test values of 0.58 pGI₅₀ unit and 0.79 , respectively. These $\text{RMSE}_{\text{test}}$ and R_0^2 test values were also found in agreement with the maximum achievable performance for the most variable profile data set (Figure .6.4).

PCM outperformed models trained exclusively on cell-lines descriptors (QCAM F), but not the ones trained exclusively on compound descriptors (QSAR F). Indeed, QCAM F displayed poor predictive power, with $\text{RMSE}_{\text{test}}$ and R_0^2 test values of 0.95 pGI₅₀ unit and 0.02 respectively, whereas the much higher predictive ability of the QSAR F model on the complete data set with $\text{RMSE}_{\text{test}}$ and R_0^2 test values of 0.449 pGI₅₀ unit and 0.780 , respectively, indicates (compared to 0.40 pGI₅₀ unit and R_{20} test = 0.83 for PCM) that the bioactivities of identical compounds are correlated on the cell-line panel. Similarly, PCM displayed higher performance than QSAR F on the variable bioactivity profile data set, with respective $\text{RMSE}_{\text{test}}$ values of 0.58 versus 0.69 pGI₅₀ unit, highlighting that PCM is more suited for modelling compounds exhibiting uncorrelated bioactivities on a cell-line panel.

Moreover, PCM significantly outperformed individual cell-line models (two-sided

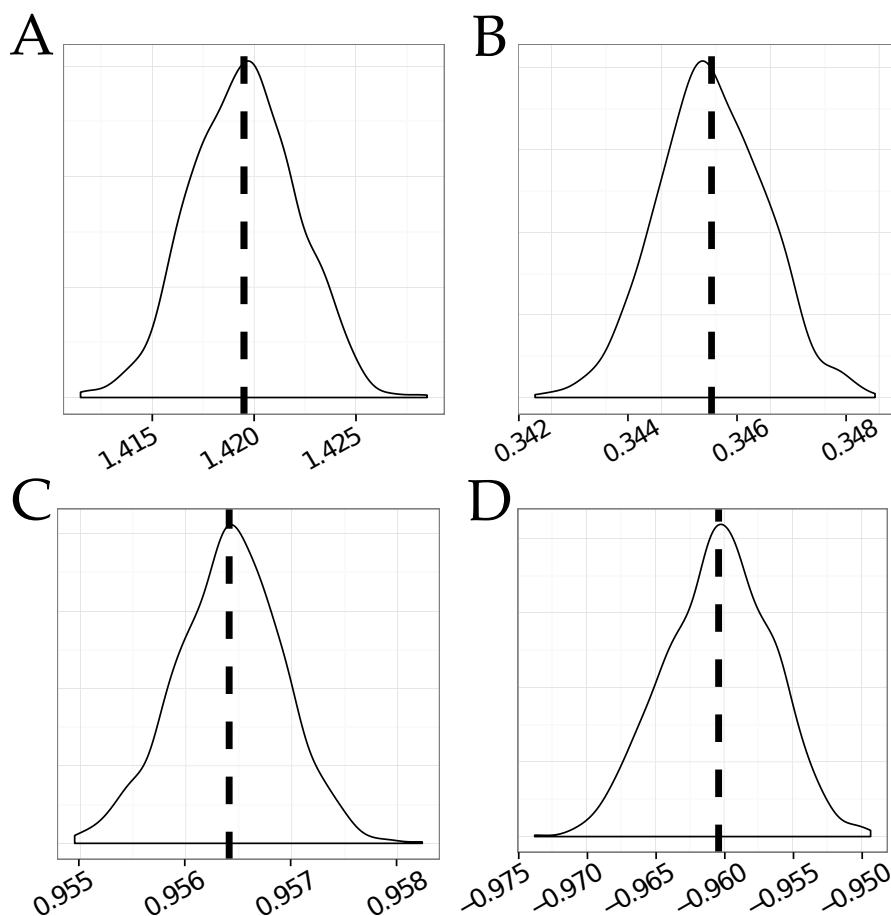


Figure .6.2: **Distribution of respective maximum and minimum $\text{RMSE}_{\text{test}}$ (A,B) and R_0^2 test (C,D) values for the complete data set.** Average maximum and minimum values of 1.42/0.35 and 0.96/-0.96, were obtained respectively for $\text{RMSE}_{\text{test}}$ / R_0^2 test with the simulated data. The performance of the **PCM** models on the test set was in agreement with the uncertainty of the experimental measurements, as mean $\text{RMSE}_{\text{test}}$ and R_0^2 test values of 0.40 ± 0.00 pGI_{50} unit and 0.83 ± 0.00 (with $n = 10$ models) were obtained. These values are between the two extreme, maximum and minimum, theoretical $\text{RMSE}_{\text{test}}$ and R_0^2 test values.

t-test, $\alpha 0.05$, $P < 0.05$), trained on the data-points corresponding to a given cell-line and using exclusively compound descriptors as input features. These individual models displayed an average $\text{RMSE}_{\text{test}}$ value of 0.73 ± 0.05 pGI_{50} unit, whereas the integration of information from different cell-lines performed by **PCM** improves drug sensitivity prediction with $\text{RMSE}_{\text{test}}$ values in the 0.40-0.58 pGI_{50} unit range.

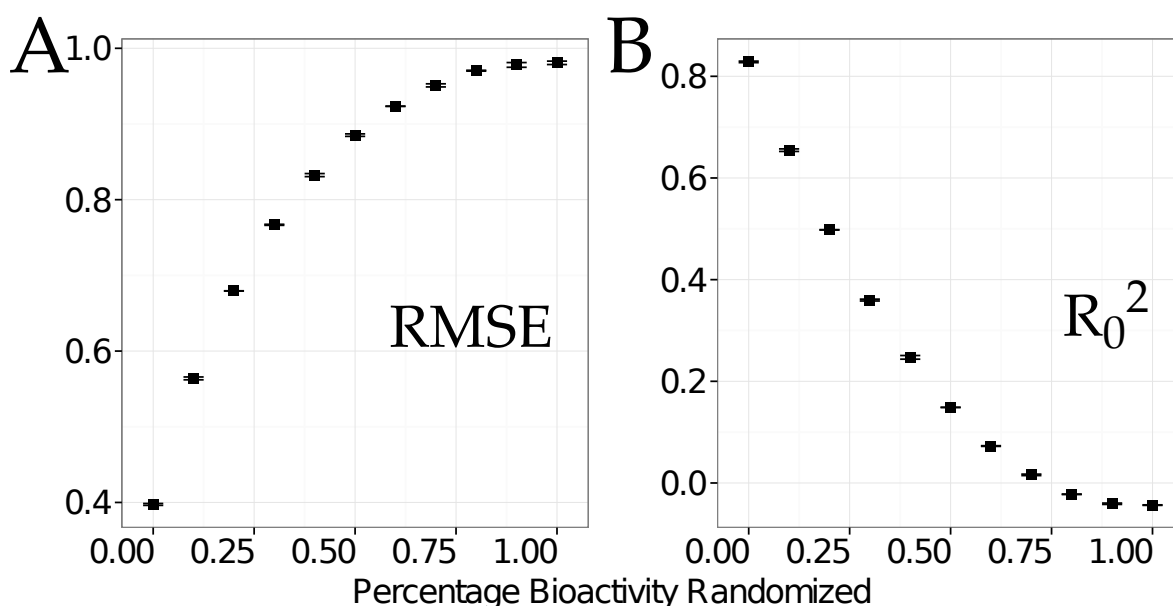


Figure .6.3: **Y-scrambling validation.** Mean RMSE_{test} (A) and R₀² test (B) values were calculated for the observed against the predicted bioactivities on the test set calculated with models trained on pGI₅₀ values increasingly randomized (n=3). R₀² test values become negative when 75% of the bioactivity values are randomized. These data suggest that the relationships established by the **PCM** models between compound and cell-line descriptors, and the pGI₅₀ values did not arise from chance correlations.

Benchmarking cell-line profiling data sets

In order to benchmark the predictive signal of the cell-line profiling data sets we used the variable bioactivity profile data set, as it contains the compounds displaying the less correlated bioactivities on the cell-line panel, and is thus more challenging to model, and crucially more likely to give specific insights into the underlying biology and mechanisms of action. For each data set view, we trained 10 models using that data set view and compound fingerprints as input features. This resulted in a total of 160 models (16 data set views x 10 replicates). An analysis of variance (**ANOVA**) on the RMSE_{test} values (Figure .6.5) yielded significant differences (P value < 1x10⁻¹⁷).

As stated previously [Costello et al. (2014)], we observed that gene transcript levels led to the highest predictive power, with median RMSE_{test} values in the 0.56-0.58 pGI₅₀ unit range (Figure .6.5A). However, the combination of transcript levels from different gene sets (Figure .6.5A) did not translate into increased performance,

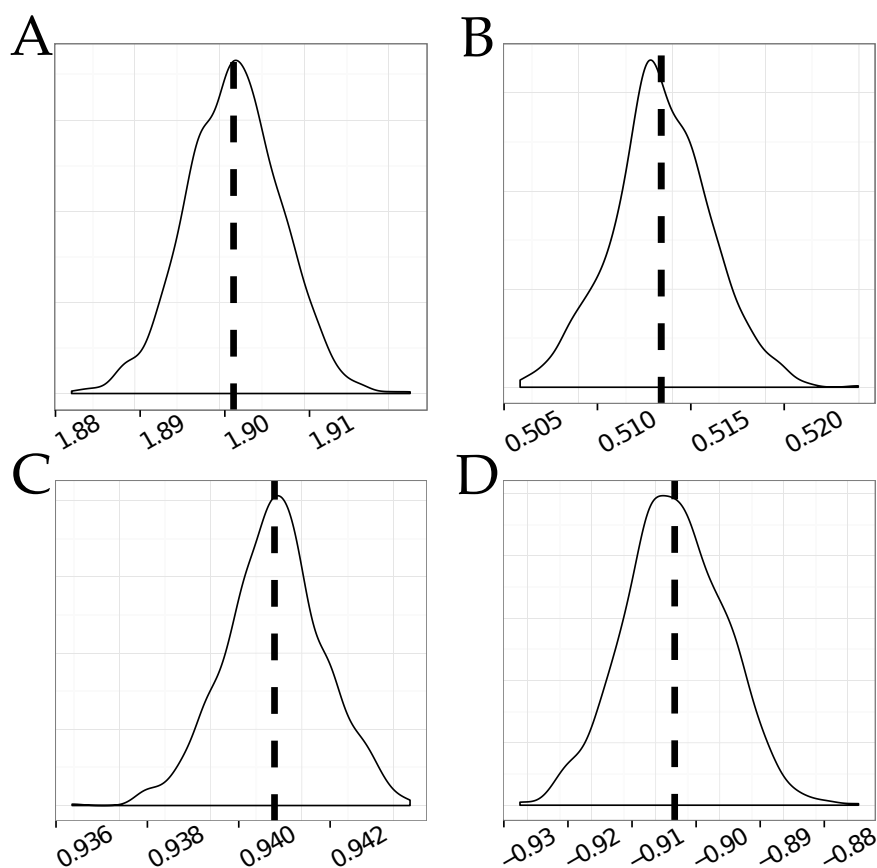


Figure .6.4: **Distribution of respective maximum and minimum $\text{RMSE}_{\text{test}}$ (A,B) and R_0^2 test (C,D) values for the variable bioactivity profile data set.** Average maximum and minimum values of 1.90/0.54 and 0.94/-0.90 were obtained respectively for $\text{RMSE}_{\text{test}}$ / R_0^2 test with the simulated data. The performance of [PCM](#) models was in agreement with the uncertainty of the experimental measurements, as mean $\text{RMSE}_{\text{test}}$ and R_0^2 test values of 0.580 pGI_{50} unit and 0.79 were obtained. These values are between the two extreme, maximum and minimum, theoretical $\text{RMSE}_{\text{test}}$ and R_0^2 test values.

suggesting that these data set views contain redundant biological information. Besides, no statistically significant differences in performance were observed between the models trained on gene set transcript levels and those trained on miRNA abundance or [Reverse Phase Protein Lysate Microarray \(RPLA\)](#) data (Tukey's Honestly Significance Difference ([HSD](#)), P value < 0.05).

Interestingly, the performance of the models trained on Copy Number Variation

(CNV) or exome sequencing information was significantly worse (Figure .6.5A), with $\text{RMSE}_{\text{test}}$ values in the 0.63-0.68 pGI_{50} unit range. This poorer performance was expected as the number of gene gain and losses corresponds to only 2.66% of the 58,020 possible cell-line-gene combinations in the descriptor matrix. Similarly, only 0.03% of the entries in the exome seq. descriptor matrix corresponded to mutations. Thus, the sparseness of these data is plausibly the reason for poor model performance.

We then tested whether PCM performance arises from Explicit Learning (EL) on cell-line descriptors or inductive transfer knowledge (IT) [Brown et al. (2014)] among cell-lines. In explicit learning, cell-lines descriptors would account for genomic differences of the cell-lines, and therefore their distance in descriptor space would be proportional to those differences, which are related to drug sensitivity. This would permit the models to explicitly learn the differences among the cell-lines. By contrast, if the descriptors do not account for genomic differences among cell-lines, they would simply act as labels, and model performance would arise from inductive transfer knowledge among cell-lines. PCM (explicit learning) outperformed IT models (Tukey's HSD, P value < 0.05) (Figure .6.5A), which highlights that the explicit inclusion of cell-line profiling data as input features improves compound sensitivity prediction.

Bioactivity interpolation to cell-lines and compound clusters in the training set

To assess the interpolation power of PCM models, we evaluated the cell-line-averaged (Figure .6.6A) and compound cluster-averaged performance (Figure .6.5B), by calculating the $\text{RMSE}_{\text{test}}$ values on subsets of the test set grouped by cell-line or compound cluster. If not otherwise indicated, the results presented in the following subsections were calculated using models trained on the variable bioactivity profile data set, with (i) compound fingerprints and (ii) the 'G.t.l. 1,000 genes' data set view as input features. Cell-line-averaged $\text{RMSE}_{\text{test}}$ values ranged from 0.41 \pm 0.01 (U251) to 0.86 \pm 0.01 pGI_{50} unit (HOP-92).

We found significant differences for tissue-averaged performance (Tukey's HSD, $P < 1 \times 10^{-16}$), with $\text{RMSE}_{\text{test}}$ values ranging from 0.48 \pm 0.01 (prostate: cyan) to 0.70 \pm 0.01 (leukemia: green) pGI_{50} unit (Figure .6.5A). Additionally, learning curves showed that $\text{RMSE}_{\text{test}}$ values of approximately twice both the replicate-averaged experimental uncertainty (0.272 pGI_{50} unit) and the maximum achievable performance (0.35 pGI_{50} unit) can be obtained when less than 10% of the data is used as training set (Figure .6.7). By contrast, an ANOVA analysis did not yield statistically significant ($P > 0.05$) differences among the 31 chemical clusters (Figure .6.6B), with observed median $\text{RMSE}_{\text{test}}$ values in the 0.48-0.65 pGI_{50} unit range.

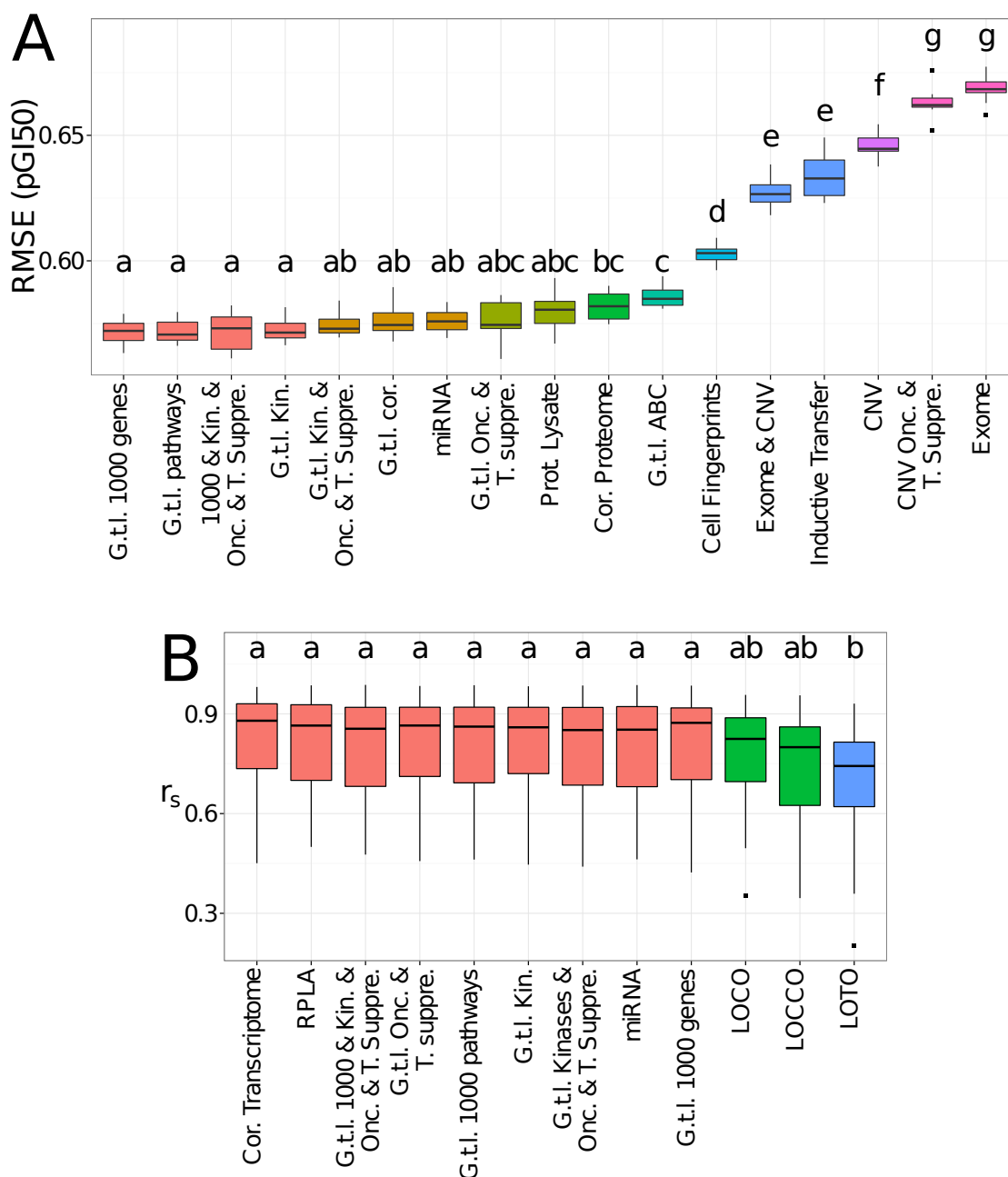


Figure .6.5: **Benchmarking of the cell-line profiling data set views for compound sensitivity prediction.** A. The predictive power of the 16 data set views (Table .6.1) was quantified by the RMSE values on the test set. For each data set view, we trained ten models on the variable bioactivity profile data set. We found significant differences among the data set views (ANOVA, $P < 0.01$). Post-hoc analyses (HSD, $\alpha 0.05$) were used to cluster the data set views according to their predictive power.

Figure .6.5: **Figure .6.5 caption continuation** Data set views sharing a letter label performed at the same level of statistical significance and are depicted in the same color. We consistently found that gene transcript levels, and the abundance of proteins and miRNA led to the most predictive models (labeled with *a*). B. The evaluation of both interpolation and extrapolation power was evaluated on the complete data set. After finding significant differences among groups (ANOVA, $P < 0.01$), we found that PCM interpolates and extrapolates to new cell-lines and tissues at the same level of statistical significance (Tukey's HSD, $\alpha 0.05$). By contrast, we found statistically significant differences in performance between extrapolation and interpolation to new chemical clusters.

Extrapolation to novel cell-lines and tissues

We further evaluated to which extent PCM extrapolates compound bioactivities to novel cell-lines and tissues with Leave-One-Cell-Line-Out (LOCO) and Leave-One-Tissue-Out (LOTO) validation using the complete data set. LOCO models exhibited mean $RMSE_{test}$ values of 0.43 ± 0.08 pGI_{50} unit on the complete data set (Figure .6.5B), with the lowest, 0.31, and the highest, 0.61, $RMSE_{test}$ values observed for cell-lines U251 and OVCAR-5, respectively. Notably, we found that LOCO and the cell-line-averaged interpolation performance are highly correlated (Spearman's Rank Correlation coefficient (r_s) = 0.92), indicating that the interpolation and the extrapolation to novel cell-lines are correlated. $RMSE_{test}$ values for LOTO models ranged between 0.35 (prostate) and 0.63 pGI_{50} unit (leukemia). Remarkably, prediction errors were similar across the entire bioactivity range (Figure .6.8).

Overall, we did not observe significant differences in performance among LOCO, LOTO (Figure .6.5B), cell-line-averaged (Figure .6.6A), and compound cluster-averaged (Figure .6.6B) results (Tukey's HSD, $P < 0.05$). We found that the RMSE value between the observed and predicted pGI_{50} values for 47 out of 81 drugs, such as Imiquimod (NSC 369100) and Bendamustine (NSC 138783), was below 0.5 pGI_{50} unit (Figure .6.9a,b). The highest RMSE values, between 1 and 1.5 pGI_{50} units, were found for 11 drugs, such as the folate antimetabolite pemetrexed (NSC 698037) and irinotecan (NSC 728073). Altogether, these data indicate that PCM models extrapolate compound bioactivities to novel cell-lines and tissues at the same level of statistical significance as for interpolation within a given cell-line or tissue.

Extrapolation to novel chemistry

A markedly different trend was observed for the ability of the models to generalize across the chemical space, assessed with Leave-One-Chemical-Cluster-Out (LOCCO)

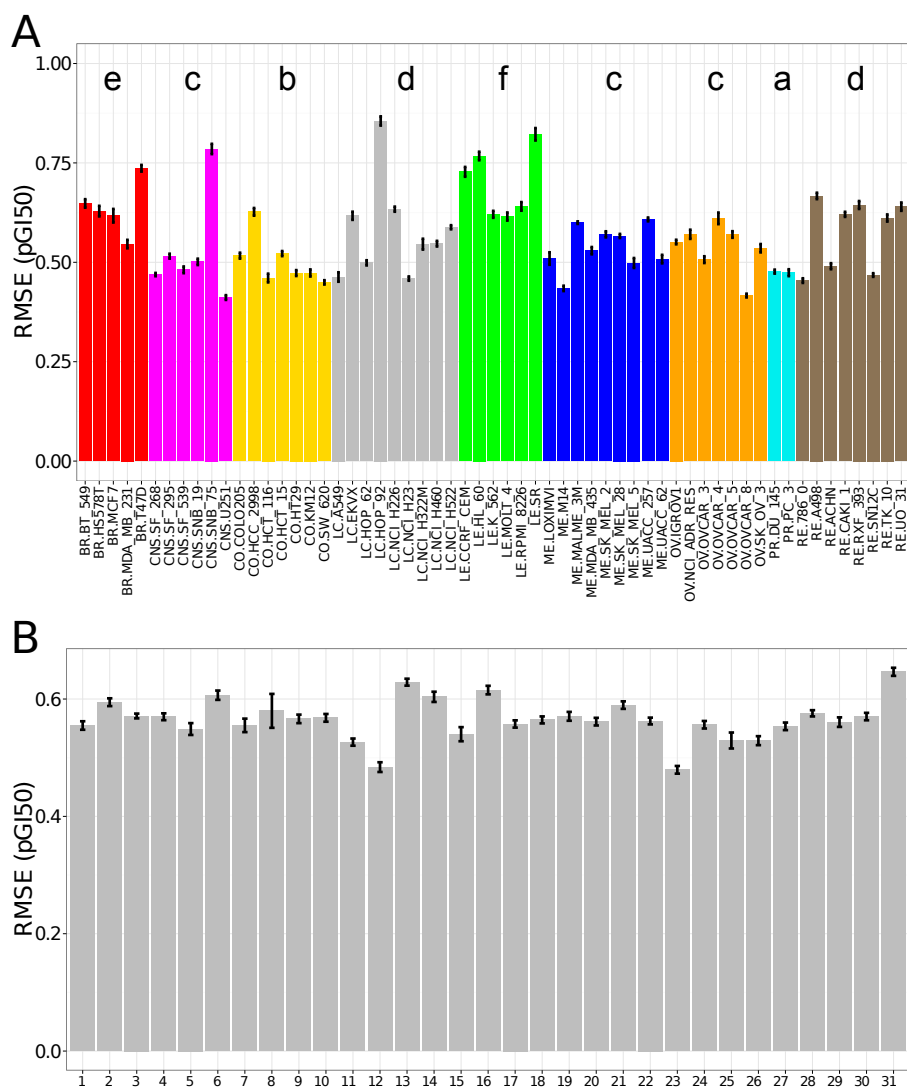


Figure .6.6: **Interpolating compound bioactivities to novel cell-lines, tissues, and chemical clusters.** A. Cell-line-averaged $\text{RMSE}_{\text{test}}$ values ranged from 0.41 ± 0.01 (U251) to 0.86 ± 0.01 pGI₅₀ unit (HOP-92). We found significant differences for tissue-averaged performance (Tukey's HSD, $P < 1 \times 10^{-16}$), with $\text{RMSE}_{\text{test}}$ values ranging from 0.48 ± 0.01 (prostate) to 0.70 ± 0.01 (leukemia) pGI₅₀ unit. Cell-lines originated from the same tissue are depicted in the same color (breast: red, central nervous system: magenta, colon: yellow, lung cancer: grey, leukemia: green, melanoma: blue, ovarian: orange, prostate: cyan, renal: brown). We did not observe significant differences in tissue-averaged performance for tissues labeled with the same letter. B. Compound-cluster averaged performance for the 31 clusters defined with SOMs.

Figure .6.6: **Figure .6.6 caption continuation** (B) One-way ANOVA among the 31 chemical clusters ($P > 0.05$), with compound cluster-averaged $\text{RMSE}_{\text{test}}$ values in the 0.48 ± 0.01 and 0.65 ± 0.01 pGI_{50} unit range. This analysis illustrates that the models do not constantly favor specific chemical clusters, thus making it possible to interpolate compound bioactivities across the chemical space covered by the data at the same level of statistical significance. By contrast, interpolating on the cell-line side depends significantly on the tissue source.

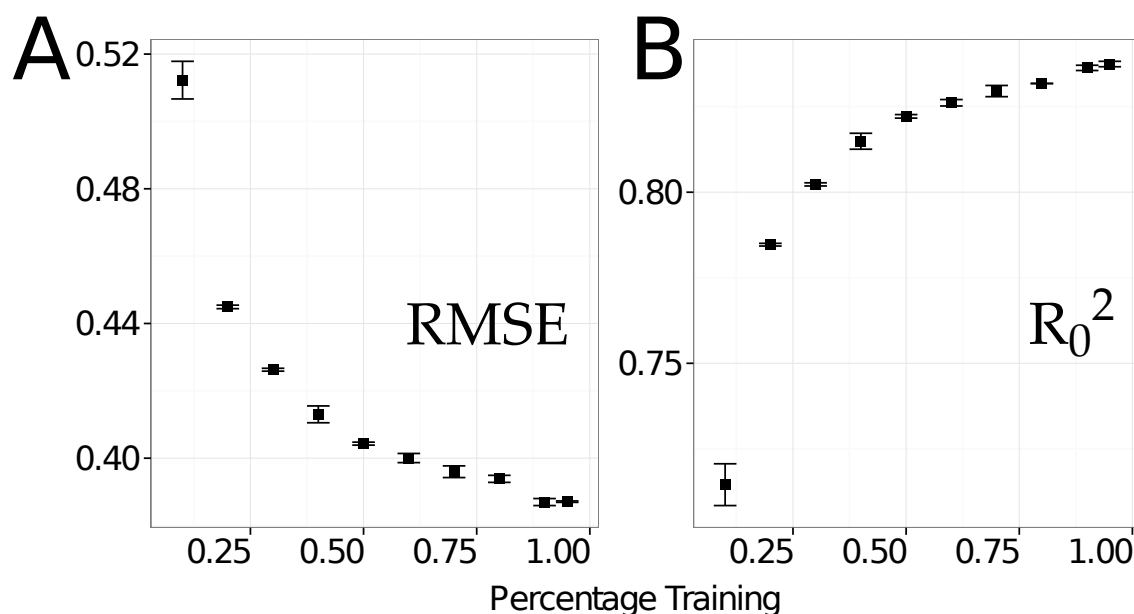


Figure .6.7: **Learning curves.** Mean $\text{RMSE}_{\text{test}}$ (A) and R_0^2 test (B) values were calculated for the observed against the predicted bioactivity values on the test set calculated with $n=3$ models obtained using training sets covering an increasingly higher fraction of the complete data set. Models trained on 5% of the data set exhibited a mean $\text{RMSE}_{\text{test}}$ value of 0.52 pGI_{50} unit, which decreased till 0.39 pGI_{50} unit when 95% of the data-points were included in the training set. These data suggest that PCM models exhibit high interpolation capabilities. In practice, the compound-cell-line interaction matrix could be completed with in silico predictions, with a $\text{RMSE}_{\text{test}}$ values of 0.39 pGI_{50} unit, without requiring further experimental testing.

validation using the complete data set. LOCCO models exhibited mean $\text{RMSE}_{\text{test}}$ values of 0.83 ± 0.17 pGI_{50} unit (Figure .6.5B), which differed significantly from

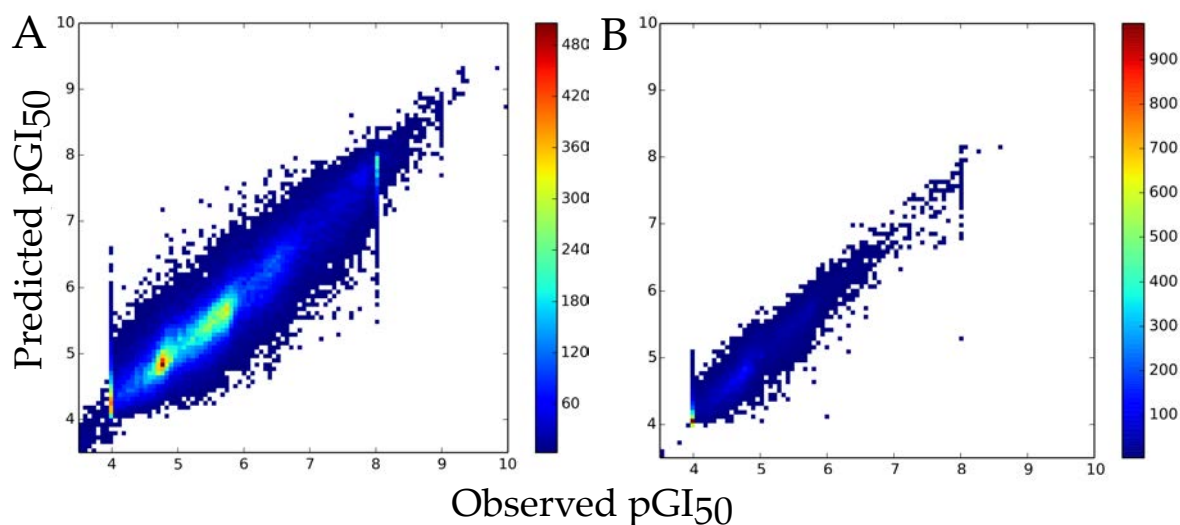


Figure .6.8: **Correlation between observed and predicted pGI₅₀ values.** Density correlation plot corresponding to the observed against predicted pGI₅₀ values on the test set for: (A) the **LOTO** model for melanoma ($RMSE_{test}$ and R_0^2 test values of 0.43 pGI₅₀ unit and 0.80), and (B) the **LOCO** model for the melanoma cell-line SK-MEL-5 ($RMSE_{test}$ and R_0^2 test values of 0.37 pGI₅₀ unit and 0.87). The color bar indicates the density of points at each region of the plot. For the rest of **LOCO** and **LOTO** models comparable results were obtained, with bioactivity values correctly predicted along the whole bioactivity range.

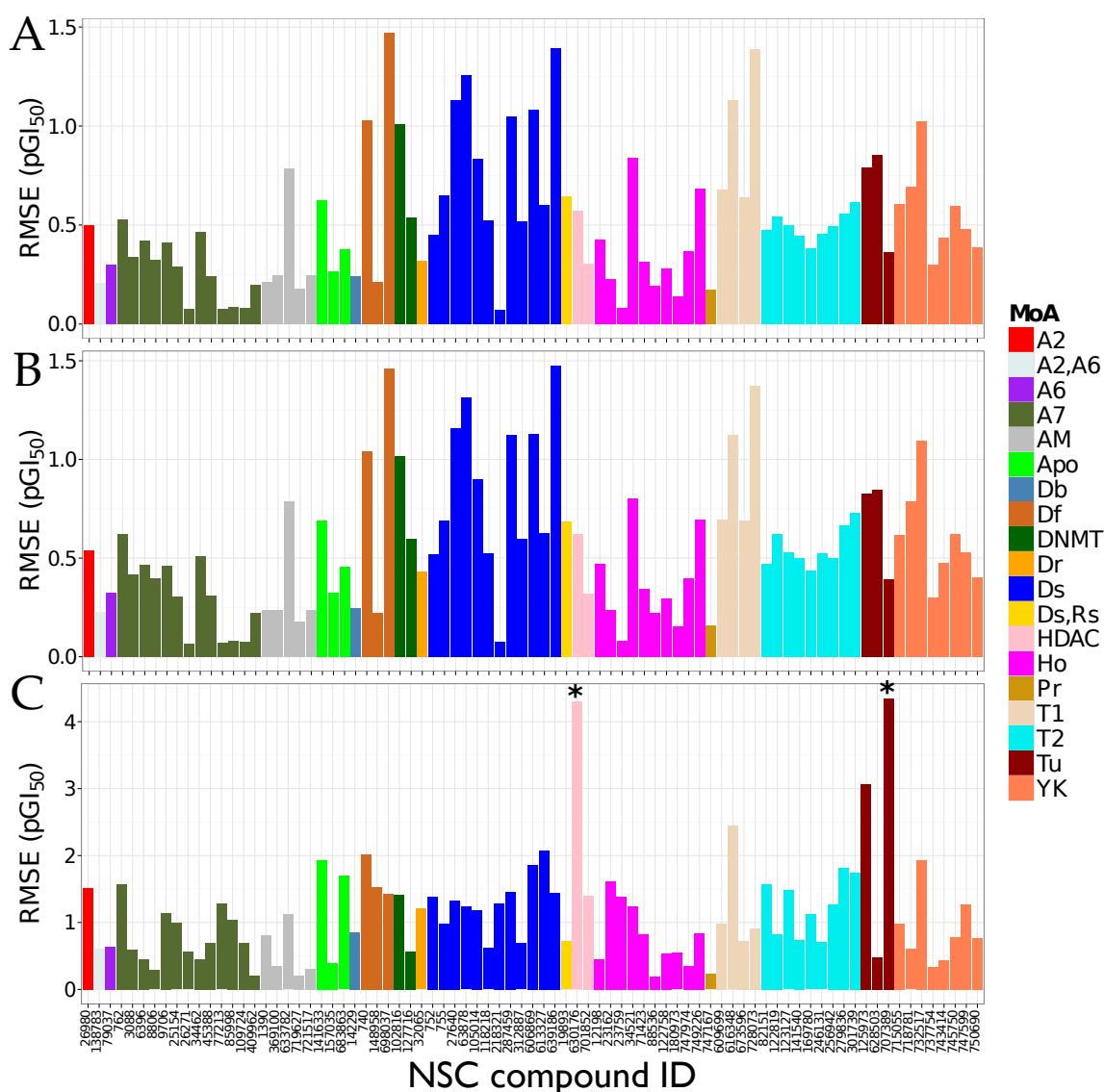


Figure 6.9: Correlation between observed and predicted pGI_{50} values for the 81 drugs present in the complete data set for the following model validation scenarios: (A) LOCO, (B) LOTO, and (C) LOCCO. The x-axis reports the drug NSC identifiers. Compounds discussed in the main text, namely NSC 630176 and NSC 707389, are marked with asterisks. Bars are colored according to drug mechanism of action (MoA). The abbreviations of the mechanisms of action are: A2: alkylating at N-2 position of guanine; A7: alkylating at N-7 position of guanine; AM: antimetabolite; Ang: angiogenesis; Apo: apoptosis inducer; Db: DNA binder; Df: antifolates; DNMT: DNA methyltransferase inhibitor; Dr: ribonucleotide reductase inhibitor; Ds: DNA synthesis inhibitor; HDAC: Histone deacetylase; Ho: hormone; P90: hsp90 binder; PI3K: PI3kinase; PKC: Protein kinase C; ROS: reactive oxygen species; RSTK: serine/threonine kinase inhibitor; T1: topoisomerase 1 inhibitor; T2: topoisomerase 2 inhibitor; Tu: tubulin-active antimitotic; YK: tyrosine kinase inhibitor.

the LOCO and LOTO results (Figure .6.5B) (Tukey's HSD, $P < 0.01$), and from the compound cluster-averaged interpolation performance (Figure .6.6B). Thus, extrapolating bioactivities to novel chemical clusters is more challenging than extrapolating to novel cell-lines and tissues. Notably, chemical diversity within compound clusters was not correlated with model performance, as low $\text{RMSE}_{\text{test}}$ values were consistently obtained for heterogeneous and homogeneous clusters (Figure .6.1B).

The lowest $\text{RMSE}_{\text{test}}$ value was obtained for cluster 24, namely 0.53 pGI_{50} unit, which contains 485 compounds presenting polycyclic ring systems, generally with no more than 3 rings fused, as well as ring assemblies linked by sulfide, sulfinyl, secondary amines, carbonyl and alkyl groups. Bigger molecules were found in cluster 16, which was modelled with the highest $\text{RMSE}_{\text{test}}$ value, namely 1.23 pGI_{50} units. Cluster 16 mainly contains molecules presenting tri- and tetracycles presenting hydroxybenzene, methoxybenzene and quinone rings in their structure. We obtained RMSE values below 0.5 pGI_{50} unit values for 15 out of 81 drugs and below 1 pGI_{50} unit for 43. The worst modelled drugs were depsipeptide (NSC 630176) and the halichondrin B analogue NSC 707389, with respective RMSE values of 4.29 and 4.35 pGI_{50} units. Taken together, these data indicate that it is possible to obtain low errors in prediction for structurally dissimilar drugs. However, the range of errors is considerably large ($> 4 \text{ pGI}_{50}$ units) and extrapolating to some compounds still remains a challenging task.

Conformal prediction provides informative confidence intervals

Conformal prediction was included in the modelling framework (section .2.8 and Figure .6.10) to provide confidence intervals (CI) for individual predictions. The CI, defined as the percentage of data-points for which the predicted value lied within different intervals of confidence, were found to be highly correlated with the size of the intervals (Spearman's $r_s > 0.99$) (.6.10). Thus, the combination of random forests and conformal prediction provides more information than individual machine learning models.

Consistency of pathway-drug associations with predicted bioactivities

To investigate whether the bioactivities predicted with PCM make it possible to identify genomic markers of drug sensitivity, we evaluated the consistency between the pathway-drug associations inferred from the experimental and from the predicted bioactivities for the 37 FDA-approved drugs and the 17 compounds in clinical trials present in the variable bioactivity profile data set. For each pathway, we fitted a linear model controlled by tissue source, where the average expression was considered as predictor of drug sensitivity.

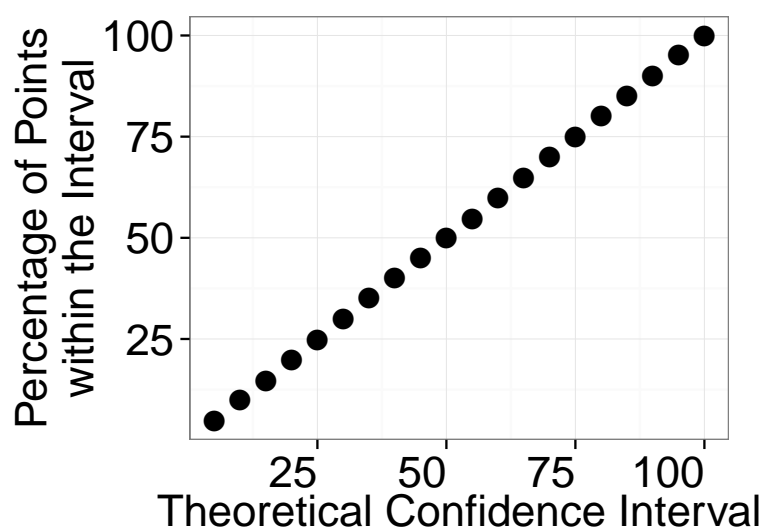


Figure .6.10: **Validation of conformal prediction.** For each confidence level (ϵ), represented in the x-axis, the number of data-points in the test set which true value lies within the predicted interval is calculated, y-axis. The high Spearman's r_s is likely due to the large size of the test set (188,366 data-points) and to the fact that the CI produced by conformal prediction are always valid [Norinder et al. (2014)]. These data indicate that the modeling framework combining PCM models and conformal prediction is more information rich than what would be possible with only point prediction algorithms.

Overall, no significant differences (Tukey's HSD, $P < 0.05$) were observed between pathway-drug associations calculated with the most predictive PCM models (Figure .6.11A) and the LOCO validation, as median Spearman's r_s values were in the 0.75-0.91 range (Figure .6.11B). Significant differences were however found among these groups and LOCCO and LOTO validation, for which median $\text{RMSE}_{\text{test}}$ values were respectively 0.63 and 0.03. We obtained similar results when considering only the pathways significantly associated to drug response (false discovery rate (FDR) $< 20\%$) (Figure .6.11C,D).

Next, we analysed whether pathway-drug associations are consistently predicted for drugs exhibiting different mechanisms of action (MoA). Out of the 56 drugs considered, 26 exhibited median Spearman's r_s values in the 0.5-0.75 range and 18 above 0.75 (Figure .6.11B). High Spearman's r_s values were obtained across the 22 distinct drug MoAs, thus indicating that no specific MoA is favoured. Notably, most Spearman's r_s values increased (Figure .6.11D) when the calculation of pathway-drug associations was restricted to the pathways significantly associated to drug response (FDR $< 20\%$). Together, we can conclude that the identification of genomic markers of drug sensitivity is significantly dependent on the presence of cell-lines originated from the same tissue and structurally similar compounds in the training set.

Prediction of growth inhibition patterns on the NCI60 panel

The experimental and predicted growth inhibition patterns determined from the experimental and predicted bioactivities with the most predictive models were fairly correlated (Figure .6.11E), with median Spearman's r_s values in the 0.53-0.58 range and higher than 0.5 for 32 out of the 56 drugs from the variable profiles data set (Figure .6.11F). LOCO, LOTO and LOCCO validation (Tukey's HSD, P value < 0.001) displayed nevertheless a marked decrease of the r_s values (Figure .6.11E) with respect to the most predictive models. Relative growth inhibition values on the NCI60 panel can be depicted in a bar plot with z-scores calculated on the predicted bioactivities.

Figure .6.12 depicts the observed and the predicted growth inhibition patterns for methotrexate (MTX), as its complex growth inhibition pattern renders this drug suitable for illustration. The predictions accounted in 55 out of 59 cases for the relative sensitivity of the cell-line. For instance, the six leukemia cell-lines (green turquoise) were predicted to be sensitive to MTX. Moreover, complex inhibition patterns for renal derived cell-lines (light magenta) were accounted by the predictions, as cell-lines TK-10, RXF-393 and A498 were predicted to be highly resistant to MTX, whereas the effect of MTX on sensitive cell-lines, namely UO-31, SN12C, CAKI-1 and ACHN, was also correctly predicted (Figure .6.12B). Taken together, these data indicate that the drug sensitivity predictions were able to account for complex patterns of cell-line growth inhibition.

Figure .6.11: **Figure .6.11 caption continuation** E. Data view-averaged Spearman's r_s coefficients for patterns of growth inhibition calculated with the experimental and the predicted values. F. Bar plot reporting the drug-averaged Spearman's r_s coefficients for the patterns of growth inhibition calculated with the observed and the predicted bioactivities. Data views sharing a letter label and color in (A,C,E) perform at the same level of statistical significance. Significance for the Spearman's r_s in (B,D,F) is represented with an asterisk if two-sided P value < 0.05, for the Spearman's r_s coefficients calculated with the predictions generated with a model trained on the 'G.t.l. 1,000 genes' data view. Bars in (B,D,F) are colored according to compound MoA. Abbreviations of mechanisms of action: MoA: Mechanism of action; A2: alkylating at N-2 position of guanine; A7: alkylating at N-7 position of guanine; AM: antimetabolite; Ang: angiogenesis; Apo: apoptosis inducer; Db: DNA binder; Df: antifolates; DNMT: DNA methyltransferase inhibitor; Dr: ribonucleotide reductase inhibitor; Ds: DNA synthesis inhibitor; HDAC: Histone deacetylase; Ho: hormone; P90: hsp90 binder; PI3K: PI3kinase; PKC: Protein kinase C; ROS: reactive oxygen species; RSTK: serine/threonine kinase inhibitor; T1: topoisomerase 1 inhibitor; T2 : topoisomerase 2 inhibitor; Tu: tubulin-active antimetabolic; YK: tyrosine kinase inhibitor.

Comparison to previous methods

To compare our results to previous studies, namely Ammad-ud-din et al. (2014); Menden et al. (2013), we applied PGM to the GDSC and CCLE data sets, using Morgan fingerprints as compound descriptors, and the gene transcript levels for the 1,000 genes displaying the highest variance across the cell-line panels to describe the cell-lines. For the GDSC data set (Table .6.2), we obtained lower mean $RMSE_{test}$ and higher R^2_{test} values, namely 0.75 ± 0.01 and 0.74 ± 0.01 , respectively, in comparison to Menden et al. (2013) ($RMSE_{test} = 0.83$; $R^2_{test} = 0.72$), and Ammad-ud-din et al. (2014) ($RMSE_{test} = 0.83 \pm 1.00$; $R^2_{test} = 0.32 \pm 0.37$). The same trend was observed for the LOTO validation (Table .6.2), with mean $RMSE_{test}$ and R^2_{test} values of 0.81 ± 0.16 and 0.72 ± 0.08 , respectively, in opposition to Menden et al. (2013) ($RMSE_{test} = 0.99$; $R^2_{test} = 0.61$). Comparable results are obtained here ($RMSE_{test} = 1.40 \pm 0.80$) and in Ammad-ud-din et al. (2014) ($RMSE_{test} = 0.85 \pm 0.41$) when extrapolating to new compounds.

We have applied our PGM methodology, using Morgan fingerprints and the transcript levels for the genes displaying the highest variance across the cell-line panel as compound and cell-line descriptors. Although this combination of descriptors

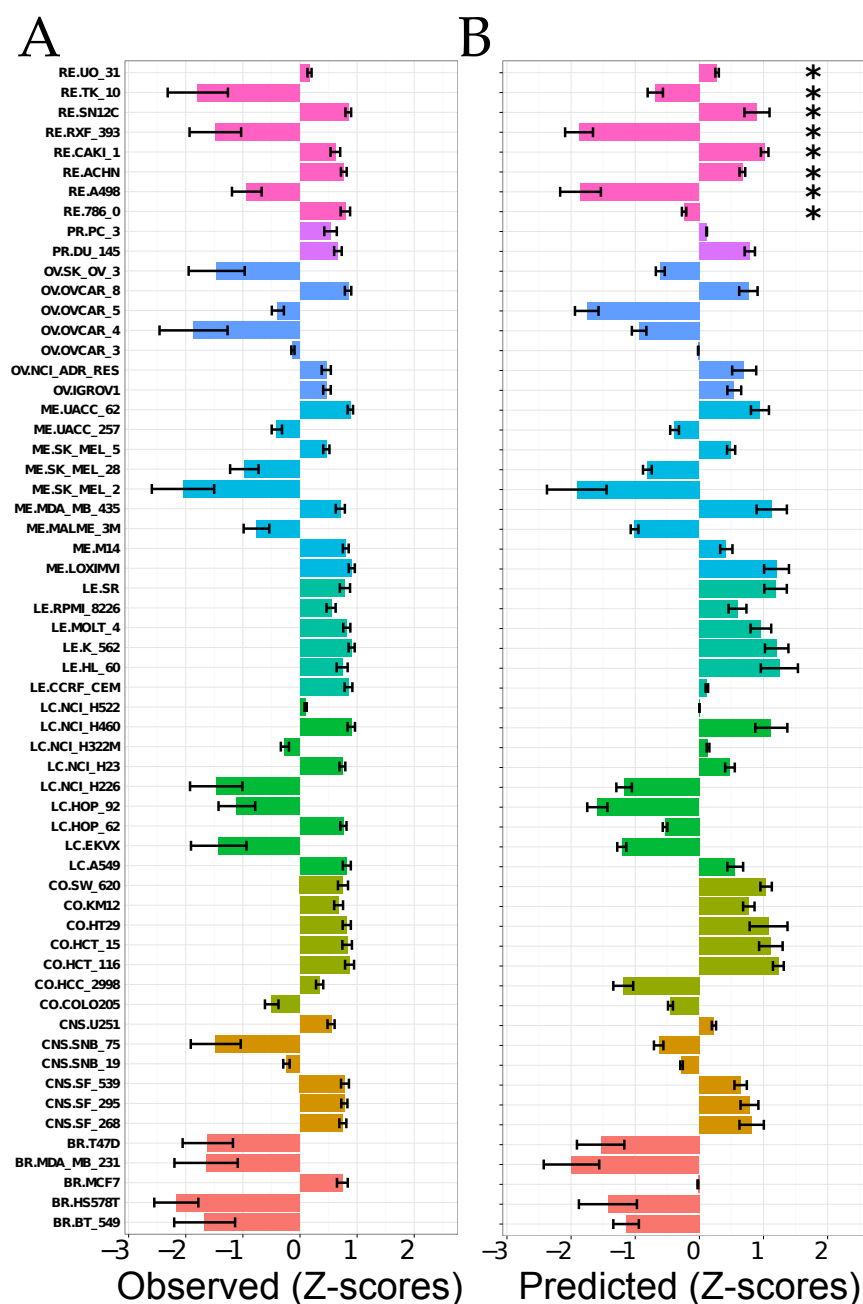


Figure 6.12: Evaluation of the predicted growth inhibition patterns for methotrexate (MTX) on the NCI60 panel. A. Relative growth inhibition pattern (z-scores) on the NCI60 panel calculated from the experimental pGI_{50} values. The experimental uncertainty of the measurements is also displayed. B. Predicted relative growth inhibition pattern of growth inhibition along with the 75% confidence interval calculated using conformal prediction. We used the predicted values on the test calculated with 10 PCM models (interpolation). Complex inhibition patterns are reflected by the predictions.

Figure .6.12: **Figure .6.12 caption continuation** For instance, renal cell-lines TK-10, RFX-393 and A498 (marked with an asterisk) were predicted to be highly resistant to MTX, whereas the effect of MTX on sensitive cell-lines, namely UO-31, SN12C, CAKI-1 and ACHN, was also correctly predicted. Cell-lines originated from the same tissue are in the same color (breast: red, central nervous system: orange, colon: olive green, lung cancer: dark green, leukemia: turquoise, melanoma: blue, ovarian: blue, prostate: purple, renal: magenta).

has led to the most predictive models on the NCI60 panel, other combinations of descriptors might be more suitable for other data sets. Thus, we advise to explore as many descriptor combinations as possible in future modeling studies on other data sets.

	GDSC		CCLE		NCI60
	Menden et al. (2013)	Ammad-ud-din et al. (ibid.)	Cortes-Ciriano, I et al. (2015)	Cortes-Ciriano, I et al. (2015)	Cortes-Ciriano, I et al. (2015)
# Cell-lines	608	650	707	480	59
# Compounds	111	116	1.8	23	17,142
# Data-points	38,930	55,619	79,902	10,630	94,231
Matrix completeness	58%	74%	81%	96.30%	93.10%
Bioactivity units	−log ₁₀ (IC ₅₀ μM)	log (IC ₅₀ μM)	−log ₁₀ (IC ₅₀ μM)	ln (IC ₅₀ μM)	−log ₁₀ (GI ₅₀ M)
Machine Learning	Neural Networks and Random Forest	Kernelized Bayesian Matrix Factorization	Random Forest	Random Forest	Random Forest and Conformal Prediction
Compound descriptors	PaDEL 1D, 2D descriptors and fingerprints	PaDEL 1D, 2D descriptors, PubChem fingerprints, and 3D Vsurf and GRIND/GRIND2	Morgan fingerprints (256) in count format	Morgan fingerprints (256) in count format	Morgan fingerprints (256) in count format
Cell-line descriptors	Microsatellite stability and CNV	Gene expression, CNV and mutation profiles	Gene transcript levels 1,000 genes	Gene transcript levels 1,000 genes	Gene transcript levels 1,000 genes
Test set (CV)	mean RMSE +/- std	0.83 +/- 1.00	0.75 +/- 0.01	1.02 +/- 0.05	0.40 +/- 0.00
	mean R ² +/- std	0.32 +/- 0.37	0.74 +/- 0.01	0.74 +/- 0.03	0.83 +/- 0.00
LOTO	mean RMSE +/- std	N/A	0.81 +/- 0.16	0.97 +/- 0.26	0.46 +/- 0.08
	mean R ² +/- std	N/A	0.72 +/- 0.08	0.75 +/- 0.12	0.78 +/- 0.05
LOCCO or Leave-One-Compound-Out	mean RMSE +/- std	0.85 +/- 0.41	1.40 +/- 0.80	1.62 +/- 1.32	0.83 +/- 0.17
	mean R ² +/- std	0.52 +/- 0.37	0.13 +/- 0.11	0.18 +/- 0.15	0.17 +/- 0.09

Table .6.2: Notes:

1. N/A refers to data not reported in the corresponding study. 2. Low R₂ values do not necessarily mean inaccurate predictions, as R² values decrease significantly, even if the predictions closely match the observations, when the range of values considered is small (see Figure .2.3). 3. The values reported for the NCI60 data set correspond to those obtained with the complete data set (see main text for details).

Abbreviations:

CCLE: Cancer Cell-Line Encyclopedia; CNV: Copy Number Variation; CV: Cross-validation; GDSC: Genomics of Drug Sensitivity in Cancer; LOCCO: Leave-One-Chemical-Cluster-Out; LOTO: Leave-One-Tissue-Out.

.6.4 Discussion

The major goal of this chapter was to capitalize on in vitro sensitivity and molecular profiling data of untreated cells to simultaneously predict compound cytotoxicity and selectivity on the NCI60 panel. Although the principles of PCM are not new, the present study represents considerable progress in the field as, to our knowledge, it is the first effort to exploit the large-scale NCI anticancer screening data and to benchmark cell-line profiling information of the NCI60 panel for drug sensitivity prediction with error bars. Unlike previous modelling studies on the NCI60 panel [Abaan et al. (2013); Paull et al. (1989); Staunton et al. (2001); Szakacs et al. (2004)], we integrate chemical information and cell-line profiling data simultaneously, which enables us to predict growth inhibition patterns and to inter- and extrapolate on the chemical and cell-line domains. Additionally, coupling conformal prediction to a machine learning algorithm, here Random Forests, enabled the definition of confidence intervals for individual predictions.

We consistently found the highest predictive signal in gene expression, miRNA and protein abundance data. The incorporation of prior biological knowledge, by including pathway information or by considering gene sets involved in cancer biology (*e.g.* oncogenes and tumour suppressors), did not improve model performance. Interestingly, predictive signals for selectivity and toxicity were present in the genes displaying the most variable transcript levels along the cell-line panel. We note in particular that the sparseness of the CNV and exome sequencing data is plausibly the reason for poor model performance, and thus anticipate that the modelling of cell-line panels with less sparse mutational data might lead to better models [Costello et al. (2014)]. A major challenge to drug sensitivity prediction is the extrapolation of compounds bioactivities to novel cell-lines and to structurally distinct compounds. We did not find significant differences in performance between interpolation and extrapolation to new cell-lines (LOCO) and tissues (LOTO), with $\text{RMSE}_{\text{test}}$ values smaller than twice the mean uncertainty value of the bioactivity measurements. This observation enables the prediction of compound activities on cancer cell-lines containing little bioactivity data, although extrapolation is improved by the presence of cell-lines from the same tissue/ontogeny in the training set. Obtaining this degree of extrapolation is notable, as the concatenation of chemical and cell-line descriptors intends to account for the whole set of complex interactions occurring in the cell upon compound administration.

Given that the compound space displays a much larger variability than the cell-line space, and that similar compounds exhibit similar growth inhibition profiles [Shivakumar and Krauthammer (2009)], it was expected that model performance would considerably decrease when extrapolating bioactivities to structurally dissimilar compounds. Nevertheless, mean $\text{RMSE}_{\text{test}}$ values of approximately three

times the average experimental uncertainty were obtained when extrapolating on the chemical space (LOCCO). Although the error in prediction should ideally be close to the experimental uncertainty, this performance is high enough to be pragmatic for experimental compound prioritization. Previous studies have shown that the addition of physicochemical descriptors or increasing the bit-string length of the Morgan fingerprints (here set to 256) leads to higher predictive power when modelling a highly diverse set of molecules [De Bruyn et al. (2013)]. Here, we did not obtain higher predictive power when increasing the bit-string length, when adding physicochemical descriptors to the compound fingerprints, or when using Morgan fingerprints in binary format [Cortes-Ciriano et al. (2015b); Murrell et al. (2014)]. These results likely arise from the fact that the compound and cell-line descriptors used here do not account for cellular events implicated in compound sensitivity, *e.g.* cell permeability. Although the modelling has been restricted to the NCI60 panel, we further anticipate that our approach could be extended to the Cancer Cell-Line Encyclopedia (CCLE) [Barretina et al. (2012)], as gene expression profiles for the 44 cell-lines shared with the NCI60 panel are highly correlated (Spearman's $r_s = 0.88$) (Figure 6.13). The extrapolation of compound bioactivities to cell-lines from the CCLE could open repurposing opportunities for the 17,142 compounds considered here, which could lead to novel cancer treatments and to testable hypotheses for the discovery of biomarkers of drug sensitivity.

Although cultured cell-lines and primary tumours might differ genetically [Borrell (2010)], investigating to which extent gene expression (and other cell-line profiling) data can be used to model *in vitro* cell-line sensitivity, can help to develop approaches for the prediction of primary tumour sensitivity from the genomic data of cancer patients.

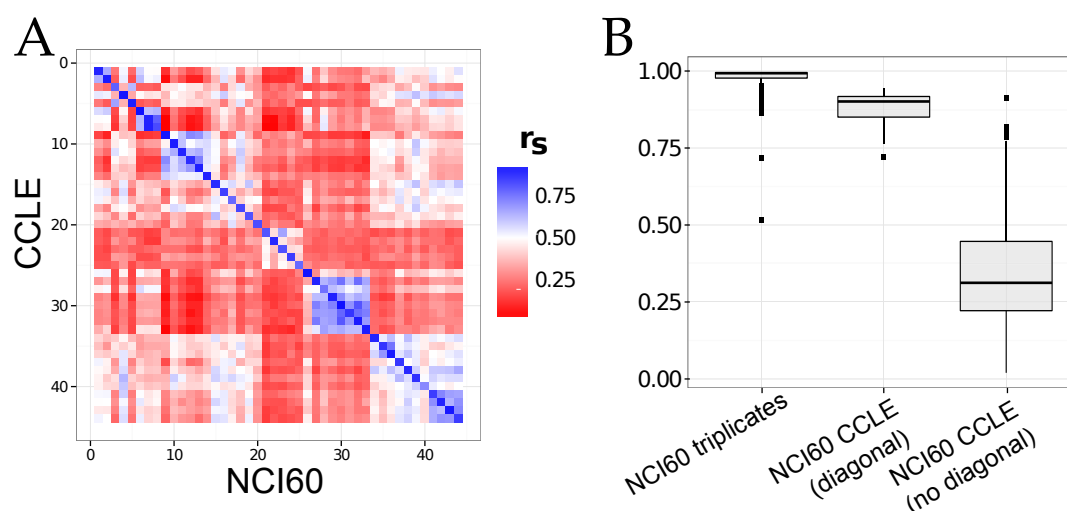


Figure .6.13: **Correlation of gene expression profiles for the 44 cell-lines present in both the NCI60 panel and the Cancer Cell Line Encyclopedia (CCLE).**

A. Pairwise Spearman's r_s correlation of the 1,000 most varying genes between the DTP-NCI60 and the CCLE data sets. Both data sets share 44 cell-lines. The correlation between the gene expression profiles of identical cell-lines is higher than 0.8 in all cases (diagonal of the matrix), with a median Spearman's r_s value close to 0.875. B. The first box plot on the left reports the Spearman's r_s correlation, above 0.98, between the gene transcript levels calculated in triplicates for the NCI60 cell-lines. The box plot in the middle corresponds to the correlation between the gene expression profiles of the cell-lines found in both the CCLE and the NCI60 data set (diagonal of the matrix in (a)). The average Spearman's r_s correlation is close to 0.875. The third boxplot reports the Spearman's r_s correlation of different cell-lines (the non-diagonal elements of the matrix in (a)). The high correlation between gene expression profiles for the cell-lines present in both the CCLE and the NCI60 cell-line panel, indicates that the PCM models reported in this study could be extended to the CCLE.

Bibliography

- Abaan, OD, EC Polley, SR Davis, YJ Zhu, S Bilke, RL Walker, M Pineda, Y Gindin, Y Jiang, WC Reinhold, SL Holbeck, RM Simon, J Doroshow, Y Pommier, and PS Meltzer (2013). "The exomes of the NCI-60 Panel: A genomic resource for cancer biology and systems pharmacology". In: *Cancer Res.* 73.14, pp. 4372–4382 (cit. on pp. 211, 237).
- Adams, J and M Kauffman (2004). "Development of the proteasome inhibitor Velcade (Bortezomib)". In: *Cancer Invest.* 22.2, pp. 304–311 (cit. on p. 207).
- Ammad-ud-din, M, E Georgii, M Gönen, T Laitinen, O Kallioniemi, K Wennerberg, A Poso, and S Kaski (2014). "Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization". In: *J. Chem. Inf. Model.* 54.8, pp. 2347–2359 (cit. on pp. 208, 215, 216, 233, 236).
- Barretina, J, G Caponigro, N Stransky, K Venkatesan, AA Margolin, S Kim, CJ Wilson, J Lehar, GV Kryukov, D Sonkin, A Reddy, M Liu, L Murray, MF Berger, JE Monahan, P Morais, J Meltzer, A Korejwa, J Jané-Valbuena, FA Mapa, J Thibault, E Bric-Furlong, P Raman, A Shipway, IH Engels, J Cheng, GK Yu, J Yu, P Aspesi, M de Silva, K Jagtap, MD Jones, L Wang, C Hatton, E Palescandolo, S Gupta, S Mahan, C Sougnez, RC Onofrio, T Liefeld, L MacConaill, W Winckler, M Reich, N Li, JP Mesirov, SB Gabriel, G Getz, K Ardlie, V Chan, VE Myer, BL Weber, J Porter, M Warmuth, P Finan, JL Harris, M Meyerson, TR Golub, MP Morrissey, WR Sellers, R Schlegel, and LA Garraway (2012). "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity". In: *Nature* 483, pp. 603–607 (cit. on pp. 207, 238).
- Borrell, B (2010). "How accurate are cancer cell lines?" In: *Nature* 463.7283, p. 858 (cit. on p. 238).
- Bouvier, G, N Duclert-Savatier, N Desdouits, D Meziane-Cherif, A Blondel, P Courvalin, M Nilges, and TE Malliavin (2014). "Functional motions modulating VanA ligand binding unraveled by Self-Organizing Maps". In: *J. Chem. Inf. Model.* 54.1, pp. 289–301 (cit. on p. 210).
- Breiman, L (Oct. 2001). "Random Forests". en. In: *Mach. Learn.* 45.1, pp. 5–32 (cit. on pp. 210, 212).
- Brown, J, Y Okuno, G Marcou, A Varnek, and D Horvath (2014). "Computational chemogenomics: Is it more than inductive transfer?" In: *J. Comput. Aided Mol. Des.* Pp. 1–22 (cit. on pp. 212, 213, 222).
- Clark, R and P Fox (2004). "Statistical variation in progressive scrambling". In: *J. Comput. Aided Mol. Des.* 18.7-9, pp. 563–576 (cit. on p. 218).

- Cortes-Ciriano, I, QU Ain, V Subramanian, EB Lenselink, O Mendez-Lucio, AP IJzerman, G Wohlfahrt, P Prusis, TE Malliavin, GJP van Westen, and A Bender (2015a). "Polypharmacology modelling using Proteochemometrics: recent developments and future prospects". In: *Med. Chem. Comm.* (Cit. on pp. 208, 213).
- Cortes-Ciriano, I, DS Murrell, GJP van Westen, A Bender, and TE Malliavin (2015b). "Prediction of the Potency of Mammalian Cyclooxygenase Inhibitors with Ensemble Proteochemometric Modeling". In: *J. Cheminf.* 7, p. 1 (cit. on p. 238).
- Cortes-Ciriano, I, van Westen, G J P, Bouvier, G, Nilges, M, Overington, J P, Bender, A, and TE Malliavin (2015). "Improved Large-scale prediction of growth inhibition patterns on the NCI60 cancer cell-line panel". In: *In revision Bioinformatics* (cit. on p. 236).
- Costello, JC, LM Heiser, E Georgii, M Gönen, MP Menden, NJ Wang, M Bansal, M Ammad-Ud-Din, P Hintsanen, SA Khan, JP Mpindi, O Kallioniemi, A Honkela, T Aittokallio, K Wennerberg, JJ Cons, D Gallahan, D Singer, J Saez-Rodriguez, S Kaski, JW Gray, and G Stolovitzky (2014). "A community effort to assess and improve drug sensitivity prediction algorithms". In: *Nat. Biotechnol.* advance on (cit. on pp. 217, 220, 237).
- De Bruyn, T, GJP van Westen, AP IJzerman, B Stieger, P de Witte, PF Augustijns, and PP Annaert (2013). "Structure-Based Identification of OATP1B1/3 Inhibitors". In: *Mol. Pharmacol.* 83.6, pp. 1257–1267 (cit. on p. 238).
- Garnett, MJ, EE Edelman, SJS Heidorn, CD Greenman, A Dastur, KW Lau, P Greninger, IR Thompson, X Luo, J Soares, Q Liu, F Iorio, L Surdez Dand Chen, RJ Milano, GR Bignell, AT Tam, H Davies, Ja Sson, S Barthorpe, SR Lutz, F Kogera, K Lawrence, A McLaren-Douglas, X Mitropoulos, T Mironenko, H Thi, L Rson, W Zhou, F Jewitt, T Zhang, P O'Brien, JL Boisvert, S Price, W Hur, W Yang, X Deng, A Butler, HG Choi, JW Chang, J Baselga, I Stamenkovic, Ja Engelman, SV Sharma, O Delattre, J Saez-Rodriguez, NS Gray, J Settleman, PA Futreal, DA Haber, MR Stratton, S Ramaswamy, U McDermott, and CH Benes (2012). "Systematic identification of genomic markers of drug sensitivity in cancer cells". In: *Nature* 483.7391, pp. 570–575 (cit. on p. 207).
- Gholami, A, H Hahne, Z Wu, F Auer, C Meng, M Wilhelm, and B Kuster (2014). "Global proteome analysis of the NCI-60 cell-line panel". In: *Cell Reports* 4.3, pp. 609–620 (cit. on p. 211).
- Golbraikh, A and A Tropsha (2002a). "Beware of q2!" In: *J. Mol. Graphics Modell.* 20.4, pp. 269–276 (cit. on p. 212).
- (2002b). "Beware of q2!" In: *J. Mol. Graphics Modell.* 20.4, pp. 269–276 (cit. on p. 212).
- Haibe-Kains, B, N El-Hachem, NJ Birkbak, AC Jin, AH Beck, HJWL Aerts, and J Quackenbush (2013). "Inconsistency in large pharmacogenomic studies". In: *Nature* 504.7480, pp. 389–93 (cit. on pp. 215, 217).
- Jacob, L and J Vert (2008). "Protein-ligand interaction prediction: an improved chemogenomics approach". In: *Bioinformatics* 24.19, pp. 2149–56 (cit. on p. 208).

- Kutalik, Z, JS Beckmann, and S Bergmann (2008). "A modular approach for integrative analysis of large-scale gene-expression and drug-response data". In: *Nature Biotechnol.* 26.5, pp. 531–9 (cit. on p. 207).
- Liberzon, A, A Subramanian, R Pinchback, H Thorvaldsdóttir, P Tamayo, and JP Mesirov (2011). "Molecular signatures database (MSigDB) 3.0". In: *Bioinformatics* 27.12, pp. 1739–1740 (cit. on p. 214).
- Lorenzi, PL, WC Reinhold, S Varma, AA Hutchinson, Y Pommier, SJ Chanock, and JN Weinstein (2009). "DNA fingerprinting of the NCI-60 cell line panel". In: *Mol. Canc. Therapeut.* 8.4, pp. 713–724 (cit. on p. 211).
- Menden, MP, F Iorio, MJ Garnett, U McDermott, CH Benes, PJ Ballester, and J Saez-Rodriguez (2013). "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties". In: *PLoS One* 8.4, e61318 (cit. on pp. 207, 215, 233, 236).
- Murrell, DS, I Cortes-Ciriano, GJP van Westen, IP Stott, TE Malliavin, A Bender, and RC Glen (2014). "Chemistry Aware Model Builder (camb): an R Package for Predictive Bioactivity Modeling". In: <http://github.com/cambDI/camb> (cit. on pp. 209, 238).
- Nishizuka, S, L Charboneau, L Young, S Major, WC Reinhold, M Waltham, H Kouros-Mehr, KJ Bussey, JK Lee, V Espina, PJ Munson, E Petricoin, LA Liotta, and JN Weinstein (2003). "Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays". In: *Proc. Natl. Acad. Sci. U.S.A.* 100.24, pp. 14229–14234 (cit. on p. 211).
- Norinder, U, L Carlsson, S Boyer, and M Eklund (2014). "Introducing Conformal Prediction in Predictive Modeling A Transparent and Flexible Alternative To Applicability Domain Determination". In: *J. Chem. Inf. Model.* 54.6, pp. 1596–1603 (cit. on pp. 214, 230).
- Paull, KD, RH Shoemaker, L Hodes, A Monks, DA Scudiero, L Rubinstein, J Plowman, and MR Boyd (1989). "Display and analysis of patterns of differential activity of drugs against human tumor cell-lines: development of mean graph and COMPARE algorithm". In: *J. Natl. Cancer Inst.* 81.14, pp. 1088–1092 (cit. on pp. 207, 237).
- Pedregosa, F, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *J. Mach. Learn. Res.* 12, pp. 2825–2830 (cit. on p. 210).
- Reinhold, WC, J Mergny, H Liu, M Ryan, TD Pfister, R Kinders, R Parchment, J Doroshow, JN Weinstein, and Y Pommier (2010). "Exon array analyses across the NCI-60 reveal potential regulation of TOP1 by transcription pausing at guanosine quartets in the first intron". In: *Cancer Res.* 70.6, pp. 2191–2203 (cit. on p. 211).
- Reinhold, WC, M Sunshine, H Liu, S Varma, KW Kohn, J Morris, J Doroshow, and Y Pommier (2012). "CellMiner: A web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell-line set". In: *Cancer Res.* 72.14, pp. 3499–3511 (cit. on p. 208).

- Riddick, G, H Song, S Ahn, J Walling, D Borges-Rivera, W Zhang, and HA Fine (2011). "Predicting in vitro drug sensitivity using Random Forests". In: *Bioinformatics* 27.2, pp. 220–4 (cit. on p. 207).
- Sheridan, RP (2013). "Using Random Forest to model the domain applicability of another Random Forest model". In: *J. Chem. Inf. Model.* 53.11, pp. 2837–2850 (cit. on pp. 210, 212).
- Shivakumar, P and M Krauthammer (2009). "Structural similarity assessment for drug sensitivity prediction in cancer". In: *BMC Bioinformatics* 10 Suppl 9.Suppl 9, S17–24 (cit. on p. 237).
- Shoemaker, RH (2006). "The NCI60 human tumour cell line anticancer drug screen". In: *Nat. Rev. Cancer.* 6.10, pp. 813–823 (cit. on p. 207).
- Staunton, JE, DK Slonim, HA Collier, P Tamayo, MJ Angelo, J Park, U Scherf, JK Lee, WO Reinhold, JN Weinstein, JP Mesirov, ES Lander, and TR Golub (2001). "Chemosensitivity prediction by transcriptional profiling". In: *Proc. Natl. Acad. Sci. USA* 98.19, pp. 10787–92 (cit. on pp. 207, 237).
- Szakacs, G, J Areau, S Lababidi, U Shankavaram, A Arciello, KJ Bussey, W Reinhold, Y Guo, GD Kruh, M Reimers, JN Weinstein, and MM Gottesman (2004). "Predicting drug sensitivity and resistance: Profiling ABC transporter genes in cancer cells". In: *Cancer Cell* 6.2, pp. 129–137 (cit. on pp. 207, 237).
- Tropsha, A and A Golbraikh (2007). "Predictive QSAR modeling workflow, model applicability domains, and virtual screening". In: *Curr. Pharm. Des.* 13.34, pp. 3494–3504 (cit. on p. 212).
- Varma, S, Y Pommier, M Sunshine, JN Weinstein, and WC Reinhold (2014). "High resolution copy number variation data in the NCI-60 cancer cell lines from whole genome microarrays accessible through CellMiner". In: *PloS ONE* 9.3. Ed. by KW Lo, e92047 (cit. on p. 211).
- Weinstein, JN (2012). "Drug discovery: Cell lines battle cancer". In: *Nature* 483.7391, pp. 544–545 (cit. on p. 207).
- Weinstein, JN, KW Kohn, MR Grever, VN Viswanadhan, LV Rubinstein, AP Monks, DA Scudiero, L Welch, AD Koutsoukos, and AJ Chiousa (1992). "Neural computing in cancer drug development: predicting mechanism of action". In: *Science* 258.5081, pp. 447–451 (cit. on p. 207).
- Weinstein, JN, TG Myers, PM O'Connor, SH Friend, AJ Fornace, KW Kohn, T Fojo, SE Bates, LV Rubinstein, NL Anderson, JK Buolamwini, WW van Osdol, AP Monks, DA Scudiero, EA Sausville, DW Zaharevitz, B Bunow, VN Viswanadhan, GS Json, RE Wittes, and KD Pl (1997). "An information-intensive approach to the molecular pharmacology of cancer". In: *Science* 275.5298, pp. 343–349 (cit. on p. 207).
- Westen, GJP van, JJ Wegner, AP IJzerman, HWT van Vlijmen, and A Bender (2011). "Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets". In: *Med. Chem. Comm.* 2.1, pp. 16–30 (cit. on p. 208).

- Wheeler, HE, ML Maitland, ME Dolan, NJ Cox, and MJ Ratain (2013). “Cancer pharmacogenomics: strategies and challenges”. In: *Nat. Rev. Genet.* 14.1, pp. 23–34 (cit. on p. 208).
- Yamanishi, Y, M Kotera, M Kanehisa, and S Goto (2010). “Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework”. In: *Bioinformatics* 26.12, pp. 1246–254 (cit. on p. 208).

Epilogue

.7 Epilogue

This thesis aimed at predicting compound activity on biomolecular targets of increasing complexity, from protein binding sites to cancer cell-lines, by integrating chemical and biological information in the frame of single machine learning models. Whereas the presented models fit reasonably well to the data, meaning that the errors in prediction on the test set (interpolation) are close to the experimental errors, extrapolating compound activity to structurally novel compounds remains a challenging task.

I was personally motivated to evaluate the influence of the experimental errors in both the training and the validation of bioactivity models. From chapters .3 and .4, it is clear that the noise in the data has a deep influence on model performance, and that the robustness to noisy bioactivity values is not constant across machine learning algorithms (and kernels) commonly used in predictive bioactivity modelling. Thus, the experimental uncertainty should: (i) guide the choice of the most suitable machine learning algorithm, and (ii) be used to validate the maximum achievable model performance given the data at hand.

These results also prompted me to evaluate the performance of algorithmically diverse methods to predict confidence intervals, and to include them in the modelling phase. The main conclusion is that conformal prediction appears as the most robust and suitable technique for this task, as (i) it is algorithm-independent, (ii) it does not require to optimize parameters beyond those of the algorithm chosen, and (iii) the confidence intervals are always valid, which means that the true value will not lie outside the confidence interval in more than a user-defined percentage of the cases.

PCM appears as a suitable technique to extrapolate bioactivities across species, specially across orthologous proteins (Chapter .5). However, the description of binding sites with amino acid descriptors could be further improved by *e.g.* considering protein dynamics. A recent paper by Brown et al. (2014) claimed that current state-of-the-art amino acid descriptors do not provide higher predictive signal than indicator (dummy) variables (which corresponds, following the terminology used in this thesis, to IT models). Although we have demonstrated that the difference in predictive power between models trained on (i) target descriptors and (ii) indicator variables is statistically significant in many cases, this difference might still be negligible from a medicinal chemistry standpoint, as it is in some cases less than half the average value

of the experimental errors. This effect might not always be due to [PCM](#) flaws, as in many available data sets (including the data set of the NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge) compound activities are highly correlated across the target panel, *i.e.* a given compound displays similar activity across the targets considered. Thus, the models are not properly challenged to predict the activity of compounds whose activities are not correlated across the considered targets, and the difference in predictive power between Family [QSAR](#) and [PCM](#) models not tested thoroughly. This evidence claims for the definition of benchmarking [PCM](#) data sets, on which new methodologies could be tested, and more importantly, the predictive power of the proposed models on novel compounds and novel targets thoroughly challenged.

Due to the mathematical formulation of [PCM](#), inherited from machine learning, protein targets need to be related, be this relationship orthology, *e.g.* [COX](#), or paralogy, *e.g.* [GPCRs](#). This originates from the fact that the covariates in a [PCM](#) model using binding site amino acid descriptors to encode the target space, need to correspond to the same biological *aspect* across the targets considered, which in this case means a given amino acid position in the binding site. The usage of full-sequence protein descriptors need to be further explored in order to integrate bioactivity data from structurally and functionally different proteins, and thus to leverage this data for the prediction of compound affinity across panels of unrelated targets¹.

On the interpretation side, [PCM](#) faces the same problems as [QSAR](#), exacerbated by the following aspects. In [QSAR](#), certain methodologies, *e.g.* [RF](#), permit to determine which covariates (descriptors) are important for the model, and, normally, also assumed to be important for compound activity. This assumption is due to the fact that models are considered to be capable to relate compound activity to compound properties or features that are *actually* implicated in activity, although in some cases, random correlations might appear. For instance, if a compound feature irrelevant for activity is always present in active compounds but not in inactive molecules. In that case, a given descriptor would be important (and useful) for the model to predict compound activity, although the property encoded by it would not be responsible for activity. In [PCM](#), data sets comprise more than one target. Therefore, identifying which descriptors are important to predict compound activity, does not reveal

¹ It is important to note that I use the term *affinity*. Other chemogenomic techniques, such as Bayesian classifiers, have been used to predict the probability of interaction between a compound and a panel of unrelated targets in the form of Bayesian scores. However, it is paramount to highlight that a high Bayesian score means that a given compound is likely to interact with a given target, and not that the compound exhibits high affinity. As a matter of fact, this affinity might be in the low, or even high, μM range, which also depends on the affinities considered in the training data [Cortes-Ciriano et al. (2013)]. We have investigated these issues in two publications, namely Cortes-Ciriano et al. (2013); Paricharak et al. (2015). In the latter, [PCM](#) and *in silico* target prediction were integrated in a novel drug discovery framework for the retrospective discovery of *Plasmodium falciparum* dihydrofolate reductase inhibitors.

whether the chemical properties encoded by them are important for pan-activity, or for activity against a particular target. Although several model interpretation methods have been proposed in this thesis, their low performance and failure in some cases [Cortes-Ciriano, I, Bender, A, and Malliavin (2015)], makes them unsuitable for real-world drug discovery campaigns, *e.g.* in a company setting, and of little help for medicinal chemists.

The most relevant part of this thesis corresponds to the prediction of cancer cell-line sensitivity from genomic data. This aim was partly fulfilled, as the errors in prediction in interpolation and extrapolation on the cell-line side were, on average, smaller than twice the average experimental error. Therefore, PCM can help (to the extent the training data allows) to (i) develop new cancer drugs, (ii) drug repurposing of existing drugs, and (iii) design tailored drug regimens on the basis of the patients genetic makeup. Nevertheless, extrapolating on the chemical domain still remains challenging, indicating that the current description of compound-cell-line interactions is far from complete and that special attention should be given to not overstep the applicability domain of the models. This is also evidenced by the fact that model predictive power reaches a *plateau* [Cortes-Ciriano, I et al. (2015); Menden et al. (2013)] irrespective of the combination of compound and cell-line descriptors used, which might indicate that many variables important for cell-line sensitivity are missed in the current modelling setting (*e.g.* cell permeability).

Haibe-Kains et al. (2013) have recently demonstrated the low concordance between drug sensitivity values between the CCLE and the GDSC, thus raising concern about the suitability of using these data in predictive modelling. However, gene transcript levels for identical cell-lines were highly correlated across institutions (see also subsection .6.4). Given the high predictive signal provided by gene expression data, PCM can build upon the vast experience gained in the last decades to standardize gene expression microarrays and their analysis, and use cell-line profiling data from different institutions as cell-line descriptors.

Another concern is how we quantify cell-line sensitivity. In this line, Fallahi-Sichani et al. (2013) applied multi-parametric analysis to a data set comprising the activity of 64 anticancer drugs on 53 breast cancer cell-lines. The results of this study indicate that the parameters of the dose-response curve vary systematically depending on the cell-line or drug class. For instance, drugs mechanism of action has a strong influence on drug efficacy (E_{\max}), potency (IC_{50}), and on the steepness of the drug response curve. Overall, this study indicates that other parameters than potency of the drug response curve might be considered in comparative studies of drug activity, as they are likely to provide crucial insight into the biology of cell-line response to drug treatment, and into drugs mechanism of action.

Overall, these studies highlight anew the complexity of predicting (or even measuring) cancer cell-line sensitivity, suggesting that predictive methods might be limited to making testable hypothesis and that prospective experimental validation is foremost to reach any biological conclusion. To date, PCM studies in the field [Ammad-ud-din et al. (2014); Cortes-Ciriano, I et al. (2015); Menden et al. (2013)] have quantified cell-line sensitivity using IC_{50} and GI_{50} values. Thus, there exists ample room for exploring new ways of: (i) encoding compound-cell-line interactions, and (ii) quantifying cancer cell-line drug sensitivity.

Bibliography

- Ammad-ud-din, M, E Georgii, M Gönen, T Laitinen, O Kallioniemi, K Wennerberg, A Poso, and S Kaski (2014). "Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization". In: *J. Chem. Inf. Model.* 54.8, pp. 2347–2359 (cit. on p. 252).
- Brown, J, Y Okuno, G Marcou, A Varnek, and D Horvath (2014). "Computational chemogenomics: Is it more than inductive transfer?" In: *J. Comput. Aided Mol. Des.* Pp. 1–22 (cit. on p. 249).
- Cortes-Ciriano, I, A Koutsoukas, O Abian, RC Glen, A Velazquez-Campoy, and A Bender (2013). "Experimental validation of in silico target predictions on synergistic protein targets". In: *Med. Chem. Commun.* 4 (1), pp. 278–288 (cit. on p. 250).
- Cortes-Ciriano, I, Bender, A, and TE Malliavin (2015). "Prediction of PARP inhibition with Proteochemometric modelling and conformal prediction". In: *In revision at Mol. Inform.* URL: <http://cran.r-project.org/package=conformal> (cit. on p. 251).
- Cortes-Ciriano, I, van Westen, G J P, Bouvier, G, Nilges, M, Overington, J P, Bender, A, and TE Malliavin (2015). "Improved Large-scale prediction of growth inhibition patterns on the NCI60 cancer cell-line panel". In: *In revision Bioinformatics* (cit. on pp. 251, 252).
- Fallahi-Sichani, M, S Honarnejad, LM Heiser, JW Gray, and PK Sorger (2013). "Metrics other than potency reveal systematic variation in responses to cancer drugs". In: *Nat. Chem. Biol.* 9.11, pp. 708–714 (cit. on p. 251).
- Haibe-Kains, B, N El-Hachem, NJ Birkbak, AC Jin, AH Beck, HJWL Aerts, and J Quackenbush (2013). "Inconsistency in large pharmacogenomic studies". In: *Nature* 504.7480, pp. 389–93 (cit. on p. 251).
- Menden, MP, F Iorio, MJ Garnett, U McDermott, CH Benes, PJ Ballester, and J Saez-Rodriguez (2013). "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties". In: *PLoS One* 8.4, e61318 (cit. on pp. 251, 252).
- Paricharak, S, I Cortes-Ciriano, AP IJzerman, TE Malliavin, and A Bender (2015). "Proteochemometric modeling coupled to in silico target prediction: an integrated approach for the simultaneous prediction of polypharmacology and binding affinity of small molecules". In: *J. Cheminf.* 7, p. 15 (cit. on p. 250).

Back Matter

Acronyms

AD

Applicability Domain

ANOVA

Analysis of Variance

ARD

Automatic Relevance Determination

AUC

Area Under the Curve

CCLE

Cancer Cell-Line Encyclopedia

CF

Collaborative Filtering

CI

Confidence Interval

CLIFP

Cell-Line Identity Fingerprints

CNV

DNA Copy Number Variation

COX

Cyclooxygenases

CV

Cross-Validation

DNA

Deoxyribonucleic Acid

DUD

Directory of Useful Decoys

ECFP

Extended Connectivity Fingerprints

EL

Explicit Learning

FDA

United States Food and Drug Administration Agency

GBM

Gradient Boosting Machine

GDSC

Genomics of Drug Sensitivity in Cancer

GP

Gaussian Process

GPCR

G Protein-Coupled Receptor

HDAC

Histone Deacetylase

HIV

Human Immunodeficiency Virus

HSD

Tukey's Honest Significant Difference Test

HTS

High-Throughput Screening

IT

Inductive Transfer

KNN

k-Nearest Neighbours

LOCCO

Leave-One-Compound-Cluster-Out

LOCO

Leave-One-Cell-Line-Out

LOTO

Leave-One-Target-Out (Leave-One-Tissue-Out in chapter .6)

MC

Monte Carlo

MDDR

MDL Drug Data Report

MS

Model Stacking

NCI

United States National Cancer Institute

NP

Normalized Polynomial Kernel

NSAID

Non-Steroidal Anti-Inflammatory Drug

OATP

Organic Anion-Transporting Polypeptide

PCA

Principal Component Analysis

PCM

Proteochemometric Modelling

PLFP

Protein-Ligand Fingerprints

PLS

Partial Least Squares

PUK

Pearson VII Function-Based Universal Kernel

QCAM

Quantitative Cell-Line-Activity Modelling

QSAM

Quantitative Sequence-Activity Modelling

QSAR

Quantitative Structure-Activity Relationship

QSPR

Quantitative Structure-Property Relationship

RBF

Radial Basis Function

RF

Random Forest

RMSE

Root Mean Squared Error

RPLA

Reverse Phase Protein Lysate Microarray

RVM

Relevance Vector Machine

SOM

Self-Organizing Map

SVM

Support Vector Machines

SVR

Support Vector Regression

TIFP

Target Identity Fingerprints

TM

Transmembrane

Author Index

- Üstün, B, 14, 47
- A Mauri, VC, 6, 35, 59, 71
- A, V, 143, 170, 176
- Abaan, OD, 189, 204, 217
- Abagyan, R, 56, 73
- Abid, A, 177
- Abraham, AK, 41
- Adams, J, 185, 217
- Aerts, HJWL, 40, 218, 227
- Agostini, F, 36
- Agostino, M, 27, 49
- Ahmad, S, 6, 38
- Ahn, S, 220
- Ain, QU, 56, 71, 131, 176, 217
- Airola, A, 74
- Aittokallio, T, 74, 176, 218
- Akella, LB, 3, 35
- Al-Lazikani, B, 39, 111, 132, 176
- Aldenberg, T, 43, 113
- Allison, KR, 178
- Alvarsson, J, 75
- Amberger, J, 40
- Ammad-Ud-Din, M, 176, 218
- Anderson, KC, 38
- Anderson, NL, 220
- Andersson, A, 75
- Andersson, CD, 25, 35
- Andersson, CR, 12, 25, 27, 35
- Andersson, P, 44
- Angelo, MJ, 220
- Annaert, PP, 38, 72, 218
- Aparoy, P, 179
- Arciello, A, 220
- Ardlie, K, 35, 217
- Areau, J, 220
- Armbruster, BN, 41
- Arnby, CH, 72, 112, 132, 176
- Arnold, FH, 80, 114
- Arrault, A, 178
- Arrowsmith, CH, 23, 35
- Artemenko, AG, 179
- Artemenko, N, 27, 35
- Aspesi, P, 35, 217
- Assefnia, S, 37
- Atteridge, CE, 41
- Auer, F, 218
- Augustijns, PF, 38, 72, 218
- Azzaoui, K, 36
- Baakman, C, 47
- Babiss, LE, 32, 46
- Bahar, I, 9, 35
- Bajorath, J, 9, 26, 47
- Ballabio, D, 65, 71, 114
- Ballester, PJ, 6, 35, 43, 73, 219, 227
- Banck, M, 44
- Bansal, M, 176, 218
- Baron, JA, 175
- Barretina, J, 30, 35, 36, 185, 205, 217
- Barthorpe, S, 39, 218
- Basak, SC, 40, 60, 72, 141, 142, 177
- Baselga, J, 39, 218
- Baskin, II, 13, 15, 41
- Basu, A, 30, 31, 36
- Bates, SE, 220

- Beare, D, 48
Beck, AH, 40, 218, 227
Beckmann, JS, 185, 218
Bellis, L, 46, 115, 133
Bellis, LJ, 39, 111, 132, 176
Bellucci, M, 32, 36
Ben-Hur, A, 13, 14, 36, 84, 111, 123, 131, 142, 175
Bender, A, 3, 36, 37, 40, 44, 48, 58, 66, 71–75, 81, 111, 112, 115, 121, 122, 131–133, 175, 176, 178, 217–220, 227
Benes, C, 48
Benes, CH, 39, 43, 73, 218, 219, 227
Bengio, Y, 38
Benigni, R, 43, 113
Bento, AP, 39, 111, 132, 176
Berenbaum, F, 177
Berger, MF, 35, 217
Bergmann, S, 185, 218
Bergsagel, PL, 38
Berman, HM, 139, 175
Betts, BJ, 45
Beukers, MW, 113
Bhat, TN, 175
Bianchi, MT, 3, 36
Bieler, M, 3, 36
Bignell, GR, 39, 218
Bilke, S, 217
Bindal, N, 48
Birkbak, NJ, 40, 218, 227
Bittker, JA, 36
Blake, EJ, 38
Blanchet, FG, 113, 179
Blaney, FE, 39, 72, 112
Globaum, AL, xxiv, 138, 169, 175, 179
Blondel, A, 42, 217
Blondel, M, 219
Bocchini, C, 40
Bock, JR, 9, 19, 20, 36
Bodycombe, NE, 36
Bohlin, L, 137, 178
Boisvert, JL, 39, 218
Bolognese, JA, 175
Bolton, E, 47, 75, 133
Borges-Rivera, D, 220
Borisy, AA, 22, 36
Borrell, B, 205, 217
Borrmann, A, 47
Bosnić, Z, 16, 36, 79, 111
Bottou, L, 37
Botzolakis, EJ, 3, 36
Bountra, C, 35, 44
Bourne, PE, 175
Bouvier, G, 37, 71, 132, 188, 217, 227
Boyce, SW, 178
Boyd, MR, 219
Boyer, S, 38, 44, 72, 74, 112, 132, 176, 219
Bracha, AL, 36
Bradner, JE, 38
Bredel, M, 4, 36
Breese, JS, 19, 36
Breiman, L, 123, 131, 142, 175, 190, 217
Brenema, C, 27, 38
Breneman, CM, 6, 27, 37
Bresalier, RS, 137, 175
Bret, G, 41
Bric-Furlong, E, 35, 217
Brideau, C, 178
Brown, J, 31, 33, 36, 64, 71, 92, 111, 141, 175, 190, 191, 196, 217, 225, 227
Brown, JB, 46
Brown, ML, 37
Brown, MT, 178
Brown, N, 47, 177
Brown, SP, 79, 111, 121, 131, 156, 175
Bruce, CL, 12, 37, 145, 180
Brucher, M, 219
Bruneau, P, 46, 115
Bruno, M, 181
Bryant, SH, 33, 42, 47, 75, 133
Bryson, BD, 38
Bukasa, A, 180

- Bule, SS, 176
 Bunow, B, 220
 Buolamwini, JK, 220
 Burden, FR, 80, 111
 Bussey, KJ, 219, 220
 Butler, A, 39, 218
 Buturovic, LJ, 73
 Buydens, LMC, 14, 47
 Byers, SW, 37
- C, A, 43
 Côté, S, 43
 Califano, A, 32, 46
 CAll, IG, 38
 Camacho, DM, 178
 Campbell, BT, 41
 Campbell, RM, 47
 Cao, D, 7, 8, 21, 37
 Cao, Z, 39, 40, 48, 111, 112, 115
 Caponigro, G, 35, 217
 Carlsson, L, 38, 44, 72, 74, 112, 120, 131–133, 176, 180, 219
 Carlstedt-Duke, J, 49
 Carpenter, AE, 38
 Carrupt, PA, 37
 Caruana, R, 143, 175
 Chae, CH, 177
 Chambers, J, 39, 111, 132, 176
 Chan, CC, 178
 Chan, KW, 41
 Chan, V, 35, 217
 Chang, JW, 39, 218
 Chanock, SJ, 219
 Charboneau, L, 219
 Chaturvedi, SC, 138, 178
 Che, D, 39
 Cheah, JH, 36
 Chen, BY, 25, 35
 Chen, CC, 38
 Chen, GC, 38
 Chen, H, 123, 131
 Chen, K, 74
 Chen, L, 138, 175
 Chen, T, 39, 111
 Chen, X, 115, 116
 Chen, Y, 40
 Chen, YZ, 74
 Cheng, F, 7, 14, 37
 Cheng, J, 35, 217
 Chennubhotla, C, 9, 35
 Cherezov, V, 112
 Chesi, M, 38
 Chiausa, AJ, 220
 Chien, EYT, 112
 Chipman, JG, 180
 Choi, HG, 39, 218
 Ciceri, P, 37, 41
 Ciriano, IC, 71
 Clark, R, 89, 111, 156, 175, 194, 217
 Clarke, SM, 38
 Clementi, S, 44
 Clemons, PA, 36
 Cohen, P, 21, 37
 Colbran, RJ, 177
 Coller, HA, 220
 Collobert, R, 15, 37
 Con, S, 178
 Cons, JJ, 176, 178, 218
 COns, PW, 177
 Consonni, V, 58, 65, 71, 75, 114
 Consortium, U, xxxii, 100, 113
 Copeland, RA, 180
 Cortes-Ciriano, I, 4, 11, 16, 17, 21, 29–31, 33, 34, 37, 44, 56–59, 61, 62, 64, 66, 71, 74, 75, 120–123, 128, 131–133, 140, 145, 156, 158, 159, 172, 175, 176, 178, 186, 192, 205, 217–219, 226, 227
 Corthals, F, 72
 Costello, JC, 170, 171, 176, 178, 193, 195, 204, 218
 Cournapeau, D, 219
 Courvalin, P, 217
 Covington, M, 180

- Cox, NJ, xxxii, 155, 180, 221
Crew, G, 175
Crichton, I, 181
Crivori, P, 37
Crofford, LJ, 138, 176
Cromlish, WA, 178
Cronin, MTD, 113
Cruciani, G, 8, 37, 44
Csányi, G, 113
Csato, L, 110, 111
Culberson, JC, 46
Curiel, RV, 138, 176
Currie, JL, 177

Dakshanamurthy, S, 7, 37
Damle, R, 47
Dancik, V, 36
Dannhardt, G, xxiv, 138, 167, 169, 171, 176
Das, 6, 27, 37
Dastur, A, 39, 218
Davies, H, 39, 218
Davies, JW, 36, 71, 111, 131
Davies, M, 39, 111, 132, 176
Davis, MI, 8, 37, 41
Davis, SR, 217
DeCaprio, D, 3, 35
Del Rosario, AM, 38
Delarge, J, 178
Delattre, O, 39, 218
Delmore, JE, 22, 38
Deng, W, 27, 38
Deng, X, 39, 218
Deng, Z, 37
Desaphy, J, 27, 38
Desdouits, N, 217
De Bruyn, T, 8, 13, 16, 38, 56, 59, 72, 205, 218
Dighe, PR, 176
Dimitrov, I, 6, 38
din, M Au, 186, 218, 226, 227
Dineen, D, 74, 179

Ding, XW, 181
Doak, AK, 43
Doddareddy, MR, 59, 72
Dogné, JM, 178
Doherty, KM, 24, 38
Dolan, ME, 221
Doroshov, J, 217, 219
Dracheva, S, 47, 75, 133
Dragos, H, 143, 170, 176
Dube, PN, 138, 176
Dubourg, V, 219
Duchesnay, E, 219
Duclert-Savatier, N, 217
Ducrot, P, 38
Dunwiddie, CT, 44
Duvenaud, D, 14, 38, 109, 111

Ebright, RY, 36
Edeen, PT, 41
Edelman, EE, 39, 218
Edelman, EJ, 48
Efron, B, xxii, xxxii, 64, 72, 125, 132, 153, 155, 176
Eklund, M, 12, 38, 41, 42, 44, 46, 74, 75, 113, 115, 125, 132, 180, 219
El-Hachem, N, 40, 218, 227
Elling, CE, 39
Elliott, PJ, 36
Emami, H, 47
Embrechts, M, 27, 38
Emmerich, M, 72
Engelman, Ja, 39, 218
Engels, IH, 35, 217
Erhan, D, 13, 19, 38
Erikson, RL, 180
Eriksson, L, 45, 114, 123, 133, 179
Eriksson, M, 131
Ernsberger, P, 41
Es, AM, 44
Espina, V, 219
Ethier, D, 178

Falcon, S, 33, 43

- Fallahi-Sichani, M, 226, 227
 Fang, H, 40
 Fang, X, 47
 Faraoni, R, 41
 Farid, R, 26, 45
 Feng, Z, 175
 Fernandez, M, 6, 38
 Feuston, BP, 46
 Fidelis, K, 46
 Fiebich, BL, 167, 176
 Finan, P, 35, 217
 Fine, HA, 220
 Fine, M, 137, 138, 176
 Fish, PV, 35
 FitzGerald, GA, 181
 Fix, E, 123, 132
 Floyd, M, 41
 Floyd, SR, 22, 38
 Foata, N, 41
 Fojo, T, 220
 Foley, MA, 36
 Foord, SM, 39, 72, 112
 Forbes, S, 48
 Fornace, AJ, 220
 Fox, P, 89, 111, 156, 175, 194, 217
 Frank, M, 41
 Franke, L, 46, 115, 133
 Fredholm, BB, 109, 111
 Friedman, JH, 123, 131, 132, 142, 176
 Frieler, RA, 179
 Friend, SH, 44, 220
 Frimurer, TM, 20, 25, 39
 Fuchs, JE, 177
 Funk, CD, 181
 Futreal, PA, 39, 48, 218
 Fydrych, A, 38

 Gönen, M, 176, 218, 227
 Gabriel, SB, 35, 217
 Galbraith, W, 180
 Gallahan, D, 176, 218
 Gallant, P, 41

 Galloway, WRJD, 40, 72, 132
 Gamble-G, J, 177
 Gans-Brangs, K, 180
 Ganzer, U, 45, 114
 Gao, J, 8, 13, 17, 19, 25, 39, 48, 109, 111, 115
 Gargallo, CJ, 137, 138, 180
 Garland, SL, 39, 72, 112
 Garnett, MJ, 30, 39, 43, 48, 73, 185, 218, 219, 227
 Garnev, P, 6, 38
 Garraway, LA, 35, 36, 217
 Gaulton, A, 3, 7, 39, 80, 81, 111, 120, 132, 139, 147, 176
 Gauthier, JY, 178
 Ge, J, 121, 132
 Ge, W, 40
 Gedeck, P, 6, 7, 27, 41, 42, 73, 112, 132, 177
 Geiss, KT, 40
 Geluykens, P, 48
 Gentleman, R, 33, 43
 Genton, MG, 84, 112
 Georgiev, V, 42
 Georgii, E, 176, 218, 227
 Geppert, H, 9, 26, 47
 Gerlach, L, 39
 Getz, G, 35, 217
 Ghahramani, Z, 38, 111
 Ghobrial, IM, 38
 Gholami, A, 189, 218
 Gibson, TJ, 74, 179
 Gilbert, JC, 36
 Gilles, M, 143, 170, 176
 Gilliland, G, 175
 GilP, T, 38
 Gilson, MK, 42
 Gindin, Y, 217
 Gindulyte, A, 47, 75, 133
 Gioria, S, 47, 75
 Giraud, E, 56, 74
 Giuliano, F, 180

- Glen, RC, 3, 36, 43, 58, 71, 73, 74, 81, 112, 122, 132, 133, 140, 176, 178, 219
- Glenn, RC, 58, 72, 112, 132, 176
- Glick, M, 36, 71, 111, 131
- Glinca, S, 25, 39
- Gloriam, DE, 25, 39, 57, 72, 81, 112
- Gohlke, H, 27, 44
- Gola, JMR, 113
- Golbraikh, A, 27, 49, 64, 65, 72, 80, 88, 112, 115, 142, 168, 176, 180, 190, 218, 220
- Golden, J, 21, 43
- Golub, TR, 35, 217, 220
- Goncalves, C, 42
- Gonzales, MJ, 45
- Gordon, R, 178
- Goto, S, 221
- Gottesman, MM, 220
- Gottfries, J, 108, 112
- Gottlieb, AS, 6, 39
- Gough, DA, 9, 19, 20, 36
- Gramatica PG, VK, 65, 75, 80, 88, 115, 142, 168, 180
- Gramatica, P, 113
- Gramatica, PO, 43
- Gramfort, A, 219
- Grant, G, 181
- Gray, JW, 176, 218, 227
- Gray, NS, 39, 218
- Greenbaum, JA, 47
- Greenberger, BA, 38
- Greenman, CD, 39, 218
- Gregori-Puigjané, E, 4, 5, 39, 43, 61, 62, 72, 74
- Greninger, P, 39, 48, 218
- Grever, MR, 220
- Griffith, MT, 112
- Grisel, O, 219
- Grosse, R, 38, 111
- Groten, JP, 40
- Grubb, M, 180
- Gruetter, M, 23, 39
- Guan, L, 29, 49
- Guay, J, 178
- Gujral, TS, 22, 40
- Guo, D, 113
- Guo, Y, 220
- Guodong, S, 114
- Gupta, GK, 138, 177
- Gupta, S, 35, 217
- Gustafsson, JA, 49
- Gustafsson, MG, 12, 25, 27, 35
- Gutcaits, A, 4, 5, 9, 42
- H, N, 45
- Högberg, T, 39
- Hüllermeier, E, 73
- Haber, DA, 39, 48, 218
- Hahn, M, 7, 45, 58, 74, 79, 81, 114, 122, 133, 140, 179
- Hahn, WC, 38
- Hahne, H, 218
- Haibe-Kains, B, 31, 40, 193, 218, 226, 227
- Haifeng, H, 114
- Hajduk, PJ, 43, 79, 111, 121, 131, 156, 175
- Hamon, J, 36, 43
- Hamosh AS, A, 6, 40
- Hamprecht, FA, 21, 42, 93, 112
- Han, L, 47, 75, 133
- Hanson, MA, 112
- Hara, T, 48, 76
- Harris, JL, 35, 217
- Harrold, JM, 41
- Hartley, ND, 177
- Hatton, C, 35, 217
- Hawkins, DM, 34, 40, 60, 72, 141, 142, 177
- Hayashi, S, 167, 177
- He, M, 37, 137, 178
- He, Y, 39, 111, 175
- Heckerman, D, 19, 36

- Heffernan, TP, 38
 Heidorn, SJS, 39, 218
 Heijne, WHM, 32, 40
 Heinrich, N, 114
 Heiser, LM, 176, 218, 227
 Hendriks, A, 48, 115
 Henley, SJ, 138, 180
 Henrotin, Y, 178
 Herrgard, S, 37, 41
 Hersey, A, 39, 111, 132, 176
 Hewes, WE, 180
 Hey-Hawkins, E, 179
 Hibert, M, 47, 75
 Higgins, DG, 74, 179
 Hinselmann, G, 179
 Hinton, GEG, 32, 40
 Hintsanen, P, 176, 218
 Hirokawa, T, 48, 76
 Hirst, JD, 37, 145, 180
 Ho, R, 38
 Ho, WS, 133
 Hocker, M, 37
 Hodes, L, 219
 Hodges, JL, 123, 132
 Hoeft, A, 178
 Hoelder, S, 47
 Hoffmann, B, 40
 Holbeck, SL, 217
 Holmes, SP, 38
 Hon, CSY, 36
 Honarnejad, S, 227
 Hong, H, 8, 40
 Honkela, A, 176, 218
 Hoof, I, 47
 Hoorn, WP v, 44
 Hopkins, AL, 44
 Hoppe, C, 26, 40
 Horgan, K, 175
 Hornik, K, 112, 132
 Horst, E vd, 9, 40, 56, 72
 Horuk, R, 29, 40
 Horvath, D, 36, 71, 111, 175, 178, 217, 227
 Hothorn, T, 123, 133
 Hoven, OO vd, 40, 72, 115
 Hoven, OOvd, 48, 75
 Howes, LG, 137, 177
 Hson, DJ, 138, 177
 Hu, Q, 37
 Huang, H, 175
 Huang, Q, 7, 38–40, 48, 92, 93, 111, 112, 115
 Huang, SL, 181
 Huang, W, 40
 Huang, Z, 23, 40
 Hubbard, A, 43
 Hunt, JP, 37, 41
 Hunter, T, 43
 Hur, W, 39, 218
 Hurst, NW, 36
 Hutchinson, AA, 219
 Hutchison, GR, 44
 Hvidsten, TR, 46
 Ida, T, 48, 76
 IJzerman, AP, 38, 40, 44, 48, 72, 74, 75, 111–113, 115, 131–133, 176, 217, 218, 220
 Ilatovskiy, AV, 56, 73
 InChI, 54, 72
 Indigo, 54, 72
 Iorio, F, 39, 43, 73, 218, 219, 227
 Irwin, JJ, 41
 Isakson, PC, 177
 Issa, GC, 38
 Issa, NT, 37
 Ito, D, 36
 J E S Wikberg, ML, 55, 72
 J, CA, 44
 Jaakola, VP, 99, 112
 Jackson, RM, 57, 73
 Jacob, L, 20, 26, 40, 186, 218

- Jacobs, HM, 38
Jacobson, KA, 111
Jacoby, E, 4, 36, 41
Jaffee, B, 180
Jagtap, K, 35, 217
Jahn, A, 179
Jané-Valbuena, J, 35, 217
Jaworska, J, 43, 113
Jenkins, JL, 36, 43, 71, 72, 111, 131
Jewitt, F, 39, 218
Jiang, H, 40, 74
Jiang, Y, 217
Jin, AC, 40, 218, 227
Jin, H, 40, 112
Jones, MD, 35, 217
Jones, R, 138, 177
Jonsson, J, 45, 114, 179
Jorissen, RN, 42
Jorrisen, RN, 42
Jouzeau, JY, 138, 177
Json, EF, 43
Json, GS, 220
Json, K, 60, 73, 123, 132, 141, 178
Junaid, M, 6, 24, 41, 45, 46, 113, 115

Kadie, C, 19, 36
Kagan, L, 29, 41
Kahn, S, 43, 113
KalantarMotamedi, Y, 73
Kalinina, O, 57, 72
Kalliokoski, T, 16, 32, 41, 42, 66, 73, 79, 89, 112, 120, 132, 139, 142, 156, 170, 177
Kallioniemi, O, 176, 218, 227
Kanehisa, M, 221
Kang, H, 40, 112
Kantor, R, 45
Kao, B, 133
Karaman, MW, 6, 8, 41
Karapetyan, K, 47, 75, 133
Karatzoglou, A, 87, 112, 123, 132
Karelson, M, 58, 73
Kargman, S, 178
Karplus, K, 74, 179
Karst, JC, 42
Karypis, G, 9, 44
Kaski, S, 176, 218, 227
Kastelein, F, 181
Kastritis, E, 38
Katz, JD, 138, 176
Katzenmeier, G, 45, 113
Kauffman, M, 185, 217
Kaufman, CG, 113
Kazi, JU, 180
Keiser, MJ, 9, 41, 43
Keith, CT, 36
Kellenberger, E, 6, 41, 46, 75
Kellis, M, 178
Kemp, GJL, 41
Kenney, JP, 180
Kerr, J, 180
Khabele, D, 36
Khan, SA, 176, 218
Khanna, IK, 166, 177
Kienhuis, AS, 40
Kifle, L, 43
Kim, H, 138, 177
Kim, S, 35, 217
Kinders, R, 219
Kindt, R, 113, 179
King, BM, 38
Kingsley, PJ, 177
Kinnings, SL, 57, 73
Kirschner, MW, 22, 40
Klauffke, W, 73
Klebe, G, 25, 39, 46, 73
Klekota, J, 101, 112, 172, 177
Klomp, J, 47
Klopman, G, 43, 113
Klotz, KN, 111
Knapp, S, 22, 41, 47
Koboldt, CM, 177
Koeppen, H, 3, 36
Kogera, F, 39, 218

- Kohn, KW, 219, 220
 Koji, T, 14, 45
 Kommenda, M, 109, 112
 Komorowski, J, 41, 46
 Kondratovich, E, 13, 15, 41
 Kong, X, 114
 Kononenko, I, 16, 36, 79, 111
 Konstam, MA, 175
 Kontijevskis, A, 24, 41
 Koonin, EV, 28, 41
 Koppisetty, CAK, 27, 41
 Korejwa, A, 35, 217
 Kostenis, E, 39
 Kotera, M, 221
 Kouros-Mehr, H, 219
 Koutsoukas, A, 58, 73, 122, 132
 Koutsoukos, AD, 220
 Kraker, J, 40
 Kramer, C, 6, 7, 16, 27, 32, 33, 41, 42, 66, 73, 79, 87–89, 110, 112, 120, 132, 142, 147, 172, 177
 Krause, A, 80, 114
 Krauthammer, M, 205, 220
 Krein, MP, 6, 27, 37
 Kronberger, G, 109, 112
 Krstajic, D, 60, 61, 73
 Kruger, FA, 4, 28, 42, 56, 73, 109, 112, 147, 178
 Kruh, GD, 220
 Kryshtafovych, A, 46
 Kryukov, GV, 35, 36, 217
 Ksikes, A, 175
 Kubinyi, H, 21, 42, 93, 112
 Kufareva, I, 56, 73
 Kuffner, R, 178
 Kuhn, D, 56, 73
 Kuhn, M, 53, 60, 73, 87, 112, 123, 132, 141, 178
 Kumar, A, 138, 177
 Kumbhare, MR, 176
 Kung, AL, 38
 Kuster, B, 218
 Kutalik, Z, 185, 218
 Kuzmin, VE, 179
 Kwanten, L, 48
 L'heureux, P, 38
 L, R, 16, 32, 41, 79, 112, 120, 132, 142, 177
 Laak, A t, 45, 114
 Lababidi, S, 220
 Ladant, D, 42
 Laine, E, 27, 42
 Laitinen, T, 218, 227
 Lam, FC, 38
 Lampa, S, 75
 Lanas, A, 137, 138, 175, 180
 Lander, ES, 220
 Landesberg, G, 181
 Landrum, G, 122, 132, 140, 178
 Lane, JR, 112
 Langford, RM, 178
 Laoui, A, 56, 74
 Lapinsh, M, 4–6, 8, 9, 14, 24, 28, 30, 41, 42, 45, 46, 55, 56, 59, 73–75, 110, 113, 115
 Lau, CK, 167, 178
 Lau, KW, 39, 218
 Laub, J, 45, 114
 Laufer, S, xxiv, 138, 167, 169, 171, 176
 Lauver, DA, 179
 Lavan, P, 43
 Lawrence, K, 39, 218
 Leahy, DE, 73
 Lee, JK, 219, 220
 Lee, JY, 179
 Lee, K, 22, 35, 44
 Lee, MJ, 38
 Lee, MS, 36
 Lee, PW, 37
 Lee, SD, 133
 Legendre, P, 113, 179
 Lehar, J, 35, 36, 217
 Lemieux, ME, 38

- Lenselink, EB, 37, 71, 131, 175, 176, 217
Lesnard, A, 42
Leval, X d, xxiv, 138, 167, 169, 171, 178
Li, C, 178
Li, F, 121, 133
Li, H, 180
Li, J, 37
Li, N, 35, 217
Li, Q, 33, 42, 113
Li, W, 37, 74, 179
Li, Y, 74, 114
Li, ZR, 74
Liang, Y, 37
Liao, H, 175
Liaw, A, 46, 123, 133
Liberzon, A, 193, 219
Liedl, KR, 177
Liefeld, T, 35, 36, 217
Light, Y, 39, 46, 111, 115, 132, 133, 176
Lightfoot, H, 48
Liland, KH, 123, 133
Lin YW, X, 42
Lin, H, 4, 42
Lin, LI, xxxii, 155, 178
Lindborg, SR, 44
Linden, J, 111
Linder, S, 3, 46
Lines, C, 175
Linusson, A, 25, 35
Liotta, LA, 219
Lipshitz, B, 113
Liu, G, 37
Liu, H, 219
Liu, K, 36
Liu, M, 35, 217
Liu, Q, 39, 40, 48, 111, 112, 115, 218
Liu, S, 37
Liu, T, 7, 8, 42
Liu, X, 40
Liu, Z, 25, 42
Lloyd, JR, 38, 111
Lockhart, DJ, 41
Lopez, R, 74, 179
Lorenzi, PL, 31, 48, 189, 219
Lounkine, E, 28, 43
Lowe, R, 15, 43, 73
Lu, Y, 47
Lucchesi, BR, 179
Lundstedt, T, 42, 44
Luo, C, 137, 178
Luo, X, 39, 74, 114, 218
Lutz, SR, 39, 218
Lv, F, 114
M G Genton, N Cristianini, J Shawe-
Taylor, RW, 14, 43
Müller, CE, 113
Müller, KR, 45
MacConaill, L, 35, 217
MacKay, DJC, 86, 113
Madhavan, S, 37
Maffa, A, 38
Mager, DE, 41
Magrane, M, xxxii, 100, 113
Mahan, S, 35, 217
Maitland, ML, 221
Major, S, 219
Malliavin, TE, 37, 42, 44, 66, 71, 74, 121,
131–133, 175, 176, 178, 217–219,
227
Manning, G, 21, 43
Mansouri, K, 114
Mapa, FA, 35, 217
Marbach, D, 170, 178
Marchant, CA, 113
Marcotte, EM, 21, 44, 60, 61, 74
Marcou, G, 36, 71, 111, 145, 175, 178,
217, 227
Marcus, D, 132
Margolin, AA, 35, 217
Mark, TD, 43
Marnett, LJ, xxiv, 138, 169, 175, 177, 179
Masin, M, 36
Mason, JS, 44

- MATLAB, 87, 113
Mauri, A, 114
Mayer, Z, 141, 178
McDermott, U, 39, 43, 48, 73, 218, 219, 227
McGlinchey, S, 39, 111, 132, 176
McHale, CM, 32, 43
McKeown, MR, 38
McKusick, VA, 40
McLaren-Douglas, A, 39, 218
McLay, I, 44
McW, H, 74, 179
Meinshausen, N, 16, 43
Melnikova, I, 21, 43
Melssen, WJ, 14, 47
Meltzer, J, 35, 217
Meltzer, PS, 217
Melville, JL, 37
Menden, MP, 7, 9, 29, 31, 33, 43, 57, 73, 176, 185, 218, 219, 226, 227
Mendez-Lucio, O, 71, 131, 176, 217
Meng, C, 218
Mente, S, 53, 73
Mer, H, 46
Mergny, J, 219
Merta, PJ, 43
Mesirov, JP, 35, 217, 219, 220
Meslamani, J, 6, 26, 43, 57, 73
Mestres, J, 3–5, 39, 43, 57, 61, 62, 72, 74
Metz, JT, 8, 43
Mevik, BH, 123, 133
Mewes, H, 46, 115
Meyerson, M, 35, 217
Mez, R, 43
Meziane-Cherif, D, 217
Michalovich, D, 39, 111, 132, 176
Michel, V, 219
Mietzner, T, 21, 42, 93, 112
Mika, S, 45, 114
Mikhailov, D, 43
Milano, RJ, 39, 218
Milanov, ZV, 41
Mills, D, 60, 72, 141, 142, 177
Minchin, PR, 113, 179
Minowa, Y, 48, 76
Mironenko, T, 39, 218
Mitchell, JA, 180
Mitchell, JBO, 6, 35, 43, 73
Mitropoulos, X, 39, 218
Mitsiades, CS, 38
Miyashiro, JM, 177
Mokale, SN, 176
Monahan, JE, 35, 217
Monks, A, 219
Monks, AP, 220
Monroe, D, 32, 46
Moore, BC, 138, 178
Morais, P, 35, 217
Morgan, M, 33, 43
Morley, C, 44
Morris, J, 219
Morrison, MJ, 41
Morrisey, MP, 35, 217
Morton, D, 175
Mpindi, JP, 176, 218
Muceniece, R, 44
Muchmore, SW, 79, 111, 121, 131, 156, 175
Mulder-Krieger, T, 48, 75, 115
Muller, P, 41
Munos, BH, 44
Munoz, B, 36
Munson, PJ, 219
Murase, A, 177
Muratov, EN, 179
Murray, L, 35, 217
Murrell, DS, 37, 53, 54, 56, 59, 60, 66, 71, 74, 122, 123, 131, 133, 140, 141, 175, 178, 188, 205, 218, 219
Mussa, HY, 43, 58, 71, 73
Mutule, I, 73
Mutulis, F, 73
Mutyala, R, 179
Myatt, G, 43, 113

- Myer, VE, 35, 217
Myers, TG, 220
Mytelka, DS, 44
MÃijller, KR, 114
- Nédélec, E, 177
Nakagawa, Y, 177
Nakatsui, M, 46
Nakka, P, 38
Nantasenamat, C, 42
Narsinghani, T, 138, 178
Neal, RM, 86, 113
Netter, P, 177
Netzeva, TI, 16, 43, 79, 113
Nickisch, H, 87, 114
Niculescu-Mizil, A, 175
Nijima, S, 7, 13, 14, 44, 46, 48, 76
Nikolova-Jeliazkova, N, 43, 113
Nilges, M, 37, 71, 132, 217, 227
Nilsson, I, 131
Nilsson, L, 9, 47
Ning, X, 9, 44
Nishizuka, S, 189, 219
Nisius, B, 27, 44
Niyomrattanakit, P, 45, 113
Norinder, U, xxviii, 31, 33, 38, 44, 49, 67, 68, 74, 131, 132, 180, 193, 210, 219
Norman, TC, 32, 44
Nussmeier, NA, 137, 178
Nyholm, PG, 41
- O'Boyle, NM, 8, 44
O'Brien, P, 39, 218
O'Connor, PM, 220
O'Hara, RB, 113, 179
Obrezanova, O, 80, 86, 110, 113
Ogawa, T, 48, 76
Oksanen, J, 83, 113, 140, 179
Okuno, Y, 6, 7, 13, 14, 20, 36, 44, 46, 48, 71, 76, 111, 175, 217, 227
Oliveira, L, 47
- Olshen, R, 131
Ommen, B v, 40
Ong, C, 13, 14, 36
Ong, CS, 111, 131, 175
Onofrio, RC, 35, 217
Oppen, M, 110, 111
Osdol, WW v, 220
Osindero, S, 32, 40
Overington, J P, 37, 71, 132, 227
Overington, JP, 4, 28, 39, 40, 42, 48, 56, 72, 73, 75, 109, 111, 112, 132, 147, 176, 178
Oxenius, B, 175
- Paciorek, CJ, 110, 113
Pacold, ME, 38
Pahikkala, T, 60, 74
Palescandolo, E, 35, 217
Pallares, G, 37, 41
Pan, W, 15, 47
Pan, Y, 114
Paolini, GV, 9, 44
Paranal, RM, 38
Parchment, R, 219
Paricharak, S, 33, 44, 56, 74, 123, 132, 133
Park, J, 220
Park, K, 177
Park, Y, 21, 44, 60, 61, 74
Parlow, JL, 178
Passos, A, 219
Pastor, M, 7, 44
Patel, HK, 41
Patel, S, 177
Patlewicz, GY, 43, 113
Patrono, C, 138, 180
Paul, SM, 3, 44
Pedregosa, F, 190, 219
Peeters, A, 48
Peeters, MC, 99, 113
Peironcely, JE, 40, 72
Perkins, R, 40, 43, 113

- Perrot, M, 219
 Persinger, CC, 44
 Peshkin, L, 22, 40
 Peters, B, 47
 Petricoin, E, 219
 Petrovska, R, 41, 45, 73, 74, 113
 Pfister, TD, 219
 Pickett, S, 44
 Pickett, SD, 37
 Pietilä, S, 74
 Pietila, L, 46, 75
 Pijl, R vd, 115
 Pijl, Rvd, 48, 75
 Pinchback, R, 219
 Pineda, M, 217
 Pirotte, B, 178
 Pl, KD, 185, 204, 219, 220
 Plowman, J, 219
 Poda, GI, 46, 115
 Polishchuk, PG, 145, 179
 Polley, EC, 217
 Pommier, Y, 217, 219, 220
 Porter, J, 35, 217
 Poso, A, 218, 227
 Post, C, 44
 Prabhat, 113
 Prachayasittikul, V, 42
 Prasit, P, 178
 Prettenhofer, P, 219
 Price, ER, 36
 Price, EV, 36
 Price, S, 39, 218
 Prill, RJ, 178
 Prinjha, RK, 22, 44
 Pritchard, S, 41
 Prusis, P, 4, 5, 7, 9, 24, 42, 44–46, 59, 73–75, 81, 88, 89, 108, 113, 131, 176, 217
 Ps, A, 123, 133
 Ps, OJ, 37
 Puntanen, S, 85, 114
 Pure, E, 181
 Qi, J, 38
 Qian, F, 40
 Qifu, Z, 92, 114
 Qin, B, 121, 133
 Quackenbush, J, 40, 218, 227
 Quan, H, 175
 R Core Team, 83, 114
 Radhakrishnan, ML, 38
 Rahl, PB, 38
 Raimbaud, E, 38
 Raman, P, 35, 217
 Ramaswamy, S, 39, 48, 218
 Rameseder, J, 38
 Ramsland, PA, 27, 49
 Rangwala, 9, 44
 Rao, HB, 56, 74
 Rasmussen, CE, 17, 45, 83, 86, 87, 109, 110, 114, 123, 133
 Ratain, MJ, 221
 Ratsch, G, 111, 131, 175
 Rault, S, 42
 Ravela, J, 45
 Reddanna, P, 179
 Reddy, A, 35, 217
 Reddy, MR, 179
 Reddy, RN, 138, 179
 Reese, J, 177
 Reich, M, 35, 217
 Reimers, M, 220
 Reindeau, D, 178
 Reinhold, W, 220
 Reinhold, WC, 188, 189, 217, 219, 220
 Reinhold, WO, 220
 Remmert, M, 74, 179
 Ren, Y, 80, 114
 Reutlinger, M, 25, 45, 80, 114
 Rhee, S, 6, 38, 45
 Ricciotti, E, 181
 Riddell, R, 175
 Riddick, G, 185, 220
 Ridgeway, G, 123, 133

- Rimon, G, 139, 179
Ripley, BD, 123, 133
Robak, P, 138, 179
Robak, T, 138, 179
Robinson, DD, 26, 45
Rodrigo, J, 46, 75
Rodrigues, T, 45, 114
Rogers, D, 7, 45, 58, 74, 79, 81, 114, 122, 133, 140, 179
Rognan, D, 5, 6, 9, 20, 25, 26, 38, 41, 43, 45–48, 57, 61, 62, 73–75
Rohrer, DC, 46, 115
Romero, PA, 80, 114
Root, DE, 38
Rosenbaum, L, 145, 179
Roth, BL, 41, 42
Roth, FP, 101, 112, 172, 177
Rs, D, 113
Rson, DH, 47
Rson, DL, 180
Rson, L, 39, 218
Rson, PG, 38
Rubin, G, 177
Rubinstein, L, 219
Rubinstein, LV, 220
Ruppin, E, 6, 39
Ryan, M, 219

Söding, J, 74, 179
Sabatini, DM, 38
Saez-Rodriguez, J, 39, 43, 73, 176, 218, 219, 227
Sahigara, F, 79, 114
Salimi, N, 47
Sandberg, M, 6–8, 45, 83, 114, 140, 179
Sanders, M, 47
Sandler, RS, 175
Sandt, JJM vd, 113
Sanschagrín, P, 46
Sarai, A, 6, 38
Sassano, MF, 42
Sausville, EA, 220

Scalia, R, 181
Schacht, AL, 44
Schaefer, GI, 36
Schalon, C, 41
Schapira, M, 35
Scheiber, J, 36, 71, 111, 131
Scheiman, J, 177
Scherf, U, 220
Schinzel, AC, 38
Schlegel, R, 35, 217
Schlkopf, B, 14, 45
Schneider, G, 45, 114
Schneider, P, 45, 114
Scholkopf, B, 14, 45, 111, 131, 175
Scholz, M, 167, 179
Schrödinger, L, 28, 45
Schreiber, SL, 36
Schroeter, T, 45, 114
Schultz, T, 113
Schwaighofer, A, 13, 17, 45, 80, 95, 110, 114
Schweppenhäuser, J, 167, 176
Scudiero, DA, 219, 220
Segall, MD, 80, 113
Seibert, K, 177
Sellers, WR, 35, 217
Serbedzija, G, 36
Seshasayee, A, 37
Sette, A, 47
Settleman, J, 39, 218
SH, W, 26, 45
Shafer, G, 67, 74
Shafer, RW, 38, 45
Shaffer, JP, 172, 179
Shakyawar, S, 74
Shamji, AF, 36
Shang, Z, 115, 116
Shankavaram, U, 220
Shapland, RHB, 44
Sharan, R, 6, 39
Sharma, NP, 179
Sharma, SV, 39, 218

- Sheinerman, FB, 56, 74
 Shen, J, 37, 56, 74
 Shen, X, 15, 47
 Sheridan, RP, 46, 142, 143, 170, 179, 190, 220
 Shi, J, 38
 Shi, L, 40
 Shi, T, 40
 Shindyalov, IN, 175
 Shipway, A, 35, 217
 Shiraishi, A, 7, 13, 14, 44
 Shiraishi, AG, 7, 25, 46
 Shivakumar, P, 205, 220
 Shoemaker, BA, 47, 75, 133
 Shoemaker, RH, 30, 46, 185, 219, 220
 Shoichet, BK, 41–43
 Shonesy, BC, 177
 Shoshan, MC, 3, 46
 Sidhu, RS, 179
 Sievers, F, 56, 74, 139, 179
 Silva, M d, 35, 217
 Simmons, DL, 138, 178, 180
 Simon, R, 60, 75
 Simon, RM, 217
 Simpson, GL, 113, 179
 Singer, D, 176, 218
 Sinz, F, 37
 Sippl, MJ, 26, 46
 Sjöström, M, 45, 114, 123, 133, 179
 Skilling, J, 86, 115
 Slonim, DK, 220
 Smith, J, 72, 112, 132, 176
 Smith, JA, 48
 Smith, MT, 43
 Smith, WL, 179
 Smola, A, 112, 132
 Smola, AJ, 14, 45
 Smolewski, P, 138, 179
 Soares, J, 39, 48, 218
 Soh, J, 138, 180
 Solé, RV, 43, 74
 Solov'ev, V, 178
 Solymos, P, 113, 179
 Somers, F, 178
 Song, H, 220
 Soni, NB, 43
 Sonkin, D, 35, 217
 Sonnenburg, S, 111, 131, 175
 Sorger, PK, 227
 Sostres, C, 137, 138, 180
 Sottriffer, C, 27, 46
 Sougnez, C, 35, 217
 Spjuth, O, 7, 24, 41, 42, 46, 75, 110, 113, 115
 Spowage, BM, 145, 180
 Spring, DR, 40, 72, 132
 Ss, MHH, 113, 179
 Ss, RC, 112
 Sson, Ja, 39, 218
 Stå lring, J, 180
 Stamenkovic, I, 39, 218
 Stanton, DW, 113
 Staunton, JE, 185, 204, 220
 Steichen, TJ, xxxii, 155, 180
 Steinbeck, C, 26, 40
 Stern, AM, 36
 Stewart, ML, 36
 Stieger, B, 38, 72, 218
 Stierum, RH, 40
 Stockwell, BR, 36
 Stolovitzky, G, 32, 46, 176, 178, 218
 Stone, C, 131
 Stott, IP, 74, 133, 178, 219
 Stoven, V, 40
 Strömbergsson, H, 12, 25–27, 35, 46
 Stransky, N, 35, 36, 217
 Stratton, MR, 39, 48, 218
 Styan, GPH, 85, 114
 Su, Z, 40
 Subramanian, A, 219
 Subramanian, V, 8, 21, 22, 26, 46, 57, 75, 131, 176, 217
 Sudarsanam, S, 43
 Sukuru, SCK, 71, 111, 131

- Sulzle, D, 45, 114
Sunshine, M, 219, 220
Surapaneni, S, 29, 49
Surdez DC, L, 39, 218
Surgand, JS, 25, 46, 56, 75
Suter, L, 32, 46
Sutherland, JJ, 47
Suzek, TO, 47, 75, 133
Svetnik, V, 16, 46
Swier, R, 48, 75
Swier, RF, 75
Szakacs, G, 185, 204, 220
Szwajda, A, 74
- T, RC, 113
T, S, 73
Takada, J, 177
Takematsu, H, 48, 76
Tam, AT, 39, 218
Tamayo, P, 219, 220
Taneishi, K, 44
Tang, J, 74
Tang, SY, 181
Tang, WJ, 42
Tang, Y, 37
Tartaglia, GG, 36
Teh, YW, 32, 40
Tenenbaum, JB, 38, 111
Terlain, B, 177
Testa, B, 37
Tetko, IV, 16, 46, 79, 115
Thérien, M, 178
Thi, H, 39, 218
Thibault, J, 35, 217
Thirion, B, 219
Thompson, IR, 39, 48, 218
Thompson, JD, 74, 179
Thompson, WJ, 180
Thorvaldsdóttir, H, 219
Thun, MJ, 138, 180
Tian, F, 115
- Tibshirani, R, xxii, xxxii, 64, 72, 125, 132, 153, 155, 176
Tiikkainen, P, 16, 33, 46, 79, 115, 120, 133
Tipping, ME, 13, 15, 47, 123, 133
Tobi, D, 9, 35
Todeschini, R, 58, 65, 71, 75, 114
Tomic, S, 9, 47
Tong, C, 46
Tong, W, 40, 113
Treiber, DK, 37, 41
Tresp, V, 110, 115
Trievel, RC, 179
Tropsha, A, 27, 49, 64, 65, 72, 75, 80, 88, 112, 115, 142, 168, 176, 180, 190, 218, 220
Trzaskos, JM, 180
Tsang, S, 121, 133
Tsujimoto, G, 44, 48, 76
Tu, Y, 121, 132
Tullio, P d, 178
- Ueno, N, 177
Uhlén, S, 73, 74
Ulven, T, 39
Urban, L, 36, 43
Uren, A, 37
Ushir, YV, 176
- Vakoc, CR, 38
Valencia, C, 47, 75
Valle, D, 40
Valverde, S, 43, 74
van Westen, G J P, 37, 71, 132, 227
Vandermeersch, T, 44
Vanderplas, J, 219
Vane, JR, 137, 180
Varkonyi, P, 131
Varma, S, 60, 75, 189, 219, 220
Varnek, A, 13, 15, 36, 41, 71, 111, 175, 178, 217, 227
Varoquaux, G, 219

- Vayer, P, 178
 Veenhuizen, AW, 177
 Vega, NM, 178
 Veiksina, S, 73
 Veith, G, 113
 Venables, WN, 123, 133
 Venkatesan, K, 35, 217
 Verburg, KM, 178
 Vereycken, I, 48
 Verhoeven, S, 47
 Vert, J, 186, 218
 Vert, JP, 14, 40, 45
 Vidler, LR, 22, 47
 Vieth, M, 4, 47
 Villa, P, 47, 75
 Visco, DM, 178
 Viswanadhan, VN, 220
 Vita, R, 6, 47
 Vlieg, J d, 47
 Vlijmen, HWT v, 40, 48, 72, 75, 115, 220
 Vojnovic, I, 180
 Vovk, V, 67, 74
 Vriend, G, 47
 Vries, H d, 115
 Vroling, B, 8, 47
 Vulpetti, A, 41, 42, 73, 112, 132, 177
 Wade, RC, 9, 47
 Wagner, H, 113, 179
 Walker, RL, 217
 Walling, J, 220
 Waltham, M, 219
 Wang, J, 15, 47, 75, 133
 Wang, L, 35, 217
 Wang, NJ, 176, 218
 Wang, Q, 40
 Wang, R, 6, 27, 47
 Wang, S, 36, 47
 Wang, Y, 3, 8, 33, 42, 47, 54, 75, 120, 133
 Warmuth, M, 35, 217
 Warner, TD, 137, 180
 Wassermann, AM, 9, 26, 47
 Wawer, M, 36
 Weber, BL, 35, 217
 Weber, E, 43
 Wegner JKaI, AP, 115
 Wegner, JJ, 48, 75, 220
 Wegner, JK, 40, 48, 72, 75, 115
 Wehrens, R, 123, 133
 Wei, W, 175
 Weier, RM, 177
 Weill, N, 6, 20, 25, 26, 47, 48, 57, 75
 Weinmann, H, 22, 41
 Weinstein, IB, 180
 Weinstein, JN, 31, 48, 185, 219, 220
 Weiss, R, 219
 Weissig, H, 175
 Welch, L, 220
 Wenlock, MC, 120, 133
 Wennerberg, K, 176, 218, 227
 Weskamp, N, 73
 Westbrook, J, 175
 Westen, GJP v, 4–7, 9, 11, 12, 14, 15, 21, 23, 24, 29, 34, 37, 38, 40, 48, 55–60, 62, 71, 72, 74, 75, 79, 81, 92, 93, 101, 108–110, 113, 115, 131, 133, 175, 176, 178, 186, 217–220
 Weston, J, 37
 Wheeldon, EB, 32, 46
 Wheeler, HE, 186, 221
 Whelton, AA, 178
 White, FM, 38
 Whitebread, S, 36, 43, 177
 Whyte, DB, 43
 Wichmann, O, 57, 72
 Wiederstein, M, 26, 46
 Wiener, M, 123, 133
 Wikberg, JES, 6, 41, 42, 44–46, 73–75, 113, 115
 Wilhelm, M, 218
 Wilkerson, WW, 166, 180
 Willett, P, 3, 48
 Willighagen, EL, 57, 75
 Wilm, A, 74, 179

- Wilson, AJ, 36
Wilson, CJ, 35, 217
Winckler, W, 35, 217
Wisse, LE, 113
Witherington, J, 22, 44
Witte, P d, 38, 72, 218
Wittes, RE, 220
Witzmann, Fa, 40
Wodicka, LM, 37, 41
Wohlfahrt, G, 26, 40, 46, 75, 131, 176, 217
Wold, S, 45, 114, 123, 133, 179
Wong, J, 40
Wong, N, 180
Wood, DJ, 143, 170, 180
Worachartcheewan, A, 42
Worth, A, 113
Worth, AP, 43
Wright, AP, 49
Ws, CKI, 17, 45, 83, 86, 109, 110, 114, 123, 133
Wu, B, 114
Wu, D, 7, 14, 23, 39, 48, 92, 111, 115
Wu, L, 42
Wu, Q, 40, 112
Wu, W, 181
Wu, Y, 116
Wu, Z, 218

Xhaard, H, 46, 75
Xia, Y, 121, 132, 133
Xiao, J, 47, 75, 133
Xiao, N, 58, 76
Xie, Q, 40
Xie, WL, 137, 180
Xu, L, 178
Xu, Q, 37, 58, 76

Y, Z, 114
Yabuuchi, H, 6, 25, 44, 48, 57, 76
Yaffe, MB, 38
Yahorau, A, 45, 113

Yahorava, S, 41, 45, 73, 113
Yamamoto, KR, 44
Yamanishi, Y, 186, 221
Yang, C, 113
Yang, GB, 74
Yang, J, 44
Yang, L, 37
Yang, Q, 114
Yang, RK, 181
Yang, W, 7, 39, 48, 218
Yap, CW, 59, 76, 140, 181
Yi, KY, 177
Yip, KY, 133
Yoo, S, 177
Young, L, 219
Young, RA, 38
Yu, GK, 35, 217
Yu, J, 35, 217
Yu, K, 74
Yu, XJ, 181
Yu, Y, 37, 137, 138, 177, 181
Yu, Z, 181
Yuan, C, 179
Yue, SY, 38
Yuriev, E, 27, 49

Zaharevitz, DW, 220
Zarebski, L, 47
Zarrinkar, P, 37, 41
Zeileis, A, 112, 132
Zell, A, 179
Zhang, D, 29, 49
Zhang, J, 40, 47, 74, 75, 133
Zhang, JBT, 121, 133
Zhang, L, 37, 43
Zhang, Q, 39, 48, 111, 115
Zhang, S, 27, 49, 138, 181
Zhang, T, 39, 218
Zhang, W, 220
Zhang, X, 42, 181
Zhang, Y, 39, 48, 111, 115
Zheng, VW, 39

Zhou, D, [181](#)
Zhou, G, [37](#)
Zhou, P, [80](#), [115](#), [116](#)
Zhou, W, [39](#), [218](#)
Zhou, Y, [37](#)
Zhou, Z, [47](#), [75](#), [133](#)
Zhu, F, [74](#)
Zhu, R, [39](#), [40](#), [48](#), [111](#), [112](#), [115](#)
Zhu, W, [74](#)
Zhu, YJ, [217](#)
Zhuo, W, [113](#)
Zilliacus, J, [9](#), [49](#)
Zimmermann, GR, [36](#)
Zou, XP, [181](#)

Subject Index

- k-Nearest Neighbours, 14, 123
- Analysis of Variance, xxvi, xxvii, 125, 127, 195, 200, 201, 207
- Applicability Domain, xix, 16, 17, 33, 79, 80, 85, 97, 98, 109
- Area Under the Curve, 14
- Automatic Relevance Determination, xx, 86, 87, 107, 108, 110
- Cancer Cell-Line Encyclopedia, xxx, 30, 31, 205, 215, 226
- Cell-Line Identity Fingerprints, 64, 191
- Collaborative Filtering, 12, 19
- Confidence Interval, xxviii, 68, 202, 210
- Cross-Validation, 60, 61, 87, 92, 123, 137, 141
- Cyclooxygenases, viii, xxi, xxii, xxiv, xxv, xxxii, 29, 123, 137–140, 145, 147–149, 152, 158, 161, 164–171
- Deoxyribonucleic Acid, 9, 22
- Directory of Useful Decoys, 15
- DNA Copy Number Variation, viii, 57, 195, 204
- Explicit Learning, 196
- Extended Connectivity Fingerprints, 81
- G Protein-Coupled Receptor, xviii, xix, xxxi, 4, 11, 17, 19–21, 25, 26, 28, 29, 33, 80, 82, 83, 87, 89, 91, 92, 94–99, 101, 104, 108, 109
- Gaussian Process, viii, xiii, xvii–xxi, 12, 16–18, 28, 34, 56, 60, 80, 83–89, 91–99, 101, 102, 108–110, 121–123, 126–129
- Genomics of Drug Sensitivity in Cancer, 30, 31, 226
- Gradient Boosting Machine, xx, xxi, 29, 122, 123, 126–129, 142, 150, 158, 168–170
- High-Throughput Screening, 19, 79
- Histone Deacetylase, 14, 15, 23
- Human Immunodeficiency Virus, 11, 23, 24, 92
- Inductive Transfer, 63, 191, 196
- Leave-One-Cell-Line-Out, xxvii, xxviii, 62, 192, 201–204, 208, 209
- Leave-One-Compound-Cluster-Out, xxviii, 62, 192, 202, 203, 205, 209
- Leave-One-Target-Out (Leave-One-Tissue-Out in chapter 6), xxvii, xxviii, 62, 192, 201–204, 208, 209
- MDL Drug Data Report, 15
- Model Stacking, 158, 159, 169, 193
- Monte Carlo, 86
- Non-Steroidal Anti-Inflammatory Drug, xxiv, 137, 138, 169
- Normalized Polynomial Kernel, xviii, xix, 32, 84, 87, 89, 92–98, 108, 109
- Organic Anion-Transporting Polypeptide, 16

- Partial Least Squares, [xx](#), [xxi](#), [12](#), [22](#), [27](#),
[59](#), [94](#), [108](#), [122](#), [123](#), [126](#), [127](#),
[129](#)
- Pearson VII Function-Based Universal
Kernel, [xix](#), [14](#), [84](#), [92](#), [95](#), [96](#)
- Principal Component Analysis, [xxi](#), [xxii](#),
[147–149](#), [151](#)
- Protein-Ligand Fingerprints, [20](#)
- Proteochemometric Modelling, [vii–ix](#),
[xvii](#), [xviii](#), [xxi–xxx](#), [4](#), [5](#), [9–12](#),
[14–17](#), [19–34](#), [53–62](#), [64](#), [79](#), [80](#),
[86](#), [89](#), [92](#), [95](#), [107–110](#), [139](#), [141](#),
[143](#), [145](#), [146](#), [148](#), [152](#), [153](#), [156](#),
[158](#), [161](#), [163–165](#), [168](#), [169](#), [171](#),
[173](#), [186–188](#), [191](#), [192](#), [194–204](#),
[207](#), [210](#), [213](#), [215](#), [226](#)
- Quantitative Cell-Line-Activity Modelling,
[191](#), [195](#)
- Quantitative Sequence-Activity Mod-
elling, [xxii](#), [53](#), [54](#), [63](#), [141](#), [153](#),
[157](#), [158](#), [168](#)
- Quantitative Structure-Activity Relation-
ship, [vii](#), [viii](#), [xvii](#), [xxii](#), [3](#), [9](#), [10](#),
[14–17](#), [19](#), [22](#), [33](#), [34](#), [53](#), [54](#), [62–](#)
[64](#), [80](#), [92](#), [94](#), [108](#), [119](#), [121](#), [122](#),
[124](#), [128](#), [138](#), [141](#), [153](#), [157](#), [158](#),
[164](#), [168](#), [190](#), [191](#), [195](#)
- Quantitative Structure-Property Rela-
tionship, [53](#), [54](#)
- Radial Basis Function, [14](#), [85](#)
- Random Forest, [xx](#), [xxi](#), [12](#), [14–16](#), [21](#),
[29](#), [122](#), [123](#), [126–129](#), [141–143](#),
[150](#), [158](#), [159](#), [161](#), [168–170](#), [190](#),
[193](#)
- Relevance Vector Machine, [xx](#), [xxi](#), [15](#),
[122](#), [123](#), [126](#), [127](#), [129](#)
- Reverse Phase Protein Lysate Microar-
ray, [195](#)
- Root Mean Squared Error, [xx](#), [xxi](#), [xxvii](#),
[19](#), [23](#), [26](#), [27](#), [29](#), [119](#), [126](#), [129](#),
[149](#), [158](#), [159](#), [192](#), [195](#), [201](#), [208](#)
- Self-Organizing Map, [xxvi](#), [206](#)
- Support Vector Machines, [xviii](#), [xx](#), [xxi](#),
[xxiii](#), [12](#), [14–17](#), [20](#), [29](#), [56](#), [60](#),
[80](#), [84](#), [93](#), [94](#), [108–110](#), [122](#), [123](#),
[126–129](#), [142](#), [143](#), [150](#), [158](#), [159](#),
[162](#), [168–170](#)
- Support Vector Regression, [19](#)
- Target Identity Fingerprints, [63](#)
- Transmembrane, [20](#), [25](#), [29](#)
- Tukey's Honest Significant Difference
Test, [xxvi](#), [195](#), [196](#), [200–203](#), [206](#)
- United States Food and Drug Adminis-
tration Agency, [31](#), [137](#), [203](#)
- United States National Cancer Institute,
[30](#), [31](#), [185](#), [204](#)